Optimal Rates of (Locally) Differentially Private Heavy-tailed Multi-Armed Bandits

Youming Tao [*1] **Yulian Wu** [*2] **Peng Zhao** [3] **Di Wang** [2]

Abstract

In this paper we investigate the problem of stochastic multi-armed bandits (MAB) in the (local) differential privacy (DP/LDP) model. Unlike previous results that assume bounded/sub-Gaussian reward distributions, we focus on the setting where each arm's reward distribution only has $(1+v)$-th moment with some $v \in (0,1]$. In the first part, we study the problem in the central ϵ-DP model. We first provide a near-optimal result by developing a private and robust Upper Confidence Bound (UCB) algorithm. Then, we improve the result via a private and robust version of the Successive Elimination (SE) algorithm. Finally, we establish the lower bound to show that the instance-dependent regret of our improved algorithm is optimal. In the second part, we study the problem in the ϵ-LDP model. We propose an algorithm that can be seen as locally private and robust version of SE algorithm, which provably achieves (near) optimal rates for both instance-dependent and instance-independent regret. Our results reveal differences between the problem of private MAB with bounded/sub-Gaussian rewards and heavy-tailed rewards. To achieve these (near) optimal rates, we develop several new hard instances and private robust estimators as byproducts, which might be used to other related problems.

1. INTRODUCTION

Multi-Armed Bandits (MAB), and its general form, bandit learning, have already been studied for more than half a century, starting from (Thompson, 1933) and (Robbins, 1952). They find numerous applications in many areas such as medicine (Gutiérrez et al., 2017), finance (Shen et al., 2015), social science (Nakayama et al., 2017), and clinical research (Press, 2009). However, due to the existence of sensitive data and their distributed nature in many applications, it is often challenging to preserve the privacy of such data.

To preserve the privacy of these sensitive data, Differential Privacy (DP) (Dwork et al., 2006) has received a great deal of attention and now has established itself as a de facto notion of privacy for data analysis. Over the past decade, differentially private bandit learning has been extensively studied from various setups including classical stochastic MAB (Mishra & Thakurta, 2015; Tossou & Dimitrakakis, 2016; Sajed & Sheffet, 2019; Ren et al., 2020; Kalogerias et al., 2020), combinatorial semi-bandits (Chen et al., 2020), and contextual bandits (Shariff & Sheffet, 2018; Hannun et al., 2019; Malekzadeh et al., 2020; Zheng et al., 2020).

However, these problems are still not well-understood. For example, all of the previous results and methods need to assume that the rewards are sampled from some bounded (or sub-Gaussian) distributions to guarantee the DP property. However, such assumptions may not hold when designing decision-making algorithms for complicated real-world systems. In particular, previous papers have shown that the rewards or the interactions in such systems often lead to heavy-tailed and power law distributions (Dubey & Pentland, 2019), such as modeling stock prices (Bradley & Taqqu, 2003), preferential attachment in social networks (Mahanti et al., 2013), and online behavior on websites (Kumar & Tomkins, 2010). Thus, it is necessary to develop new methods to deal with these heavy-tailed rewards in the private bandit learning.

To address the above issue, in this paper, we focus on the most fundamental bandit model, *i.e.,* multi-armed bandits, with heavy-tailed rewards. We conduct a comprehensive and the first study on MAB with heavy-tailed rewards in both central and local DP models. Our contributions are summarized as follows.

- In the first part (Section 2), we consider the problem in the central ϵ-DP model. Specifically, we first pro-

[*]Equal contribution [1]School of Computer Science, Shandong University, Qingdao, China [2]Division of Computer, Electrical and Mathematical Sciences and Engineering, King Abdullah University of Science and Technology, Jeddah, Saudi Arabia [3]School of Artificial Intelligence,Nanjing University, Nanjing, China. Correspondence to: Di Wang <di.wang@kaust.edu.sa>.

ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland, USA, 2022.

pose a method based on a robust version of the Upper Confidence Bound (UCB) algorithm, and also design a new mechanism that could be seen as an adaptive version of the Tree-based mechanism (Dwork et al., 2010). To further improve the result, we then develop a private and robust version of the Successive Elimination (SE) algorithm and show that the (expected) regret bound is improved by a factor of $\log^{1.5+\frac{1}{v}} T$, where T is the number of rounds. Moreover, we establish the lower bound and show that the instance-dependent regret bound of $O\left(\frac{\log T}{\epsilon}\sum_{\Delta_a>0}(\frac{1}{\Delta_a})^{\frac{1}{v}} + \max_a \Delta_a\right)$ achieved by our second algorithm is optimal (up to $\text{poly}(\log\log\frac{1}{\Delta_a})$ factors), where Δ_a is the mean gap.

- In the second part (Section 3), we study the problem in the ϵ-LDP model. We first develop a LDP version of the SE algorithm which achieves an instance-dependent regret bound of $O\left(\frac{\log T}{\epsilon^2}\sum_{\Delta_a>0}(\frac{1}{\Delta_a})^{\frac{1}{v}} + \max_a \Delta_a\right)$ and an $\widetilde{O}\left(\left(\frac{K}{\epsilon^2}\right)^{\frac{v}{1+v}} T^{\frac{1}{1+v}}\right)$ instance-independent bound. Then, we show that the above instance-dependent regret bound is optimal and the instance-independent regret bound is near-optimal (up to $\text{poly}(\log T)$ factors).

- All of our results also reveal the differences between the problem of private MAB with bounded/sub-Guassian rewards and that with heavy-tailed rewards. To achieve these (near) optimal results, we develop several new hard instances, mechanisms and private robust estimators as byproducts, which could be used to other related problems, such as private contextual bandits (Shariff & Sheffet, 2018) or private reinforcement learning (Vietri et al., 2020).

2. DP HEAVY-TAILED MAB

In the classical setting where the rewards follow some bounded distributions, the most commonly used approach is using the Tree-based mechanism to privately calculate the sum of rewards and then modify the Upper Confidence Bound (UCB) algorithm (Auer et al., 2002), such as (Mishra & Thakurta, 2015; Tossou & Dimitrakakis, 2016). However, their methods cannot be directly generalized to the heavy-tailed setting, since now the reward is unbounded. Thus, the most natural idea is to first preprocess the rewards to make them bounded and then use the Tree-based mechanism and UCB algorithm, see Algorithm 1 for details.

Lemma 1 ((Adaptive) Tree-based Mechanism)**.** Given a stream σ such that $\sigma(t) \in [-B_t, B_t]$ for $\forall t \in [T]$, where B_t is non-decreasing with t, we want to privately and continually release the sum of the stream $S(t) \triangleq \sum_{i=1}^{t}\sigma(i)$ for each $t \in [T]$. Tree-based Mechanism (Algorithm 2) outputs

Algorithm 1 DP Robust Upper Confidence Bound

Input: time horizon T, parameters ϵ, v, u.
Create an empty tree $tree_a$ for each arm $a \in [K]$.
Initialize pull number $n_a \leftarrow 0$ for each arm $a \in [K]$.
Denote B_n as $(\frac{\epsilon u n}{\log^{1.5} T})^{1/(1+v)}$ for any $n \in \mathbb{N}^+$.
for $t = 1, \ldots, K$ **do**
  Pull arm t and observe a reward x_t.
  Update the pull number $n_t \leftarrow n_t + 1$.
  Truncate the reward by $\widetilde{x}_t \leftarrow x_t \cdot \mathbb{I}_{|x_t| \le B_{n_t}}$.
  Insert $\widetilde{x}_t$ into $tree_t$.
end for
for $t = K+1, \ldots, T$ **do**
  Obtain $\widehat{S}_a(t)$ for each $a \in [K]$ via Algorithm 2.
  Pull arm

$$a_t = \arg\max_a \frac{\widehat{S}_a(t)}{n_a} + 18u^{\frac{1}{1+v}}\left(\frac{\log(2t^4)\log^{1.5+\frac{1}{v}} T}{n_a \epsilon}\right)^{\frac{v}{1+v}}$$

  and observe the reward x_t.
  Update the pull number $n_{a_t} \leftarrow n_{a_t} + 1$.
  Truncate the reward by $\widetilde{x}_t \leftarrow x_t \cdot \mathbb{I}_{|x_t| \le B_{n_{a_t}}}$.
  Insert $\widetilde{x}_t$ into $tree_{a_t}$.
end for

