No-Regret Algorithms for Heavy-Tailed Linear Bandits

Andres Munoz Medina AMMEDINA@GOOGLE.COM
Google Research, 111 8th Av, New York, NY 10011

Scott Yang YANGS@CIMS.NYU.EDU
Courant Institute, 251 Mercer Street, New York, NY 10012

Abstract

We analyze the problem of linear bandits under heavy tailed noise. Most of the work on linear bandits has been based on the assumption of bounded or sub-Gaussian noise. This assumption however is often violated in common scenarios such as financial markets. We present two algorithms to tackle this problem: one based on dynamic truncation and one based on a median of means estimator. We show that, when the noise admits only a $1+\epsilon$ moment, these algorithms are still able to achieve regret in $\widetilde{O}(T^{\frac{2+\epsilon}{2(1+\epsilon)}})$ and $\widetilde{O}(T^{\frac{1+2\epsilon}{1+3\epsilon}})$ respectively. In particular, they guarantee sublinear regret as long as the noise has finite variance. We also present empirical results showing that our algorithms achieve a better performance than the current state of the art for bounded noise when the L_∞ bound on the noise is large yet the $1+\epsilon$ moment of the noise is small.

1. Introduction

Sequential decision-making under limited feedback has become a classic topic in machine learning. Dating as far back as the classical work of (Robbins, 1952), "bandit problems" are a prime example of the exploration-exploitation trade-off that comes up in machine learning, and they have been analyzed, extended, and applied in many forms. In particular, bandit algorithms have been successfully used in tasks such as medical diagnosis, job scheduling, computational advertising, and repeated games.

In the original stochastic setting of (Robbins, 1952), at each time t, the learner selects an action out of a set of K possibilities. Each action i makes the learner incur a loss of $l_{i,t}$ which is a random variable with mean μ_i. The learner can only observe the loss for the chosen action and his objective is to minimize his regret with respect to the best action $i^* = \text{argmin}_i \mu_i$. That is, the seller minimizes the following quantity $\text{Reg} = \mathbb{E}[\sum_{t=1}^{T} \mu_{I_t} - \mu_i^*]$, where I_t is the action chosen by the learner.

There have been a plethora of extensions to the stochastic MAB problem of (Robbins, 1952), including to that of infinite action sets (Auer, 2002; Kleinberg, 2004), adversarial loss functions (Auer et al., 2003), and various types of additional structure (e.g. switching costs as in (Cesa-Bianchi et al., 2013) and contexts as in (Auer et al., 2003; Beygelzimer et al., 2011; Agarwal et al., 2014)). We cannot attempt to do a comprehensive literature review, so we refer the interested reader to the work of (Bubeck and Cesa-Bianchi, 2012) for an exposition of many of the latest advances.

Despite the large body of work in this field, one aspect of bandit problems that has remained largely ignored is the setting of heavy-tailed loss functions. In nearly all papers and settings proposed, the authors assume that the losses incurred by the learner are either bounded or at worst, sub-Gaussian (i.e. $\exists C \geq 0$ constant such that $\forall \lambda \in \mathbb{R}$, $\ln\left(\mathbb{E}\left[e^{\lambda(l-\mathbb{E}[l])}\right]\right) \leq \frac{1}{2}C\lambda^2$). The motivation behind this assumption is largely technical, as it is the most convenient relaxation of bounded loss functions that still allows the use of standard concentration of measure techniques (e.g. Hoeffding's inequality (Hoeffding, 1963)).

However, many real-life sequential decision-making problems do not exhibit bounded or sub-Gaussian losses. A prime example is that of financial markets, where heavy-tailed price fluctuations occur far more frequently than Gaussian models would predict (see e.g. (Rachev, 2003; Hull, 2012)). Another example can be found in auctions run to sell online advertisement where unusually large bids are seen, albeit with low probability.

Since the standard techniques cannot be directly applied, it is not clear whether one can attain equivalent regret bounds for such scenarios, or whether sublinear regret algorithms of any form are even possible at all.

A remarkable analysis of this scenario was given by (Bubeck et al., 2013) who consider the classic stochastic MAB setting with heavy-tailed noise. In their work, the

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

authors show that by using statistical estimators that are more robust than the empirical mean for each specific action, one can attain regret bounds of the same order as in the bounded-loss setting. This is an exceptional result since only a second moment is required to achieve this regret bound.

In this paper, we consider heavy-tailed losses in the scenario of stochastic linear bandits, a natural yet non-trivial extension of the MAB problem. Unlike the MAB problem, the linear bandit setting forces the learner to choose among and compare himself against an infinite number of actions. This makes the presence of heavy-tailed losses more difficult to manage.

A large body of work analyzing the linear bandit problem exists, most of which is based on the seminal work of (Auer, 2002). In this paper, the author builds confidence regions for the true model parameter and then optimistically selects the action minimizing the loss over these sets. However, the so called "associative reinforcement learning" scenario of (Auer, 2002) is actually a bit different from the linear bandit problem we present, since the learner there is constrained to only a finite number of actions. This is similarly the case for a number of other bandit problems with linear payoff functions, in particular for the "LinUCB" algorithm (Li et al., 2010; Chu et al., 2011). The setup we define allows for infinite action sets and is based upon the stochastic linear bandit framework of (Dani et al., 2008), (Rusmevichientong and Tsitsiklis, 2010), and (Abbasi-Yadkori et al., 2011). In order to tackle this problem , (Abbasi-Yadkori et al., 2011) show that by using ridge regression to estimate the model parameter, one can build an algorithm such that with probability at least $1-\delta$, the learner will attain a regret bound of $\tilde{O}\big(Rn\sqrt{T\log(1/\delta)}\big)$, where R is the sub-Gaussian constant and can be as large as the L^∞ norm of the losses. Finally, (Liu and Zhao, 2012) analyze the linear bandit problem in the context of adaptive shortest path algorithms for both light and heavy tailed losses. Their estimates are based on sample means which do not have the desired exponential concentration required for optimal learning in bandit problems (Hsu and Sabato, 2014).

We present two algorithms based on more sophisticated statistical estimators than ridge regression. We begin by introducing notation and our exact setup in Section 2. Then, we describe a general confidence region algorithm and explain how, if one were able to produce robust estimators, one can attain sublinear regret using this general algorithm. In the subsequent two sections, 4 and 5, we construct such estimators and derive the resulting regret bounds. Our first algorithm will demonstrate that if the loss functions exhibit $(1+\epsilon)$ finite moments, then sublinear regret is attainable of order $\tilde{O}\big(nT^{\frac{1}{2}+\frac{1}{2(1+\epsilon)}}\big)$. It thereby upper bounds the price of heavy-tailed loss distributions in this scenario as $\tilde{O}(T^{\frac{1}{2(1+\epsilon)}})$. When the loss functions have infinitely-many moments, we recover the known regret bounds. Our second algorithm will provide even better regret guarantees of $\tilde{O}\left(T^{\frac{1+2\epsilon}{1+3\epsilon}}\right)$ when the losses have minimal regularity (i.e. when $\epsilon < 1$), but it will not achieve the asymptotically optimal rate of $O(T^{1/2})$ as $\epsilon \to \infty$.

2. Notation and Preliminaries

Throughout the paper, we will denote vectors by lower case boldface letters, e.g. $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, scalars by Roman characters, e.g. $a, b \in \mathbb{R}$, and matrices by upper case boldface characters, e.g. $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times m}$. We will also use the notation $x_{1:t}$ to denote the sequence $(x_1, \ldots, x_t)$.

