An Optimal Private Stochastic-MAB Algorithm Based on an Optimal Private Stopping Rule

Touqir Sajed [*1] **Or Sheffet** [*1]

Abstract

We present a provably optimal differentially private algorithm for the stochastic multi-arm bandit problem, as opposed to the private analogue of the UCB-algorithm (Mishra and Thakurta, 2015; Tossou and Dimitrakakis, 2016) which doesn't meet the recently discovered lower-bound of $\Omega\left(K\log(T)/\epsilon\right)$ (Shariff and Sheffet, 2018). Our construction is based on a different algorithm, Successive Elimination (Even-Dar et al., 2002), that repeatedly pulls all remaining arms until an arm is found to be suboptimal and is then eliminated. In order to devise a private analogue of Successive Elimination we visit the problem of private *stopping rule*, that takes as input a stream of i.i.d samples from an unknown distribution and returns a *multiplicative* $(1 \pm \alpha)$-approximation of the distribution's mean, and prove the optimality of our private stopping rule. We then present the private Successive Elimination algorithm which meets both the non-private lower bound (Lai and Robbins, 1985) and the above-mentioned private lower bound. We also compare empirically the performance of our algorithm with the private UCB algorithm.

1. Introduction

The well-known *stochastic multi-armed bandit* (MAB) is a sequential decision-making task in which a learner repeatedly chooses an action (or arm) and receives a noisy reward. The learner's objective is to maximize cumulative reward by *exploring* the actions to discover optimal ones (having the highest expected reward), balanced with *exploiting* them. The problem, originally stemming from experiments in medicine (Robbins, 1952), has applications in fields such as ranking (Kveton et al., 2015), recommendation systems (collaborative filtering) (Caron and Bhagat, 2013), investment portfolio design (Hoffman et al., 2011) and online advertising (Schwartz et al., 2017), to name a few. Such applications, relying on sensitive data, raise privacy concerns.

Differential privacy (Dwork et al., 2006) has become in recent years the gold-standard for privacy preserving data-analysis alleviating such concerns, as it requires that the output of the data-analysis algorithm has a limited dependency on any single datum. Differentially private variants of online learning algorithms have been successfully devised in various settings (Smith and Thakurta, 2013), including a private UCB-algorithm for the MAB problem (details below) (Mishra and Thakurta, 2015; Tossou and Dimitrakakis, 2016) as well as UCB variations in the linear (Kannan et al., 2018) and contextual (Shariff and Sheffet, 2018) settings.

More formally, in the MAB problem at every round t the learner selects an arm a out of K available arms, pulls it, and receives a random reward $r_{a,t}$ drawn i.i.d from a distribution $\mathcal{P}_a$ — of support $[0,1]$ and unknown mean μ_a. The Upper Confidence Bound (UCB) algorithm for the MAB problem was developed in a series of works (Berry and Fristedt, 1985; Agrawal, 1995) culminating into (Auer et al., 2002a), which is provably optimal for the MAB problem (Lai and Robbins, 1985). The UCB algorithm maintains a time-dependent high-probability upper-bound $B_{a,t}$ for each arm's mean, and at each timestep optimistically pulls the arm with the highest bound. The above-mentioned ε-differentially private (ε-DP) analogues of the UCB-algorithm follow the same procedure except for maintaining noisy estimations $\widetilde{B_{a,t}}$ using the "tree-based mechanism" (Chan et al., 2010; Dwork et al., 2010). This mechanism continuously releases aggregated statistics over a stream of T observations, introducing only $\mathrm{poly}\log(T)/\varepsilon$ noise in each timestep. The details of this poly-log factor are the focus of this work.

It was recently shown (Shariff and Sheffet, 2018) that

[*]Equal contribution [1]Department of Computing Science, University of Alberta, Edmonton AB, Canada. Correspondence to: Touqir Sajed <touqir@ualberta.ca>, Or Sheffet <osheffet@ualberta.ca>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

any ε-DP stochastic MAB algorithm[1] must incur an added pseudo regret of $\Omega(K\log(T)/\varepsilon)$. However, it is commonly known that any algorithm that relies on the tree-based mechanism must incur an added pseudo regret of $\omega(K\log(T)/\varepsilon)$. Indeed, the tree-based mechanism maintains a binary tree over the T streaming observations, a tree of depth $\log_2(T)$, where each node in this tree holds an i.i.d sample from a $\mathsf{Lap}(\frac{\log_2(T)}{\varepsilon})$ distribution. At each timestep t, the mechanism outputs the sum of the first t observations added to the sum of the $\log_2(T)$ nodes on the root-to-tth-leaf path in the binary tree. As a result, the variance of the added noise at *each* timestep is $\Theta(\frac{\log^3(T)}{\varepsilon^2})$, making the noise per timestep $\omega(\log(T)/\varepsilon)$. (In fact, most analyses[2][3] of the tree-based mechanism rely on the union bound over all T timesteps, obtaining a bound of $\log^{5/2}(T)/\varepsilon$; consequentially the added-regret bound of the DP-UCB algorithm is $O(\frac{K\log^{2.5}(T)}{\varepsilon})$.) Thus, in a setting where each of the K tree-mechanisms (one per arm) is run over $\mathrm{poly}(T)$ observations (say, if all arms have suboptimality gap of $T^{-0.1}$), the private UCB-algorithm must unavoidably obtain an added regret of $\omega(K\log(T)/\varepsilon)$ (on top of the regret of the UCB-algorithm). It is therefore clear that the challenge in devising an *optimal* DP algorithm for the MAB problem, namely an algorithm with added regret of $O(K\log(T)/\varepsilon)$, is *algorithmic* in nature — we must replace the tree-based mechanism with a different, simpler, mechanism.

