Bandits With Heavy Tail

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Gábor Lugosi, *Member, IEEE*

***Abstract*—The stochastic multiarmed bandit problem is well understood when the reward distributions are sub-Gaussian. In this paper, we examine the bandit problem under the weaker assumption that the distributions have moments of order $1+\varepsilon$, for some $\varepsilon \in (0,1]$. Surprisingly, moments of order 2 (i.e., finite variance) are sufficient to obtain regret bounds of the same order as under sub-Gaussian reward distributions. In order to achieve such regret, we define sampling strategies based on refined estimators of the mean such as the truncated empirical mean, Catoni's M-estimator, and the median-of-means estimator. We also derive matching lower bounds that also show that the best achievable regret deteriorates when $\varepsilon < 1$.**

***Index Terms*—Heavy-tailed distributions, regret bounds, robust estimators, stochastic multi-armed bandit.**

I. Introduction

In this paper, we investigate the classical stochastic multi-armed bandit problem introduced by [1] and described as follows: an agent facing K actions (or bandit arms) selects one arm at every time step. With each arm $i \in \{1,\dots,K\}$ there is an associated probability distribution ν_i with finite mean μ_i. These distributions are unknown to the agent. At each round $t = 1,\dots,n$, the agent chooses an arm I_t, and observes a reward drawn from ν_{I_t} independently from the past given I_t. The goal of the agent is to minimize the *regret*

$$R_n = n \max_{i=1,\dots,K} \mu_i - \sum_{t=1}^{n} \mathbb{E}\, \mu_{I_t}.$$

We refer the reader to [2] for a survey of the extensive literature of this problem and its variations. The vast majority of authors assume that the unknown distributions ν_i are sub-Gaussian, that is, the moment generating function of each ν_i is such that if X is a random variable drawn according to the distribution ν_i, then for all $\lambda \geq 0$,

$$\ln \mathbb{E}\, e^{\lambda(X-\mathbb{E}X)} \leq \frac{v\lambda^2}{2} \quad \text{and} \quad \ln \mathbb{E}\, e^{\lambda(\mathbb{E}X-X)} \leq \frac{v\lambda^2}{2} \tag{1}$$

where $v > 0$, the so-called "variance factor" is a parameter that is usually assumed to be known. In particular, if rewards take values in $[0,1]$, then by Hoeffding's lemma, one may take $v = 1/4$. Similarly to the asymptotic bound of [3, Th. 4.10], this moment assumption was generalized in [2, Ch. 2] by assuming that there exists a convex function $\psi : \mathbb{R}_+ \to \mathbb{R}$ such that, for all $\lambda \geq 0$,

$$\ln \mathbb{E}\, e^{\lambda(X-\mathbb{E}X)} \leq \psi(\lambda) \text{ and } \ln \mathbb{E}\, e^{\lambda(\mathbb{E}X-X)} \leq \psi(\lambda). \tag{2}$$

Then, one can show that the so-called ψ-UCB strategy (a variant of the basic UCB strategy of [4]) satisfies the following regret guarantee. Let $\Delta_i = \max_{j=1,\dots,K} \mu_j - \mu_i$, and ψ^* the Legendre–Fenchel transform of ψ, defined by

$$\psi^*(\varepsilon) = \sup_{\lambda \in \mathbb{R}} \big(\lambda\varepsilon - \psi(\lambda)\big).$$

Then, ψ-UCB[1] satisfies

$$R_n \leq \sum_{i\,:\,\Delta_i > 0} \left(\frac{4\Delta_i}{\psi^*(\Delta_i/2)} \ln n + 2 \right).$$

In particular, when the reward distributions are sub-Gaussian, the regret bound is of the order $\sum_i (\log n)/\Delta_i$, which is known to be optimal even for bounded reward distributions, see [4].

While this result shows that assumptions weaker than sub-Gaussian distributions may suffice for a logarithmic regret, it still requires the distributions to have finite moment generating function. Another disadvantage of the bound above is that the dependence on the gaps Δ_i deteriorates as the tail of the distributions become heavier. In fact, as we show it in this paper, the bound is suboptimal when the tails are heavier than sub-Gaussian.

Such heavy-tailed reward distributions naturally arise in various contexts where bandit algorithms have been used in the past. A prominent example is the distribution of delays in end-to-end network routing [5], a typical application domain for bandits—see, e.g., [2]. Another interesting example is the distribution of running times of heuristics for solving hard combinatorial problems [6], where bandit algorithms have been used to select the heuristics [7].