Given any norm $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$, we will denote its dual by $\|\cdot\|_*$. That is, $\|\mathbf{x}\|_* = \sup_{\|\mathbf{y}\|\le 1} \mathbf{x}\cdot\mathbf{y}$. We let $\|\cdot\|_2$ be the Euclidean norm, and given any symmetric positive definite (SPSD) matrix $\mathbf{A}$, we define the norm $\|\mathbf{x}\|_\mathbf{A} = \sqrt{\mathbf{x}^T\mathbf{A}\mathbf{x}}$. Finally, we let $B_1 = \{\mathbf{x} \in \mathbb{R}^n | \|\mathbf{x}\|_2 \le 1\}$ be the unit ball.

We consider the standard stochastic online linear optimization scenario with bandit feedback presented in (Dani et al., 2008). Specifically, let $\mathcal{X} \subset \mathbb{R}^n$ be a compact set representing the action space, and without loss of generality, assume that $\sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_2 \le 1$ (this can always be achieved by re-scaling). At each round t, the learner chooses an action $\mathbf{x}_t \in \mathcal{X}$ and observes the loss $l_t = l_t(\mathbf{x}_t)$. We assume the existence of a mean vector $\boldsymbol{\mu} \in B_1$ such that $l_t = \boldsymbol{\mu}^\top\mathbf{x}_t + \eta_t$, where η_t is a random variable satisfying $\mathbb{E}[\eta_t|\mathcal{F}_{t-1}] = 0$. Here, $\mathcal{F}_t := \sigma(\eta_{1:t})$ is the σ-algebra of events up to time t. We will consider deterministic algorithms where the choice of $\mathbf{x}_t$ is decided by $\eta_1, \ldots, \eta_{t-1}$ and therefore is $\mathcal{F}_{t-1}$-measurable. The goal of the learner is to minimize the (pseudo) regret: $\mathrm{Reg}_T = \sum_{t=1}^T \boldsymbol{\mu}^\top\mathbf{x}_t - \boldsymbol{\mu}^\top\mathbf{x}^*$, where $\mathbf{x}^* = \mathrm{argmin}_{\mathbf{x}\in\mathcal{X}}\boldsymbol{\mu}^\top\mathbf{x}$.

The standard assumption in the linear stochastic bandit scenario is that the random variable η_t is bounded or sub-Gaussian. In contrast, we will only require that the $(1+\epsilon)$-th moment exists: $\mathbb{E}[|\eta_t|^{1+\epsilon}|\mathcal{F}_{t-1}] \le v < \infty$.

Throughout the paper we will denote by $\mathbf{X}_t \in \mathbb{R}^{n\times t}$ the matrix $(\mathbf{x}_1, \ldots, \mathbf{x}_t)$ whose columns are the played actions. Similarly, we denote by $\boldsymbol{\eta}_t = (\eta_1, \ldots, \eta_t)^\top \in \mathbb{R}^t$ the vector of noise and $\mathbf{l}_t = (l_1, \ldots, l_t)^\top \in \mathbb{R}^t$ the vector of observed losses.

3. Confidence Region Algorithms

The underlying idea behind many regret minimization algorithms for stochastic linear bandits (e.g. (Dani et al., 2008; Abbasi-Yadkori et al., 2011)) is that of confidence regions and optimism in the face of uncertainty. More pre-

Algorithm 1 ConfidenceRegion

Input: Confidence function β, Estimate function $\mathcal{E}$, number of steps T.
$C_0 = B_1$
for $t = 1, \dots, T$: **do**
  Let $(\mathbf{x}_t, \overline{\boldsymbol{\mu}}_t) = \text{argmin}_{(\mathbf{x},\boldsymbol{\nu}) \in \mathcal{X} \times C_{t-1}} \boldsymbol{\nu}^\top \mathbf{x}$.
  Play $\mathbf{x}_t$ and suffer loss l_t.
  Estimate $\boldsymbol{\mu}_t = \mathcal{E}(\mathbf{x}_{1:t}, l_{1:t})$
  Update confidence region: $C_t = \{\boldsymbol{\nu} \mid \|\boldsymbol{\nu} - \boldsymbol{\mu}_t\|_{\mathbf{V}_t} \leq \beta(t)\}$
end for

cisely, at every time t, the learner keeps track of an ellipsoid C_t centered at a current estimate of the mean $\boldsymbol{\mu}_t$, in which he believes the true mean vector $\boldsymbol{\mu}$ lies. The learner then optimistically chooses action $\mathbf{x}_{t+1}$ via $(\mathbf{x}_{t+1}, \overline{\boldsymbol{\mu}}_{t+1}) = \text{argmin}_{(\mathbf{x},\boldsymbol{\nu}) \in \mathcal{X} \times C_t} \boldsymbol{\nu}^\top \mathbf{x}$. In order to achieve small regret, the ellipsoid must shrink quickly around $\boldsymbol{\mu}$. (Abbasi-Yadkori et al., 2011) construct these ellipsoids by setting $\boldsymbol{\mu}_t = (\mathbf{I} + \mathbf{X}_t \mathbf{X}_t^\top)^{-1} \mathbf{X}_t \mathbf{l}_t$, i.e. the ridge regression estimate of the mean obtained using the observed actions and rewards. They then specify the ellipsoid to be $C_t = \{\boldsymbol{\nu} | \|\boldsymbol{\nu} - \boldsymbol{\mu}_t\|_{\mathbf{V}_t} \leq \beta(t)\}$, where they define $\mathbf{V}_t = \mathbf{I} + \mathbf{X}_t \mathbf{X}_t^\top$ and $\beta(t)$ is in $O(\sqrt{\log t})$. The authors show that with high probability, $\boldsymbol{\mu} \in C_t$ for all t, and they are able to provide a regret bound in $O(n\sqrt{T \log T})$, which is known to be tight up to a logarithmic factor.

The algorithm of (Abbasi-Yadkori et al., 2011) cannot be used in our scenario, as it is known that that the ridge regression estimate of the mean is not robust under heavy-tailed noise (Hsu and Sabato, 2014). Nevertheless, we can apply the same confidence region techniques: find an estimate $\boldsymbol{\mu}_t$ of the mean and define $C_t = \{\boldsymbol{\nu} \mid \|\boldsymbol{\nu} - \boldsymbol{\mu}_t\|_{\mathbf{V}_t} \leq \beta(t)\}$. For an appropriate choice of radius function β, we will show that there are estimates $\boldsymbol{\mu}_t$ such that with high probability, $\boldsymbol{\mu}$ remains in C_t. This master algorithm, which we call ConfidenceRegion, is defined in Algorithm 1 and its regret is analyzed in the following proposition.

