Mitigating Bias in Adaptive Data Gathering via Differential Privacy

Seth Neel [1] **Aaron Roth** [2]

Abstract

Data that is gathered adaptively — via bandit algorithms, for example — exhibits bias. This is true both when gathering simple numeric valued data — the empirical means kept track of by stochastic bandit algorithms are biased downwards — and when gathering more complicated data — running hypothesis tests on complex data gathered via contextual bandit algorithms leads to false discovery. In this paper, we show that this problem is mitigated if the data collection procedure is differentially private. This lets us both bound the bias of simple numeric valued quantities (like the empirical means of stochastic bandit algorithms), and correct the p-values of hypothesis tests run on the adaptively gathered data. Moreover, there exist differentially private bandit algorithms with near optimal regret bounds: we apply existing theorems in the simple stochastic case, and give a new analysis for linear contextual bandits. We complement our theoretical results with experiments validating our theory[1].

1. Introduction

Many modern data sets consist of data that is gathered *adaptively*: the choice of whether to collect more data points of a given type depends on the data already collected. For example, it is common in industry to conduct "A/B" tests to make decisions about many things, including ad targeting, user interface design, and algorithmic modifications, and this A/B testing is often conducted using "bandit learning algorithms" (Bubeck et al., 2012), which adaptively select treatments to show to users in an effort to find the best treatment as quickly as possible. Similarly, sequential clinical trials may halt or re-allocate certain treatment groups due to preliminary results, and empirical scientists may initially try and test multiple hypotheses and multiple treatments, but then decide to gather more data in support of certain hypotheses and not others, based on the results of preliminary statistical tests.

Unfortunately, as demonstrated by (Nie et al., 2017), the data that results from adaptive data gathering procedures will often exhibit substantial *bias*. As a result, subsequent analyses that are conducted on the data gathered by adaptive procedures will be prone to error, unless the bias is explicitly taken into account. This can be difficult. (Nie et al., 2017) give a selective inference approach: in simple stochastic bandit settings, if the data was gathered by a specific stochastic algorithm that they design, they give an MCMC based procedure to perform maximum likelihood estimation to recover de-biased estimates of the underlying distribution means. In this paper, we give a related, but orthogonal approach whose simplicity allows for a substantial generalization beyond the simple stochastic bandits setting. We show that in very general settings, if the data is gathered by a differentially private procedure, then we can place strong bounds on the bias of the data gathered, without needing any additional de-biasing procedure. Via elementary techniques, this connection implies the existence of simple stochastic bandit algorithms with nearly optimal worst-case regret bounds, with very strong bias guarantees. By leveraging existing connections between differential privacy and adaptive data analysis (Dwork et al., 2015c; Bassily et al., 2016; Rogers et al., 2016), we can extend the generality of our approach to bound not just bias, but to correct for effects of adaptivity on arbitrary statistics of the gathered data. Since the data being gathered will generally be useful for some as yet unspecified scientific analysis, rather than just for the narrow problem of mean estimation, our technique allows for substantially broader possibilities compared to past approaches.

1.1. Our Results

This paper has three main contributions:

1. Using elementary techniques, we provide explicit bounds on the bias of empirical arm means maintained by bandit algorithms in the simple stochastic

[1]Department of Statistics, The Wharton School, University of Pennsylvania [2]Department of Computer Science, University of Pennsylvania. Correspondence to: Seth Neel <sethneel93@gmail.com>, Aaron Roth <aaroth@cis.upenn.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

[1]This extended abstract is missing many details, proofs, and results that can be found in the full version (Neel & Roth, 2018).

setting that make their selection decisions as a differentially private function of their observations. Together with existing differentially private algorithms for stochastic bandit problems, this yields an algorithm that obtains an essentially optimal worst-case regret bound, and guarantees minimal bias (on the order of $O(1/\sqrt{K \cdot T})$) for the empirical mean maintained for every arm. In the full version (Neel & Roth, 2018), we also extend our results to the linear contextual bandit problem, proving new bounds for a private linear UCB algorithm along the way.

2. We then make a general observation, relating adaptive data gathering to an adaptive analysis of a fixed dataset (in which the choice of which query to pose to the dataset is adaptive). This lets us apply the large existing literature connecting differential privacy to adaptive data analysis. In particular, it lets us apply the max-information bounds of (Dwork et al., 2015b; Rogers et al., 2016) to our adaptive data gathering setting. This allows us to give much more general guarantees about the data collected by differentially private collection procedures, that extend well beyond bias. For example, it lets us correct the p-values for arbitrary hypothesis tests run on the gathered data.

3. Finally, we run a set of experiments that measure the bias incurred by the standard UCB algorithm in the stochastic bandit setting, contrast it with the low bias obtained by a private UCB algorithm, and show that there are settings of the privacy parameter that simultaneously can make bias statistically insignificant, while having competitive empirical regret with the non-private UCB algorithm. We also demonstrate in the linear contextual bandit setting how failing to correct for adaptivity can lead to false discovery when applying t-tests for non-zero regression coefficients on an adaptively gathered dataset.

1.2. Related Work

This paper bridges two recent lines of work. Our starting point is two recent papers: (Villar et al., 2015) empirically demonstrate in the context of clinical trials that a variety of simple stochastic bandit algorithms produce biased sample mean estimates (Similar results have been empirically observed in the context of contextual bandits (Dimakopoulou et al., 2017)). (Nie et al., 2017) prove that simple stochastic bandit algorithms that exhibit two natural properties (satisfied by most commonly used algorithms, including UCB and Thompson Sampling) result in empirical means that exhibit negative bias. They then propose a heuristic algorithm which computes a maximum likelihood estimator for the sample means from the empirical means gathered by a modified UCB algorithm which adds Gumbel noise to the decision statistics. (Deshpande et al., 2017) propose a debiasing procedure for ordinary least-squares estimates computed from adaptively gathered data that trades off bias for variance, and prove a central limit theorem for their method. In contrast, the methods we propose in this paper are quite different. Rather than giving an ex-post debiasing procedure, we show that if the data were gathered in a differentially private manner, no debiasing is necessary. The strength of our method is both in its simplicity and generality: rather than proving theorems specific to particular estimators, we give methods to correct the p-values for *arbitrary* hypothesis tests that might be run on the adaptively gathered data.