In this paper, we investigate the behavior of the regret when the distributions are heavy tailed, and might not have a finite moment generating function. We show that under significantly weaker assumptions, regret bounds of the same form as in the sub-Gaussian case may be achieved. In fact, the only condition we need is that the reward distributions have a finite variance. Moreover, even if the variance is infinite but the distributions have finite moments of order $1+\varepsilon$ for some $\varepsilon > 0$, one may still achieve a regret logarithmic in the number n of rounds through the dependency on the Δ_is worsens as ε gets smaller. For instance, for distributions with moment of order $1+\varepsilon$ bounded by 1 we derive a strategy that satisfies

$$R_n \leq \sum_{i\,:\,\Delta_i > 0} \left(8 \left(\frac{4}{\Delta_i} \right)^{\frac{1}{\varepsilon}} \log n + 5\Delta_i \right).$$

The key to this result is to replace the empirical mean by more refined robust estimators of the mean and construct "upper confidence bound" strategies. Note that algorithms based on the empirical mean estimator have been previously applied to heavy-

Manuscript received September 08, 2012; accepted May 12, 2013. Date of publication August 08, 2013; date of current version October 16, 2013.

S. Bubeck is with the Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08540 USA (e-mail: sebastien.bubeck@princeton.edu).

N. Cesa-Bianchi is with Dipartimento di Informatica, Università degli Studi di Milano, Milano 20135, Italy (e-mail: nicolo.cesa-bianchi@unimi.it).

G. Lugosi is with the Department of Economics, ICREA and Universitat Pompeu Fabra, Barcelona, Spain (e-mail: gabor.lugosi@gmail.com).

Communicated by V. Saligrama, Associate Editor for Signal Processing.

Digital Object Identifier 10.1109/TIT.2013.2277869

[1]More precisely, (α,ψ)-UCB with $\alpha = 4$.

tailed stochastic bandits in [8] but they obtained polynomial regret bounds while we show logarithmic regret bounds.

We also prove matching lower bounds that show that the proposed strategies are optimal up to constant factors. In particular, the dependency in $1/\Delta_i^{1/\varepsilon}$ is unavoidable.

In the following, we start by defining a general class of sampling strategies that are based on the availability of estimators of the mean with certain performance guarantees. Then, we examine various estimators for the mean. For each estimator, we describe their performance (in terms of concentration to the mean) and deduce the corresponding regret bound.

II. Robust Upper Confidence Bound Strategies

The rough idea behind upper confidence bound (UCB) strategies (see [3], [4], [9]) is that one should choose an arm for which the sum of its estimated mean and a confidence interval is highest. When the reward distributions all satisfy the sub-Gaussian condition (1) for a common variance factor v, then such a confidence interval is easy to obtain. Suppose that at a certain time instance arm i has been sampled s times and the observed rewards are $X_{i,1},\dots,X_{i,s}$. Then, the $X_{i,r}$, $r=1,\dots,s$ are i.i.d. random variables with mean $\mathbb{E}\,X_{i,r}=\mu_i$ and by a simple Chernoff bound, for any $\delta\in(0,1)$, the empirical mean $(1/s)\sum_{r=1}^{s}X_{i,r}$ satisfies with probability at least $1-\delta$,

$$\frac{1}{s}\sum_{r=1}^{s}X_{i,r}\le\mu_i+\sqrt{\frac{2v\log(1/\delta)}{s}}.$$

This property of the empirical mean turns out to be crucial in order to achieve a regret of optimal order. However, when the sub-Gaussian assumption does not hold, one cannot expect the empirical mean to have such an accuracy. In fact, if one only knows, say, that the variance of each $X_{i,r}$ is bounded, then the best possible confidence intervals are significantly wider, deteriorating the performance of standard UCB strategies. (See Appendix A for properties of the empirical mean under distributions of heavy tails.)

The key to successful handling heavy-tailed reward distributions is to replace the empirical mean with other, more robust estimators of the mean. All we need is a performance guarantee like the one shown above for the empirical mean. More precisely, we need a mean estimator with the following property.