Our Contribution and Organization. In this work, we present an optimal algorithm for the stochastic MAB-problem, which meets both the non-private lower-bound of (Lai and Robbins, 1985) and the private lower-bound of (Shariff and Sheffet, 2018). Our algorithm is a DP variant of the Successive Elimination (SE) algorithm (Even-Dar et al., 2002), a different optimal algorithm for stochastic MAB. SE works by pulling all arms sequentially, maintaining the same confidence interval around the empirical average of each arm's reward (as all remaining arms are pulled the exact same number of times); and when an arm is found to be noticeably suboptimal in comparison to a different arm, it is then eliminated from the set of viable arms (all arms are viable initially). To design a DP-analogue of SE we first consider the case of 2 arms and ask ourselves — what is the optimal way to privately discern whether the gap between the mean rewards of two arms is positive or negative? This motivates the study of private *stopping rules* which take as input a stream of i.i.d observations from a distribution of support $[-R, R]$ and unknown mean μ, and halt once they obtain a $(1\pm\alpha)$-approximation of μ with confidence of at least $1-\beta$. Note that due to the multiplicative nature of the required approximation, it is impossible to straight-forwardly use the Hoeffding or Bernstein bounds; rather a stopping rule must alter its halting condition with time. (Domingo et al., 2002) proposed a stopping rule known as the Nonmonotonic Adaptive Sampling (NAS) algorithm that relies on the Hoeffding's inequality to maintain a confidence interval at each timestep. They showed a sample complexity bound of $O\left(\frac{R^2}{\alpha^2\mu^2}\left(\log(\frac{R}{\beta\cdot\alpha|\mu|})\right)\right)$, later improved slightly by (Mnih et al., 2008) to $O\left(\frac{R^2}{\alpha^2\mu^2}\left(\log(\frac{1}{\beta})+\log\log(\frac{R}{\alpha|\mu|})\right)\right)$. The work of (Dagum et al., 2000) shows an essentially matching sample complexity lower-bound. Stopping Rules have also been applied to Reinforcement Learning and Racing algorithms (See Sajed et al. (2018); Mnih et al. (2008)).

In this work we introduce a ε-DP analogue of the NAS algorithm that is based on the *sparse vector technique* (SVT), with added sample complexity of (roughly) $O(\frac{R\log(1/\beta)}{\varepsilon\alpha|\mu|})$. Moreover, we show that this added sample complexity is optimal in the sense that any ε-DP stopping rule has a matching sample complexity lower-bound. After we introduce preliminaries in Section 2, we present the private NAS in Section 3. We then turn our attention to the design of the private SE algorithm. Note that straight-forwardly applying K private stopping rules yields a suboptimal algorithm whose regret bound is proportional to K^2. Instead, we *partition* the algorithm's arm-pulls into epochs, where epoch e is set to eliminate all arms with suboptimality-gaps greater than 2^{-e}. By design each epoch must be at least twice as long as the previous epoch, and so we can reset (compute empirical means from fresh reward samples) the algorithm in-between epochs while incurring only a constant-factor increase to the regret bound. Details appear in Section 4. We also assess the empirical performance of our algorithm in comparison to the DP-UCB baseline and show that the improvement in analysis (despite the use of large constants) is empirically evident; details provided in Section 5. Lastly, future directions for this work are discussed in Section 6.

Discussion. Some may find the results of this work underwhelming — after all the improvement we put forth is solely over $\mathrm{poly}\log$-factors, and admittedly they are already subsumed by the non-private regret bound of the algorithm under many "natural" settings of parameters. Our reply to these is two-fold. First, our experiments (see Section 5) show a significantly improved performance empirically, which is only due to the different algorithmic approach. Second, as the designers of privacy-preserving learning algorithms it is our "moral duty" to quantify the *added* cost of privacy on top of the already existing cost, and push this added cost to its absolute lowest.

[1] In this work, we focus on pure ε-DP, rather than (ε,δ)-DP.

[2] (Tossou and Dimitrakakis, 2016) claim a $O(\log(T)/\varepsilon)$ bound, but (i) rely on (ε,δ)-DP rather than pure-DP and more importantly (ii) "sweep under the rug" several factors that are themselves on the order of $\log(T)$.

[3] (Mishra and Thakurta, 2015) shows a bound of $O(\log^3(T)/\varepsilon)$

We would also like to emphasize a more philosophical point arising from this work. Both the UCB-algorithm and the SE-algorithm are provably optimal for the MAB problem in the non-private setting, and are therefore equivalent. But the UCB-algorithm makes in each timestep an input-dependent choice (which arm to pull); whereas the SE-algorithm input-dependent choices are reflected only in $K-1$ special timesteps in which it declares "eliminate arm a" (in any other timestep it chooses the next viable arm). In that sense, the SE-algorithm is *simpler* than the UCB-algorithm, making it the less costly to privatize between the two. In other words, differential privacy gives quantitative reasoning for preferring one algorithm to another because "simpler is better." While not a full-fledged theory (yet), we believe this narrative is of importance to anyone who designs differentially private data-analysis algorithms.

2. Preliminaries

Stopping Rules. In the *stopping rule* problem, the input consists of a stream of i.i.d samples $\{X_t\}_{t\geq 1}$ drawn from a distribution over an a-priori known support $[-R,R]$ and with unknown mean μ. Given $\alpha,\beta\in(0,1)$, the goal of the stopping rule is to halt after seeing as few samples as possible while releasing a $(1\pm\alpha)$-approximation of μ at halting time. Namely, a (α,β)-*stopping rule* halts at some time t^* and releases $\hat{\mu}$ such that $\Pr[|\hat{\mu}-\mu|\geq\alpha|\mu|]\leq\beta$. (It should be clear that the halting time t^* increases as $|\mu|$ decreases.) During any timestep t, we denote $X_{1:t}\stackrel{\text{def}}{=}\sum_{i=1}^{t}X_i$ and $\overline{X_t}\stackrel{\text{def}}{=}X_{1:t}/t$.

Stochastic MAB and its optimal bounds. The formal description of the stochastic MAB problem was provided in the introduction. Formally, the bound maintained by the UCB-algorithm for each arm a at a given round t is $B_{a,t}\stackrel{\text{def}}{=}\overline{\mu}_a+\sqrt{2\log(t)/t_a}$ with $\overline{\mu}_a$ denoting the empirical mean reward from pulling arm a and t_a denoting the number of times a was pulled thus far. We use a^* to denote the leading arm, namely, an arm of highest mean reward: $\mu_{a^*}=\max_{a=1}^{K}\{\mu_a\}$. Given any arm a we denote the mean-gap as $\Delta_a\stackrel{\text{def}}{=}\mu_{a^*}-\mu_a$, with $\Delta_{a^*}=0$ by definition. Additionally we denote the horizon with T - the number of rounds that a MAB algorithm will be run for. An algorithm that chooses arm a_t at timestep t incurs an *expected regret* or *pseudo-regret* of $\sum_t\Delta_{a_t}$. It is well-known (Lai and Robbins, 1985) that any consistent[4] regret-minimization algorithm must incur a pseudo-regret of $\Omega(\sum_{a\neq a^*}\frac{\log(T)}{\Delta_a})$; and indeed the UCB-algorithm meets this bound and has pseudo-regret of $O(\sum_{a\neq a^*}\frac{\log(T)}{\Delta_a})$. However, the minimax regret bound of the UCB-algorithm is $O(\sqrt{KT\log(T)})$, obtained by an adversary that knows T and sets all suboptimal arms' gaps to $\sqrt{K\log(T)/T}$, whereas the minimax lower-bound of any algorithm is slightly smaller: $\Omega(\sqrt{KT})$ (Auer et al., 2002a).