The second line of work is the recent literature on *adaptive data analysis* (Dwork et al., 2015c;b; Hardt & Ullman, 2014; Steinke & Ullman, 2015; Russo & Zou, 2016; Wang et al., 2016; Bassily et al., 2016; Hardt & Blum, 2015; Cummings et al., 2016; Feldman & Steinke, 2017a;b) which draws a connection between differential privacy (Dwork et al., 2006) and generalization guarantees for adaptively chosen statistics. The adaptivity in this setting is dual to the setting we study in the present paper: In the adaptive data analysis literature, the dataset itself is fixed, and the goal is to find techniques that can mitigate bias due to the adaptive selection of analyses. In contrast, here, we study a setting in which the data gathering procedure is itself adaptive, and can lead to bias even for a fixed set of statistics of interest. However, we show that adaptive data gathering can be re-cast as an adaptive data analysis procedure, and so the results from the adaptive data analysis literature can be ported over.

2. Preliminaries

2.1. Simple Stochastic Bandit Problems

In a simple stochastic bandit problem, there are K unknown distributions P_i over the unit interval [0,1], each with (unknown) mean μ_i. Over a series of rounds $t \in \{1, \ldots, T\}$, an algorithm $\mathcal{A}$ chooses an arm $i_t \in [K]$, and observes a reward $y_{i_t,t} \sim P_{i_t}$. Given a sequence of choices $i_1, \ldots, i_T$, the pseudo-regret of an algorithm is defined to be:

$$\mathrm{Regret}((P_1, \ldots, P_K), i_1, \ldots, i_T) = T \cdot \max_i \mu_i - \sum_{t=1}^{T} \mu_{i_t}$$

We say that regret is bounded if we can put a bound on the quantity $\mathrm{Regret}((P_1, \ldots, P_K), i_1, \ldots, i_T)$ in the worst case over the choice of distributions $P_1, \ldots, P_K$, and with high probability or in expectation over the randomness of the algorithm and of the reward sampling.

As an algorithm $\mathcal{A}$ interacts with a bandit problem, it generates a *history* Λ , which records the sequence of actions

taken and rewards observed thus far: $\Lambda_t = \{(i_\ell, y_{i_\ell,\ell})\}_{\ell=1}^{t-1}$. We denote the space of histories of length T by $\mathcal{H}^T = ([K] \times \mathbb{R})^T$.

The definition of an algorithm $\mathcal{A}$ induces a sequence of T (possibly randomized) selection functions $f_t : \mathcal{H}^{t-1} \to [K]$, which map histories onto decisions of which arm to pull at each round.

2.2. Contextual Bandit Problems

In the contextual bandit problem, decisions are endowed with observable features. Our algorithmic results in this paper focus on the *linear* contextual bandit problem, but our general connection between adaptive data gathering and differential privacy extends beyond the linear case. For simplicity of exposition, we specialize to the linear case here.

There are K arms i, each of which is associated with an unknown d-dimensional linear function represented by a vector of coefficients $\theta_i \in \mathbb{R}^d$ with $||\theta_i||_2 \leq 1$. In rounds $t \in \{1, \ldots, T\}$, the algorithm is presented with a *context* $x_{i,t} \in \mathbb{R}^d$ for each arm i with $||x_{i,t}||_2 \leq 1$, which may be selected by an adaptive adversary as a function of the past history of play. We write x_t to denote the set of all K contexts present at round t. As a function of these contexts, the algorithm then selects an arm i_t, and observes a reward $y_{i_t,t}$. The rewards satisfy $\mathbb{E}[y_{i,t}] = \theta_i \cdot x_{i,t}$ and are bounded to lie in $[0, 1]$.

In the contextual setting, histories incorporate observed context information as well: $\Lambda_t = \{(i_\ell, x_\ell, y_{i_\ell,\ell})\}_{\ell=1}^{t-1}$.

Again, the definition of an algorithm $\mathcal{A}$ induces a sequence of T (possibly randomized) selection functions $f_t : \mathcal{H}^{t-1} \times \mathbb{R}^{d \times K} \to [K]$, which now maps both a history and a set of contexts at round t to a choice of arm at round t.

2.3. Data Gathering in the Query Model

Above we've characterized a bandit algorithm $\mathcal{A}$ as *gathering* data adaptively using a sequence of selection functions f_t, which map the observed history $\Lambda_t \in \mathcal{H}^{t-1}$ to the index of the next arm pulled. In this model only after the arm is chosen is a reward drawn from the appropriate distribution. Then the history is updated, and the process repeats.

In this section, we observe that whether the reward is drawn after the arm is "pulled," or in advance, is a distinction without a difference. We cast this same interaction into the setting where an analyst asks an adaptively chosen sequence of queries to a fixed dataset, representing the arm rewards. The process of running a bandit algorithm $\mathcal{A}$ up to time T can be formalized as the adaptive selection of T queries against a single database of size T - fixed in advance. The formalization consists of observing two things. First, by the principle of deferred randomness, we can view any (simple or contextual) bandit algorithm as operating in a setting in the rewards available for every arm at every time step have been sampled before the start of the algorithm, rather than online as the algorithm makes its selections. Second, the choice of arm pulled at time t by the bandit algorithm can be viewed as the answer to an adaptively selected query against this fixed dataset of rewards.