Assumption 1: Let $\varepsilon\in(0,1]$ be a positive parameter and let c,v be positive constants. Let $X_1,\dots,X_n$ be i.i.d. random variables with finite mean μ. Suppose that for all $\delta\in(0,1)$ there exists an estimator $\widehat{\mu}=\widehat{\mu}(n,\delta)$ such that, with probability at least $1-\delta$,

$$\widehat{\mu}\le\mu+v^{1/(1+\varepsilon)}\left(\frac{c\log(1/\delta)}{n}\right)^{\varepsilon/(1+\varepsilon)}$$

and also, with probability at least $1-\delta$,

$$\mu\le\widehat{\mu}+v^{1/(1+\varepsilon)}\left(\frac{c\log(1/\delta)}{n}\right)^{\varepsilon/(1+\varepsilon)}.$$

Robust UCB:

Parameter: $\varepsilon\in(0,1]$, mean estimator $\widehat{\mu}(t,\delta)$.

For arm i, define $\widehat{\mu}_{i,s,t}$ as the estimate $\widehat{\mu}(s,t^{-2})$ based on the first s observed values $X_{i,1},\dots,X_{i,s}$ of the rewards of arm i. Define the index

$$B_{i,s,t}=\widehat{\mu}_{i,s,t}+v^{1/(1+\varepsilon)}\left(\frac{c\log t^2}{s}\right)^{\varepsilon/(1+\varepsilon)},$$

for $s,t\ge1$ and $B_{i,0,t}=+\infty$.

At time t, draw an arm maximizing $B_{i,T_i(t-1),t}$.

Fig. 1. Robust UCB policy.

For example, if the distribution of the X_t satisfies the sub-Gaussian condition (1), then Assumption 1 is satisfied for $\varepsilon=1$, $c=2$, and variance factor v. Interestingly, the assumption may be satisfied for significantly more general distributions by using more sophisticated mean estimators. We recall some of these estimators in the following sections, where we also show how they satisfy Assumption 1. As we shall see, the basic requirement for Assumption 1 to be satisfied is that the distribution of the X_t has a finite moment of order $1+\varepsilon$.

We are now ready to define our generalized robust UCB strategy, described in Fig. 1. We denote by $T_i(t)$ the (random) number of times arm i is selected up to time t.

The following proposition gives a performance bound for the robust UCB policy provided that the reward distributions and the mean estimator used by the policy jointly satisfy Assumption 1. Below we exhibit several mean estimators that, under various moment assumptions, lead to regret bounds of optimal order.

Proposition 1: Let $\varepsilon\in(0,1]$ and let $\widehat{\mu}(s,\delta)$ be a mean estimator. Suppose that the distributions $\nu_1,\dots,\nu_K$ are such that the mean estimator satisfies Assumption 1 for all $i=1,\dots,K$. Then, the regret of the Robust UCB policy satisfies

$$R_n\le\sum_{i\,:\,\Delta_i>0}\left(2c\left(\frac{v}{\Delta_i}\right)^{\frac{1}{\varepsilon}}\log n+5\Delta_i\right). \tag{3}$$

Also, if n is such that $\log n\ge\max_i\left(5\Delta_i^{(1+\varepsilon)/\varepsilon}/(2cv^{1/\varepsilon})\right)$, then

$$R_n\le n^{\frac{1}{1+\varepsilon}}\left(4Kc\log n\right)^{\frac{\varepsilon}{1+\varepsilon}}v^{1/(1+\varepsilon)}. \tag{4}$$

Note that a regret of at least $\sum_i\Delta_i$ is suffered by any strategy that pulls each arm at least once. Thus, the interesting term in (3) is the one of the order of $\sum_{i\,:\,\Delta_i>0}\left(v/\Delta_i\right)^{\frac{1}{\varepsilon}}\log n$. We show below in Theorem 2 that this term is of optimal order under a moment assumption on the reward distributions. We also show in Theorem 2 that the gap-independent inequality (4) is optimal up to a logarithmic factor.

Proof: Both proofs of (3) and (4) rely on bounding the expected number of pulls for a suboptimal arm. More precisely, in the first two steps of the proof we prove that, for any i such that $\Delta_i>0$,

$$\mathbb{E}\,T_i(n)\le2c\frac{v^{1/\varepsilon}}{\Delta_i^{(1+\varepsilon)/\varepsilon}}\log n+5. \tag{5}$$

To lighten notation, we introduce $u = \left\lceil 2c\frac{v^{1/\varepsilon}}{\Delta_i^{(1+\varepsilon)/\varepsilon}}\log n \right\rceil$. Note that, up to rounding, (5) is equivalent to $\mathbb{E}\,T_i(n) \le u + 4$.