[4] A regret minimization algorithm is called consistent if its regret is sub-polynomial, namely in $o(n^p)$ for any $p>0$.

Differential Privacy. In this work, we preserve *event-level* privacy under continuous observation (Dwork et al., 2010). Formally, we say two streams are neighbours if they differ on a single entry in a single timestep t, and are identical on any other timestep. An algorithm $\mathcal{M}$ is ε-differentially private if for any two neighboring streams S and S' and for any set $\mathcal{O}$ of decisions made from timestep 1 through T, it holds that $\Pr[\mathcal{M}(S)\in\mathcal{O}]\leq e^{\varepsilon}\Pr[\mathcal{M}(S')\in\mathcal{O}]$. Note that much like its input, the output $\mathcal{M}(S)$ is also released in a stream-like fashion, and the requirement should hold for all decisions made by $\mathcal{M}$ in all timesteps.

In this work, we use two mechanisms that are known to be ε-DP. The first is the Laplace mechanism (Dwork et al., 2006). Given a function f that takes as input a stream S and releases an output in $\mathbb{R}^d$, we denote its global sensitivity as $GS(f)=\max_{S,S'}||f(S)-f(S')||_1$; and the Laplace mechanism adds a random (independent) sample from $\mathsf{Lap}(GS(f)/\varepsilon)$ to each coordinate of $f(S)$. The other mechanism we use is the *sparse-vector technique* (SVT), that takes in addition to S a sequence of queries $\{q_i\}_i$ (each query has a global sensitivity $\leq GS$), and halts with the very first query whose value exceeds a given threshold. The SVT works by adding a random noise sampled i.i.d from $\mathsf{Lap}(3GS/\varepsilon)$ to both to the threshold and to each of the query-values. See (Dwork et al., 2014) for more details.

Concentration bounds. A Laplace r.v. $X\sim\mathsf{Lap}(\lambda)$ is sampled from a distribution with $\mathsf{PDF}(x)\propto e^{-|x|/\lambda}$. It is known that $\mathbb{E}[X]=0$, $\mathrm{Var}[X]=2\lambda^2$ and that for any $\tau>0$ it holds that $\Pr[|X|>\tau]=e^{-\tau/\lambda}$.

Throughout this work we also rely on the Hoeffding inequality (Hoeffding, 1963). Given a collection $\{X_t\}_{t=1}^{T}$ of i.i.d random variables that take value in a finite interval of length R with mean μ, it holds that $\Pr\left[|\overline{X_t}-\mu|\geq\alpha\right]\leq 2\exp\left(-2\alpha^2T/R^2\right)$.

Miscellaneous. We emphasize we made no effort to minimize constants throughout this work. We use $\log(x)$ to denote the base-e logarithm of x. Given two distributions $\mathcal{P}$ and $\mathcal{Q}$, we denote their *total-variation* distance as $d_{\mathrm{TV}}(\mathcal{P},\mathcal{Q})=\sup_S\left(\left|\Pr_{X\sim\mathcal{P}}[X\in S]-\Pr_{X\sim\mathcal{Q}}[X\in S]\right|\right)$. We also rely on the following fact (proof omitted).

Fact 2.1. *Fix any $a>1$ and $0<b<1/16$. Then for any $e<x<\frac{\log(a\log(1/b))}{b}$ it holds that $\frac{\log(a\log(x))}{x}>b$, and for any $x>\frac{2\log(a\log(1/b))}{b}$ it holds that $\frac{\log(a\log(x))}{x}<b$.*

3. Differentially Private Stopping Rule

In this section, we derive a differentially private stopping rule algorithm, DP-NAS, which is based on the non-private NAS (Nonmonotonic Adaptive Sampling). The non-private NAS is rather simple. Given β, denote h_t as confidence interval derived by the Hoeffding bound with confidence $1 - \beta/2t^2$ for t iid random samples bounded in magnitude by R; thus, w.p. $\geq 1-\beta$ it holds that $\forall t, |\overline{X_t} - \mu| \leq h_t$. The NAS algorithm halts at the first t for which $|\overline{X_t}| \geq h_t\left(\frac{1}{\alpha}+1\right)$. Indeed, such a stopping rule assures that $|\overline{X_t} - \mu| \leq h_t \leq \alpha(|\overline{X_t}| - h_t) \leq \alpha|\mu|$, the last inequality follows from $\left||\overline{X_t}| - |\mu|\right| \leq |\overline{X_t} - \mu| \leq h_t$.

In order to make NAS differentially private we use the sparse vector technique, since the algorithm is basically asking a series of threshold queries: $q_t \stackrel{\text{def}}{=} |\overline{X_t}| - h_t\left(\frac{1}{\alpha}+1\right) \stackrel{?}{\geq} 0$. Recall that the sparse-vector technique adds random noise both to the threshold and to the answer of each query, and so we must adjust the naïve threshold of 0 to some c_t in order to make sure that $\overline{X_t}$ is sufficiently close to μ when the threshold is crossed. Lastly, since our goal is to provide a private approximation of the distribution mean, we also apply the Laplace mechanism to $\overline{X_t}$ to assert the output is differentially private. Details appear in Algorithm 1.

Algorithm 1 DP-NAS

Set $\sigma_1 \leftarrow 12R/\varepsilon$, $\sigma_2 \leftarrow 12R/\varepsilon$, $\sigma_3 \leftarrow 4R/\varepsilon$.
Sample $B \sim \mathsf{Lap}(\sigma_1)$.
Initialize $t \leftarrow 0$.
repeat
  $t \leftarrow t+1$
  $A_t \sim \mathsf{Lap}(\sigma_2)$
  Get a new sample X_t and update the mean $\overline{X_t}$.
  $h_t \leftarrow R\sqrt{\frac{2}{t}\log(\frac{16t^2}{\beta})}$
  $c_t \leftarrow \sigma_1 \log(4/\beta) + \sigma_2 \log(8t^2/\beta) + \frac{\sigma_3}{\alpha}\log(4/\beta)$
until $|\overline{X_t}| \geq h_t(1+\frac{1}{\alpha}) + \frac{c_t+B+A_t}{t}$
Sample $L \sim \mathsf{Lap}(\sigma_3)$.
return $\overline{X_t} + \frac{L}{t}$

Theorem 3.1. *Algorithm 1 is a ε-DP (α,β)-stopping rule.*

Proof. First, we argue that Algorithm 1 is ε-differentially private. This follows immediately from the fact that the algorithm is a combination of the sparse-vector technique with the Laplace mechanism. The first part of the algorithm halts when $|\sum_{i=1}^{t} X_i| - h_t \cdot t(\frac{1}{\alpha}+1) - c_t \geq A_t + B$. Indeed, this is the sparse-vector mechanism for a sum-query of sensitivity of no more than $2R$. It follows that sampling both the threshold-noise B and the query noise A_t from $\mathsf{Lap}(3\cdot\frac{2}{\varepsilon}\cdot 2R)$ suffices to maintain $\frac{\varepsilon}{2}$-DP. Similarly, adding a sample from $\mathsf{Lap}(\frac{2}{t\varepsilon}\cdot 2R)$ suffices to release the mean with $\frac{\varepsilon}{2}$-DP at the very last step of the algorithm.