Algorithm 2 (Adaptive) Tree-based Mechanism

Input: time horizon T, privacy budget ϵ, a stream σ.
Output: A private version $\widehat{S}(t)$ for $S(t) = \sum_{i=1}^{t}\sigma(i)$ at each $t \in [T]$
Initialize each α_i and noisy $\widehat{\alpha}_i$ to 0.
$\epsilon' \leftarrow \epsilon / \log T$.
for $t = 1, \ldots, T$ **do**
  Express t in binary form: $t = \sum_j \text{Bin}_j(t) \cdot 2^j$.
  $i \leftarrow \min\{j : \text{Bin}_j(t) \neq 0\}$.
  $\alpha_i \leftarrow \sum_{j<i}\alpha_j + \sigma(t)$.
  for $j = 0, \ldots, i-1$ **do**
    $\alpha_j \leftarrow 0, \widehat{\alpha}_j \leftarrow 0$.
  end for
  $\widehat{\alpha}_i \leftarrow \alpha_i + \text{Lap}(2B_t/\epsilon')$.
  $\hat{S}(t) \leftarrow \sum_{j:\text{Bin}_j(t)=1}\widehat{\alpha}_j$.
end for

an estimation $\widehat{S}(t)$ for $S(t)$ at each $t \in [T]$ such that $\widehat{S}(t)$ preserves ϵ-differential privacy and guarantees the following noise bound with probability at least $1-\delta$ for any $\delta > 0$,

$$\left|\widehat{S}(t) - S(t)\right| \le \frac{2B_t}{\epsilon} \cdot \log^{1.5} T \cdot \log\frac{1}{\delta}. \tag{1}$$

When $B_t = B$, Algorithm 2 will be the same as the original one. Theorem 1 presents the privacy guarantee of overall algorithm (Algorithm 1 and Algorithm 2).

Theorem 1. For any $\epsilon > 0$, the overall algorithm (Algorithm 1 and Algorithm 2) is ϵ-differentially private.

In fact, the $\widehat{S}_a(t)/n_a$ term in step 12, which is denoted by $\widehat{\mu}_a(n_a, t)$, could be seen as a robust and private estimator of the mean μ_a after total n_a pulls of arm a till time t. Our selection strategy in step 12 is based on the following estimation error between $\widehat{\mu}_a(n_a, t)$ and μ_a, which is also a key lemma that will be used to bound the regret of Algorithm 1.

Lemma 2. In Algorithm 1, for a fixed arm a and t, we have the following estimation error with probability at least $1 - t^{-4}$,

$$\widehat{\mu}_a(n_a, t) \leq \mu_a + 18u^{\frac{1}{1+v}} \left(\frac{\log(2t^4) \log^{1.5+\frac{1}{v}} T}{n_a \epsilon} \right)^{\frac{v}{1+v}}. \tag{2}$$

We have the following instance-dependent regret bound by the proof of Theorem 1 in (Auer et al., 2002).

Theorem 2. Under our assumptions, for any $0 < \epsilon \leq 1$ the instance-dependent expected regret of Algorithm 1 satisfies

$$\mathcal{R}_T \leq O\left(\sum_{a:\Delta_a > 0} \left(\frac{\log^{2.5+\frac{1}{v}} T}{\epsilon} \left(\frac{u}{\Delta_a} \right)^{\frac{1}{v}} + \Delta_a \right) \right). \tag{3}$$

Compared with the $O(\sum_{a:\Delta_a>0} [\log T (\frac{u}{\Delta_a})^{\frac{1}{v}} + \Delta_a])$ optimal rate of the regret in the non-private version (Bubeck et al., 2013), we can see that there is an additional factor of $\frac{\log^{1.5+\frac{1}{v}} T}{\epsilon}$ in the private case. Thus, a natural question arises here is *whether it is possible to further improve the regret*. We answer this question affirmatively by designing an optimal algorithm, see Algorithm 3 for details.

Theorem 3. For any $\epsilon > 0$, Algorithm 3 is ϵ-differentially private.

