# Optimal Algorithms for Lipschitz Bandits with Heavy-tailed Rewards

**Shiyin Lu** [1] **Guanghui Wang** [1] **Yao Hu** [2] **Lijun Zhang** [1]

## Abstract

We study Lipschitz bandits, where a learner repeatedly plays one arm from an infinite arm set and then receives a stochastic reward whose expectation is a Lipschitz function of the chosen arm. Most of existing work assume the reward distributions are bounded or at least sub-Gaussian, and thus do not apply to heavy-tailed rewards arising in many real-world scenarios such as web advertising and financial markets. To address this limitation, in this paper we relax the assumption on rewards to allow arbitrary distributions that have finite $(1+\epsilon)$-th moments for some $\epsilon \in (0, 1]$, and propose algorithms that enjoy a sublinear regret of $\widetilde{O}\left(T^{(d_z\epsilon+1)/(d_z\epsilon+\epsilon+1)}\right)$ where $T$ is the time horizon and $d_z$ is the zooming dimension. The key idea is to exploit the Lipschitz property of the expected reward function by adaptively discretizing the arm set, and employ upper confidence bound policies with robust mean estimators designed for heavy-tailed distributions. Furthermore, we provide a lower bound for Lipschitz bandits with heavy-tailed rewards, and show that our algorithms are optimal in terms of $T$. Finally, we conduct numerical experiments to demonstrate the effectiveness of our algorithms.

## 1. Introduction

The multi-armed bandits (MAB) is a powerful framework for modeling sequential decision-making under uncertainty, and has found applications in various areas such as medical trials (Robbins, 1952), news recommendation (Li et al., 2010), and network routing (Kveton et al., 2015). In the classic stochastic MAB, there are $K$ independent arms and a learner. In each round, the learner chooses one of $K$ arms to play and then obtains a reward drawn i.i.d. over time from a fixed but unknown probability distribution associated with the chosen arm. In order to maximize his gain, the learner has to balance the trade-off between exploration and exploitation, i.e., pulling the less pulled arms to acquire more information while playing the seemingly optimal arms to obtain more reward. For this problem, the standard metric is regret, defined as the difference between the cumulative reward of the learner and that of the best arm in hindsight. In their seminal paper, Lai & Robbins (1985) established an $\Omega(K \log T)$ lower bound on regret, and various algorithms matching this lower bound have been developed (Lai & Robbins, 1985; Agrawal, 1995b; Auer et al., 2002a).

One limitation of the stochastic MAB is that the lower bound scales linearly with the number of arms $K$, and thus deteriorates as $K$ goes large and becomes vacuous when the arm set is infinite. Another limitation is the fully independent setting on arms which is too pessimistic, since in many real-world scenarios the expected rewards of different arms could be related. To address these limitations, Kleinberg et al. (2008a) and Bubeck et al. (2009) introduced a variant of the stochastic MAB—Lipschitz bandits, which admits infinite arm set and models the relation between the expected rewards of different arms through a Lipschitz function. More precisely, in this setting the arm set $\mathcal{X}$ can be from any metric space $(\mathcal{X}, \mathcal{D})$, and each time after pulling an arm $x \in \mathcal{X}$, the learner receives a reward $y$ sampled independently from some distribution $\mathbb{P}_x$ satisfying

$$\mathbb{E}[y|x] = \mathbb{E}_{\mathbb{P}_x}[y] = \mu(x)$$

where $\mu$ is called the expected reward function and is Lipschitz with respect to the metric $\mathcal{D}$, i.e.,

$$|\mu(u) - \mu(v)| \leq \mathcal{D}(u, v), \ \forall u, v \in \mathcal{X}.$$

The goal of the learner is to minimize the (pseudo) regret:

$$R(T) = T \max_{x \in \mathcal{X}} \mu(x) - \sum_{t=1}^{T} \mu(x_t)$$

where $x_t$ is the arm chosen by the learner in round $t$.

While the Lipschitz bandits has been extensively studied in the literature (Bubeck et al., 2011; Kleinberg et al., 2013; Slivkins, 2014; Magureanu et al., 2014; Locatelli & Carpentier, 2018), most of existing work either assume the rewards

[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China [2]YouKu Cognitive and Intelligent Lab, Alibaba Group, Beijing 100102, China. Correspondence to: Lijun Zhang <zhanglj@lamda.nju.edu.cn>.

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

are bounded or require the sub-Gaussian properties of reward distributions , i.e., there exists a constant $a > 0$ such that for all $x \in \mathcal{X}$,

$$\mathbb{E}[e^{\lambda(y-\mathbb{E}[y|x])}] \leq e^{a\lambda^2/2}, \ \forall \lambda > 0.$$

However, in many real-life problems, such as financial markets (Cont & Bouchaud, 2000) and web advertising (Park et al., 2013), the rewards fluctuate sharply and do not behave bounded or sub-Gaussian but follow heavy-tailed distributions (Foss et al., 2011), i.e.,

$$\lim_{a\to\infty} \Pr(y > a|x) \cdot e^{\lambda a} = \infty, \ \forall \lambda > 0.$$

Till now, we have very limited knowledge on Lipschitz bandits with heavy-tailed rewards. One was given by Kleinberg et al. (2008b), who demonstrated that their algorithm enjoys a regret bound of $\widetilde{O}\left(T^{(3d_z+5)/(3d_z+6)}\right)$[1] under the assumption that the heavy-tailed rewards have finite third moments, where $d_z$ is the zooming dimension to be defined in Section 3.1. However, this result suffers the following limitations. First, the assumption on rewards is too stringent since many heavy-tailed distributions have infinite variance and hence do not admit finite third moments, such as Pareto distributions with shape parameter $\alpha \in (1, 2]$ and Student's $t$-distributions with degrees-of-freedom parameter $\gamma \in (1, 2]$. Second, the algorithm is far away from optimal as there exists a large gap between the regret bound and the lower bound established in this paper.

To address these drawbacks, we only assume the reward distributions have finite $(1 + \epsilon)$-th moments for some $\epsilon \in (0, 1]$ , i.e., there exists a constant $\nu$ such that for all $x \in \mathcal{X}$,

$$\mathbb{E}[|y|^{1+\epsilon}|x] \leq \nu \tag{1}$$

which is a common assumption in bandits learning with heavy-tailed rewards (Bubeck et al., 2013; Medina & Yang, 2016; Shao et al., 2018). Under this milder assumption, we propose two algorithms that attain a sublinear regret of $\widetilde{O}\left(T^{(d_z\epsilon+1)/(d_z\epsilon+\epsilon+1)}\right)$, and derive a lower bound matching our upper bound. To the best of our knowledge, they are the first optimal algorithms for Lipschitz bandits with heavy-tailed rewards. Both algorithms adopt adaptive discretization procedures to exploit the Lipschitz property of the expected reward function. To handle heavy-tailed rewards, one of our algorithms conducts a dynamic truncation on the observed reward, and the other algorithm makes ues of a median of means estimator.