Proposition 1. *Let $r_t = \boldsymbol{\mu}^\top \mathbf{x}_t - \boldsymbol{\mu}^\top \mathbf{x}^*$ denote the instantaneous regret of our algorithm. If $\boldsymbol{\mu} \in C_t$ for all t, then*

$$r_t \leq 2\beta(t-1) \|\mathbf{x}_t\|_{\mathbf{V}_{t-1}^{-1}}. \tag{1}$$

Proof. Using the definition of $(\mathbf{x}_t, \overline{\boldsymbol{\mu}})$ as minimizers, we have

$$\begin{aligned} r_t &= \boldsymbol{\mu}^\top \mathbf{x}_t - \boldsymbol{\mu}^\top \mathbf{x}^* \leq \boldsymbol{\mu}^\top \mathbf{x}_t - \overline{\boldsymbol{\mu}}_t^\top \mathbf{x}_t \\ &= (\boldsymbol{\mu} - \overline{\boldsymbol{\mu}}_t)^\top \mathbf{x}_t = (\boldsymbol{\mu} - \boldsymbol{\mu}_{t-1})^\top \mathbf{x}_t + (\boldsymbol{\mu}_{t-1} - \overline{\boldsymbol{\mu}}_t)^\top \mathbf{x}_t \\ &\leq \|\boldsymbol{\mu} - \boldsymbol{\mu}_{t-1}\|_{\mathbf{V}_{t-1}} \|\mathbf{x}_t\|_{\mathbf{V}_{t-1}^{-1}} \\ &\quad + \|\boldsymbol{\mu}_{t-1} - \overline{\boldsymbol{\mu}}_t\|_{\mathbf{V}_{t-1}} \|\mathbf{x}_t\|_{\mathbf{V}_{t-1}^{-1}}, \end{aligned}$$

where the last inequality follows from the definition of dual norm and the fact that $\|\cdot\|_{\mathbf{V}_{t-1}}^* = \|\cdot\|_{\mathbf{V}_{t-1}^{-1}}$. The result is then true since $\boldsymbol{\mu}, \overline{\boldsymbol{\mu}}_t \in C_{t-1}$. □

Corollary 1. *If $\boldsymbol{\mu} \in C_t$ for all t and $\beta(t) > 1$, then the regret of Algorithm 1 can be bounded as:*

$$Reg_T \leq \sqrt{\sum_{t=1}^T \beta(t-1)^2 \sum_{t=1}^T 4 \min\left(\|\mathbf{x}_t\|_{\mathbf{V}_{t-1}^{-1}}^2, 1\right)}.$$

Proof. Since $\|\mathbf{x}\|_2, \|\boldsymbol{\mu}\|_2 \leq 1$, we have that $r_t \leq 2$. Therefore, by the previous proposition as well as the Cauchy-Schwarz inequality, we can bound the total regret as:

$$\begin{aligned} \sum_{t=1}^T r_t &\leq \sum_{t=1}^T 2 \min(\beta(t-1) \|\mathbf{x}_t\|_{\mathbf{V}_{t-1}^{-1}}, 1) \\ &\leq \sqrt{\sum_{t=1}^T \beta(t-1)^2 \sum_{t=1}^T 4 \min(\|\mathbf{x}_t\|_{\mathbf{V}_{t-1}^{-1}}^2, 1)}. \end{aligned}$$

□

The previous corollary shows that we need to control both terms $\sqrt{\sum_{t=1}^T \beta(t-1)^2}$ and $\sqrt{\sum_{t=1}^T 4 \min(\|\mathbf{x}_t\|_{\mathbf{V}_{t-1}^{-1}}^2, 1)}$. Toward this end, we must use the crucial fact that the matrix $\mathbf{V}_t$ is closely related to the vectors $\mathbf{x}_{1:t}$ via the equation $\mathbf{V}_t = \mathbf{I} + \mathbf{X}_t \mathbf{X}_t^\top$. The following proposition appears in (Dani et al., 2008) and we include the proof in the appendix for completeness.

Proposition 2. *Let $\mathbf{V}_0 = \mathbf{I}$ and suppose $\mathbf{V}_t = \mathbf{V}_{t-1} + \mathbf{x}_t \mathbf{x}_t^\top$. Then*

$$\begin{aligned} \sum_{t=1}^T \min(\|\mathbf{x}_t\|_{\mathbf{V}_{t-1}^{-1}}^2, 1) &\leq 2 \log\left(\det \mathbf{V}_T\right) \\ &\leq 2n \log\left(1 + \frac{T}{n}\right) \end{aligned}$$

In view of this result, we need only be concerned with defining $\beta(t)$ in a way such that with high probability, $\mu \in C_t$ for all t and such that $\sqrt{\sum_{t=1}^T \beta(t-1)^2}$ is in $o(T)$. For bounded losses, this can be achieved directly through the standard ridge regression estimate:

$$\mathcal{E}(\mathbf{x}_{1:t}, l_{1:t}) = (\mathbf{I} + \mathbf{X}_t \mathbf{X}_t^\top)^{-1} \mathbf{X}_t \mathbf{l}_t. \tag{2}$$

yielding the following form of $\beta(t)$ from (Abbasi-Yadkori et al., 2011):

$$\beta(t) = R\sqrt{2 \log\left(\frac{\det(V_t)^{1/2}}{\delta}\right)} + 1 \tag{3}$$

Algorithm 2 Estimate by Truncation

Input: $\mathbf{x}_{1:t}, l_{1:t}, \alpha_{1:t}$ with $\alpha_s < \alpha_{s+1}\ \forall s \in [1, t-1]$.
Let $\widehat{l}_t = l_t \mathbb{1}_{|l_t| \le \alpha_t}$ and $\widehat{\mathbf{l}}_t = (\widehat{l}_1, \dots, \widehat{l}_t)$
Return $\mathbf{V}_t^{-1} \mathbf{X}_t \widehat{\mathbf{l}}_t$.

However, this estimate is no longer valid when the noise exhibits heavy fluctuations. In the following two sections, we show how one can still derive good estimates on the confidence region using more sophisticated constructions of the subroutine $\mathcal{E}$ in Algorithm 1.

4. Truncation

One way to counteract the effect of heavy-tailed noise is to try truncating losses that become too large. More precisely, one can consider the implementation of $\mathcal{E}$ for Algorithm 1 defined in Algorithm 2.

Since truncation biases the distribution, we cannot truncate our losses at a fixed level uniformly over all time. Instead, the estimation algorithm we construct must dynamically adjust the truncation level as the rounds progress. In particular, we use an increasing sequence $\alpha_{1:T}$ of truncation levels as input to Algorithm 2. This sequence will be tuned optimally later on. We now provide a bound on the distance between the truncation estimate $\boldsymbol{\mu}_t$ and $\boldsymbol{\mu}$.

Lemma 1. *Let $\boldsymbol{\mu}_t = \mathbf{V}_t^{-1} \mathbf{X}_t \widehat{\mathbf{l}}_t$, and define $\widehat{\eta}_t = \widehat{l}_t - \boldsymbol{\mu}^\top \mathbf{x}_t$ as the truncated noise. Denote $\widehat{\boldsymbol{\eta}}_t = (\widehat{\eta}_1, \dots, \widehat{\eta}_t)$. Then*

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}\|_{\mathbf{V}_t} \le \|\mathbf{X}_t \widehat{\boldsymbol{\eta}}_t\|_{\mathbf{V}_t^{-1}} + \|\boldsymbol{\mu}\|_2. \tag{4}$$

Lemma 1 implies that to define β appropriately, it suffices to bound $\|\mathbf{X}_t \widehat{\boldsymbol{\eta}}_t\|_{\mathbf{V}_t^{-1}}$. Our monotonicity constraint $\alpha_t < \alpha_{t+1}$ in Algorithm 2 makes η_t uniformly bounded on $[1, T]$. Thus, by considering the stochastic process restricted to the time interval $[1, T]$, we can modify the concentration inequality in (Abbasi-Yadkori et al., 2011) and (Peña et al., 2009), derived using the theory of self-normalized processes, to arrive at the following result.