First Step: We show that if $I_t = i$, then one of the following three inequalities is true: either

$$B_{i^*,T_{i^*}(t-1),t} \le \mu^*, \tag{6}$$

or

$$\widehat{\mu}_{i,T_i(t-1),t} > \mu_i + v^{1/(1+\varepsilon)}\left(\frac{c\log t^2}{T_i(t-1)}\right)^{\varepsilon/(1+\varepsilon)} \tag{7}$$

or

$$T_i(t-1) < 2c\frac{v^{1/\varepsilon}}{\Delta_i^{(1+\varepsilon)/\varepsilon}}\log n. \tag{8}$$

Indeed, assume that all three inequalities are false. Then, we have

$$\begin{aligned}
&B_{i^*,T_i(t-1),t}\\
&> \mu^*\\
&= \mu_i + \Delta_i\\
&\ge \mu_i + 2v^{1/(1+\varepsilon)}\left(\frac{c\log t^2}{T_i(t-1)}\right)^{\varepsilon/(1+\varepsilon)}\\
&\ge \widehat{\mu}_{i,T_i(t-1),t} + v^{1/(1+\varepsilon)}\left(\frac{c\log t^2}{T_i(t-1)}\right)^{\varepsilon/(1+\varepsilon)}\\
&= B_{i,T_i(t-1),t}
\end{aligned}$$

which implies, in particular, that $I_t \ne i$.

Second Step: Here, we first bound the probability that (6) or (7) hold. By Assumption 1 as well as an union bound over the value of $T_{i^*}(t-1)$ and $T_i(t-1)$ we obtain

$$\mathbb{P}(\text{(6) or (7) is true}) \le 2\sum_{s=1}^{t}\frac{1}{s^4} \le \frac{2}{t^3}.$$

Now using the first step, we obtain

$$\begin{aligned}
\mathbb{E}T_i(n) &= \mathbb{E}\sum_{t=1}^{n}\mathbb{1}_{I_t=i}\\
&\le u + \mathbb{E}\sum_{t=u+1}^{n}\mathbb{1}_{I_t=i \text{ and (8) is false}}\\
&\le u + \mathbb{E}\sum_{t=u+1}^{n}\mathbb{1}_{\text{(6) or (7) is true}}\\
&\le u + \sum_{t=u+1}^{n}\frac{2}{t^3}\\
&\le u + 4.
\end{aligned}$$

This concludes the proof of (5).

Third Step: Using that $R_n = \sum_{i=1}^{K}\Delta_i\mathbb{E}T_i(n)$ and (5), we directly obtain (3). On the other hand, for (4) we use Hölder's inequality to obtain

$$\begin{aligned}
&R_n\\
&= \sum_{i:\Delta_i>0}\Delta_i(\mathbb{E}T_i(n))^{\frac{\varepsilon}{1+\varepsilon}}(\mathbb{E}T_i(n))^{\frac{1}{1+\varepsilon}}\\
&\le \sum_{i:\Delta_i>0}\Delta_i(\mathbb{E}T_i(n))^{\frac{1}{1+\varepsilon}}\left(2c\frac{v^{1/\varepsilon}}{\Delta_i^{(1+\varepsilon)/\varepsilon}}\log n + 5\right)^{\frac{\varepsilon}{1+\varepsilon}}\\
&\le \sum_{i:\Delta_i>0}\Delta_i(\mathbb{E}T_i(n))^{\frac{1}{1+\varepsilon}}\left(4c\frac{v^{1/\varepsilon}}{\Delta_i^{(1+\varepsilon)/\varepsilon}}\log n\right)^{\frac{\varepsilon}{1+\varepsilon}}\\
&\quad\text{(by assumption on } n)\\
&\le K^{\frac{\varepsilon}{1+\varepsilon}}\left(\sum_{i:\Delta_i>0}\mathbb{E}T_i(n)\right)^{\frac{1}{1+\varepsilon}}(4c)^{\frac{\varepsilon}{1+\varepsilon}}v^{1/(1+\varepsilon)}(\log n)^{\frac{\varepsilon}{1+\varepsilon}}\\
&\quad\text{(by Hölder's inequality)}\\
&\le n^{\frac{1}{1+\varepsilon}}\left(4Kc\log n\right)^{\frac{\varepsilon}{1+\varepsilon}}v^{1/(1+\varepsilon)}.
\end{aligned}$$

■

In the next sections, we show how Proposition 1 may be applied, with different mean estimators, to obtain optimal regret bounds for possibly heavy-tailed reward distributions.