Theorem 4 (DP Upper Bound)**.** If we set $\beta = \frac{1}{T}$ in Algorithm 3, then for sufficiently large T and any $\epsilon \in (0, 1]$, the instance-dependent expected regret of Algorithm 3 satisfies

$$\mathcal{R}_T \leq O\left(\frac{u^{\frac{1}{1+v}} \log T}{\epsilon} \sum_{\Delta_a > 0} \left(\frac{1}{\Delta_a} \right)^{\frac{1}{v}} + \max_a \Delta_a \right). \tag{4}$$

Moreover, the instance-independent expected regret of Algorithm 3 satisfies

$$\mathcal{R}_T \leq O\left(u^{\frac{v}{(1+v)^2}} \left(\frac{K \log T}{\epsilon} \right)^{\frac{v}{1+v}} T^{\frac{1}{1+v}} \right), \tag{5}$$

where the $O(\cdot)$-notation omits $\log \log \frac{1}{\Delta_a}$ terms.

From Theorem 4 we can see that compared with the regret bound $O(\frac{\log^{2.5} T}{\epsilon} \sum_{\Delta_a>0} (\frac{1}{\Delta_a})^{\frac{1}{v}})$ in Theorem 2, we achieve an improved bound of $O(\frac{\log T}{\epsilon} \sum_{\Delta_a>0} (\frac{1}{\Delta_a})^{\frac{1}{v}})$. We show the instance-dependent regret bound presented in Theorem 4 is **optimal**. The lower bound of the instance-independent regret is still unclear, and we leave it as an open problem.

Algorithm 3 DP Robust Successive Elimination

Input: confidence β, parameters ϵ, v, u.

1: $\mathcal{S} \leftarrow \{1, \cdots, K\}$
2: Initialize: $t \leftarrow 0$, $\tau \leftarrow 0$.
3: **repeat**
4: $\quad \tau \leftarrow \tau + 1$.
5: $\quad$ Set $\bar{\mu}_a = 0$ for all $a \in \mathcal{S}$.
6: $\quad r \leftarrow 0$, $D_\tau \leftarrow 2^{-\tau}$.
7: $\quad R_\tau \leftarrow \left\lceil u^{\frac{1}{v}} (\frac{24^{(1+v)/v} \log(4|\mathcal{S}|\tau^2/\beta)}{\epsilon D_\tau^{(1+v)/v}}) + 1 \right\rceil$.
8: $\quad B_\tau \leftarrow (\frac{u R_\tau \epsilon}{\log(4|\mathcal{S}|\tau^2/\beta)})^{1/(1+v)}$.
9: $\quad$ **while** $r < R_\tau$ **do**
10: $\quad\quad r \leftarrow r + 1$.
11: $\quad\quad$ **for** $a \in \mathcal{S}$ **do**
12: $\quad\quad\quad t \leftarrow t + 1$.
13: $\quad\quad\quad$ Sample a reward $x_{a,r}$.
14: $\quad\quad\quad \widetilde{x}_{a,r} \leftarrow x_{a,r} \cdot \mathbb{I}_{\{|x_{a,r}| \leq B_\tau\}}$.
15: $\quad\quad$ **end for**
16: $\quad$ **end while**
17: $\quad$ For each $a \in \mathcal{S}$, compute $\bar{\mu}_a \leftarrow (\sum_{l=1}^{R_\tau} \widetilde{x}_{a,l})/R_\tau$.
18: $\quad$ Set $\widetilde{\mu}_a \leftarrow \bar{\mu}_a + \mathrm{Lap}(\frac{2B_\tau}{R_\tau \epsilon})$ for all $a \in \mathcal{S}$.
19: $\quad \widetilde{\mu}_{\max} \leftarrow \max_{a \in \mathcal{S}} \widetilde{\mu}_a$.
20: $\quad err_\tau \leftarrow u^{1/(1+v)} (\frac{\log(4|\mathcal{S}|\tau^2/\beta)}{R_\tau \epsilon})^{v/(1+v)}$.
21: $\quad$ **for** all viable arm a **do**
22: $\quad\quad$ **if** $\widetilde{\mu}_{\max} - \widetilde{\mu}_a > 12 err_\tau$ **then**
23: $\quad\quad\quad$ Remove arm a from $\mathcal{S}$.
24: $\quad\quad$ **end if**
25: $\quad$ **end for**
26: **until** $|\mathcal{S}| = 1$
27: Pull the arm in $\mathcal{S}$ in all remaining $T - t$ rounds.