Lemma 2. *Let $\eta_1, \dots, \eta_T$ be random variables such that $|\eta_t| \le R$ for all $t \in [1, T]$. Denote by $\mathcal{F}_t = (\eta_1, \dots, \eta_t)$ the σ-algebra generated by these variables up to time t. Let $\mathbf{x}_t \in \mathbb{R}^n$ be $\mathcal{F}_{t-1}$-measurable random vectors and $\mathbf{V}_t = \mathbf{I} + \sum_{s=1}^t \mathbf{x}_t \mathbf{x}_t^\top$. Then for any $\delta > 0$, with probability at least $1 - \delta$, $\forall t \in [1, T]$,*

$$\|\mathbf{X}_t \boldsymbol{\eta}_t\|_{\mathbf{V}_t^{-1}} \le R \sqrt{2 \log\left(\frac{\det \mathbf{V}_t^{1/2}}{\delta}\right)}.$$

By leveraging Lemmas 1 and 2, we can derive an upper bound on the magnitude of $\mathbf{X}_t \widehat{\boldsymbol{\eta}}_t$ uniformly across all rounds.

Proposition 3. *Let $\boldsymbol{\mu}_t$ denote the estimate returned by Algorithm 2. Then for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds uniformly over t:*

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}\|_{\mathbf{V}_t} \le \alpha_T \sqrt{8 \log\left(\frac{\det(\mathbf{V}_t)^{1/2}}{\delta}\right)} + v \sqrt{\sum_{s=1}^t \frac{1}{\alpha_s^{2\epsilon}}} + \|\boldsymbol{\mu}\|_2.$$

Proof. We can decompose $\|\mathbf{X}_t \widehat{\boldsymbol{\eta}}_t\|_{\mathbf{V}_t^{-1}}$ as a sum of a bias and a variance term:

$$\|\mathbf{X}_t \widehat{\boldsymbol{\eta}}_t\|_{\mathbf{V}_t^{-1}} \le \|\mathbf{X}_t (\widehat{\boldsymbol{\eta}}_t - \mathbb{E}[\widehat{\boldsymbol{\eta}}_t | \mathcal{F}_{t-1}])\|_{\mathbf{V}_t^{-1}} + \|\mathbf{X}_t \mathbb{E}[\widehat{\boldsymbol{\eta}}_t | \mathcal{F}_{t-1}]\|_{\mathbf{V}_t^{-1}} \tag{5}$$

Let $\xi_t = \widehat{\eta}_t - \mathbb{E}[\eta_t | \mathcal{F}_{t-1}]$, and notice that $|\xi_t| \le 2\alpha_T$ uniformly over t. Therefore, by Lemma 2, with probability at least $1 - \delta$, we have that for all $t \le T$:

$$\|\mathbf{X}_t (\widehat{\boldsymbol{\eta}}_t - \mathbb{E}[\widehat{\boldsymbol{\eta}}_t | \mathcal{F}_{t-1}])\|_{\mathbf{V}_t^{-1}} \le \alpha_T \sqrt{8 \log\left(\frac{\det(\mathbf{V}_t)^{1/2}}{\delta}\right)}. \tag{6}$$

We proceed to bound the second term

$$\begin{aligned}
&\|\mathbf{X}_t \mathbb{E}[\widehat{\boldsymbol{\eta}}_t | \mathcal{F}_{t-1}]\|_{\mathbf{V}_t^{-1}}^2 \\
&= \left\| \sum_{s=1}^t \mathbf{x}_s \left(\mathbb{E}[l_s \mathbb{1}_{|l_s| \le \alpha_s} | \mathcal{F}_{s-1}] - \mathbf{x}_s^\top \boldsymbol{\mu} \right) \right\|_{\mathbf{V}_t^{-1}}^2 \\
&= \left\| \sum_{s=1}^t \mathbf{x}_s \mathbb{E}[l_s \mathbb{1}_{\{|l_s| > \alpha_s\}} | \mathcal{F}_{s-1}] \right\|_{\mathbf{V}_t^{-1}}^2 \\
&= \widetilde{\mathbf{l}}_t^\top \mathbf{X}_t^\top \mathbf{V}_t^{-1} \mathbf{X}_t \widetilde{\mathbf{l}}_t,
\end{aligned}$$

where $\widetilde{\mathbf{l}}_t = \left(\mathbb{E}[l_1 \mathbb{1}_{l_1 > \alpha_1} | \mathcal{F}_0], \dots, \mathbb{E}[l_t \mathbb{1}_{l_t > \alpha_t} | \mathcal{F}_{t-1}] \right)^\top$. Now, we can use the definition of $\mathbf{V}_t$ as well as the matrix identity $\mathbf{X}(\mathbf{I} + \mathbf{X}^\top \mathbf{X})^{-1} = (\mathbf{I} + \mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{X}$ to bound the last term as

$$\begin{aligned}
\widetilde{\mathbf{l}}_t^\top \mathbf{X}_t^\top \mathbf{V}_t^{-1} \mathbf{X}_t \widetilde{\mathbf{l}}_t &= \widetilde{\mathbf{l}}_t^\top \mathbf{X}_t^\top (\mathbf{I} + \mathbf{X}_t \mathbf{X}_t^\top)^{-1} \mathbf{X}_t \widetilde{\mathbf{l}}_t \\
&= \widetilde{\mathbf{l}}_t^\top \mathbf{X}_t^\top \mathbf{X}_t (\mathbf{I} + \mathbf{X}_t^\top \mathbf{X}_t)^{-1} \widetilde{\mathbf{l}}_t \\
&\le \|\widetilde{\mathbf{l}}_t\|_2^2,
\end{aligned}$$

where for the last inequality we have used the fact that the eigenvalues of the matrix $\mathbf{X}_t^\top \mathbf{X}_t (\mathbf{I} + \mathbf{X}_t^\top \mathbf{X}_t)^{-1}$ are less than 1. Finally, we have

$$\begin{aligned}
\|\widetilde{\mathbf{l}}_t\|_2^2 &\le \sum_{s=1}^t \mathbb{E}[|l_s| \mathbb{1}_{|l_s| > \alpha_s} | \mathcal{F}_{s-1}]^2 \\
&\le \sum_{s=1}^t \mathbb{E}\left[\frac{|l_s|^{1+\epsilon}}{\alpha_s^\epsilon} \middle| \mathcal{F}_{s-1} \right]^2 \le v^2 \sum_{s=1}^t \frac{1}{\alpha_s^{2\epsilon}}.
\end{aligned}$$

In view of inequalities (4), (5), and (6), as well as the last inequality, we see that with probability at least $1-\delta$, the following holds for all $t \le T$:

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}\|_{\mathbf{V}_t} \le \alpha_T \sqrt{8 \log\left(\frac{\det(\mathbf{V}_t)^{1/2}}{\delta}\right)} + v\sqrt{\sum_{s=1}^{t} \frac{1}{\alpha_s^{2\epsilon}}} + \|\boldsymbol{\mu}\|_2.$$

□

The previous quantitive estimate on the confidence region motivates the choice of $\alpha_t = t^{\frac{1}{2(1+\epsilon)}}$. A regret bound for this truncation scheme can now be readily derived.

Corollary 2. *Let $\alpha_t = t^{\frac{1}{2(1+\epsilon)}}$, and let $\boldsymbol{\mu}_t$ be the estimate returned by Algorithm 2. Then for any $\delta > 0$, with probability at least $1-\delta$, the following holds uniformly over all $t \ge 3$:*

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}\|_{\mathbf{V}_t} \le 2\sqrt{2t^{\frac{1}{(1+\epsilon)}} \log\left(\frac{\det(\mathbf{V}_t)^{1/2}}{\delta}\right)} + \sqrt{2v^2 \log t} + 1.$$

Proof. The result follows by combining Proposition 3, the fact that $\boldsymbol{\mu} \in B_1$, and the following computation:

$$\sum_{s=1}^{t} \frac{1}{s^{\frac{\epsilon}{1+\epsilon}}} \le 1 + \int_1^t \frac{1}{s^{\frac{\epsilon}{1+\epsilon}}} ds \le 2t^{\frac{1}{1+\epsilon}} \log t.$$

The last inequality follows from Lemma 4 in the appendix, where we set $u = \frac{1}{\epsilon}$. □

Combining the results of Corollary 1, Proposition 2, and Corollary 2 yields the following regret bound for Algorithm 1 using the subroutine of Algorithm 2 for the parameter estimate.