A. Truncated Empirical Mean

In this section, we consider the simplest of the proposed mean estimators, a truncated version of the empirical mean. This estimator is similar to the "winsorized mean" and "trimmed mean" of Tukey, see [10].

The following lemma shows that if the $(1+\varepsilon)$th raw moment is bounded, then the truncated mean satisfies Assumption 1.

Lemma 1: Let $\delta \in (0,1)$, $\varepsilon \in (0,1]$, and $u > 0$. Consider the truncated empirical mean $\widehat{\mu}_T$ defined as

$$\widehat{\mu}_T = \frac{1}{n}\sum_{t=1}^{n}X_t\mathbb{1}_{\left\{|X_t|\le\left(\frac{ut}{\log(\delta^{-1})}\right)^{\frac{1}{1+\varepsilon}}\right\}}.$$

If $\mathbb{E}|X|^{1+\varepsilon} \le u$, then with probability at least $1-\delta$,

$$\widehat{\mu}_T \le \mu + 4u^{\frac{1}{1+\varepsilon}}\left(\frac{\log(\delta^{-1})}{n}\right)^{\frac{\varepsilon}{1+\varepsilon}}.$$

Proof: Let $B_t = \left(\frac{ut}{\log(\delta^{-1})}\right)^{\frac{1}{1+\varepsilon}}$. From Bernstein's inequality for bounded random variables, noting that $\mathbb{E}\left(X^2\mathbb{1}_{|X|\le B}\right) \le uB^{1-\varepsilon}$, we have, with probability at least $1-\delta$,

$$\begin{aligned}
&\mathbb{E}X - \frac{1}{n}\sum_{t=1}^{n}X_t\mathbb{1}_{|X_t|\le B_t}\\
&= \frac{1}{n}\sum_{t=1}^{n}\left(\mathbb{E}X - \mathbb{E}\left(X\mathbb{1}_{|X|\le B_t}\right)\right)\\
&\qquad + \frac{1}{n}\sum_{t=1}^{n}\left(\mathbb{E}\left(X\mathbb{1}_{|X|\le B_t}\right) - X_t\mathbb{1}_{|X_t|\le B_t}\right)\\
&= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left(X\mathbb{1}_{|X|>B_t}\right) + \frac{1}{n}\sum_{t=1}^{n}\left(\mathbb{E}\left(X\mathbb{1}_{|X|\le B_t}\right) - X_t\mathbb{1}_{|X_t|\le B_t}\right)\\
&\le \frac{1}{n}\sum_{t=1}^{n}\frac{u}{B_t^{\varepsilon}} + \sqrt{\frac{2B_n^{1-\varepsilon}u\log(\delta^{-1})}{n}} + \frac{B_n\log(\delta^{-1})}{3n},
\end{aligned}$$

where in the last line we also used $\mathbb{E}\left(X\mathbb{1}_{|X|>B_t}\right) \leq \mathbb{E}\left(|X|^{1+\varepsilon}/B_t^{\varepsilon}\right) = u/B_t^{\varepsilon}$. An easy computation concludes the proof. ■

The following is now a straightforward corollary of Proposition 1 and Lemma 1.

Theorem 1: Let $\varepsilon \in (0,1]$ and $u > 0$. Assume that the reward distributions $\nu_1, \ldots, \nu_K$ satisfy

$$\mathbb{E}_{X\sim\nu_i}|X_i|^{1+\varepsilon} \leq u \qquad \forall i \in \{1,\ldots,K\}. \tag{9}$$

Then, the regret of the Robust-UCB policy used with the truncated mean estimator defined above satisfies

$$R_n \leq \sum_{i\,:\,\Delta_i>0}\left(8\left(\frac{4u}{\Delta_i}\right)^{\frac{1}{\varepsilon}}\log n + 5\Delta_i\right).$$

When $\varepsilon = 1$, the only assumption of the theorem above is that each reward distribution has a finite variance. In this case, the obtained regret bound is of the order of $\sum_i(\log n)/\Delta_i$, which is known to be not improvable in general, even when the rewards are bounded—note, however, that the KL-UCB algorithm of [11] is never worse than Robust-UCB in case of bounded rewards. We find it remarkable that regret of this order may be achieved under the only assumption of finite variance and one cannot improve the order by imposing stronger tail conditions.