## 2. Related Work

In this section, we briefly review the related work.

[1]We use the $\widetilde{O}$ notation to hide constant factors as well as polylogarithmic factors in $T$.

### 2.1. Lipschitz Bandits

In his seminal work, Agrawal (1995a) studied a special case of Lipschitz bandits under the name of "continuum-armed bandits" in which the arm set is an interval on the line (say, $\mathcal{X} = [0, 1]$), and developed an allocation-based algorithm that achieves an $O(T^{3/4})$ regret. Along this line of study, Kleinberg & Leighton (2003) established an $\Omega(T^{1/2})$ lower bound. Later, Kleinberg (2005) improved the lower bound to $\Omega(T^{2/3})$, and proposed a discretization-based method that enjoys a nearly optimal regret of $\widetilde{O}(T^{2/3})$.

One of the first papers that investigated Lipschitz bandits on general metric spaces was due to Kleinberg et al. (2008a). They proposed the zooming algorithm which enjoys a regret of $\widetilde{O}\left(T^{(d_z+1)/(d_z+2)}\right)$ for bounded rewards. In the full version of this paper (Kleinberg et al., 2008b), they extended the analysis to classes of heavy-tailed reward distributions with finite third moments and derived an $\widetilde{O}\left(T^{(3d_z+5)/(3d_z+6)}\right)$ regret. Another concurrent and independent work was given by Bubeck et al. (2009). They proposed a tree-based algorithm and demonstrated it attains a regret of $\widetilde{O}\left(T^{(d_z+1)/(d_z+2)}\right)$ for sub-Gaussian rewards. By further assuming the expected reward function has a finite number of maxima and is smooth around each maxima, they derived a regret bound of $\widetilde{O}(\sqrt{T})$. Similar regret bounds in which the order on $T$ is dimensionality-free also exist if the expected reward function is endowed with some additional curvatures. This line of research, ranging from specific to general, include linear bandits (Auer, 2002; Dani et al., 2008; Abbasi-Yadkori et al., 2011), generalized linear bandits (Filippi et al., 2010; Zhang et al., 2016; Jun et al., 2017), and convex bandits (Agarwal et al., 2011).

### 2.2. Learning with Heavy-tailed Distributions

There exists a rich body of work on learning with heavy-tailed distribution (Audibert et al., 2011; Catoni, 2012; Hsu & Sabato, 2014; Brownlees et al., 2015; Hsu & Sabato, 2016; Zhang & Zhou, 2018). For brevity, below we only discuss the related work in bandits literature. Liu & Zhao (2011) firstly relaxed the sub-Gaussian assumption on rewards to allow heavy-tailed reward distributions that have finite moments of order $1 + \epsilon$ for some $\epsilon \in (0, 1]$. They studied the stochastic MAB setting and proposed an algorithm based on a deterministic sequencing of exploration and exploitation, which attains a polynomial regret of $O(T^{1/(1+\epsilon)})$. Later, Bubeck et al. (2013) derived the first logarithmic regret of $O\left(\sum_{\Delta_i>0} \Delta_i^{-1/\epsilon} \log T\right)$ through combining UCB policies with robust mean estimators, where $\Delta_i$ is the gap between the expected reward of the $i$-th arm and that of the optimal arm. They also provided an matching lower bound. Medina & Yang (2016) extended the analysis to linear bandits and derived sublinear regret bounds, which are subsequently improved to be nearly optimal by Shao et al. (2018).

# 3. Lipschitz Bandits with Heavy-tailed Rewards

In this section, we first review necessary preliminaries, then present our algorithms as well as their theoretical guarantees, and finally propose a matching lower bound.

## 3.1. Preliminaries

Following previous studies (Kleinberg et al., 2013; Slivkins, 2014), we introduce some notions which are crucial to designing algorithms and analyzing lower bounds for Lipschitz bandits. Let $\mathcal{X}$ be an arm set and $\mathcal{D}$ be a metric on it. We assume $\mathcal{X}$ is compact and without loss of generality the diameter of $\mathcal{X}$ is not more than 1, i.e.,

$$\sup_{u,v\in\mathcal{X}} \mathcal{D}(u,v) \le 1. \tag{2}$$

Let $\mathcal{B}(x_0, r_0)$ denote a (closed) ball of radius $r_0 > 0$ centered at $x_0 \in \mathcal{X}$, defined by

$$\mathcal{B}(x_0, r_0) = \{x \in \mathcal{X} : \mathcal{D}(x, x_0) \le r_0\}.$$

We say a collection of balls is an $r$-covering of $\mathcal{X}$ if the radius of each ball is not more than $r$ and the union of these balls covers $\mathcal{X}$. The $r$-covering number of $\mathcal{X}$ is defined as the minimal number of balls in an $r$-covering of $\mathcal{X}$:

$$N_c(r) = \min\{|S| : S \text{ is an } r\text{-covering of } \mathcal{X}\}.$$

The covering dimension of $\mathcal{X}$ is defined as the minimal $d \ge 0$ such that $N_c(r)$ is $O(r^{-d})$ for any $r > 0$:

$$d_c = \min\{d \ge 0 : \exists a > 0, N_c(r) \le ar^{-d}, \forall r > 0\}. \tag{3}$$

We also define the covering constant $C$ of $\mathcal{X}$ as the minimal $a$ to make the above inequality true, i.e.,

$$C = \min\{a \ge 0 : N_c(r) \le ar^{-d_c}, \forall r > 0\}. \tag{4}$$

Another useful notion is the doubling constant of $\mathcal{X}$ which is denoted by $M$ and defined as the smallest $k$ such that any ball of radius $r$ in $\mathcal{X}$ can be covered by not more than $k$ balls of half the radius.