Theorem 5 (DP Instance-dependent Lower Bound)**.** There exists a heavy-tailed K-armed bandit instance with $u \leq 1$, $\mu_a \leq \frac{1}{6}$ and $\Delta_a \in (0, \frac{1}{12})$, such that for any ϵ-DP ($0 < \epsilon \leq 1$) algorithm $\mathcal{A}$ whose expected regret is at most $T^{\frac{3}{4}}$, we have

$$\mathcal{R}_T \geq \Omega\left(\frac{\log T}{\epsilon} \sum_{\Delta_a > 0} \left(\frac{1}{\Delta_a} \right)^{\frac{1}{v}} \right). \tag{6}$$

Remark 1. In the MAB with bounded rewards case, it has been shown that the optimal rate of the expected rate is $O(\frac{K \log T}{\epsilon} + \sum_{\Delta_a>0} \frac{\log T}{\Delta_a})$ (Sajed & Sheffet, 2019). Compared with the optimal rate $O(\frac{\log T}{\epsilon} \sum_{\Delta_a>0} (\frac{1}{\Delta_a})^{\frac{1}{v}})$ in the heavy-tailed case, we can see there is a huge difference. First, the dependency on $\frac{1}{\Delta_a}$ now becomes to $(\frac{1}{\Delta_a})^{\frac{1}{v}}$. Secondly, the price of privacy in the bounded rewards case is an additional term of $O(\frac{K \log T}{\epsilon})$ compared with the non-private rate, while in the heavy-tailed case, there is an additional factor of $\frac{1}{\epsilon}$ compared with the non-private one.

3. LDP HEAVY-TAILED MAB

Unlike the Tree-based mechanism in the central model, the LDP version of the UCB algorithm will introduce a huge amount of error to estimate the mean. To achieve a better utility, we propose an ϵ-LDP version of the SE algorithm, see Algorithm 4 for details. The following theorems provide the privacy and utility guarantees for Algorithm 4.

Theorem 6. For any $\epsilon > 0$, Algorithm 4 is ϵ-local differentially private.

Theorem 7 (LDP Upper Bound)**.** Set $\beta = \frac{1}{T}$ in Algorithm 4. For any $\epsilon \in (0, 1]$ and sufficiently large T, the instance-dependent expected regret of Algorithm 4 satisfies

$$\mathcal{R}_T \leq O\left(\frac{u^{\frac{2}{v}} \log T}{\epsilon^2} \sum_{\Delta_a > 0} \left(\frac{1}{\Delta_a}\right)^{\frac{1}{v}} + \max_a \Delta_a\right). \quad (7)$$

Moreover, the instance-independent expected regret of Algorithm 4 satisfies

$$\mathcal{R}_T \leq O\left(u^{\frac{2}{1+v}} \left(\frac{K \log T}{\epsilon^2}\right)^{\frac{v}{1+v}} T^{\frac{1}{1+v}}\right), \quad (8)$$

where the $O(\cdot)$-notations omit $\log\log \frac{1}{\Delta_a}$ terms.

In the following, we derive both instance-dependent and instance-independent lower bounds for heavy-tailed MAB in the ϵ-LDP model.

Theorem 8 (LDP Instance-dependent Lower Bound)**.** There exists a heavy-tailed K-armed bandit instance with $u \leq 1$ and $\Delta_a \triangleq \mu_1 - \mu_a \in (0, \frac{1}{5})$, such that for any ϵ-LDP ($0 < \epsilon \leq 1$) algorithm whose regret $\leq o(T^\alpha)$ for any $\alpha > 0$, the regret satisfies

$$\liminf_{T\to\infty} \frac{\mathcal{R}_T}{\log T} \geq \Omega\left(\frac{1}{\epsilon^2} \sum_{\Delta_a > 0} \left(\frac{1}{\Delta_a}\right)^{\frac{1}{v}}\right).$$

Remark 2. The attained bound in Theorem 7 is optimal. Compared with the optimal rate $O(\frac{1}{\Delta_a^{\frac{1}{v}}})$ in the non-private case, we can see the price of privacy is an additional factor of $\frac{1}{\epsilon^2}$, which is similar to other MAB with bounded/sub-Gaussian rewards problems in the LDP model (Zhou & Tan, 2021; Ren et al., 2020).