When the variance is infinite but moments of order $1+\varepsilon$ are available, we still have a regret that depends only logarithmically on n. The bound deteriorates slightly as the dependency on $1/\Delta_i$ is replaced by $1/\Delta_i^{1/\varepsilon}$. We show next that this dependency is inevitable.

Theorem 2: For any $\Delta \in (0, 1/4)$, there exist two distributions ν_1 and ν_2 satisfying (9) with $u = 1$ and with $\mu_1 - \mu_2 = \Delta$, such that the following holds. Consider an algorithm such that for any two-armed bandit problem satisfying (9) with $u = 1$ and with arm 2 being suboptimal, one has $\mathbb{E}\,T_2(n) = o(n^a)$ for any $a > 0$. Then, on the two-armed bandit problem with distributions ν_1 and ν_2, the algorithm satisfies

$$\liminf_{n\to+\infty}\frac{R_n}{\log n} \geq \frac{0.4}{\Delta^{\frac{1}{\varepsilon}}}. \tag{10}$$

Furthermore, for any fixed n, there exists a set of K distributions satisfying (9) with $u = 1$ and such that for any algorithm, one has

$$R_n \geq 0.01\,K^{\frac{\varepsilon}{1+\varepsilon}}n^{\frac{1}{1+\varepsilon}}. \tag{11}$$

Proof: To prove (10), we take $\nu_1 = (1-\gamma^{1+\varepsilon})\delta_0 + \gamma^{1+\varepsilon}\delta_{1/\gamma}$ with $\gamma = (2\Delta)^{\frac{1}{\varepsilon}}$, and $\nu_2 = (1+\Delta\gamma-\gamma^{1+\varepsilon})\delta_0 + (\gamma^{1+\varepsilon}-\Delta\gamma)\delta_{1/\gamma}$ (with δ_x being the Dirac distribution on x). It is easy to see that ν_1 and ν_2 are well defined, and they satisfy (9) with $u = 1$ and $\mu_1 - \mu_2 = \Delta$. Now clearly, the two-armed bandit problem with these two distributions is equivalent to the two-armed bandit problem with two Bernoulli distributions with parameters $\gamma^{1+\varepsilon}$ and $\gamma^{1+\varepsilon} - \Delta\gamma$, respectively. Slightly more formally, we could define a new algorithm $\mathcal{A}'$ that on $\mathrm{Ber}(\gamma^{1+\varepsilon}), \mathrm{Ber}(\gamma^{1+\varepsilon} - \Delta\gamma)$ behaves equivalently to the original algorithm $\mathcal{A}$ on ν_1 and ν_2. Therefore, we can use [12, Th. 2.7] to directly obtain the following lower bound for $\mathcal{A}'$:

$$\liminf_{n\to+\infty}\frac{\mathbb{E}\,T_2(n)}{\log n} \geq \frac{1}{\mathrm{KL}\Big(\mathrm{Ber}(\gamma^{1+\varepsilon}-\Delta\gamma), \mathrm{Ber}(\gamma^{1+\varepsilon})\Big)}$$

where KL denotes Kullback–Leibler divergence. This implies the following lower bound for the original algorithm $\mathcal{A}$:

$$\liminf_{n\to+\infty}\frac{R_n}{\log n} \geq \frac{\Delta}{\mathrm{KL}\Big(\mathrm{Ber}(\gamma^{1+\varepsilon}-\Delta\gamma), \mathrm{Ber}(\gamma^{1+\varepsilon})\Big)}.$$

Equation (10) then follows directly by using $\mathrm{KL}(\mathrm{Ber}(p), \mathrm{Ber}(q)) \leq \frac{(p-q)^2}{q(1-q)}$ along with straightforward computations.