Theorem 1. *Fix $\delta > 0$, and denote $A_{T,\mathbf{V}_T,\delta} = \log\left(\frac{\det(\mathbf{V}_T)^{1/2}}{\delta}\right)$, $\widetilde{A}_{T,\delta} = \log\left(\frac{(1+\frac{T}{n})^{1/2}}{\delta}\right)$. Then with probability at least $1-\delta$, the regret of Algorithm 1 using the truncation estimate with $\alpha_t = t^{\frac{1}{2(1+\epsilon)}}$ and $\beta(t) = 2\sqrt{2t^{\frac{1}{(1+\epsilon)}}\left(A_{T,\mathbf{V}_T,\delta} + v^2 \log T\right)} + 1$ is bounded by:*

$$\begin{aligned} \mathit{Reg}_T &\le C\sqrt{T^{\frac{2+\epsilon}{1+\epsilon}}\left(A_{T,\mathbf{V}_T,\delta} + v^2 \log T\right) \log(\det \mathbf{V}_T)} \\ &\le Cn\sqrt{T^{\frac{2+\epsilon}{1+\epsilon}}\left(\widetilde{A}_{T,\delta} + v^2 \log T\right) \log\left(1+\frac{T}{n}\right)}, \end{aligned}$$

where C is a universal constant.

Algorithm 3 Mini-Batch ConfidenceRegion

Input: confidence function β, estimate function $\mathcal{E}$, confidence level δ, number of batches m, number of steps T.
$C_0 = B_1$
for $i = 1, \dots, m$: **do**
  Let $r = \lceil \frac{T}{m} \rceil$ and set $t_i = ir$
  Let $(\mathbf{x}_i, \overline{\boldsymbol{\mu}}_i) = \operatorname{argmin}_{(\mathbf{x},\boldsymbol{\nu}) \in \mathcal{X} \times C_{i-1}} \boldsymbol{\nu}^\top \mathbf{x}$.
  Play $\mathbf{x}_i$ r times, and suffer losses $l_{t_i}, l_{t_i+1}, \dots, l_{t_i+r}$.
  Let $\widetilde{l}_i = \mathrm{MoM}(l_{t_i+1}, \dots, l_{t_i+r}, \delta)$ and $\widetilde{\mathbf{l}}_i = (\widetilde{l}_1, \dots, \widetilde{l}_i)$
  Estimate $\boldsymbol{\mu}_i = \mathbf{V}_i^{-1} \mathbf{X}_i \widetilde{\mathbf{l}}_i$
  Update confidence region: $C_i = \{\boldsymbol{\nu} \mid \|\boldsymbol{\nu} - \boldsymbol{\mu}_i\|_{\mathbf{V}_i} \le \beta(i)\}$
end for

Algorithm 4 Median of Means (MoM)

Input: Observed losses $l_1, \dots, l_r$, confidence δ
Set $k = \lfloor 8 \log(2/\delta) \rfloor$, $N = \lfloor r/k \rfloor$
Set $\widehat{l}_j = \frac{1}{N} \sum_{t=(j-1)N+1}^{jN} l_t$ for $j = 1, \dots, k$.
return Median$(\widehat{l}_1, \dots, \widehat{l}_k)$.

Thus, in view of Proposition 2, the regret of the truncation algorithm is in $\widetilde{O}(vT^{\frac{2+\epsilon}{2(1+\epsilon)}})$. Recall that the regret of the standard bandit algorithm is in $\widetilde{O}(RT^{1/2})$ (e.g. (Abbasi-Yadkori et al., 2011)), where R is the sub-Gaussian constant. Thus, our algorithm not only admits a non-trivial guarantee for losses that are not sub-Gaussian, but it can also be more favorable than the standard bandit algorithm when the sub-Gaussian constant of the noise is much larger than the $(1+\epsilon)$-th moment (i.e. $vT^{\frac{1}{2(1+\epsilon)}} < R$). This fact is empirically verified in the appendix.

5. Median of Means

In this section we present an algorithm with a better regret guarantee than the truncation algorithm for $\epsilon \in (0, 1)$. We consider Algorithm 3 together with the subroutine shown in Algorithm 4, which depends on the median of means estimator of (Alon et al., 1999) and follows the confidence region guidelines of Section 3. However, instead of recomputing a new estimate of the model parameter at every time step, our new algorithm runs in m batches.

At each stage i, beginning at time t_i, it will choose an action x_i and play this action for $r = \frac{T}{m}$ rounds, After observing losses $\{l_t\}_{t=t_i}^{t_i+r}$, Algorithm 4 combines them to create a surrogate loss $\widetilde{l}_i$ associated with stage i. This surrogate loss is calculated using the median of means estimator. One may view this modified algorithm as an instance of Algo-

rithm 1 run on the data $\big((\mathbf{x}_1, \widetilde{l}_1), \dots (\mathbf{x}_m, \widetilde{l}_m)\big)$, where the regret accumulated at stage i must now be accounted for r times.

A natural concern raised by this algorithm is that the modified losses $\widetilde{l}_i$ are not an unbiased estimate of the true loss. That is, $\mathbb{E}[\widetilde{l}_i] \neq \boldsymbol{\mu}^\top \mathbf{x}_i$; nevertheless we can still derive the following concentration bound:

Proposition 4. *Let $l_1, \dots, l_r$ be random variables satisfying $\mathbb{E}[l_i|\mathcal{F}_{i-1}] = \mu$ for all i and $\mathbb{E}[|l_i - \mu|^{1+\epsilon} \mid \mathcal{F}_{i-1}] \leq v$. If $\widetilde{l}$ denotes the median of means estimate of Algorithm 4, then for any $\delta > 0$, with probability at least $1 - \delta$,*

$$|\widetilde{l} - \mu| \leq (12v)^{\frac{1}{1+\epsilon}} \left(\frac{8\log(2/\delta)}{r}\right)^{\frac{\epsilon}{1+\epsilon}}.$$

To prove this proposition, we first derive the following tail bound on the sum of martingale differences.

Lemma 3. *Let $X_1, \dots, X_n$ be random variables satisfying $\mathbb{E}[X_i|\mathcal{F}_{i-1}] = 0$, and $\mathbb{E}[|X_i|^{1+\epsilon}|\mathcal{F}_{i-1}] \leq v$. Where $\mathcal{F}_i = \sigma(\mathcal{G} \cup \sigma(X_1, \dots, X_i))$ and $\mathcal{G}$ is an arbitrary sigma-algebra. Then*

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n X_i\Big| > t \Big| \mathcal{G}\Big) \leq \frac{3v}{n^\epsilon t^{1+\epsilon}}$$

Proof. For notation simplicity we will let $\mathbb{P}$ and $\mathbb{E}$ denote the probability and expectation conditioned on $\mathcal{G}$. We can draw inspiration from the ideas of (Bubeck et al., 2013) and bound the desired probability as:

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| > t\right) \\
&\leq \mathbb{P}(\exists i \; s.t |X_i| > a) + \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n X_i \mathbb{1}_{|X_i|\leq a}\Big| > t\Big).
\end{aligned}$$