Since $\sum_{t\geq 1} t^{-2} \leq 2$, under the assumption that all $\{X_t\}$ are i.i.d samples from a distribution of mean μ, the Hoeffding-bound and union-bound give that $\Pr[\exists t,\ |\overline{X_t} - \mu| > h_t] \leq \beta/4$. Standard concentration bound on the Laplace distribution give that $\Pr[|B| > \sigma_1 \log(4/\beta)] \leq \beta/4$, $\Pr[\exists t,\ |A_t| > \sigma_2 \log(8t^2/\beta)] \leq \beta/4$, and $\Pr[|L| > \sigma_3 \log(4/\beta)] \leq \beta/4$. It follows that w.p. $\geq 1-\beta$ none of these events happen, and so $\forall t,\ c_t \geq |B| + |A_t| + |L|/\alpha$.

It follows that at the time we halt we have that

$$|\overline{X_t} - \mu| \stackrel{\text{Hoeffding}}{\leq} h_t \leq \alpha(|\overline{X_t}| - h_t) - \frac{\alpha}{t}(c_t + A_t + B)$$
$$\stackrel{(*)}{\leq} \alpha|\mu| - \frac{\alpha}{t}(c_t + A_t + B) \leq \alpha|\mu| - \frac{|L|}{t}$$

where $(*)$ is due to $\left||\overline{X_t}| - |\mu|\right| \leq |\overline{X_t} - \mu| \leq h_t$. Therefore, we have that $|\overline{X_t} + \frac{L}{t} - \mu| \leq |\overline{X_t} - \mu| + \frac{|L|}{t} \leq \alpha|\mu|$. □

Rather than analyzing the utility of Algorithm 1, namely, the high-probability bounds on its stopping time, we now turn our attention to a slight modification of the algorithm and analyze the revised algorithm's utility. The modification we introduce, albeit technical and non-instrumental in the utility bounds, plays a conceptual role in the description of later algorithms. We introduce Algorithm 2 where we exponentially reduce the number of SVT queries using standard doubling technique. Instead of querying the magnitude of the average at each timestep, we query it at exponentially growing intervals, thus paying no more than a constant factor in the utility guarantees while still reducing the number of SVT queries dramatically.

Algorithm 2 DP exponential NAS

Set $\sigma_1 \leftarrow 12R/\varepsilon$, $\sigma_2 \leftarrow 12R/\varepsilon$, $\sigma_3 \leftarrow 4R/\varepsilon$.
Sample $B \sim \mathsf{Lap}(\sigma_1)$
Initialize $k \leftarrow 0$ and $t \leftarrow 0$.
repeat
  $k \leftarrow k+1$
  repeat
    $t \leftarrow t+1$
    Sample X_t and update $\overline{X_t}$.
  until $t = 2^k$
  $A_t \sim \mathsf{Lap}(\sigma_2)$
  $c_t \leftarrow \sigma_1 \log(4/\beta) + \sigma_2 \log(8k^2/\beta) + \frac{\sigma_3}{\alpha}\log(4/\beta)$
  $h_t \leftarrow R\sqrt{\frac{2}{t}\log(\frac{16k^2}{\beta})}$
until $|\overline{X_t}| \geq h_t(1+\frac{1}{\alpha}) + \frac{c_t+B+A_t}{t}$
$L \sim \mathsf{Lap}(\sigma_3)$
return $\overline{X_t} + \frac{L}{t}$

Corollary 3.2. *Algorithm 2 is a ε-DP (α,β)-stopping rule.*

Proof. The only difference between Algorithms 1 and 2 lies in checking the halting condition at exponentially increasing time-intervals, namely during times $t = 2^k$ for $k \in \mathbb{N}$. The privacy analysis remains the same as in the proof of Theorem 3.1, and the algorithm correctness analysis is modified by considering only the timesteps during which we checking for the halting condition. Formally, we denote $\mathcal{E}$ as the event where (i) $\forall k,\ |\overline{X_{2^k}} - \mu| \le h_{2^k}$, (ii) $|B| \le \sigma_1 \log(4/\beta)$, (iii) $\forall k,\ |A_{2^k}| \le \sigma_2 \log(8k^2/\beta)$, and (iv) $|L| \le \sigma_3 \log(4/\beta)$. Analogous to the proof of Theorem 3.1 we bound $\Pr[\mathcal{E}] \ge 1-\beta$ and the result follows. □

Theorem 3.3. *Fix $\beta \le 0.08$ and $\mu \neq 0$. Let $\{X_t\}_t$ be an ensemble of i.i.d samples from any distribution over the range $[-R,R]$ and with mean μ. Denote $t_0 \stackrel{\text{def}}{=} \frac{R^2 \log((1/\beta)\cdot\log(\frac{R}{\alpha|\mu|}))}{\alpha^2\mu^2}$, $t_1 \stackrel{\text{def}}{=} \frac{R\log((1/\beta)\cdot\log(\frac{R}{\alpha|\mu|}))}{\varepsilon|\mu|}$, $t_2 \stackrel{\text{def}}{=} \frac{R\log(1/\beta)}{\varepsilon\alpha|\mu|}$. Then with probability at least $1-\beta$, Algorithm 2 halts by timestep $t_U = 2000(t_0+t_1+t_2)$.*

Proof. Recall the event $\mathcal{E}$ from the proof of Corollary 3.2 and its four conditions. We assume $\mathcal{E}$ holds and so the algorithm releases a $(1\pm\alpha)$-approximation of μ. To prove the claim, we show that under $\mathcal{E}$, at time t_U it must hold that $|\overline{X_t}| \ge h_t(1+\frac{1}{\alpha}) + \frac{c_t+B+A_t}{t}$.

Under $\mathcal{E}$ we have that $|\overline{X_t}| \ge |\mu| - h_t$ and $\frac{c_t+B+A_t}{t} \le \frac{2\sigma_1}{t}\log(4/\beta) + \frac{2\sigma_2}{t}\log(8k^2/\beta) + \frac{\sigma_3}{\alpha t}\log(4/\beta)$; and so it suffices to show that $|\mu| \ge h_t(2+\frac{1}{\alpha}) + \frac{24R\log(4/\beta)}{\varepsilon t} + \frac{24R\log(8k^2/\beta)}{\varepsilon t} + \frac{4R\log(4/\beta)}{\alpha\varepsilon t}$. In fact, since $\alpha < 1$ we show something slightly stronger: that at time t_U we have $|\mu| \ge \frac{3h_t}{\alpha} + \frac{48R\log(8k^2/\beta)}{\varepsilon t} + \frac{4R\log(4/\beta)}{\alpha\varepsilon t}$. This however is an immediate corollary of the following three facts.