Adaptive data analysis is formalized as an interaction in which a data analyst $\mathcal{A}$ performs computations on a dataset D, observes the results, and then may choose the identity of the next computation to run as a function of previously computed results (Dwork et al., 2015c;a). A sequence of recent results shows that if the queries are differentially private in the dataset D, then they will not in general overfit D, in the sense that the distribution over results induced by computing $q(D)$ will be "similar" to the distribution over results induced if q were run on a new dataset, freshly sampled from the same underlying distribution (Dwork et al., 2015c;a; Bassily et al., 2016; Dwork et al., 2015b; Rogers et al., 2016). We will be more precise about what these results say in Section 4.

Recall that histories Λ record the choices of the algorithm, in addition to its observations. It will be helpful to introduce notation that separates out the choices of the algorithm from its observations. In the simple stochastic setting and the contextual setting, given a history Λ_t, an *action history* $\Lambda_t^{\mathcal{A}} = (i_1, \ldots, i_{t-1}) \in [K]^{t-1}$ denotes the portion of the history recording the actions of the algorithm.

In the simple stochastic setting, a *bandit tableau* is a $T \times K$ matrix $D \in \left([0,1]^K\right)^T$. Each row D_t of D is a vector of K real numbers, intuitively representing the rewards that would be available to a bandit algorithm at round t for each of the K arms. In the contextual setting, a bandit tableau is represented by a pair of $T \times K$ matrices: $D \in \left([0,1]^K\right)^T$ and $C \in \left((\mathbb{R}^d)^K\right)^T$. Intuitively, C represents the *contexts* presented to a bandit algorithm $\mathcal{A}$ at each round: each row C_t corresponds to a set of K contexts, one for each arm. D again represents the rewards that would be available to the bandit algorithm at round t for each of the K arms.

We write Tab to denote a bandit tableau when the setting has not been specified: implicitly, in the simple stochastic case, $\mathrm{Tab} = D$, and in the contextual case, $\mathrm{Tab} = (D, C)$.

Given a bandit tableau and a bandit algorithm $\mathcal{A}$, we have the following interaction:

We denote the subset of the reward tableau D corresponding to rewards that would have been revealed to a bandit algorithm $\mathcal{A}$ given action history $\Lambda_t^{\mathcal{A}}$, by $\Lambda_t^{\mathcal{A}}(D)$. Concretely if $\Lambda_t^{\mathcal{A}} = (i_1, \ldots, i_{t-1})$ then $\Lambda_t^{\mathcal{A}}(D) = \{(i_\ell, y_{i_\ell,\ell})\}_{\ell=1}^{t-1}$. Given a selection function f_t and an action history $\Lambda_t^{\mathcal{A}}$, de-

Interact
Inputs: Time horizon T, bandit algorithm $\mathcal{A}$, and bandit tableau Tab (D in the simple stochastic case, (D, C) in the contextual case)

1: **for** $t = 1$ to T **do**
2: (contextual case) Show$\mathcal{A}$ contexts $C_{t,1}, \ldots, C_{t,K}$
3: Let $\mathcal{A}$ play action i_t
4: Show $\mathcal{A}$ reward D_{t,i_t}
5: **end for**
6: **Return:** $(i_1, \ldots, i_T)$

fine the query $q_{\Lambda_t^{\mathcal{A}}}$ as $q_{\Lambda_t^{\mathcal{A}}}(D) = f_t(\Lambda_t^{\mathcal{A}}(D))$.

We now define Algorithms **Bandit** and **InteractQuery**. **Bandit** is a standard contextual bandit algorithm defined by selection functions f_t, and **InteractQuery** is the **Interact** routine that draws the rewards in advance, and at time t selects action i_t as the result of query $q_{\Lambda_t^{\mathcal{A}}}$. With the above definitions in hand, it is straightforward to show that the two Algorithms are equivalent, in that they induce the same joint distribution on their outputs. In both algorithms for convenience we assume we are in the linear contextual setting, and we write η_{i_t} to denote the i.i.d. error distributions of the rewards, conditional on the contexts.

Bandit
Inputs: $T, k, \{x_{it}\}, \{\theta_i\}, f_t, \Lambda_0 = \emptyset$

1: **for** $t = 1, \ldots, T$: **do**
2: Let $i_t = f_t(\Lambda_{t-1})$
3: Draw $y_{i_t,t} \sim x_{i_t,t} \cdot \theta_{i_t} + \eta_{i_t}$
4: Update $\Lambda_t = \Lambda_{t-1} \cup (i_t, y_{i_t,t})$
5: **end for**
6: **Return:** Λ_T

InteractQuery
Inputs: $T, k, D : D_{it} = \theta_i \cdot x_{it} + \eta_{it}, f_t$

1: **for** $t = 1, \ldots, T$: **do**
2: Let $q_t = q_{\Lambda_{t-1}^{\mathcal{A}}}$
3: Let $i_t = q_t(D)$
4: Update $\Lambda_t^{\mathcal{A}} = \Lambda_{t-1}^{\mathcal{A}} \cup i_t$
5: **end for**
6: **Return:** $\Lambda_T^{\mathcal{A}}$

Claim 1. *Let $P_{1,t}$ be the joint distribution induced by Algorithm **Bandit** on Λ_t at time t, and let $P_{2,t}$ be the joint distribution induced by Algorithm **InteractQuery** on $\Lambda_t = \Lambda_t^{\mathcal{A}}(D)$. Then $\forall t$ $P_{1,t} = P_{2,t}$.*

The upshot of this equivalence is that we can import existing results that hold in the setting in which the dataset is fixed, and queries are adaptively chosen. There are a large collection of results of this form that apply when the queries are differentially private (Dwork et al., 2015c; Bassily et al., 2016; Rogers et al., 2016) which apply directly to our setting. In the next section we formally define differential privacy in the simple stochastic and contextual bandit setting, and leave the description of the more general transfer theorems to Section 4.

2.4. Differential Privacy

We will be interested in algorithms that are differentially private. In the simple stochastic bandit setting, we will require differential privacy with respect to the rewards. In the contextual bandit setting, we will also require differential privacy with respect to the rewards, but *not necessarily* with respect to the contexts.