The first term can be bounded by using the union bound and Markov's inequality: $\sum_{i=1}^n \mathbb{P}(|X_i| > a) \leq \frac{\sum_{i=1}^n \mathbb{E}[|X_i|^{1+\epsilon}]}{a^{1+\epsilon}} = \frac{nv}{a^{1+\epsilon}}$. On the other hand, we can bound the second term using Tchebyschev's inequality:

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i \mathbb{1}_{|X_i|\leq a}\right| > t\right) \\
&\leq \frac{\mathbb{E}\left[\sum_{i=1}^n (X_i \mathbb{1}_{|X_i|\leq a})^2\right]}{n^2 t^2} \\
&= \frac{\mathbb{E}\left[\sum_{i=1}^n X_i^2 \mathbb{1}_{|X_i|\leq a}\right]}{n^2 t^2} \\
&+ \frac{\mathbb{E}\left[\sum_{i=1}^n \sum_{j<i} X_i \mathbb{1}_{|X_i|\leq a} X_j \mathbb{1}_{|X_j|\leq a}\right]}{n^2 t^2}. \quad (7)
\end{aligned}$$

We now proceed to bound the cross terms $\mathbb{E}[X_i \mathbb{1}_{|X_i|\leq a} X_j \mathbb{1}_{|X_j|\leq a}]$. Let $m_i = \mathbb{E}[X_i \mathbb{1}_{|X_i|\leq a}|\mathcal{F}_{i-1}]$. Then, using the fact that $j > i$, we have

$$\begin{aligned}
&\mathbb{E}[X_i \mathbb{1}_{|X_i|\leq a} X_j \mathbb{1}_{|X_j|\leq a}] \\
&= \mathbb{E}[(X_i \mathbb{1}_{|X_i|\leq a} - m_i) X_j \mathbb{1}_{|X_j|\leq a}] + \mathbb{E}[m_i X_j \mathbb{1}_{|X_j|\leq a}] \\
&= \mathbb{E}[m_i X_j \mathbb{1}_{|X_j|\leq a}],
\end{aligned}$$

where the last equality can be derived from the tower property of conditional expectation. Now, using Hölders inequality for conditional expectation with parameters $1+\epsilon$ and $1 + \frac{1}{\epsilon}$ as well as the fact that $\mathbb{E}[X_i \mathbb{1}_{|X_i|\leq a}|\mathcal{F}_{i-1}] = -\mathbb{E}[X_i \mathbb{1}_{|X_i|>a}|\mathcal{F}_{i-1}]$, we have

$$\begin{aligned}
m_i &\leq \mathbb{E}[|X_i| \mathbb{1}_{|X_i|>a}|\mathcal{F}_{i-1}] \\
&\leq \mathbb{E}\left[|X_i|^{1+\epsilon}|\mathcal{F}_{i-1}\right]^{\frac{1}{1+\epsilon}} \mathbb{P}(|X_i| > a|\mathcal{F}_{i-1})^{\frac{\epsilon}{1+\epsilon}}.
\end{aligned}$$

Using the conditional form of Markov's inequality yields

$$m_i \leq v^{\frac{1}{1+\epsilon}} \frac{v^{\frac{\epsilon}{1+\epsilon}}}{a^\epsilon} = \frac{v}{a^\epsilon}.$$

A similar argument then shows that

$$\mathbb{E}[m_i X_j \mathbb{1}_{|X_j|\leq a}] \leq \frac{v^2}{a^{2\epsilon}}. \quad (8)$$

On the other hand, a simple manipulation shows that $\mathbb{E}[X_i^2 \mathbb{1}_{|X_i|\leq a}] \leq a^{1-\epsilon} v$. Replacing these estimates in (7) shows that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n |X_i| \mathbb{1}_{|X_i|\leq a} > t\right) \leq \frac{a^{1-\epsilon} v}{nt^2} + \frac{v^2}{t^2 a^{2\epsilon}}.$$

We have thus shown that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n |X_i| > t\right) \leq \frac{nv}{a^{1+\epsilon}} + \frac{a^{1-\epsilon} v}{nt^2} + \frac{v^2}{t^2 a^{2\epsilon}}.$$

Choosing $a = nt$ yields

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n |X_i| > t\right) \leq \frac{2v}{n^\epsilon t^{1+\epsilon}} + \left(\frac{v}{n^\epsilon t^{1+\epsilon}}\right)^2.$$

If $\frac{v}{n^\epsilon t^{1+\epsilon}} > 1$, the desired inequality trivially holds. Otherwise, it follows from the previous bound. □

Proof of Proposition 4. Using the notation of Algorithm 4, we know by Lemma 3 that $\mathbb{P}(|\widehat{l}_j - \mu| > \eta|\mathcal{G}_{j-1}) \leq \frac{3v}{N^\epsilon \eta^{1+\epsilon}}$, where $\mathcal{G}_j = \sigma(l_1, \dots, l_{jN})$. Define the random variable $X_j = \mathbb{1}_{\widehat{l}_j - \mu > \eta}$ and notice that $p_j := \mathbb{P}(X_j = 1|\mathcal{G}_{j-1}) \leq \frac{1}{4}$ when $\eta = (12v)^{\frac{1}{1+\epsilon}} \left(\frac{1}{N}\right)^{\frac{\epsilon}{1+\epsilon}}$. Furthermore, the random variables $X_j - p_j$ form a martingale difference sequence with respect to $\mathcal{G}_j$. Therefore by Azuma-Hoeffding's inequality for bounded random variables we obtain

$$\begin{aligned}
\mathbb{P}\Big(\sum_{j=1}^k X_j \geq k/2\Big) &= \mathbb{P}\Big(\sum_{j=1}^k X_j - p_j \geq k/4\Big) \leq e^{-k/8} \\
&= \delta/2.
\end{aligned}$$

On the other hand, the median $\widetilde{l}$ satisfies $\widetilde{l} > \mu + \eta$ if and only if at least half of the estimates $\widehat{l}_j$ are above μ which happens if and only if $\sum_{j=1}^{k} X_j \geq k/2$. Therefore $\widetilde{l} > \mu + \eta$ with probability at most $\delta/2$, and a similar argument shows that $\widetilde{l} < \mu - \eta$ with probability at most $\delta/2$. Thus, by the union bound, with probability at least $1 - \delta$,

$$|\widetilde{l} - \mu| \leq (12v)^{\frac{1}{1+\epsilon}} \left(\frac{8\log(2/\delta)}{r}\right)^{\frac{\epsilon}{1+\epsilon}}.$$

□

Proposition 4 suggests choosing $\boldsymbol{\mu}_i$ at stage i as $\boldsymbol{\mu}_i = \mathbf{V}_i^{-1}\mathbf{X}_i\widetilde{\mathbf{l}}_i$, where $\mathbf{V}_i$, $\mathbf{X}_i$ and $\widetilde{\mathbf{l}}_i$ have analogous definitions to the ones used in the previous section. By selecting $\boldsymbol{\mu}_i$ in this manner, we can bound the regret of our algorithm using the same line of reasoning as Corollaries 1 and 2, with the only difference being that the regret of stage i must be accounted for r times.