1. For any $t \ge 1000t_0$ we have $\frac{\log(4\log_2(t)/\beta)}{t} \le \left(\frac{\alpha|\mu|}{2\cdot3\cdot3\cdot R}\right)^2$, implying $\frac{|\mu|}{3} \ge \frac{3h_t}{\alpha}$.
2. For any $t \ge 1000t_1$ we have $\frac{\log(4\log_2(t)/\beta)}{t} \le \frac{\varepsilon|\mu|}{3\cdot2\cdot48\cdot R}$, implying $\frac{|\mu|}{3} \ge \frac{2\cdot48R\log(4k/\beta)}{\varepsilon t} \ge \frac{48R\log(8k^2/\beta)}{\varepsilon t}$.
3. For any $t \ge 48t_2$ we have $\frac{|\mu|}{3} \ge \frac{4R\log(4/\beta)}{\alpha\varepsilon t}$.

where the first two rely on Fact 2.1. It follows that at time $1000(t_0+t_1+t_2)$ all three conditions hold. Therefore by time $t_u = 2000(t_0+t_1+t_2)$ Algorithm 2 reaches some t which is a power of 2, on which it poses a query for the SVT mechanism and halts. □

3.1. Private Stopping Rule Lower bounds

We turn our attention to proving the (near) optimality of Algorithm 2. A non-private lower bound was proven in (Dagum et al., 2000), who showed that no stopping rule algorithm can achieve a sample complexity better than $\Omega\left(\frac{\max\{\sigma^2, R\alpha|\mu|\}}{\alpha^2\mu^2}\log(1/\beta)\right)$ (with σ^2 denoting the variance of the underlying distribution). In this section, we prove the following lower bound on the additional sample complexity that *any* ε-DP stopping rule algorithm must incur.

Theorem 3.4. *Any ε-differentially private (α,β)-stopping rule whose input consists of a steam of i.i.d samples from a distribution over support $[-R,R]$ and with mean $\mu \neq 0$, must have an additional sample complexity of $\Omega\left(R\log(1/\beta)/\varepsilon\alpha|\mu|\right)$.*

Proof. Fix $\varepsilon,\alpha,\beta > 0$ such that $\alpha < 1$ and $\beta < 1/4$, and fix R and $\mu > 0$. We define two distributions $\mathcal{P},\mathcal{Q}$ over a support consisting of two discrete points: $\{-R,R\}$. Setting $\Pr_{\mathcal{P}}[R] = \frac{1}{2} + \frac{\mu}{2R}$ we have that $\mathbb{E}_{X\sim\mathcal{P}}[X] = \mu$. Set μ' as any number infinitesimally below the threshold of $\frac{1-\alpha}{1+\alpha}\mu$, so that we have $(1+\alpha)\mu' < (1-\alpha)\mu$; we set the parameters of $\mathcal{Q}$ s.t. $\Pr_{\mathcal{Q}}[R] = \frac{1}{2} + \frac{\mu'}{2R}$ so $\mathbb{E}_{X\sim\mathcal{Q}}[X] = \mu'$. By definition, the total variation distance $\mathrm{d_{TV}}(\mathcal{P},\mathcal{Q}) = \frac{\mu-\mu'}{2R} = \frac{2\alpha\mu}{2R(1+\alpha)} < \frac{\alpha\mu}{R}$.

Let $\mathcal{M}$ be any ε-differentially private (α,β)-stopping rule. Denote $n = \frac{R\log(1/\beta)}{12\alpha\mu\varepsilon}$. Let $\mathcal{E}$ be the event "after seeing at most n samples, $\mathcal{M}$ halts and outputs a number in the interval $\big[(1-\alpha)\mu,(1+\alpha)\mu\big]$." We now apply the following, very elegant, lemma from (Karwa and Vadhan, 2018), stating that the group privacy loss of a differentially privacy mechanism taking as input n i.i.d samples either from a distributions $\mathcal{D}$ or from a distribution $\mathcal{D}'$ scales effectively as $O(\varepsilon n\cdot\mathrm{d_{TV}}(\mathcal{D},\mathcal{D}'))$.

Lemma 3.5 (Lemma 6.1 from (Karwa and Vadhan, 2018))**.** *Let $\mathcal{M}$ be any ε-differentially private mechanism, fix a natural n and fix two distributions $\mathcal{D}$ and $\mathcal{D}'$, and let $\bar{S}$ and $\bar{S}'$ denote an ensemble of n i.i.d samples taken from $\mathcal{D}$ and $\mathcal{D}'$ resp. Then for any possible set of outputs O it holds that $\Pr[\mathcal{M}(S)\in O] \le e^{6\varepsilon n\cdot\mathrm{d_{TV}}(\mathcal{D},\mathcal{D}')}\Pr[\mathcal{M}(S')\in O]$.*

And so, applying $\mathcal{M}$ over n i.i.d samples taken from $\mathcal{Q}$, we must have that $\Pr_{\mathcal{M},S\sim\mathcal{Q}^n}[\mathcal{E}] \le \beta$, since $(1-\alpha)\mu > (1+\alpha)\mu'$. Applying Lemma 3.5 to our setting, we get

$$\begin{aligned}\Pr_{\mathcal{M},S\sim\mathcal{P}^n}[\mathcal{E}] &\le e^{6\varepsilon n\cdot\mathrm{d_{TV}}(\mathcal{P},\mathcal{Q})}\Pr_{\mathcal{M},S\sim\mathcal{Q}^n}[\mathcal{E}]\\ &\le \beta\cdot\exp(6\varepsilon n\cdot\frac{\alpha\mu}{R})\\ &= \beta\cdot\exp(\frac{6\varepsilon\alpha\mu}{R}\cdot\frac{R\log(1/\beta)}{12\varepsilon\alpha\mu}) = \frac{\beta}{\sqrt{\beta}} < \frac{1}{2}\end{aligned}$$

since $\beta < 1/4$. Since, by definition, we have that the probability of the event $\mathcal{E}'$ "after seeing at most n samples, $\mathcal{M}$ halts and outputs a number *outside* the interval $\big[(1-\alpha)\mu,(1+\alpha)\mu\big]$" over n i.i.d samples from $\mathcal{P}$ is at

most β, then it must be that $\mathcal{M}$ halts after seeing strictly more than n samples w.p. $> 1 - (1/2 + \beta) > 1/4$. □

Combining the non-private lower bound of (Dagum et al., 2000) and the bound of Theorem 3.4, we immediately infer the overall sample complexity bound, which follows from the fact that the variance of the distribution $\mathcal{P}$ used in the proof of Theorem 3.4 has variance of $\Theta(R^2)$.