We now define the neighboring relation we need to define bandit differential privacy:

Definition 1. In the simple stochastic setting, two bandit tableau's D, D' are *reward neighbors* if they differ in at most a single row: i.e. if there exists an index ℓ such that for all $t \neq \ell$, $D_t = D'_t$.

In the contextual setting, two bandit tableau's $(D, C), (D', C')$ are *reward neighbors* if $C = C'$ and D and D' differ in at most a single row: i.e. if there exists an index ℓ such that for all $t \neq \ell$, $D_t = D'_t$.

Note that changing a *context* does not result in a neighboring tableau: this neighboring relation will correspond to privacy for the rewards, but not for the contexts.

Remark 1. Note that we could have equivalently defined reward neighbors to be tableaus that differ in only a single entry, rather than in an entire row. The distinction is unimportant in a bandit setting, because a bandit algorithm will be able to observe only a single entry in any particular row.

Definition 2. A bandit algorithm $\mathcal{A}$ is (ϵ, δ) reward differentially private if for every time horizon T and every pair of bandit tableau $\mathrm{Tab}, \mathrm{Tab}'$ that are reward neighbors, and every subset $S \subseteq [K]^T$:

$$\mathbb{P}\left[\mathbf{Interact}(T, \mathcal{A}, \mathrm{Tab}) \in S\right] \leq$$

$$e^{\epsilon}\mathbb{P}\left[\mathbf{Interact}(T, \mathcal{A}, \mathrm{Tab}') \in S\right] + \delta$$

If $\delta = 0$, we say that $\mathcal{A}$ is ϵ-differentially private.

3. Privacy Reduces Bias in Stochastic Bandit Problems

We begin by showing that differentially private algorithms that operate in the stochastic bandit setting compute empirical means for their arms that are nearly unbiased. Together

with known differentially private algorithms for stochastic bandit problems, the result is an algorithm that obtains a nearly optimal (worst-case) regret guarantee while also guaranteeing that the collected data is nearly unbiased. We could (and do) obtain these results by combining the reduction to answering adaptively selected queries given by Theorem 1 with the standard generalization theorems in adaptive data analysis (e.g. Corollary 2 in its most general form), but we first prove these de-biasing results from first principles to build intuition.

Theorem 1. *Let $\mathcal{A}$ be an (ϵ, δ)-differentially private algorithm in the stochastic bandit setting. Then, for all $i \in [K]$, and all t, we have:*

$$\left|\mathbb{E}\left[\hat{Y}_i^t - \mu_i\right]\right| \leq (e^\epsilon - 1 + T\delta)\mu_i$$

Remark 2. Note that since $\mu_i \in [0,1]$, and for $\epsilon \ll 1$, $e^\epsilon \approx 1+\epsilon$, this theorem bounds the bias by roughly $\epsilon + T\delta$. Often, we will have $\delta = 0$ and so the bias will be bounded by roughly ϵ.

Proof. First we fix some notation. Fix any time horizon T, and let $(f_t)_{t\in[T]}$ be the sequence of selection functions induced by algorithm $\mathcal{A}$. Let $\mathbb{1}_{\{f_t(\Lambda_t)=i\}}$ be the indicator for the event that arm i is pulled at time t. We can write the random variable representing the sample mean of arm i at time T as

$$\hat{Y}_i^T = \sum_{t=1}^{T} \frac{\mathbb{1}_{\{f_t(\Lambda_t)=i\}}}{\sum_{t'=1}^{T} \mathbb{1}_{\{f_{t'}(\Lambda_{t'})=i\}}} y_{it}$$

where we recall that $y_{i,t}$ is the random variable representing the reward for arm i at time t. Note that the numerator $(f_t(\Lambda_t) = i)$ is by definition independent of $y_{i,t}$, but the denominator $(\sum_{t'=1}^{T} \mathbb{1}_{\{f_{t'}(\Lambda_{t'})=i\}})$ is not, because for $t' > t$ $\Lambda_{t'}$ depends on $y_{i,t}$. It is this dependence that leads to bias in adaptive data gathering procedures, and that we must argue is mitigated by differential privacy.

We recall that the random variable N_i^T represents the number of times arm i is pulled through round T: $N_i^T = \sum_{t'=1}^{T} \mathbb{1}_{\{f_{t'}(\Lambda_{t'})=i\}}$. Using this notation, we write the sample mean of arm i at time T, as:

$$\hat{Y}_i^T = \sum_{t=1}^{T} \frac{\mathbb{1}_{\{f_t(\Lambda_t)=i\}}}{N_i^T} \cdot y_{it}$$

We can then calculate:

$$\begin{aligned}
\mathbb{E}[\hat{Y}_i^t] &= \sum_{t=1}^{T} \mathbb{E}[\frac{\mathbb{1}_{\{f_t(\Lambda_t)=i\}}}{N_i^T} y_{it}] \\
&= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{y_{it}\sim P_i}[y_{it} \cdot \mathop{\mathbb{E}}_{\mathcal{A}}[\frac{\mathbb{1}_{\{f_t(\Lambda_t)=i\}}}{N_i^T}|y_{it}]]
\end{aligned}$$

where the first equality follows by the linearity of expectation, and the second follows by the law of iterated expectation.

Our goal is to show that the conditioning in the inner expectation does not substantially change the value of the expectation. Specifically, we want to show that all t, and any value y_{it}, we have

$$\mathbb{E}[\frac{\mathbb{1}_{\{f_t(\Lambda_t)=i\}}}{N_i}|y_{it}] \geq e^{-\epsilon}\,\mathbb{E}[\frac{\mathbb{1}_{\{f_t(\Lambda_t)=i\}}}{N_i^T}] - \delta$$

If we can show this, then we will have

$$\begin{aligned}
\mathbb{E}[\hat{Y}_i^T] &\geq (e^{-\epsilon} \sum_{t=1}^{T} \mathbb{E}[\frac{\mathbb{1}_{\{f_t(\Lambda_t)=i\}}}{N_i^T}] - T\delta) \cdot \mu_i \\
&= (e^{-\epsilon}\,\mathbb{E}[\frac{N_i^T}{N_i^T}] - T\delta) \cdot \mu_i = (e^{-\epsilon} - T\delta) \cdot \mu_i
\end{aligned}$$

which is what we want (The reverse inequality is symmetric).