While the above notions only focus on the arm set $\mathcal{X}$, the following notions, originated from Kleinberg et al. (2008a), take the expected reward function $\mu$ into account as well. Let $x_* \in \arg\max_{x\in\mathcal{X}} \mu(x)$ be an optimal arm and denote by $\mathcal{X}_r$ the $r$-optimal region, defined as

$$\mathcal{X}_r = \{x \in \mathcal{X} : r/2 < \mu(x_*) - \mu(x) \le r\}.$$

Then, we define the $r$-zooming number of $(\mathcal{X}, \mu)$ as the minimal number of balls of radius not more than $r/18$ required to cover $\mathcal{X}_r$, denoted by $N_z(r)$. Here the constant 18 is due to technical reasons (see Appendix G). Next, we define the zooming dimension $d_z$ and the zooming constant $Z$ of $(\mathcal{X}, \mu)$ in the same way as in (3) and (4) respectively:

$$d_z = \min\{d \ge 0 : \exists a > 0, N_z(r) \le ar^{-d}, \forall r > 0\}; \tag{5}$$

$$Z = \min\{a \ge 0 : N_z(r) \le ar^{-d_z}, \forall r > 0\}. \tag{6}$$

Finally, we would like to discuss the relation between the covering dimension and the zooming dimension. On one hand, since the $r$-optimal region $\mathcal{X}_r$ is a subset of $\mathcal{X}$, it can be covered by not more than $N_c(r)$ balls of radius at most $r$ and hence $M^5 N_c(r)$ balls of radius at most $r/18$, which implies $N_z(r) \le M^5 N_c(r)$ and $d_z \le d_c$. On the other hand, in some benign cases, the zooming dimension can be much smaller than the covering dimension, as indicated by Kleinberg et al. (2008a). For example, let $\mathcal{X}$ be a $d$-dimensional ball of diameter 1 in the Euclidean space and set $\mu(x) = 1 - \|x - x_*\|_2$, where $x_* \in \mathcal{X}$ is the optimal arm. Then, for any $r > 0$, the $r$-optimal region $\mathcal{X}_r = \{x \in \mathcal{X} : r/2 < \|x - x_*\|_2 \le r\}$ can be covered by not more than $M^5$ balls of radius at most $r/18$, which implies the zooming dimension of $(\mathcal{X}, \mu)$ is 0, whereas the covering dimension of $\mathcal{X}$ is $d$ as shown by Engelking (1978).

## 3.2. Warm-up: Static Discretization

The basic idea behind existing algorithms for Lipschitz bandits is to exploit the Lipschitz property of the expected reward function by discretizing the arm set, and there exist two types of discretization schemes, namely static discretization and adaptive discretization. While our proposed algorithms that achieve optimal regret bound are inspired by the zooming algorithm (Kleinberg et al., 2008a) in which an adaptive discretization procedure is employed, to help understanding, we start with a suboptimal but simple algorithm that is based on static discretization.

Recall the definition of the $r$-covering number $N_c(r)$, which tells us that for any $r > 0$, the arm set $\mathcal{X}$ can be covered by not more than $N_c(r)$ balls of radius at most $r$. While the implementation of finding such balls depends on the specific geometry of the arm set, we assume access to an oracle which takes input of an arm set $\mathcal{X}$ and a resolution parameter $r$ and outputs $K$ balls of radius not more than $r$ covering $\mathcal{X}$ with $K \le N_c(r)$. Equipped with this oracle, we are now ready to describe the algorithm. Firstly, we query the oracle with $\mathcal{X}$ and $r$ and receive $K$ balls $\mathcal{B}_1, \ldots, \mathcal{B}_K$. Then, we partition the arm set $\mathcal{X}$ by constructing

$$\mathcal{X}_i = \mathcal{B}_i \cap \mathcal{X} - \cup_{s\in[i-1]} \mathcal{X}_s, i \in [K].^2$$

In this way, we can ensure that $\mathcal{X}_1, \ldots, \mathcal{X}_K$ are mutually disjoint and their union is exactly $\mathcal{X}$. Next, for each $\mathcal{X}_i$, we pick an arbitrary arm $\bar{x}_i$ from it. We refer to these arms

[2] We denote $\{1, 2, \ldots, n\}$ by $[n]$ for $n \in \mathbb{N}_+$ and use the convention that $[0] = \varnothing$.

$\bar{x}_1, \ldots, \bar{x}_K$ as skeleton arms, as they essentially constitute a skeleton of the arm set $\mathcal{X}$ in the sense that for any arm $x \in \mathcal{X}$, there must exist an $\mathcal{X}_i$ such that $x \in \mathcal{X}_i \subseteq \mathcal{B}_i$ and thus $\mathcal{D}(x, \bar{x}_i) \leq 2r$. Finally, we reduce the problem to a $K$-armed bandits problem on the skeleton arms and employ algorithms for multi-armed bandits.

However, since the rewards of these skeleton arms follow distributions that are not necessarily sub-Gaussian, classic UCB algorithms that are built upon on the empirical mean estimator do not apply. To address this problem, we adopt UCB policies with the truncated mean estimator, which converges to the mean of reward even under heavy-tailed distributions (Bubeck et al., 2013). Specifically, in the first $K$ rounds, we play each skeleton arm once. After that, let $n_t(\bar{x}_i)$ be the count of times the skeleton arm $\bar{x}_i$ has been pulled up to round $t$ and denote by $y_{i,s}$ the reward observed after the $s$-th playing of $\bar{x}_i$. In each round $t > K$, we first compute the average truncated reward of each skeleton arm $\bar{x}_i, i \in [K]$ as follows

$$\widehat{\mu}_t(\bar{x}_i) = \frac{1}{n_{t-1}(\bar{x}_i)} \sum_{s=1}^{n_{t-1}(\bar{x}_i)} y_{i,s} \mathbb{1}_{|y_{i,s}| \leq \left(\frac{\nu s}{2 \log t}\right)^{1/(1+\epsilon)}} \quad (7)$$

where $\nu$ is defined in (1) and the computation can be performed incrementally to reduce the time complexity. Then, following the principle of "optimism in the face of uncertainty", we play arm $x_t$ that achieves the highest average truncated reward plus a confidence term:

$$x_t = \arg\max_{\bar{x}_i} \widehat{\mu}_t(\bar{x}_i) + 4\nu^{\frac{1}{1+\epsilon}} \left(\frac{2 \log t}{n_{t-1}(\bar{x}_i)}\right)^{\frac{\epsilon}{1+\epsilon}} \quad (8)$$

in which ties are broken arbitrarily. The above procedure is summarized in Algorithm 1, and is referred to as Static Discretization with Truncated Mean (SDTM).