Theorem 9 (LDP Instance-independent Lower Bound)**.** There exists a heavy-tailed K-armed bandit instance with the $(1 + v)$-th bounded moment of each reward distribution is bounded by 1. Moreover, if T is large enough, for any the ϵ-LDP algorithm $\mathcal{A}$ with $\epsilon \in (0, 1]$, the expected regret must satisfy

$$\mathcal{R}_T \geq \Omega\left(\left(\frac{K}{\epsilon^2}\right)^{\frac{v}{1+v}} T^{\frac{1}{1+v}}\right).$$

Algorithm 4 LDP Robust Successive Elimination

Input: Confidence β, parameters ϵ, v, u.
$\mathcal{S} \leftarrow \{1, \cdots, K\}$
Initialize: $t \leftarrow 0$, $\tau \leftarrow 0$.
repeat
  $\tau \leftarrow \tau + 1$.
  Set $\bar{\mu}_a = 0$ for all $a \in \mathcal{S}$.
  $r \leftarrow 0$, $D_\tau \leftarrow 4^{-\tau}$.
  $R_\tau \leftarrow \left\lceil u^{\frac{2}{v}} \left(\frac{28^{2(1+v)/v} \log(8|\mathcal{S}|\tau^2/\beta)}{\epsilon^2 D_\tau^{2(1+v)/v}}\right) + \log\left(\frac{8|\mathcal{S}|\tau^2}{\beta}\right)\right\rceil$.
  $B_\tau \leftarrow \left(\frac{u\sqrt{R_\tau}\epsilon}{\sqrt{\log(8|\mathcal{S}|\tau^2/\beta)}}\right)^{1/(1+v)}$.
  while $r < R_\tau$ **do**
    $r \leftarrow r + 1$.
    for $a \in \mathcal{S}$ **do**
      $t \leftarrow t + 1$.
      Sample a reward $x_{a,r}$ for each arm $a \in \mathcal{S}$.
      $\widetilde{x}_{a,r} \leftarrow x_{a,r} \cdot \mathbb{I}_{\{|x_{a,r}| \leq B_\tau\}}$.
      $\widehat{x}_{a,r} \leftarrow \widetilde{x}_{a,r} + \mathrm{Lap}(\frac{2B_\tau}{\epsilon})$
    end for
  end while
  For each $a \in \mathcal{S}$, compute $\bar{\mu}_a \leftarrow \left(\sum_{l=1}^{R_\tau} \widehat{x}_{a,l}\right)/R_\tau$.
  $\bar{\mu}_{\max} \leftarrow \max_{a\in\mathcal{S}} \bar{\mu}_i$.
  $err_\tau \leftarrow u^{1/(1+v)} \left(\frac{\sqrt{\log(8|\mathcal{S}|\tau^2/\beta)}}{R_\tau \epsilon}\right)^{v/(1+v)}$.
  for all viable arm a **do**
    if $\widetilde{\mu}_{\max} - \widetilde{\mu}_a > 14 err_\tau$ **then**
      Remove arm a from $\mathcal{S}$.
    end if
  end for
until $|\mathcal{S}| = 1$
Pull the arm in $\mathcal{S}$ in all remaining $T - t$ rounds.

Remark 3. From Theorem 9, we can see the upper bound (8) of Algorithm 4 is nearly optimal. However, compared with instance-independent lower bound, there is still a poly$(\log T)$ factor gap. We conjecture this factor could be removed by using some more advanced robust estimator, such as the estimator in (Lee et al., 2020) and we will leave it as an open problem. For MAB with bounded rewards in the LDP model, (Basu et al., 2019) shows that its instance-dependent regret bound is always at least $\Omega(\frac{\sqrt{KT}}{\epsilon})$, *i.e.,* there is an additional factor of $\frac{1}{\epsilon}$ compared with the non-private case. However, for heavy-tailed MAB, compared with the lower bound of $\Omega(K^{\frac{v}{1+v}} T^{\frac{1}{1+v}})$ in the non-private case, from Theorem 9 we can observe that the difference is a factor of $(\frac{1}{\epsilon^2})^{\frac{v}{1+v}}$. Thus, combining with Remark 1, we can conclude that heavy-tailed MAB and bounded MAB are quite different in both central and local differential privacy models.