This is what we now show to complete the proof. Observe that for all t, i, the quantity $\frac{\mathbb{1}_{\{f_t(\Lambda_t)=i\}}}{N_i}$ can be derived as a post-processing of the sequence of choices $(f_1(\Lambda_1), \ldots, f_T(\Lambda_T))$, and is therefore differentially private in the observed reward sequence. Observe also that the quantity $\frac{\mathbb{1}_{\{f_t(\Lambda_t)=i\}}}{N_i^T}$ is bounded in $[0,1]$. Hence (by a lemma in the full version) for any pair of values y_{it}, y'_{it}, we have $\mathbb{E}[\frac{\mathbb{1}_{\{f_t(\Lambda_t)=i\}}}{N_i^T}|y_{it}] \geq e^{-\epsilon}\,\mathbb{E}[\frac{\mathbb{1}_{\{f_t(\Lambda_t)=i\}}}{N_i^T}|y'_{it}] - \delta$. All that remains is to observe that there must exist some value y'_{it} such that $\mathbb{E}[\frac{\mathbb{1}_{\{f_t(\Lambda_t)=i\}}}{N_i}|y'_{it}] \geq \mathbb{E}[\frac{\mathbb{1}_{\{f_t(\Lambda_t)=i\}}}{N_i}]$. (Otherwise, this would contradict $\mathbb{E}_{y'_{it}\sim P_i}[\mathbb{E}[\frac{\mathbb{1}_{\{f_t(\Lambda_t)=i\}}}{N_i}|y'_{it}]] = \mathbb{E}[\frac{\mathbb{1}_{\{f_t(\Lambda_t)=i\}}}{N_i^T}]$). Fixing any such y'_{it} implies that for all y_{it}

$$\begin{aligned}
\mathbb{E}[\frac{\mathbb{1}_{\{f_t(\Lambda_t)=i\}}}{N_i}|y_{it}] &\geq e^{-\epsilon}\,\mathbb{E}[\frac{\mathbb{1}_{\{f_t(\Lambda_t)=i\}}}{N_i^T}|y'_{i,t}] - \delta \\
&\geq e^{-\epsilon}\,\mathbb{E}[\frac{\mathbb{1}_{\{f_t(\Lambda_t)=i\}}}{N_i^T}] - \delta
\end{aligned}$$

as desired. The upper bound on the bias follows symmetrically. □

3.1. A Private UCB Algorithm

There are existing differentially private variants of the classic UCB algorithm ((Auer et al., 2002; Agrawal, 1995; Lai & Robbins, 1985)), which give a nearly optimal tradeoff between privacy and regret (Mishra & Thakurta, 2014; Tossou & Dimitrakakis, 2017; 2016). For completeness, we give a simple version of a private UCB algorithm in the full version which we use in our experiments. Here, we simply quote the relevant theorem, which is a consequence of a theorem in (Tossou & Dimitrakakis, 2016):

Theorem 2. *(Tossou & Dimitrakakis, 2016) There is an ϵ-differentially private algorithm that obtains expected regret bounded by:*

$$O\left(\max\left(\frac{\ln T}{\epsilon}\cdot(\ln\ln(T)+\ln(1/\epsilon)),\sqrt{kT\log T}\right)\right)$$

Thus, we can take ϵ to be as small as $\epsilon = O(\frac{\ln^{1.5} T}{\sqrt{kT}})$ while still having a regret bound of $O(\sqrt{kT\log T})$, which is nearly optimal in the worst case (over instances) (Audibert & Bubeck, 2009).

Combining the above bound with Theorem 1, and letting $\epsilon = O(\frac{\ln^{1.5} T}{\sqrt{kT}})$, we have:

Corollary 1. *There exists a simple stochastic bandit algorithm that simultaneously guarantees that the bias of the empirical average for each arm i is bounded by $O(\mu_i \cdot \frac{\ln^{1.5} T}{\sqrt{kT}})$ and guarantees expected regret bounded by $O(\sqrt{kT\log T})$.*

Of course, other tradeoffs are possible using different values of ϵ. For example, the algorithm of (Tossou & Dimitrakakis, 2016) obtains sub-linear regret so long as $\epsilon = \omega(\frac{\ln^2 T}{T})$. Thus, it is possible to obtain non-trivial regret while guaranteeing that the bias of the empirical means remains as low as $\mathrm{polylog}(T)/T$.

4. Max Information & Arbitrary Hypothesis Tests

Up through this point, we have focused our attention on showing how the private collection of data mitigates the effect that adaptivity has on *bias*, in both the stochastic and (in the full version) contextual bandit problems. In this section, we draw upon more powerful results from the adaptive data analysis literature to go substantially beyond bias: to correct the p-values of hypothesis tests applied to adaptively gathered data. These p-value corrections follow from the connection between differential privacy and a quantity called *max information*, which controls the extent to which the dependence of selected test on the dataset can distort the statistical validity of the test (Dwork et al., 2015b; Rogers et al., 2016). We briefly define max information, state the connection to differential privacy, and illustrate how max information bounds can be used to perform adaptive analyses in the private data gathering framework.