It remains to tune the value of the parameter $r$ of SDTM. To this end, we analyze the relation between the regret of the algorithm and the value of $r$ as follows.

**Theorem 1** *Assume (1) and (2) hold. For sufficiently large $T$ such that*

$$\log T \geq \frac{5}{8}(4\nu)^{-\frac{1}{\epsilon}}$$

*the regret of SDTM with parameter $r > 0$ satisfies*

$$\mathbb{E}[R(T)] \leq 2rT + (4\nu T)^{\frac{1}{1+\epsilon}} (16 N_c(r) \log T)^{\frac{\epsilon}{1+\epsilon}}$$

*where $N_c(r)$ is the $r$-covering number of the arm set $\mathcal{X}$.*

Here, the first term in the regret bound stems from the gap between the expected reward of the optimal skeleton arm and that of the optimal arm in $\mathcal{X}$, and the second term is incurred by not playing the optimal skeleton arm. It is easy to see that as one term falls the other term rises. To obtain a tight bound, we substitute $N_c(r) \leq C r^{-d_c}$ into the RHS of the second inequality in Theorem 1 and minimize it over $r$, from which we derive an optimal configuration of $r$:

$$r = \left(\frac{(4\nu T)^{1/(1+\epsilon)} (16 C \log T)^{\epsilon/(1+\epsilon)} d_c \epsilon}{2T(1+\epsilon)}\right)^{\frac{\epsilon+1}{d_c\epsilon+\epsilon+1}} \quad (9)$$

and the following corollary:

**Corollary 1** *Let $r$ be configured as in (9). We have*

$$\mathbb{E}[R(T)] \leq \widetilde{O}\left(T^{\frac{d_c\epsilon+1}{d_c\epsilon+\epsilon+1}}\right)$$

*in which $d_c$ is the covering dimension of the arm set $\mathcal{X}$, defined in (3).*

**Algorithm 1** Static Discretization with Truncated Mean (SDTM)

**Require:** resolution parameter $r > 0$
Query the oracle with $\mathcal{X}$ and $r$
Receive $K$ balls $\mathcal{B}_1, \ldots, \mathcal{B}_K$ from the oracle
**for** $i = 1, 2, \ldots, K$ **do**
  Construct $\mathcal{X}_i = \mathcal{B}_i \cap \mathcal{X} - \cup_{s \in [i-1]} \mathcal{X}_s$
  Pick an arbitrary arm $\bar{x}_i$ from $\mathcal{X}_i$
**end for**
**for** $t = 1, 2, \ldots, K$ **do**
  Play arm $\bar{x}_t$
  Observe reward $y_t$
**end for**
**for** $t = K+1, K+2, \ldots, T$ **do**
  Compute truncated means of $\bar{x}_i, i \in [K]$ as in (7)
  Play arm $x_t$ defined by (8)
  Observe reward $y_t$
**end for**

### 3.3. Improved Method: Adaptive Discretization

While SDTM is simple, it has an obvious limitation: the arm set is discretized before the learning process and the discretization is fixed during the execution of the algorithm. Intuitively, an improvement can be obtained by adaptively discretizing the arm set based on the rewards observed over time. As we shall see, such adaptive approaches, while still suffering an $\widetilde{O}\left(T^{(d_c\epsilon+1)/(d_c\epsilon+\epsilon+1)}\right)$ regret in the worst case (i.e., $d_z = d_c$), can attain much tighter regret bound for benign cases (i.e., $d_z < d_c$).

#### 3.3.1. Adaptive Discretization with Truncated Mean

Our first adaptive algorithm is called Adaptive Discretization with Truncated Mean (ADTM) and is outlined in Algorithm 2, where $\delta$ is a confidence parameter. Inspired by the zooming algorithm (Kleinberg et al., 2008a), we maintain

**Algorithm 2** Adaptive Discretization with Truncated Mean (ADTM)

**Require:** confidence parameter $\delta \in (0, 1/2)$
Initialize $\mathcal{A} = \varnothing$
**for** $t = 1, 2, \ldots, T$ **do**
  **if** $\mathcal{X} \nsubseteq \cup_{x\in\mathcal{A}} \mathcal{B}(x, r_t(x))$ **then**
    Pick an arbitrary arm $x$ from the uncovered region $\mathcal{X} - \cup_{x\in\mathcal{A}} \mathcal{B}(x, r_t(x))$
    Add $x$ into $\mathcal{A}$
    Initialize $\widehat{\mu}(x) = 0$ and $n(x) = 0$
    Play arm $x_t = x$
  **else**
    Play arm $x_t = \arg\max_{x\in\mathcal{A}} \widehat{\mu}(x) + 2r_t(x)$
  **end if**
  Observe reward $y_t$
  Update

$$\widehat{\mu}(x_t) = \frac{n(x_t)\widehat{\mu}(x_t) + y_t \cdot \mathbb{1}_{|y_t| \le \left(\frac{\nu n(x_t)+\nu}{\log(T^2/\delta)}\right)^{\frac{1}{(1+\epsilon)}}}}{n(x_t)+1}$$

  and $n(x_t) = n(x_t) + 1$
**end for**

an active arm set $\mathcal{A} \subseteq \mathcal{X}$ as the discretization of $\mathcal{X}$, which is initialized to be $\varnothing$ and is updated in the beginning of each round. Each arm $x$ in this set is called active arm and is associated with a time varying confidence radius:

$$r_t(x) = 4\bar{\nu}^{\frac{1}{1+\epsilon}} \left(\frac{\log(T^2/\delta)}{n(x)}\right)^{\frac{\epsilon}{1+\epsilon}} \tag{10}$$

in which

$$\bar{\nu} = \max\left(\nu, \left(12\sqrt{2}\right)^{-(1+\epsilon)}\right) \tag{11}$$

with $\nu$ defined in (1), and $n(x)$ is the count of times an arm $x$ has been played before round $t$.[3] During the whole learning process, we only play active arms.

We now describe the algorithm in detail. In each round $t$, we first check whether the union of balls defined by pairs of active arms and their confidence radius covers the arm set, i.e., $\mathcal{X} \subseteq \cup_{x\in\mathcal{A}} \mathcal{B}(x, r_t(x))$. If it is false, we pick an arbitrary arm $x$ from the uncovered region $\mathcal{X} - \cup_{x\in\mathcal{A}} \mathcal{B}(x, r_t(x))$, designate it as an active arm (i.e., add $x$ into $\mathcal{A}$), and play this arm. Its average truncated reward $\widehat{\mu}(x)$ and the count of times it has been played $n(x)$ are both initialized to be 0. Otherwise, we select arm $x_t$ from the active arm set $\mathcal{A}$ with the highest sum of average truncated reward and double confidence radius to play. Finally, after observing reward $y_t$, we update the average truncated reward of $x_t$ and the count of times it has been played.