Corollary 3.6. *There exists a distribution $\mathcal{P}$ for which any ε-differentially private (α, β)-stopping rule algorithm has a sample complexity of* $\Omega\left(\frac{R^2 \log(1/\beta)}{\alpha^2\mu^2} + \frac{R\log(1/\beta)}{\varepsilon\alpha|\mu|}\right)$.

Discussion. How optimal is Algorithm 2? The sample complexity bound in Theorem 3.3 can be interpreted as the sum of the non-private and private parts. The non-private part is $\Omega\left(\frac{R^2}{\alpha^2\mu^2}\left(\log(1/\beta) + \log\log\frac{R}{\alpha|\mu|}\right)\right)$ and the private part is $\Omega\left(\frac{R}{\varepsilon|\mu|}\left(\log(1/\beta) + \log\log\frac{R}{\alpha|\mu|}\right) + \frac{R\log(1/\beta)}{\varepsilon\alpha|\mu|}\right)$. If we add in the assumption that $\log(\frac{R}{\alpha|\mu|}) \leq 1/\beta$ we get that the upper-bound of Theorem 3.3 matches the lower-bound in Corollary 3.6.

How benign is this assumption? Much like in (Mnih et al., 2008), we too believe it is a very mild assumption. Specifically, in the next section, where we deal with finite sequences of length T, we set β as proportional to $1/T$. Since over finite-length sequence we can only retrieve an approximation of μ if $\frac{|\mu|}{R} \gg \frac{1}{T}$, requiring $\frac{R}{|\mu|} < 2^T$ is trivial. However, we cannot completely disregard the possibility of using a private stopping rule in a setting where, for example, both α, β are constants whereas $\frac{|\mu|}{R}$ is a sub-constant. In such a setting, $\log(\frac{R}{\alpha|\mu|})$ may dominate $1/\beta$, and there it might be possible to improve on the performance of Algorithm 2 (or tighten the bound).

4. An Optimal Private MAB Algorithm

In this section, our goal is to devise an optimal ε-differentially private algorithm for the stochastic K-arms bandit problem, in a setting where all rewards are between $[0, 1]$. We denote the mean reward of each arm as μ_a, the best arm as a^*, and for any $a \neq a^*$ we refer to the gap $\Delta_a = \mu_{a^*} - \mu_a$. We seek in the optimal algorithm in the sense that it should meet both the non-private instance-dependent bound of (Lai and Robbins, 1985) and the lower bound of (Shariff and Sheffet, 2018); namely an algorithm with an instance-dependent pseudo-regret bound of $O\left(\frac{K\log(T)}{\varepsilon} + \sum_{a\neq a^*}\frac{\log(T)}{\Delta_a}\right)$. The algorithm we devise is a differentially private version of the Successive Elimination (SE) algorithm (Even-Dar et al., 2002). SE initializes by setting all K arms as viable options, and iteratively pulls all viable arms maintaining the same confidence interval around the empirical average of each viable arm's reward. Once some viable arm's upper confidence bound is strictly smaller than the lower confidence bound of some other viable arm, the arm with the lower empirical reward is eliminated and is no longer considered viable. It is worth while to note that the classical UCB algorithm and the SE algorithm have the same asymptotic pseudo-regret. To design the differentially private analouge of SE, one can use our results from the previous section regarding stopping rules. After all, in the special case where we have $K = 2$ arms, we can straight-forwardly use the private stopping-rule to assess the mean of the difference between the arms up to a constant α (say $\alpha = 0.5$). The question lies in applying this algorithm in the $K > 2$ case.

Here are a few failed first-attempts. The most straightforward ideas is to apply $\binom{K}{2}$ stopping rules / SVTs for all pairs of arms; but since a reward of a single pull of any single arm plays a role in $K - 1$ SVT instantiations, it follows we would have to scale down the privacy-loss of each SVT to $\Theta(\varepsilon/K)$ resulting in an added regret scaled up by a factor of K. In an attempt to reduce the number of SVT-instantiations, we might consider asking for each arm whether *there exists* an arm with a significantly greater reward, yet it still holds that the reward from a single pull of the *leading* arm a^* plays a role in K SVT-instantiations. Next, consider merging all queries into a single SVT, posing in each round K queries (one per arm) and halting once we find that a certain arm is suboptimal; but this results in a single SVT that may halt $K-1$ times, causing us yet again to scale ε by a factor of K.

In order to avoid scaling down ε by a factor of K, our solution leverages on the combination of parallel decomposition and geometrically increasing intervals. Namely we partition the arm pulls of the algorithm into *epochs* of geometrically increasing lengths, where in epoch e we eliminate *all* arms of optimality-gap $\geq 2^{-e}$. In fact, it turns out we needn't apply the SVT at the end of each epoch[5] but rather just test for a noticeably underperforming arm using a private histogram. The key point is that at the beginning of each new epoch we nullify all counters (i.e delete all prior rewards) and start the mean-reward estimation completely anew (over the remaining set of viable arms) — and so a single reward plays a role in only one epoch, allowing for ε-DP mean-estimation in each epoch (rather than ε/K). Yet due to the fact that the epochs are of exponentially growing lengths the total number of pulls for any suboptimal arm is proportional to the length of the epoch in which it eliminated, resulting in only a constant factor increase to the regret. The full-fledged details appear in

[5] We thank the anonymous referee for this elegant observation.

Algorithm 3.

Algorithm 3 DP Successive Elimination

input K arms, confidence β, privacy-loss ε.
Let $S \leftarrow \{1, \ldots, K\}$.
Initialize: $t \leftarrow 0$, $epoch \leftarrow 0$.
repeat
  Increment $epoch \leftarrow epoch + 1$.
  Set $r \leftarrow 0$
  Zero all means: $\forall i \in S$ set $\bar{\mu}_i \leftarrow 0$
  Set $\Delta_e \leftarrow 2^{-epoch}$
  Set $R_e \leftarrow \max\left(\frac{32\log(8|S|epoch^2/\beta)}{\Delta_e^2}, \frac{8\log(4|S|epoch^2/\beta)}{\varepsilon\Delta_e}\right) + 1$
  while $r < R_e$ **do**
    Increment $r \leftarrow r + 1$.
    foreach $i \in S$
      Increment $t \leftarrow t + 1$
      Sample reward of arm i, update mean $\bar{\mu}_i$.
  end while
  Set $h_e \leftarrow \sqrt{\frac{\log\left(8|S|\cdot epoch^2/\beta\right)}{2R_e}}$
  Set $c_e \leftarrow \frac{\log\left(4|S|\cdot epoch^2/\beta\right)}{R_e\varepsilon}$
  foreach $i \in S$ set $\widetilde{\mu}_i \leftarrow \bar{\mu}_i + \mathsf{Lap}(1/\varepsilon r)$
  Let $\widetilde{\mu}_{\max} = \max_{i\in S} \widetilde{\mu}_i$
  Remove all arm j from S such that:
    $\widetilde{\mu}_{\max} - \widetilde{\mu}_j > 2h_e + 2c_e$
until $|S| = 1$
Pull the arm in S in all remaining rounds.