Definition 3 (Max-Information (Dwork et al., 2015b).)**.** Let X, Z be jointly distributed random variables over domain $(\mathcal{X}, \mathcal{Z})$. Let $X \otimes Z$ denote the random variable that draws independent copies of X, Z according to their marginal distributions. The β-approximate max-information between X, Z, denoted $I_\beta(X, Z)$, is defined as:

$$I_\beta(X,Z) = \log \sup_{\mathcal{O}\subset(\mathcal{X}\times\mathcal{Z}),\ \mathbb{P}[(X,Z)\in\mathcal{O}]>\beta} \frac{\mathbb{P}\left[(X,Z)\in\mathcal{O}\right]-\beta}{\mathbb{P}\left[X\otimes Z\in\mathcal{O}\right]}$$

Following (Rogers et al., 2016), define a test statistic $t : \mathcal{D} \to \mathbb{R}$, where $\mathcal{D}$ is the space of all datasets. For $D \in \mathcal{D}$, given an output $a = t(D)$, the p-value associated with the test t on dataset D is $p(a) = \mathbb{P}_{D\sim\mathbb{P}_0}\left[t(D) \geq a\right]$, where P_0 is the null hypothesis distribution. Consider an algorithm $\mathcal{A}$, mapping a dataset to a test statistic.

Definition 4 (Valid p-value Correction Function (Rogers et al., 2016).)**.** A function $\gamma : [0,1] \to [0,1]$ is a valid p-value correction function for $\mathcal{A}$ if the procedure:

1. Select a test statistic $t = \mathcal{A}(D)$
2. Reject the null hypothesis if $p(t(D)) \leq \gamma(\alpha)$

has probability at most α of rejection, when $D \sim P_0$.

Then the following theorem gives a valid p-value correction function when $(D, A(D))$ have bounded β-approximate max information.

Theorem 3 ((Rogers et al., 2016).)**.** *Let $\mathcal{A}$ be a data-dependent algorithm for selecting a test statistics such that $I_\beta(X, \mathcal{A}(X)) \leq k$. Then the following function γ is a valid p-value correction function for $\mathcal{A}$: $\gamma(\alpha) = \max(\frac{\alpha-\beta}{2^k}, 0)$*

Finally, we can connect max information to differential privacy, which allows us to leverage private algorithms to perform arbitrary valid statistical tests.

Theorem 4 (Theorem 20 from (Dwork et al., 2015b).)**.** *Let $\mathcal{A}$ be an ϵ-differentially private algorithm, let P be an arbitrary product distribution over datasets of size n, and let $D \sim P$. Then for every $\beta > 0$:*

$$I_\beta(D, \mathcal{A}(D)) \leq \log(e)(\epsilon^2 n/2 + \epsilon\sqrt{n\log(2/\beta)/2})$$

Remark 3. We note that a hypothesis of this theorem is that the data is drawn from a product distribution. In the contextual bandit setting, this corresponds to rows in the bandit tableau being drawn from a product distribution. This will be the case if contexts are drawn from a distribution at each round, and then rewards are generated as some fixed stochastic function of the contexts. Note that contexts (and even rewards) can be correlated with one another within a round, so long as they are selected independently across rounds.

We now formalize the process of running a hypothesis test against an adaptively collected dataset. A bandit algorithm $\mathcal{A}$ generates a history $\Lambda_T \in \mathcal{H}^T$. Let the reward portion of the gathered dataset be denoted by $D_\mathcal{A}$. We define an *adaptive test statistic selector* as follows.

Definition 5. Fix the reward portion of a bandit tableau D and bandit algorithm $\mathcal{A}$. An adaptive test statistic selector is a function s from action histories to test statistics such that $s(\Lambda_T^{\mathcal{A}})$ is a real-valued function of the adaptively gathered dataset $D_{\mathcal{A}}$.

Importantly, the selection of the test statistic $s(\Lambda_T^{\mathcal{A}})$ can depend on the sequence of arms pulled by $\mathcal{A}$ (and in the contextual setting, on all contexts observed), but not otherwise on the reward portion of the tableau D. For example, $t_{\mathcal{A}} = s(\Lambda_T^{\mathcal{A}})$ could be the t-statistic corresponding to the null hypothesis that the arm i^* which was pulled the greatest number of times has mean μ: $t_{\mathcal{A}}(D_{\mathcal{A}}) = \frac{\sum_{t=1}^{N_{i^*}^T} y_{i^* t} - \mu}{\sqrt{N_{i^*}^T}}$ By virtue of Theorems 3 and 4, and our view of adaptive data gathering as adaptively selected queries, we get the following corollary:

Corollary 2. *Let $\mathcal{A}$ be an ϵ reward differentially private bandit algorithm, and let s be an adaptive test statistic selector. Fix $\beta > 0$, and let $\gamma(\alpha) = \frac{\alpha}{2^{\log(e)(\epsilon^2 T/2 + \epsilon\sqrt{T\log(2/\beta)/2})}}$, for $\alpha \in [0,1]$. Then for any adaptively selected statistic $t_{\mathcal{A}} = s(\Lambda_T^{\mathcal{A}})$, and any product distribution P corresponding to the null hypothesis for $t_{\mathcal{A}}$*

$$\mathbb{P}_{D\sim P,\mathcal{A}}\left[p(t_{\mathcal{A}}(D)) \leq \gamma(\alpha)\right] \leq \alpha$$

If we set $\epsilon = O(1/\sqrt{T})$ in Corollary 2, then $\gamma(\alpha) = O(\alpha)$– i.e. a valid p-value correction that only scales α by a constant.

5. Experiments

We first validate our theoretical bounds on bias in the simple stochastic bandit setting. As expected the standard UCB algorithm underestimates the mean at each arm, while the private UCB algorithm of (**?**) obtains very low bias. While using the ϵ suggested by the theory effectively reduces bias and achieves near optimal asymptotic regret, the resulting private algorithm only achieves non-trivial regret for large T due to large constants and logarithmic factors in our bounds. This motivates a heuristic choice of ϵ that provides no theoretical guarantees on bias reduction, but leads to regret that is comparable to the non-private UCB algorithm. We find empirically that even with this large choice of ϵ we achieve an 8 fold reduction in bias relative to UCB. This is consistent with the observation that our guarantees hold in the worst-case, and suggests that there is room for improvement in our theoretical bounds — both improving constants in the worst-case bounds on bias and on regret, and for proving instance specific bounds. Finally, we show that in the linear contextual bandit setting collecting data adaptively with a linear UCB algorithm and then conducting t-tests for regression coefficients yields incorrect inference (absent a p-value correction). These findings confirm the necessity of our methods when drawing conclusions from adaptively gathered data.