[3] Here $n(x)$ in fact varies with $t$, and the reason for using $n(x)$ rather than $n_t(x)$ is to be consistent with the pseudocode of Algorithm 2.

It would be helpful to compare ADTM with the simple static algorithm SDTM in Section 3.2. Both algorithms employ the truncated mean estimator and adopt the UCB policy, and the main difference between them lies in the discretization procedure. While SDTM conducts static discretization by querying an oracle, ADTM discretizes the arm set adapting to the rewards observed over time. This adaptive discretization approach leads to a tighter regret bound for ADTM, which depends on the zooming dimension $d_z$ instead of the covering dimension $d_c$, as shown by the following theoretical results.

**Theorem 2** *Assume (1) and (2) hold. With probability at least $1 - 2\delta$, the regret of ADTM satisfies*

$$R(T) \le \inf_{r_0\in(0,1)} \left( r_0 T + 17(34\bar{\nu})^{\frac{1}{\epsilon}} \log(T^2/\delta) \sum_{r=2^{-i}:i\in\mathbb{N},r\ge r_0} \frac{N_z(r)}{r^{1/\epsilon}} \right)$$

*where $\bar{\nu}$ is defined in (11) and $N_z(r)$ is the $r$-zooming number of $(\mathcal{X}, \mu)$.*

Substituting $N_z(r) \le Zr^{-d_z}$ into the above inequality, we obtain the following corollary in which the order on $T$ in regret bound is explicitly given.

**Corollary 2** *We have*

$$\sum_{r=2^{-i}:i\in\mathbb{N},r\ge r_0} \frac{N_z(r)}{r^{1/\epsilon}} \le O\left(r_0^{-(d_z+1/\epsilon)}\right)$$

*and thus*

$$R(T) \le O\left(\inf_{r_0\in(0,1)} \left(r_0 T + \log T \cdot r_0^{-(d_z+1/\epsilon)}\right)\right) \le \widetilde{O}\left(T^{\frac{d_z\epsilon+1}{d_z\epsilon+\epsilon+1}}\right)$$

*where $d_z$ is the zooming dimension of $(\mathcal{X}, \mu)$, defined in (5).*

### 3.3.2. Adaptive Discretization with Median of Means

Our second adaptive algorithm is called Adaptive Discretization with Median of Means (ADMM), which follows the general framework of the first one but takes a different median of means estimator (MME). As outlined in Algorithm 3, given $n$ observed rewards of an arm $x$, MME divides these rewards into $M$ groups, computes empirical mean within each group, and returns the median of these empirical means. For MME, we only require that for some $\epsilon \in (0, 1]$, the central $(1 + \epsilon)$-th moments of the reward distributions are bounded, i.e., there exists a constant $\sigma > 0$ such that

$$\mathbb{E}[|y - \mathbb{E}[y]|^{1+\epsilon}|x] \le \sigma, \ \forall x \in \mathcal{X}. \tag{12}$$

The theoretical analysis of MME (Bubeck et al., 2013) indicates that the concentration bound of MME holds only for $n \geq 16 \log\left(e^{1/8}T^2/\delta\right)$.

To this end, as shown in Algorithm 4, we introduce a flag variable called replay and keep it true until an arm has been played at least $16 \log\left(e^{1/8}T^2/\delta\right)$ times. Furthermore, for the sake of computing MME, for each active arm $x$ in $\mathcal{A}$ we use $\mathcal{H}(x)$ to store historical rewards of $x$. Finally, due to the difference between the concentration bound of MME and that of the truncated mean estimator, we modify the definition of the confidence radius $r_t(x)$ in (10) slightly:

$$r_t(x) = (12\bar{\sigma})^{\frac{1}{1+\epsilon}} \left(\frac{16 \log\left(e^{1/8}T^2/\delta\right)}{n(x)}\right)^{\frac{\epsilon}{1+\epsilon}}$$

in which

$$\bar{\sigma} = \max\left(\sigma, \left(36\sqrt{2}\right)^{-1}\right) \tag{13}$$

with $\sigma$ defined in (12), and $n(x)$ is the count of times an arm $x$ has been pulled before round $t$.

Similar to the regret bounds of ADTM in Theorem 2 and Corollary 2, we have the following theorem regarding the regret of ADMM.

**Theorem 3** *Assume (2) and (12) hold. With probability at least $1-2\delta$, the regret of ADMM satisfies*

$$R(T) \leq \inf_{r_0 \in (0,1)} \Bigg( r_0 T + 68(102\bar{\sigma})^{\frac{1}{\epsilon}} \log\left(e^{1/8}T^2/\delta\right) \sum_{r=2^{-i}: i\in\mathbb{N}, r\geq r_0} \frac{N_z(r)}{r^{1/\epsilon}} \Bigg)$$

*where $\bar{\sigma}$ is defined in (13) and $N_z(r)$ is the $r$-zooming number of $(\mathcal{X}, \mu)$. Furthermore, by the first inequality in Corollary 2 we have*

$$R(T) \leq \widetilde{O}\left(T^{\frac{d_z\epsilon+1}{d_z\epsilon+\epsilon+1}}\right)$$

*where $d_z$ is the zooming dimension of $(\mathcal{X}, \mu)$, defined in (5).*

**Remark.** The regret bound of ADMM depends on the central $(1+\epsilon)$-th moments of the reward distributions, at the cost of a worse constant factor in the leading term compared to Theorem 2. The central moments are insensitive to constant shift changes in distributions and, in some cases, could be much smaller than the raw moments in the regret bound of ADTM in Theorem 2. However, while ADTM is an efficient algorithm, ADMM requires storing the learning history and is thus inefficient.

### 3.3.3. Comparison with the Zooming Algorithm

While our two adaptive algorithms are inspired by the zooming algorithm (Kleinberg et al., 2008a), we would like to emphasize a significant difference between our algorithms and the zooming algorithm. Specifically, once a new arm is added into the active arm set, our algorithms will play this arm immediately, which is not the case in the zooming algorithm. In the analysis of the zooming algorithm, at each round, one has to assume that the mean rewards of all arms are upper bounded by a known constant to prove Claim 2.2 in Kleinberg et al. (2008a) for those active arms that have never been played before this round. Note that this assumption is valid only for bounded rewards. If we directly extend the zooming algorithm to unbounded rewards (even if the rewards are sub-Gaussian), we need this additional assumption (e.g., Theorem 4.14 in Kleinberg et al. 2008b). In contrast, our algorithms ensure that at the beginning of each round, all active arms have been played at least once so as to avoid this assumption.