The proof of (11) follows the same scheme. We use the same distributions as above and we consider the multiarmed bandit problem where one arm has distribution ν_1, and the $K-1$ remaining arms have distribution ν_2. Furthermore, we set $\Delta = (K/n)^{\frac{\varepsilon}{1+\varepsilon}}$ for this part of the proof. Now we can use the same proof as for [12, Th. 2.6] on the modified algorithm $\mathcal{A}'$ that runs on the Bernoulli distributions corresponding to ν_1 and ν_2. We leave the straightforward details to the reader. ■

B. Median of Means

The truncated mean estimator and the corresponding bandit strategy are not entirely satisfactory as they are not translation invariant in the sense that the arms selected by the strategy may change if all reward distributions are shifted by the same constant amount. The reason for this is that the truncation is centered, quite arbitrarily, around zero. If the raw moments $\mathbb{E}_{X\sim\nu_i}|X|^{1+\varepsilon}$ are small, then the strategy has a small regret. However, it would be more desirable to have a regret bound in terms of the centered moments $\mathbb{E}_{X\sim\nu_i}|X-\mu_i|^{1+\varepsilon}$. This is indeed possible if one replaces the truncated mean estimator by more sophisticated estimators of the mean. We show one such possibility, the "median-of-means" estimator in this section. In the next section, we discuss Catoni's M-estimator, a quite different alternative.

The median-of-means estimator was proposed by [13] and [14]. The simple idea is to divide the data into various disjoint blocks. Within each block one calculates the standard empirical mean and takes a median value of these empirical means. The next lemma shows that for certain block size the estimator has the property required by our robust UCB strategy.

Lemma 2: Let $\delta \in (0,1)$ and $\varepsilon \in (0,1]$ and $n \geq 16\log(1/\delta) + 2$. Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean $\mathbb{E}\,X = \mu$ and centered $(1+\varepsilon)$th moment $\mathbb{E}|X-\mu|^{1+\varepsilon} = u$. Let $k = \lfloor 8\log(e^{1/8}/\delta) \wedge n/2 \rfloor$ and $N = \lfloor n/k \rfloor$. Let

$$\hat{\mu}_1 = \frac{1}{N}\sum_{t=1}^{N} X_t\ ,\ \ldots\ ,\ \hat{\mu}_k = \frac{1}{N}\sum_{t=(k-1)N+1}^{kN} X_t$$

Modified robust UCB–median of means:

For arm i, define $\widehat{\mu}_{i,s,t}$ as the median-of-means estimate $\widehat{\mu}_M(s,t^{-2})$ based on the first s observed values $X_{i,1},\dots,X_{i,s}$ of the rewards of arm i. Define the index

$$B_{i,s,t} = \widehat{\mu}_{i,s,t} + (12v)^{1/(1+\varepsilon)} \left(\frac{16\log(e^{1/8}t^2)}{s}\right)^{\varepsilon/(1+\varepsilon)},$$

for $s,t \geq 1$ such that $s \geq 32\log t + 2$ and $B_{i,s,t} = +\infty$ otherwise.

At time t, draw an arm maximizing $B_{i,T_i(t-1),t}$.

Fig. 2. Modified robust UCB policy with the median-of-means estimator.

be k empirical mean estimates, each one computed on N data points. Consider a median $\widehat{\mu}_M$ of these empirical means. Then, with probability at least $1-\delta$,

$$\widehat{\mu}_M \leq \mu + (12v)^{\frac{1}{1+\varepsilon}} \left(\frac{16\log(e^{1/8}\delta^{-1})}{n}\right)^{\frac{\varepsilon}{1+\varepsilon}}.$$

Proof: Let $\eta > 0$ and $Y_\ell = \mathbb{1}_{\widehat{\mu}_\ell > \mu+\eta}$ for $\ell \in \{1,\dots,k\}$. According to (12) in the Appendix, Y_ℓ has a Bernoulli distribution with parameter

$$p \leq \frac{3v}{N^\varepsilon \eta^{1+\varepsilon}}.$$

Note that for

$$\eta = (12v)^{\frac{1}{1+\varepsilon}} \left(\frac{1}{N}\right)^{\frac{\varepsilon}{1+\varepsilon}}$$

we have $p \leq 1/4$. Thus, using Hoeffding's inequality for the tail of a binomial distribution, we obtain

$$\begin{aligned}\mathbb{P}(\widehat{\mu}_M > \mu+\eta) &= \mathbb{P}\left(\sum_{\ell=1}^k Y_\ell \geq k/2\right) \\ &\leq \exp\left(-2k(1/2-p)^2\right) \\ &\leq \exp(-k/8) \leq \delta.\end{aligned}$$