5.1. Stochastic Multi-Armed Bandit

In our first stochastic bandit experiment we set $K = 20$ and $T = 500$. The K arm means are equally spaced between 0 and 1 with gap $\Delta = .05$, with $\mu_0 = 1$. We run UCB and ϵ-private UCB for T rounds with $\epsilon = .05$, and after each run compute the difference between the sample mean at each arm and the true mean. We repeat this process $10,000$ times, averaging to obtain high confidence estimates of the bias at each arm. The average absolute bias over all arms for private UCB was $.00176$, with the bias for every arm being statistically indistinguishable from 0 at 95% confidence (see Figure 1 for confidence intervals) while the average absolute bias (over arms) for UCB was $.0698$, or over 40 times higher. The most biased arm had a measured bias of roughly 0.14, and except for the top 4 arms, the bias of each arm was statistically significant. It is worth noting that private UCB achieves bias significantly lower than the $\epsilon = .05$ guaranteed by the theory, indicating that the theoretical bounds on bias obtained from differential privacy are conservative. Figures 1, 2 show the bias at each arm for private UCB vs. UCB, with 95% confidence intervals around the bias at each arm. Not only is the bias for private UCB an order of magnitude smaller on average, it does not exhibit the systemic negative bias evident in Figure 2.

Noting that the observed reduction in bias for $\epsilon = .05$ exceeded that guaranteed by the theory, we run a second experiment with $K = 5, T = 100000, \Delta = .05$, and $\epsilon = 400$, averaging results over 1000 iterations. Figure 5 shows that private UCB achieves sub-linear regret comparable with UCB. While $\epsilon = 400$ provides no meaningful theoretical guarantee, the average absolute bias at each arm mean obtained by the private algorithm was $.0015$ (statistically indistinguishable from 0 at 95% confidence for each arm), while the non-private UCB algorithm obtained average bias $.011$, 7.5 times larger. The bias reduction for the arm with the smallest mean (for which the bias is the worst with the non private algorithm) was by more than a factor of 10. Figures 3,4 show the bias at each arm for the private and non-private UCB algorithms together with 95% confidence intervals; again we observe a negative skew in the bias for UCB, consistent with the theory in (Nie et al., 2017).

5.2. Linear Contextual Bandits

Our second experiment confirms that adaptivity leads to bias in the linear contextual bandit setting in the context of hypothesis testing – and in particular can lead to false discovery in testing for non-zero regression coefficients. The set up is as follows: for $K = 5$ arms, we observe rewards

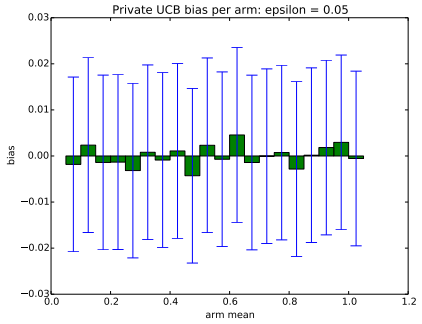

Figure 1: Private UCB Bias per Arm (experiment 1)

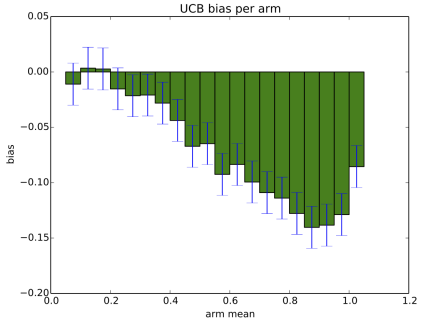

Figure 2: UCB Bias per Arm (experiment 1)

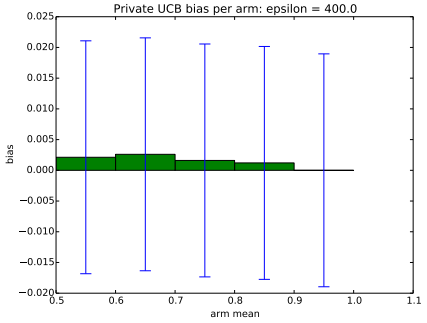

Figure 3: Private UCB Bias per Arm (experiment 2)

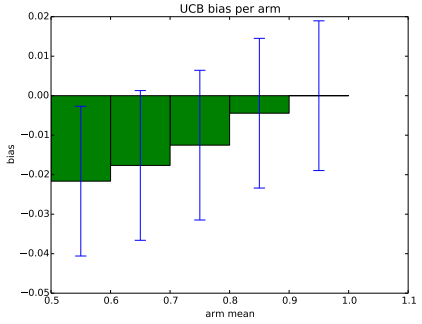

Figure 4: UCB Bias per Arm (experiment 2)

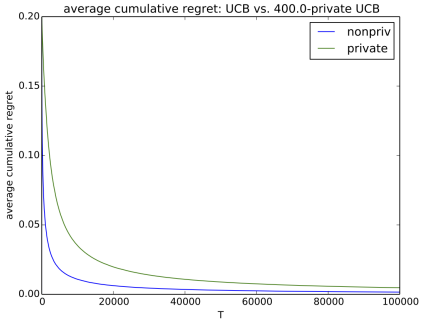

Figure 5: Average Regret: UCB vs. Private UCB

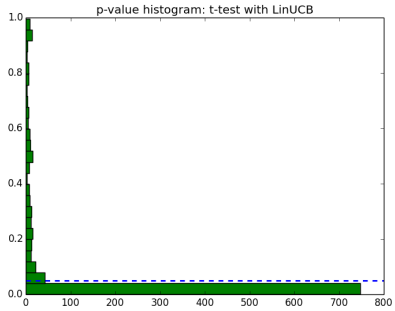

Figure 6: Histogram of p-values for z-test under the null hypothesis. $K = d = 5, T = 500$.