**Algorithm 3** Median of Means Estimator (MME)

**Require:** observed rewards $y_1, y_2, \ldots, y_n$, confidence $\frac{\delta}{T^2}$
Set $M = \lfloor 8 \log\left(e^{1/8}T^2/\delta\right) \rfloor$ and $B = \lfloor n/M \rfloor$
**for** $m = 1, 2, \ldots, M$ **do**
  Compute $\widehat{y}_m = \frac{1}{B}\sum_{i=(m-1)B+1}^{mB} y_i$
**end for**
Return the median of $(\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_M)$

**Algorithm 4** Adaptive Discretization with Median of Means (ADMM)

**Require:** confidence parameter $\delta \in (0, 1/2)$
Initialize $\mathcal{A} = \varnothing$, $replay = false$
**for** $t = 1, 2, \ldots, T$ **do**
  **if** $replay$ **then**
    Play arm $x_t = x_{t-1}$
  **else**
    **if** $\mathcal{X} \not\subseteq \cup_{x\in\mathcal{A}} \mathcal{B}(x, r_t(x))$ **then**
      Pick an arbitrary arm $x$ from the uncovered region $\mathcal{X} - \cup_{x\in\mathcal{A}} \mathcal{B}(x, r_t(x))$
      Add $x$ into $\mathcal{A}$
      Initialize $n(x) = 0$, $\mathcal{H}(x) = \varnothing$
      Play arm $x_t = x$
    **else**
      Play arm $x_t = \arg\max_{x\in\mathcal{A}} \widehat{\mu}(x) + 2r_t(x)$
    **end if**
  **end if**
  Observe reward $y_t$ and add $y_t$ into $\mathcal{H}(x_t)$
  Update $n(x_t) = n(x_t) + 1$
  **if** $n(x_t) < 16 \log\left(e^{1/8}T^2/\delta\right)$ **then**
    Set $replay = true$
  **else**
    Set $replay = false$
    Update $\widehat{\mu}(x_t) = \text{MME}(\mathcal{H}(x_t), \delta/T^2)$
  **end if**
**end for**

### 3.4. Lower Bound

Finally, we show that our adaptive algorithms are optimal in terms of $T$ by establishing the following lower bound.

**Theorem 4** *Fix an arm set $\mathcal{X}$ with diameter 1 and a parameter of moment $\epsilon \in (0,1]$. Define $\kappa = \frac{2^{1/\epsilon}\cdot\epsilon}{\log 2}$ and*

$$R_c(T) = \inf_{r_0\in(0,1)} \left( r_0 T + \log T \sum_{r=2^{-i}:i\in\mathbb{N},r\geq r_0} \frac{N_c(r)}{r^{1/\epsilon}} \right)$$

*where $N_c(r)$ is the $r$-covering number of $\mathcal{X}$. Then, for any $T > 2$ and any positive number $R \leq R_c(T)$, there exists a set $\mathcal{I}$ of problem instances on $\mathcal{X}$ such that*
(i) *for each problem instance $I \in \mathcal{I}$, define*

$$R_z(T) = \inf_{r_0\in(0,1)} \left( r_0 T + \log T \sum_{r=2^{-i}:i\in\mathbb{N},r\geq r_0} \frac{N_z(r)}{r^{1/\epsilon}} \right)$$

*in which $N_z(r)$ is the $r$-zooming number of $I$. We have $R_z(T) \leq 3R/(8\kappa \log T)$.*
(ii) *for any algorithm $\mathcal{A}$, there exists at least one problem instance $I \in \mathcal{I}$ on which the expected regret of $\mathcal{A}$ satisfies $\mathbb{E}[R(T)] \geq R/(2560\kappa \log T)$.*

**Remark.** The above theorem essentially establishes an $\Omega(R_z(T))$ lower bound on expected regret suffered by any algorithm, which matches the $O(R_z(T))$ regret bounds in Theorems 2 and 3 up to constant factors. While there exist lower bounds of Lipschitz bandits for sub-Gaussian rewards (Slivkins, 2014), to the best of our knowledge, this is the first lower bound for heavy-tailed rewards.

## 4. Analysis

Due to the limitation of space, we only prove Theorem 2 and the omitted proofs can be found in the supplementary material.

### 4.1. Proof of Theorem 2

**Notations.** Note that the active arm set $\mathcal{A}$ is time varying, and for each active arm $x \in \mathcal{A}$, its average truncated reward $\widehat{\mu}(x)$ and the count of times it has been pulled $n(x)$ also change after being played. For convenience, we use $\mathcal{A}_t$, $\widehat{\mu}_t(x)$, and $n_t(x)$ to denote the value of $\mathcal{A}$, $\widehat{\mu}(x)$, and $n(x)$ in the end of the $t$-th round respectively.

Following Kleinberg et al. (2008a), we first propose to bound the distance between the mean of reward $\mu(x)$ and the average truncated reward $\widehat{\mu}_t(x)$ for each active arm $x$.

**Lemma 1** *With probability at least $1-2\delta$, for all rounds $t \in [T]$ and all active arms $x \in \mathcal{A}_t$, we have*

$$|\widehat{\mu}_t(x) - \mu(x)| \leq r_{t+1}(x).$$

Let $x_* \in \arg\max_{x\in\mathcal{X}} \mu(x)$ be an optimal arm and $\Delta(x) = \mu(x_*) - \mu(x)$ be the gap between the expected reward of an arm $x$ and that of the optimal arm. The following lemma shows that this gap can be upper bounded by the confidence radius of $x$ up to constant factor.

**Lemma 2** *With probability at least $1-2\delta$, for all rounds $t \in [T]$ and all active arms $x \in \mathcal{A}_t$, we have*

$$\Delta(x) \leq 3\sqrt{2} r_{t+1}(x).$$

**Remark.** Recall that our Adaptive Discretization with Truncated Mean algorithm uses balls centered at active arms to cover the arm set. The above lemma tells us that the radius of the ball centered at active arm $x$ is lower bounded by $\Delta(x)/(3\sqrt{2})$. Roughly speaking, this implies that our algorithm uses more balls (of smaller radius) to discretize near-optimal arms and less balls (of larger radius) to cover poor arms.