Proposition 5. *If $\boldsymbol{\mu} \in C_i$ for all i and $\beta(i) > 1$, then the regret of Algorithm 1 can be bounded as:*

$$\begin{aligned} \mathit{Reg} &\leq r\sqrt{\sum_{i=1}^{m}\beta(i)^2\sum_{i=1}^{m} 4\min\left(\|\mathbf{x}_i\|^2_{\mathbf{V}_{i-1}^{-1}}, 1\right)} \\ &\leq r\beta(m)\sqrt{m\log\left(\frac{\det(\mathbf{V}_m)^{1/2}}{\delta}\right)}. \end{aligned}$$

As before, to ensure $\boldsymbol{\mu} \in C_i$ we must bound the value of $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}\|_{\mathbf{V}_i}$. To do so, let $\widetilde{\eta}_i = \widetilde{l}_i - \boldsymbol{\mu}^\top \mathbf{x}_i$, then by Lemma 1 we must have $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}\|^2_{\mathbf{V}_i} \leq \|\mathbf{X}_i\widetilde{\boldsymbol{\eta}}_i\|_{\mathbf{V}_i^{-1}} + \|\boldsymbol{\mu}\|_2$. We now proceed to bound $\|\mathbf{X}_i\widetilde{\boldsymbol{\eta}}_i\|_{\mathbf{V}_i^{-1}}$.

Proposition 6. *For any $\delta > 0$, with probability at least $1 - 2\delta$ the following bound holds uniformly over $i \leq m$:*

$$\begin{aligned} \|\mathbf{X}_i\widetilde{\boldsymbol{\eta}}_i\|_{\mathbf{V}_i^{-1}} \leq (12v)^{\frac{1}{1+\epsilon}} &\left(\frac{8\log\left(\frac{2m}{\delta}\right)}{r}\right)^{\frac{\epsilon}{1+\epsilon}} \\ &\left(\sqrt{2\log\left(\frac{m\det(\mathbf{V}_i)^{1/2}}{\delta}\right)} + \sqrt{i}\right). \end{aligned}$$

Proof. Let $c = (12v)^{\frac{1}{1+\epsilon}}\left(\frac{8\log\left(\frac{2m}{\delta}\right)}{r}\right)^{\frac{\epsilon}{1+\epsilon}}$ and define the following variables $\eta_i^{(1)} = \widetilde{\eta}_i\mathbb{1}_{|\widetilde{\eta}_i|\leq c} - \mathbb{E}[\widetilde{\eta}_i\mathbb{1}_{|\widetilde{\eta}_i|\leq c}|\mathcal{F}_{i-1}]$, $\eta_i^{(2)} = \mathbb{E}[\widetilde{\eta}_i\mathbb{1}_{|\widetilde{\eta}_i|\leq c}|\mathcal{F}_{i-1}]$ and $\eta_i^{(3)} = \widetilde{\eta}_i\mathbb{1}_{|\widetilde{\eta}_i|>c}$. We can decompose the desired expression as

$$\|\mathbf{X}_i\boldsymbol{\eta}_i\|_{\mathbf{V}_i^{-1}} \leq \|\mathbf{X}_i\boldsymbol{\eta}_i^{(1)}\|_{\mathbf{V}_i^{-1}} + \|\mathbf{X}_i\boldsymbol{\eta}_i^{(2)}\|_{\mathbf{V}_i^{-1}} + \|\mathbf{X}_i\boldsymbol{\eta}_i^{(3)}\|_{\mathbf{V}_i^{-1}},$$

where for $j = 1, 2, 3$, $\boldsymbol{\eta}_i^{(j)}$ is the vector with entries $(\eta_1^{(j)}, \ldots, \eta_i^{(j)})$. Since $\eta_i^{(1)}$ is bounded and $\mathbb{E}[\eta_i^{(1)}|\mathcal{F}_{i-1}] = 0$, we can apply Lemma 2 and the union bound to see that with probability at least $1 - \delta$, for all i

$$\|\mathbf{X}_i\boldsymbol{\eta}_i^{(1)}\|_{\mathbf{V}_i^{-1}} \leq c\sqrt{2\log\left(\frac{m\det(\mathbf{V}_i)^{1/2}}{\delta}\right)}.$$

In the same way as in the proof of Proposition 3, we also see that: $\|\mathbf{X}_i\boldsymbol{\eta}_i^{(2)}\|^2_{\mathbf{V}_i^{-1}} + \|\mathbf{X}_i\boldsymbol{\eta}_i^{(3)}\|^2_{\mathbf{V}_i^{-1}} \leq \sum_{j=1}^{i}(\eta_j^{(2)})^2 + (\eta_j^{(3)})^2$.

Moreover, by definition of c, Proposition 4, and the union bound, we have that with probability at least $1 - \delta$, $(\eta_j^{(3)})^2 = 0$ for all $j \in \{1, \ldots, m\}$. On the other hand, $\eta^{(2)} \leq c$, therefore

$$\|\mathbf{X}_i\widetilde{\boldsymbol{\eta}}_i\|_{\mathbf{V}_i^{-1}} \leq c\sqrt{2\log\left(\frac{m\det(\mathbf{V}_i)^{1/2}}{\delta}\right)} + c\sqrt{i}.$$

Substituting in the value of c yields the result. □

The previous proposition implies that by setting $\beta(i) = (12v)^{\frac{1}{1+\epsilon}}\left(\frac{8\log(2m/\delta)}{r}\right)^{\frac{\epsilon}{1+\epsilon}}\left(\sqrt{2\log\left(\frac{m\det(\mathbf{V}_i)^{1/2}}{\delta}\right)} + \sqrt{i}\right) + 1$, we can ensure that $\boldsymbol{\mu} \in C_i$ with probability at least $1 - 2\delta$. Moreover, in view of Proposition 5 and Proposition 2 again, we can provide the following regret guarantee for our median of means algorithm.

Theorem 2. *Let $T^{\frac{\epsilon}{1+\epsilon}} > n$ and set $\beta(i) = (12v)^{\frac{1}{1+\epsilon}}\left(\frac{8\log(2m/\delta)}{r}\right)^{\frac{\epsilon}{1+\epsilon}}\left(\sqrt{2\log\left(\frac{m\det(\mathbf{V}_i)^{1/2}}{\delta}\right)} + \sqrt{i}\right) + 1$ and $m = T^{\frac{2\epsilon}{1+3\epsilon}}$. Then with probability at least $1 - 3\delta$, the regret of Algorithms 3 and 4 together is bounded by:*

$$\begin{aligned} \mathit{Reg}_T \leq Cv^{\frac{1}{1+\epsilon}}T^{\frac{1+2\epsilon}{1+3\epsilon}}&\sqrt{n\log\left(\frac{2T^{\frac{\epsilon}{1+\epsilon}}}{n\delta}\right)}\Bigg[\log^{\frac{\epsilon}{1+\epsilon}}\left(\frac{2T}{\delta}\right) \\ &\left(T^{-\frac{\epsilon}{1+3\epsilon}}\sqrt{2n\log\left(\frac{2T^{\frac{2\epsilon}{1+\epsilon}}}{n\delta}\right)} + 1\right) + 1\Bigg]. \end{aligned}$$

for some universal constant C.

The details of the computation are provided in the appendix.

Theorem 2 tells us that the regret of Algorithms 3 and 4 together is in $\widetilde{O}(T^{\frac{1+2\epsilon}{1+3\epsilon}})$. Note that the regret bound of this algorithm compares favorably against that of the truncation algorithm (which was in $\widetilde{O}(T^{\frac{2+\epsilon}{2(1+\epsilon)}})$) for $\epsilon < 1$, but has a worse asymptotic rate when $\epsilon > 1$.

Another interesting point is that as $\epsilon \to \infty$ (i.e. when all moments exist), the regret of the median of means algorithm is in $\widetilde{O}(vT^{2/3})$ and the truncation algorithm is in $\widetilde{O}(vT^{1/2})$. Thus, the median of means algorithm does not converge to the asymptotically optimal rate.