Theorem 4.1. *Algorithm 3 is ε-differentially private.*

Proof. Consider two streams of arm-rewards that differ on the reward of a single arm in a single timestep. This timestep plays a role in a single epoch e. Moreover, let a be the arm whose reward differs between the two neighboring streams. Since the reward of each arm is bounded by $[0, 1]$ it follows that the difference of the mean of arm a between the two neighboring streams is $\leq 1/R_e$. Thus, adding noise of $\mathsf{Lap}(1/\varepsilon R_e)$ to $\bar{\mu}_a$ guarantees ε-DP. □

To argue about the optimality of Algorithm 3, we require the following lemma, a key step in the following theorem that bounds the pseudo-regret of the algorithm. Its proof is deferred to the full version.

Lemma 4.2. *Fix any instance of the K-MAB problem, and denote a^* as its optimal arm (of highest mean), and the gaps between the mean of arm a^* and any suboptimal arm $a \neq a^*$ as Δ_a. Fix any horizon T. Then w.p. $\geq 1 - \beta$ it holds that Algorithm 3 pulls each suboptimal arm $a \neq a^*$ for a number of timesteps upper bounded by*

$$\min\{T,\ O\left((\log(K/\beta) + \log\log(1/\Delta_a))\left(\frac{1}{\Delta_a^2} + \frac{1}{\varepsilon\Delta_a}\right)\right)\}$$

Based on Lemma 4.2 we can now reason about the pseudo-regret of the DP-SE algorithm.

Theorem 4.3. *Under the same notation as in Lemma 4.2 and for sufficiently large T, the expected regret of Algorithm 3 is at most $O\left(\left(\sum_{a\neq a^*}\frac{\log(T)}{\Delta_a}\right) + \frac{K\log(T)}{\varepsilon}\right)$.*

The proof of Theorem 4.3, which is a straight-forward calculation once one sets $\beta = 1/T$, is deferred to the full version. It is worth noting yet again that the expected regret of Algorithm 3 meets both the (instance dependent) non-private lower bound (Lai and Robbins, 1985) of $\Omega\left(\sum_{a\neq a^*}\frac{\log(T)}{\Delta_a}\right)$ and the private lower bound (Shariff and Sheffet, 2018) of $\Omega\left(K\log(T)/\varepsilon\right)$.

Minimax Regret Bound. The bound of Theorem 4.3 is an instance-dependent bound, and so we turn our attention to the minimax regret bound of Algorithm 3 — Given horizon bound T, how should an adversary set the gaps between the different arms as to maximize the expected regret of Algorithm 3? We next show that in any setting of the gaps, the following is an instance independent bound on the expected regret of Algorithm 3.

Theorem 4.4. *(Instance Independent Bound) The pseudo regret of Algorithm 3 is $O\left(\sqrt{TK\log(T)} + K\log(T)/\varepsilon\right)$.*

Again, we comment on the optimality of the bound in Theorem 4.4. The non-private minimax bound (Auer et al., 2002b) is known to be $\Omega(\sqrt{TK})$ and combining it with the private bound of $\Omega(K\log(T)/\varepsilon)$ we see that the above minimax bound is just $\sqrt{\log(T)}$-factor away from being optimal. The full proof of Theorem 4.4 is deferred to the full version.

5. Empirical Evaluation

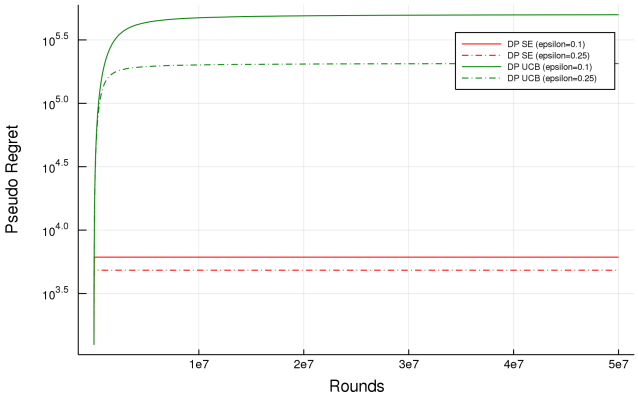
In this section, we empirically compare the existing DP-UCB algorithm to our DP-SE algorithm (Algorithm 3). Our goal is to assert that indeed our DP-SE algorithm outperforms (achieves smaller pseudo regret than) the DP-UCB baseline under a wide range of parameters. Afterall, the improvement we introduce is over $\mathrm{poly}\log(T)$ factors and does incur an increase in the constants repressed by the big-O notation.

In our experiments we set the default values of $T = 5\times10^7$, $\varepsilon = 0.25$ and $K = 5$. We assume T is a-priori known to both algorithms and set $\beta = 1/T$. Due to space constraints, we only present our empirical results on one instance where the reward means of the arms are linearly spaced in the range $[0.25, 0.75]$[6]. Namely in our instance, for $K = 5$,

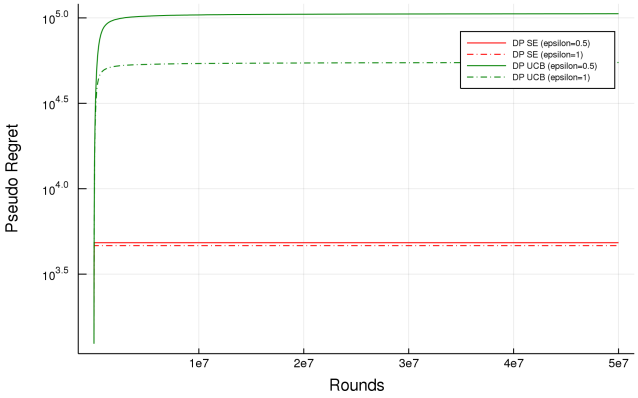
[6] Constraining the means within $[0.25, 0.75]$ ensures the variance of the arms are similar (upto a constant of $4/3$)

the mean rewards are in $\{0.75, 0.625, 0.5, 0.375, 0.25\}$. (In the full version of this work we experiment with three additional settings, varying both ε and K.) We then vary $\varepsilon \in \{0.1, 0.25, 0.5, 1\}$. Using a^* to denote the optimal arm, we measure the algorithms' performances in terms of their pseudo regret, so upon pulling a suboptimal arm $a \neq a^*$ each algorithm incurs a cost $\Delta_a = \mu_{a^*} - \mu_a$. For each setting, 30 runs of the algorithms were carried out and their average pseudo regrets are plotted. The results, presented in *log-scale*, are given in Figure 1.