We proceed to prove Theorem 2 and partition the set comprised of all active arms in $\mathcal{A}_T$ that are suboptimal as follows

$$\{x \in \mathcal{A}_T : \Delta(x) > 0\} = \cup_{i=0}^{\infty} \bar{\mathcal{A}}_T(i)$$

in which

$$\bar{\mathcal{A}}_T(i) = \{x \in \mathcal{A}_T \mid 2^{-(i+1)} < \Delta(x) \leq 2^{-i}\}. \tag{14}$$

Then, we establish upper bounds for the cardinality of $\bar{\mathcal{A}}_T(i)$ by making use of Lemma 2.

**Lemma 3** *With probability at least $1-2\delta$, for all $i \in \mathbb{N}$,*

$$|\bar{\mathcal{A}}_T(i)| \leq N_z(2^{-i}).$$

Based on Lemmas 2 and 3, we can further bound the regret incurred by playing arms in $\bar{\mathcal{A}}_T(i)$.

**Lemma 4** *With probability at least $1-2\delta$, for all $i \in \mathbb{N}$,*

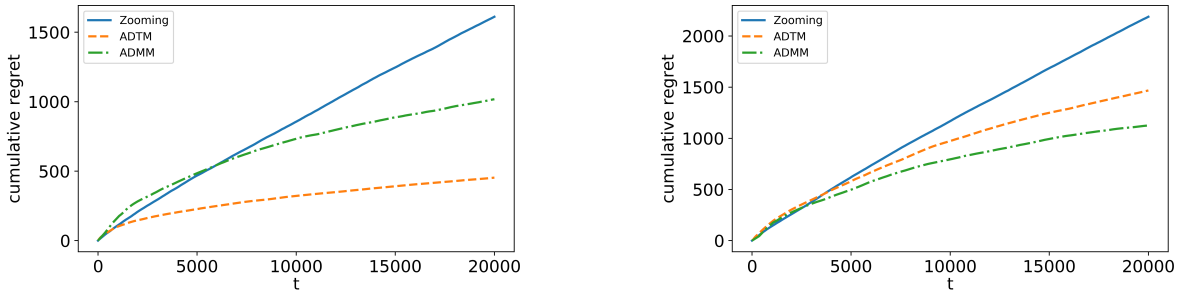
$$\sum_{x\in\bar{\mathcal{A}}_T(i)} n_T(x)\Delta(x) \leq 2^{\frac{i+1}{\epsilon}} \cdot 17^{\frac{\epsilon+1}{\epsilon}} \bar{\nu}^{\frac{1}{\epsilon}} \log(T^2/\delta) N_z(2^{-i}).$$

We are now ready to prove the theorem. For any $r_0 \in (0,1)$, we have

$$R(T) = T\mu(x_*) - \sum_{t=1}^{T} \mu(x_t) = \sum_{x\in\mathcal{A}_T} n_T(x)\Delta(x)$$
$$= \underbrace{\sum_{x\in\mathcal{A}_T:\Delta(x)\leq r_0} n_T(x)\Delta(x)}_{A} + \underbrace{\sum_{x\in\mathcal{A}_T:\Delta(x)>r_0} n_T(x)\Delta(x)}_{B}.$$

In the following, we bound $A$ and $B$ separately.
(i) $A$ can be bounded easily:

$$A \leq r_0 \sum_{x\in\mathcal{A}_T:\Delta(x)\leq r_0} n_T(x) \leq r_0 T. \tag{15}$$




(a) $a = 0$ (b) $a = -2$

*Figure 1.* Comparison of our algorithms versus the zooming algorithm for heavy-tailed rewards

(ii) To bound $B$, we make use of the definition of $\bar{\mathcal{A}}_T(i)$ in (14) and apply Lemma 4.

$$\begin{aligned} B \leq & \sum_{i\in\mathbb{N}:2^{-i}\geq r_0} \sum_{x\in\bar{\mathcal{A}}_T(i)} n_T(x)\Delta(x) \\ \leq & 17(34\bar{\nu})^{\frac{1}{\epsilon}} \log\left(T^2/\delta\right) \sum_{i\in\mathbb{N}:2^{-i}\geq r_0} N_z(2^{-i})2^{\frac{i}{\epsilon}} \\ = & 17(34\bar{\nu})^{\frac{1}{\epsilon}} \log\left(T^2/\delta\right) \sum_{r=2^{-i}:i\in\mathbb{N},r\geq r_0} \frac{N_z(r)}{r^{1/\epsilon}}. \end{aligned} \tag{16}$$

Combining (15) and (16), we finish the proof. □

## 5. Experiments

In this section, we provide numerical experiments to illustrate the performance of our proposed algorithms: ADTM and ADMM. For comparison, we adopt a type of zooming algorithm proposed in Section 4.5 of Kleinberg et al. (2008b) as the baseline, which requires the existence of finite third moments of reward distributions.

Following Magureanu et al. (2014), we set $\mathcal{X} = [0, 1]$ with $\mathcal{D}$ being the Euclidean metric on it, and choose

$$\mu(x) = a - \min(|x - 0.4|, |x - 0.8|)$$

as the expected reward function in which $a$ is a constant. To generate heavy-tailed rewards, we adopt the Pareto distribution with shape parameter $\alpha = 3.1$ so that the third moments of rewards are bounded and thus the zooming algorithm can apply. Specifically, each round after playing an arm $x$, the learner receives a stochastic reward $y$ satisfying $y = \mu(x) + \eta - \frac{\alpha}{\alpha-1}$ in which

$$\Pr(\eta|x) = \begin{cases} \frac{\alpha}{\eta^{\alpha+1}}, & \eta \geq 1 \\ 0, & \eta < 1 \end{cases}.$$

By straightforward calculations, we can bound the moments of different orders of $y$ as follows and configure confidence radius of each tested algorithm accordingly:

$$\mathbb{E}[y|x] = \mu(x) \leq \bar{a}, \qquad \mathbb{E}[|y|^3|x] \leq \bar{a}^3 + 1.92\bar{a} + 24.96$$
$$\mathbb{E}[|y|^2|x] \leq \nu = \bar{a}^2 + 0.64, \mathbb{E}[|y - \mu(x)|^2|x] \leq \sigma = 0.64.$$

where $\bar{a} = \max(|a|, |a - 0.4|)$. Then, following common practice (Zhang et al., 2016; Jun et al., 2017), we scale the confidence radius by a factor $c$ searched within $[1e - 2, 1]$.