References

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

Basu, D., Dimitrakakis, C., and Tossou, A. Privacy in multi-armed bandits: Fundamental definitions and lower bounds. *arXiv preprint arXiv:1905.12298*, 2019.

Bradley, B. O. and Taqqu, M. S. Financial risk and heavy tails. In *Handbook of heavy tailed distributions in finance*, pp. 35–103. Elsevier, 2003.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59 (11):7711–7717, 2013.

Chen, X., Zheng, K., Zhou, Z., Yang, Y., Chen, W., and Wang, L. (Locally) differentially private combinatorial semi-bandits. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 1757–1767, 2020.

Dubey, A. and Pentland, A. Thompson sampling on symmetric α-stable bandits. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5715–5721, 2019.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Theory of Cryptography Conference (TCC)*, pp. 265–284, 2006.

Dwork, C., Naor, M., Pitassi, T., and Rothblum, G. N. Differential privacy under continual observation. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing (STOC)*, pp. 715–724, 2010.

Gutiérrez, B., Peter, L., Klein, T., and Wachinger, C. A multi-armed bandit to smartly select a training set from big medical data. In *Proceedings of the 20th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 38–45, 2017.

Hannun, A., Knott, B., Sengupta, S., and van der Maaten, L. Privacy-preserving multi-party contextual bandits. *arXiv preprint arXiv:1910.05299*, 2019.

Kalogerias, D. S., Nikolakakis, K. E., Sarwate, A. D., and Sheffet, O. Best-arm identification for quantile bandits with privacy. *arXiv preprint arXiv:2006.06792*, 2020.

Kumar, R. and Tomkins, A. A characterization of online browsing behavior. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pp. 561–570, 2010.

Lee, K., Yang, H., Lim, S., and Oh, S. Optimal algorithms for stochastic multi-armed bandits with heavy tailed rewards. In *Proceedings of the 34th Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Mahanti, A., Carlsson, N., Mahanti, A., Arlitt, M., and Williamson, C. A tale of the tails: Power-laws in internet measurements. *IEEE Network*, 27(1):59–64, 2013.

Malekzadeh, M., Athanasakis, D., Haddadi, H., and Livshits, B. Privacy-preserving bandits. In *Proceedings of the 3rd Conference on Machine Learning and Systems (MLSys)*, 2020.

Mishra, N. and Thakurta, A. (Nearly) optimal differentially private stochastic multi-arm bandits. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 592–601, 2015.

Nakayama, K., Hisakado, M., and Mori, S. Nash equilibrium of social-learning agents in a restless multiarmed bandit game. *Scientific reports*, 7(1):1–8, 2017.

Press, W. H. Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research. *Proceedings of the National Academy of Sciences*, 106(52):22387–22392, 2009.

Ren, W., Zhou, X., Liu, J., and Shroff, N. B. Multi-armed bandits with local differential privacy. *arXiv preprint arXiv:2007.03121*, 2020.

Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Sajed, T. and Sheffet, O. An optimal private stochastic-mab algorithm based on optimal private stopping rule. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 5579–5588, 2019.

Shariff, R. and Sheffet, O. Differentially private contextual linear bandits. *arXiv preprint arXiv:1810.00068*, 2018.

Shen, W., Wang, J., Jiang, Y.-G., and Zha, H. Portfolio choices with orthogonal bandit learning. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

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

Tossou, A. and Dimitrakakis, C. Algorithms for differentially private multi-armed bandits. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 2087–2093, 2016.

Vietri, G., Balle, B., Krishnamurthy, A., and Wu, S. Private reinforcement learning with pac and regret guarantees. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 9754–9764, 2020.

Zheng, K., Cai, T., Huang, W., Li, Z., and Wang, L. Locally differentially private (contextual) bandits learning. In *Proceedings of the 34th Conference on Advances in Neural Information Processing Systems (NeurIPS)*, pp. 12300–12310, 2020.

Zhou, X. and Tan, J. Local differential privacy for bayesian optimization. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 11152–11159, 2021.