The results conclusively show that DP-SE outperforms DP-UCB — subject to the caveat that our experiments are proof-of-concept only and we did not conduct a thorough investigation of the entire hyper-parameter space, we *could not find even a single setting where DP-UCB is even comparable to our DP-SE*. I.e. in *all* settings we tested (the one presented here and the additional ones presented in the full version), we outperform DP-UCB by at least 5 times. We also comment as to the difference in the shape of the two pseudo-regret curves — while the DP-UCB curve is smooth (attesting to the fact it pulls suboptimal arms even for fairly large values of T), the DP-SE is piece-wise linear (exhibiting the fact that at some point is eliminates all suboptimal arms). Note also that changing ε affects the performance of DP-UCB much more than DP-SE due to the $\text{poly}\log T/\varepsilon$ factor.



(a) $\varepsilon = 0.1$ and 0.25



(b) $\varepsilon = 0.5$ and 1

Figure 1: Expected regret with $K = 5$ arms of mean rewards $\{0.75, 0.625, 0.5, 0.375, 0.25\}$ and $T = 5 \times 10^7$

6. Future Directions

While it seems this work "closes the book" on the private stochastic-MAB problem, we want to point out a few future research directions. First, the MAB problem has actually multiple lower-bounds, where even low-order terms in the lower bound have been devised under different setting (see for example Bubeck et al. (2013)); so studying the lower-order terms of the bounds on the private MAB problem may be of importance. Secondly, much of the work on stopping rules is devoted to the case where the variance σ^2 of the distribution is significantly smaller than its range. E.g. Mnih et al. (2008) gave an algorithm whose sample complexity is actually $O\left(\max\{\frac{\sigma^2}{\alpha^2\mu^2}, \frac{R}{\alpha|\mu|}\}(\log(1/\beta + \log\log(R/\alpha|\mu|))\right)$. Note that the lower-bound in Theorem 3.4 deals with a distribution of variance $\Theta(R^2)$, so by restricting our attention to distributions with much smaller variance we may bypass this lower-bound. We leave the problem of designing privacy-preserving analogues of the Bernstein stopping rule (Mnih et al., 2008) as an interesting open-problem.

Also, note that our entire analysis is restricted to ε-DP. While our results extend to the more-recent notion of concentrated differential privacy (Bun and Steinke, 2016), we do not know how to extend them to (ε, δ)-DP, as we do not know the lower-bounds for this setting. Similarly, we do not know the concrete privacy-utility bounds of the MAB problem in the local-model of DP. Lastly, it would be interesting to see if the overall approach of private Successive Elimination is applicable, and yields better bounds than currently known, for natural extensions of the MAB, such as in the linear and contextual settings. Even-Dar et al. (2002) themselves motivated their work by various applications in a Markov-chain related setting. It is an interesting open problem of adjusting this work to such applications.

Acknowledgements

We gratefully acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC) for supporting O.S. with grant #201706701. O.S. is also an unpaid collaborator on NSF grant #1565387. We thank the anonymous referee for helpful advice as to simplifying our original version of the DP-SE algorithm.

References

Rajeev Agrawal. *Sample mean based index policies with O(log n) regret for the multi-armed bandit problem.*, volume 27, pages 1054–1078. Applied Probability Trust, 1995.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *JMLR*, 47(2-3):235–256, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Donald A Berry and Bert Fristedt. Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall*, 5:71–87, 1985.

Sbastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. In *Proceedings of the 26th Annual Conference on Learning Theory*, pages 122–134. PMLR, 2013.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography - 14th International Conference, TCC 2016-B, Beijing, China, October 31 - November 3, 2016, Proceedings, Part I*, pages 635–658, 2016.

Stéphane Caron and Smriti Bhagat. Mixing bandits: a recipe for improved cold-start recommendations in a social network. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, page 11, 2013.

T.-H. Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. In *Automata, Languages and Programming*, Lecture Notes in Computer Science, pages 405–417, 2010.

Paul Dagum, Richard Karp, Michael Luby, and Sheldon Ross. An optimal algorithm for monte carlo estimation. *SIAM Journal on computing*, 29(5):1484–1496, 2000.

Carlos Domingo, Ricard Gavaldà, and Osamu Watanabe. Adaptive sampling methods for scaling up knowledge discovery algorithms. *Data Mining and Knowledge Discovery*, 6(2):131–152, 2002.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, Lecture Notes in Computer Science, pages 265–284. Springer, Berlin, Heidelberg, 2006.

Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N. Rothblum. Differential privacy under continual observation. In *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, STOC '10, pages 715–724, 2010.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, pages 255–270. Springer, 2002.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Matthew Hoffman, Eric Brochu, and Nando de Freitas. Portfolio allocation for bayesian optimization. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 327–336, 2011.

Sampath Kannan, Jamie H. Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 2231–2241, 2018.

Vishesh Karwa and Salil Vadhan. Finite Sample Differentially Private Confidence Intervals. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, volume 94, pages 44:1–44:9, 2018.

Branislav Kveton, Csaba Szepesvári, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 767–776, 2015.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Nikita Mishra and Abhradeep Thakurta. (nearly) optimal differentially private stochastic multi-arm bandits. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 592–601. AUAI Press, 2015.

Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical bernstein stopping. In *Proceedings of the 25th international conference on Machine learning*, pages 672–679. ACM, 2008.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58(5):527–535, 09 1952.

Touqir Sajed, Wesley Chung, and Martha White. High-confidence error estimates for learned value functions. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 683–692, 2018.

Eric M. Schwartz, Eric T. Bradlow, and Peter S. Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4): 500–522, July 2017.

Roshan Shariff and Or Sheffet. Differentially private contextual linear bandits. In *Advances in Neural Information Processing Systems*, pages 4301–4311, 2018.

Adam Smith and Abhradeep Thakurta. (Nearly) optimal algorithms for private online learning in full-information and bandit settings. In *NIPS*, pages 2733–2741, 2013.

Aristide CY Tossou and Christos Dimitrakakis. Algorithms for differentially private multi-armed bandits. In *AAAI*, pages 2087–2093, 2016.