We consider two cases: $a = 0$ and $a = -2$. For each case, we run 40 independent repetitions and report the average cumulative regret of each tested algorithm in Figure 1. As can be seen, not surprisingly, our adaptive algorithms outperform the zooming algorithm in both cases. In the first case, ADTM achieves the smallest regret which is expected, since compared to ADMM it has a more favorable constant factor in regret bound, and in this case the upper bound of second raw moments $\nu$ is only 1.25 times as large as that of second central moments $\sigma$. By contrast, in the second case, due to the significant ratio of $\nu/\sigma = 10$, ADTM suffers a larger regret and ADMM behaves the best. Finally, the curves of ADMM in both cases are nearly the same, which is consistent with Theorem 3 and supports the claim that ADMM is insensitive to constant shift changes in rewards.

## 6. Conclusion and Future Work

We have proposed two adaptive algorithms for Lipschitz bandits with heavy-tailed rewards. Our algorithms only require the existence of finite $(1 + \epsilon)$-th moments of rewards for some $\epsilon \in (0, 1]$, and are optimal in the sense that the regret bounds match the lower bound established in this paper. One of our algorithms is efficient but its regret deteriorates as the absolute bound of the expected reward function $\max_x |\mu(x)|$ increases. The other one enjoys a regret bound that depends on the central moments and is thus insensitive to constant shift changes in rewards but is inefficient. Therefore, a natural question arises: is there an efficient algorithm with regret bound depending on central moments? Obtaining such algorithm seems highly non-trivial even for the multi-armed setting, and we leave it as a future work.

## Acknowledgements

This work was partially supported by the NSFC-NRF Joint Research Project (61861146001), NSFC (61603177), JiangsuSF (BK20160658), and YESS (2017QNRC001). We thank the anonymous reviewers for their constructive suggestions.

## References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pp. 2312–2320, 2011.

Agarwal, A., Foster, D. P., Hsu, D. J., Kakade, S. M., and Rakhlin, A. Stochastic convex optimization with bandit feedback. In *Advances in Neural Information Processing Systems 24*, pp. 1035–1043, 2011.

Agrawal, R. The continuum-armed bandit problem. *SIAM Journal on Control and Optimization*, 33(6):1926–1951, 1995a.

Agrawal, R. Sample mean based index policies by $o(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995b.

Audibert, J.-Y., Catoni, O., et al. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a.

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

Brownlees, C., Joly, E., Lugosi, G., et al. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015.

Bubeck, S., Stoltz, G., Szepesvári, C., and Munos, R. Online optimization in $\mathcal{X}$-armed bandits. In *Advances in Neural Information Processing Systems 22*, pp. 201–208, 2009.

Bubeck, S., Stoltz, G., and Yu, J. Y. Lipschitz bandits without the lipschitz constant. In *Proceedings of the 22nd International Conference on Algorithmic Learning Theory*, pp. 144–158, 2011.

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

Catoni, O. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et Statistiques*, volume 48, pp. 1148–1185, 2012.

Cont, R. and Bouchaud, J.-P. Herd behavior and aggregate fluctuations in financial markets. *Macroeconomic Dynamics*, 4(2):170–196, 2000.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 1991.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Conference on Learning Theory*, pp. 355–366, 2008.

Engelking, R. *Dimension theory*. North-Holland Publishing Company Amsterdam, 1978.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems 23*, pp. 586–594, 2010.

Foss, S., Korshunov, D., and Zachary, S. *An introduction to heavy-tailed and subexponential distributions*, volume 6. Springer, 2011.

Hsu, D. and Sabato, S. Heavy-tailed regression with a generalized median-of-means. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 37–45, 2014.

Hsu, D. and Sabato, S. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(1):543–582, 2016.

Jun, K.-S., Bhargava, A., Nowak, R., and Willett, R. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems 30*, pp. 99–109, 2017.

Kleinberg, R. and Leighton, T. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, pp. 594–605, 2003.

Kleinberg, R., Slivkins, A., and Upfal, E. Multi-armed bandits in metric spaces. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pp. 681–690, 2008a.

Kleinberg, R., Slivkins, A., and Upfal, E. Multi-armed bandits in metric spaces. *arXiv preprint arXiv:0809.4882*, 2008b.

Kleinberg, R., Slivkins, A., and Upfal, E. Bandits and experts in metric spaces. *arXiv preprint arXiv:1312.1277*, 2013.

Kleinberg, R. D. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems 18*, pp. 697–704, 2005.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. Combinatorial cascading bandits. In *Advances in Neural Information Processing Systems 28*, pp. 1450–1458, 2015.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1): 4–22, 1985.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670, 2010.

Liu, K. and Zhao, Q. Multi-armed bandit problems with heavy-tailed reward distributions. In *Proceedings of the 49th Annual Allerton Conference on Communication, Control and Computing*, pp. 485–492, 2011.

Locatelli, A. and Carpentier, A. Adaptivity to smoothness in $\mathcal{X}$-armed bandits. In *Proceedings of the 31st Conference on Learning Theory*, pp. 1463–1492, 2018.

Magureanu, S., Combes, R., and Proutiere, A. Lipschitz bandits: Regret lower bounds and optimal algorithms. In *Proceedings of the 27th Conference on Learning Theory*, pp. 975–999, 2014.

Medina, A. M. and Yang, S. No-regret algorithms for heavy-tailed linear bandits. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 1642–1650, 2016.

Park, J. Y., Lee, K.-W., Kim, S. Y., and Chung, C.-W. Ads by whom? ads about what?: exploring user influence and contents in social advertising. In *Proceedings of the 1st ACM conference on Online Social Networks*, pp. 155–164, 2013.

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

Shao, H., Yu, X., King, I., and Lyu, M. R. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Advances in Neural Information Processing Systems 31*, pp. 8430–8439, 2018.

Slivkins, A. Contextual bandits with similarity information. *Journal of Machine Learning Research*, 15(1):2533–2568, 2014.

Zhang, L. and Zhou, Z.-H. $\ell_1$-regression with heavy-tailed distributions. In *Advances in Neural Information Processing Systems 31*, pp. 1076–1086, 2018.

Zhang, L., Yang, T., Jin, R., Xiao, Y., and Zhou, Z.-h. Online stochastic linear optimization under one-bit feedback. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 392–401, 2016.