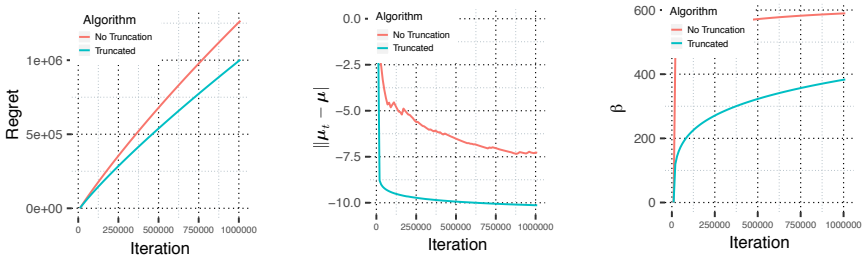

Figure 1. Comparison of the truncation algorithm versus the vanilla algorithm for bounded losses. (a) Mean regret over 20 replicas of the same experiment, the error bars are too small to be noticed in the plots. (b) Distance $\|\boldsymbol{\mu}_t - \boldsymbol{\mu}\|$ for one realization of the experiment. The y-axis is in logarithmic scale. (c) Magnitude of confidence radius $\beta(t)$ for one experiment realization.

It is known (Bubeck et al., 2013) that for the multi-armed bandit problem, the optimal regret for sub-Gaussian noise can be recovered with only a second moment assumption. In our case, we have only ensured an $O(T^{3/4})$ regret for both algorithms. It is unclear if this $O(T^{1/4})$ gap between heavy-tailed noise and sub-Gaussian noise is only due to our choice of estimator or if it is a general problem for linear bandits with heavy tails. While one could try other robust estimates such as the technique of (Hsu and Sabato, 2014), that algorithm is difficult to apply, because the regression matrix $\mathbf{V}_t$ could change at every round. In that case the recurrence $\mathbf{V}_t = \mathbf{V}_{t-1} + \mathbf{x}_t \mathbf{x}_t^\top$,which was crucial in the proof of Proposition 2, would no longer be satisfied.

Finally, our analysis relied on a concentration bound for self-normalized processes (Lemma 2). This bound is of the Hoeffding type and depends on the L_∞ norm of the random variables. We believe that recovering the optimal $O(T^{1/2})$ regret would be possible under the truncation algorithm if this concentration bound depended instead on the second moment in a way similar to Freedman's inequality. However, no such bound exists in the literature, and it seems highly non-trivial to craft such an extension.

Both of the algorithms presented require a priori knowledge of ϵ. It is reasonable to inquire whether this can be avoided. Unfortunately, this issue appears in all confidence-bound based algorithms in the bandit literature, including the unbounded MAB case in (Bubeck et al., 2013). In fact, an a priori upper bound is required even in the bounded noise case (Bubeck and Cesa-Bianchi, 2012).

6. Experiments

We now present empirical results showing that the truncation algorithm benefits from a better regret than the vanilla linear bandit algorithm of (Abbasi-Yadkori et al., 2011). Our experimental setup is as follows: we let $d = 50$ and $\boldsymbol{\mu} = \frac{1}{\sqrt{n}}\mathbf{1} \in \mathbb{R}^n$, where $\mathbf{1}$ is a vector with all entries set to 1. For every $\mathbf{x} \in B_1$ the reward function is given by $\mathbf{x} \mapsto \boldsymbol{\mu}^\top \mathbf{x} + \eta$, where η is a random variable taking values $-\gamma$ with probability $1 - \gamma^2$ and $\frac{1}{\gamma}$ with probability γ^2 where $\gamma = \frac{1}{\sqrt{40T}}$. Notice that in this scenario the L_∞ norm of the noise is in $O(\sqrt{T})$ while the second moment of the noise is equal to $v = 1$. In order to make the algorithm of (Abbasi-Yadkori et al., 2011) as competitive as possible we set the parameter R, defining the function β in (3), to the optimal sub-Gaussian constant. That is, $R = \inf\{r | \, \mathbb{E}[e^{t\eta}] \leq e^{\frac{r^2 t^2}{2}} \; \forall t\}$.

Figure 1(a) shows the mean regret over 20 replicas achieved by both the truncation algorithm and the algorithm of (Abbasi-Yadkori et al., 2011) for $T = 10^6$. Notice that, not only does our algorithm achieves much better regret (we show a 25% improvement), but it is also more stable. In particular, Figure 1(b) shows the effect of noisy observations on each algorithm. Whereas the distance of the truncation algorithm estimates to the true hypothesis consistently approach zero, the same estimates of the vanilla algorithm seem to vary more and converge slower. Finally, notice that the confidence radius $\beta(t)$ remains consistently smaller for our algorithm, which, in combination with the previous statement makes the choice of $\overline{\boldsymbol{\mu}}_t$ closer to $\boldsymbol{\mu}$.

7. Conclusion

We provided the first known sublinear regret bounds for stochastic linear bandits with heavy-tailed losses. Instead of assuming bounded or sub-Gaussian noise, our algorithms only require the existence of the $(1+\epsilon)$-th moment. One of our algorithms has a regret bound that converges to the known optimal rate when infinitely many moments exist, and the other one has a more favorable rate when $\epsilon < 1$. An interesting question is whether one can modify the median of means algorithm, so that it, too, converges to the optimal rate for infinitely many moments. Finally, our analysis poses the non-trivial question of whether an optimal regret in $O(T^{1/2})$ can even be achieved when the noise has only a second moment.

References

Abbasi-Yadkori, Y., D. Pál, and C. Szepesvári (2011). Improved algorithms for linear stochastic bandits. In *NIPS*, pp. 2312–2320.

Agarwal, A., D. Hsu, S. Kale, J. Langford, L. Li, and R. E. Schapire (2014). Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of ICML 2014*, pp. 1638–1646.

Alon, N., Y. Matias, and M. Szegedy (1999). The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci. 58*(1), 137–147.

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

Auer, P., N. Cesa-Bianchi, Y. Freund, and R. E. Schapire (2003, January). The nonstochastic multiarmed bandit problem. *SIAM J. Comput. 32*(1), 48–77.

Beygelzimer, A., J. Langford, L. Li, L. Reyzin, and R. E. Schapire (2011). Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of AISTATS 2011*, pp. 19–26.

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

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

Cesa-Bianchi, N., O. Dekel, and O. Shamir (2013). Online learning with switching costs and other adaptive adversaries. In *NIPS*, pp. 1160–1168.

Chu, W., L. Li, L. Reyzin, and R. E. Schapire (2011). Contextual bandits with linear payoff functions. In *AISTATS*, pp. 208–214.

Dani, V., T. P. Hayes, and S. M. Kakade (2008). Stochastic linear optimization under bandit feedback. In *COLT*, pp. 355–366.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*(301), 13–30.

Hsu, D. and S. Sabato (2014). Heavy-tailed regression with a generalized median-of-means. In *Proceedings of ICML*, pp. 37–45.

Hull, J. C. (2012). *Risk Management and Financial Institutions*. Wiley Finance. Wiley.

Kleinberg, R. D. (2004). Nearly tight bounds for the continuum-armed bandit problem. In *Proceedings of NIPS 2004*, pp. 697–704.

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

Liu, K. and Q. Zhao (2012). Adaptive shortest-path routing under unknown and stochastically varying link states. In *Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, pp. 232–237.

Peña, V. H., T. L. Lai, and Q.-M. Shao (2009). *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media.

Rachev, S. T. (2003). *Handbook of Heavy Tailed Distributions in Finance: Handbooks in Finance*. Handbooks in Finance. Elsevier Science.

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

Rusmevichientong, P. and J. N. Tsitsiklis (2010). Linearly parameterized bandits. *Math. Oper. Res. 35*(2), 395–411.