$y_{i,t} \sim \mathcal{N}(\theta_i' x_{it}, 1)$, where $\theta_i, x_{it} \in \mathbb{R}^5, ||\theta_i|| = ||x_{it}|| = 1$. For each arm i, we set $\theta_{i1} = 0$. Subject to these constraints, we pick the θ parameters uniformly at random (once per run), and select the contexts x uniformly at random (at each round). We run a linear UCB algorithm (OFUL (**?**)) for $T = 500$ rounds, and identify the arm i^* that has been selected most frequently. We then conduct a z-test for whether the first coordinate of θ_{i^*} is equal to 0. By construction the null hypothesis $H_0 : \theta_{i^*1} = 0$ of the experiment is true, and absent adaptivity, the p-value should be distributed uniformly at random. In particular, for any value of α the probability that the corresponding p-value is less than α is exactly α. We record the observed p-value, and repeat the experiment 1000 times, displaying the histogram of observed p-values in Figure 6. As expected, the adaptivity of the data gathering process leads the p-values to exhibit a strong downward skew. The dotted blue line demarcates $\alpha = .05$. Rather than probability .05 of falsely rejecting the null hypothesis at 95% confidence, we observe that 76% of the observed p-values fall below the .05 threshold. This shows that a careful p-value correction in the style of Section 2.3 is essential even for simple testing of regression coefficients, lest bias lead to false discovery.

References

Agrawal, R. Sample mean based index policies with o(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995. ISSN 00018678. URL http://www.jstor.org/stable/1427934.

Audibert, J.-Y. and Bubeck, S. Minimax policies for adversarial and stochastic bandits. In *COLT*, pp. 217–226, 2009.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, May 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

Bassily, R., Nissim, K., Smith, A., Steinke, T., Stemmer, U., and Ullman, J. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 1046–1059. ACM, 2016.

Bubeck, S., Cesa-Bianchi, N., et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Cummings, R., Ligett, K., Nissim, K., Roth, A., and Wu, Z. S. Adaptive learning with robust generalization guarantees. In *Conference on Learning Theory*, pp. 772–814, 2016.

Deshpande, Y., Mackey, L., Syrgkanis, V., and Taddy, M. Accurate inference for adaptive linear models. *arXiv preprint arXiv:1712.06695*, 2017.

Dimakopoulou, M., Athey, S., and Imbens, G. Estimation considerations in contextual bandits. *arXiv preprint arXiv:1711.07077*, 2017.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pp. 265–284, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-32731-2, 978-3-540-32731-8. doi: 10.1007/11681878_14. URL http://dx.doi.org/10.1007/11681878_14.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015a.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. Generalization in adaptive data analysis and holdout reuse. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pp. 2350–2358, Cambridge, MA, USA, 2015b. MIT Press. URL http://dl.acm.org/citation.cfm?id=2969442.2969502.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. L. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pp. 117–126, New York, NY, USA, 2015c. ACM. ISBN 978-1-4503-3536-2. doi: 10.1145/2746539.2746580. URL http://doi.acm.org/10.1145/2746539.2746580.

Feldman, V. and Steinke, T. Generalization for adaptively-chosen estimators via stable median. In *Conference on Learning Theory*, pp. 728–757, 2017a.

Feldman, V. and Steinke, T. Calibrating noise to variance in adaptive data analysis. *arXiv preprint arXiv:1712.07196*, 2017b.

Hardt, M. and Blum, A. The ladder: a reliable leaderboard for machine learning competitions. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pp. 1006–1014. JMLR. org, 2015.

Hardt, M. and Ullman, J. Preventing false discovery in interactive data analysis is hard. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pp. 454–463. IEEE, 2014.

Lai, T. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6(1):4–22, March 1985. ISSN 0196-8858. doi: 10.1016/0196-8858(85)90002-8. URL http://dx.doi.org/10.1016/0196-8858(85)90002-8.

Mishra, N. and Thakurta, A. Private stochastic multi-arm bandits: From theory to practice. In *ICML Workshop on Learning, Security, and Privacy*, 2014.

Neel, S. and Roth, A. Mitigating bias in adaptive data gathering via differential privacy. *arXiv preprint arXiv:1806.02329*, 2018.

Nie, X., Tian, X., Taylor, J., and Zou, J. Why adaptively collected data have negative bias and how to correct for it. *ArXiv e-prints*, August 2017.

Rogers, R. M., Roth, A., Smith, A. D., and Thakkar, O. Max-information, differential privacy, and post-selection hypothesis testing. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pp. 487–494, 2016. doi: 10.1109/FOCS.2016.59. URL https://doi.org/10.1109/FOCS.2016.59.

Russo, D. and Zou, J. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pp. 1232–1240, 2016.

Steinke, T. and Ullman, J. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Conference on Learning Theory*, pp. 1588–1628, 2015.

Tossou, A. C. Y. and Dimitrakakis, C. Algorithms for differentially private multi-armed bandits. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pp. 2087–2093. AAAI Press, 2016. URL http://dl.acm.org/citation.cfm?id=3016100.3016190.

Tossou, A. C. Y. and Dimitrakakis, C. Achieving privacy in the adversarial multi-armed bandit. *CoRR*, abs/1701.04222, 2017. URL http://arxiv.org/abs/1701.04222.

Villar, S. S., Bowden, J., and Wason, J. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

Wang, Y.-X., Lei, J., and Fienberg, S. E. A minimax theory for adaptive data analysis. *arXiv preprint arXiv:1602.04287*, 2016.