■

This bound has the same form as in Assumption 1, though it only holds with the additional requirement that $n \geq 16\log(1/\delta)+2$ and therefore it does not formally fit in the framework of the robust UCB strategy as described in Section II. However, by a simple modification, one may define a strategy that incorporates such a restriction. In Fig. 2, we describe a policy based on the median-of-means estimator. Then, by a simple modification of the proof of Proposition 1, and using Lemma 2, we obtain the performance bound below. In some situations it significantly improves on Theorem 1 as the bound depends on the centered moments of order $1+\varepsilon$ rather than on raw moments. However, a term of the order $\sum_i \Delta_i \log n$ appears due to the restricted range of validity of the median-of-means estimator.

Theorem 3: Let $\varepsilon \in (0,1]$ and $v > 0$. Assume that the reward distributions $\nu_1,\dots,\nu_K$ satisfy

$$\mathbb{E}_{X\sim\nu_i}|X-\mu_i|^{1+\varepsilon} \leq v \qquad \forall i \in \{1,\dots,K\}.$$

Then, the regret of the Robust-UCB policy used with the median-of-means mean estimator defined in Lemma 2 satisfies

$$R_n \leq \sum_{i\,:\,\Delta_i>0} \left(32\left(\frac{12v}{\Delta_i}\right)^{\frac{1}{\varepsilon}} \log n + 32\Delta_i \log n + 7\Delta_i\right).$$

C. Catoni's M Estimator

Finally, we consider an elegant mean estimator introduced by [15]. As we will see, this estimator has similar performance guarantees as the median-of-means estimator but with better, near optimal, numerical constants. However, we only have a good guarantee in terms of the variance. Thus, in this section we assume that the variance is finite and we do not consider the case $\varepsilon < 1$.

Catoni's mean estimator is defined as follows: let $\psi : \mathbb{R} \to \mathbb{R}$ be a continuous strictly increasing function satisfying

$$-\log(1-x+x^2/2) \leq \psi(x) \leq \log(1+x+x^2/2).$$

Let $\delta \in (0,1)$ be such that $n > 2\log(1/\delta)$ and introduce

$$\alpha_\delta = \sqrt{\frac{2\log(1/\delta)}{n(v+\frac{2v\log(1/\delta)}{n-2\log(1/\delta)})}}.$$

If $X_1,\dots,X_n$ be i.i.d. random variables, then Catoni's estimator is defined as the unique value $\widehat{\mu}_C = \widehat{\mu}_C(n,\delta)$ such that

$$\sum_{i=1}^n \psi\big(\alpha_\delta(X_i - \widehat{\mu}_C)\big) = 0.$$

Catoni [15] proves that if $n \geq 4\log(1/\delta)$ and the X_i have mean μ and variance at most v, then, with probability at least $1-\delta$,

$$\widehat{\mu}_C \leq \mu + 2\sqrt{\frac{v\log\delta^{-1}}{n}}$$

and a similar bound holds for the lower tail.

Similarly to the case of the median-of-means estimator, here we also have an additional requirement that $n \geq 4\log(1/\delta)$ and the general estimator described at the beginning of Section II needs to be slightly modified. The policy described in Fig. 3 assumes that there is a known upper bound v for the largest variance of any reward distribution. Then, by a simple modification of the proof of Proposition 1, we obtain the following performance bound.

Modified robust UCB–Catoni's estimator:

For arm i, define $\widehat{\mu}_{i,s,t}$ as Catoni's mean estimate $\widehat{\mu}_C(s,t^{-2})$ based on the first s observed values $X_{i,1},\dots,X_{i,s}$ of the rewards of arm i. Define the index

$$B_{i,s,t} = \widehat{\mu}_{i,s,t} + \left(\frac{4v\log t^2}{s}\right)^{1/2},$$

for $s,t \geq 1$ such that $s \geq 8\log t$ and $B_{i,s,t} = +\infty$ otherwise.

At time t, draw an arm maximizing $B_{i,T_i(t-1),t}$.

Fig. 3. Modified robust UCB policy with Catoni's estimator.

Theorem 4: Let $v > 0$. Assume that the reward distributions $\nu_1,\dots,\nu_K$ satisfy

$$\mathbb{E}_{X\sim\nu_i}|X-\mu_i|^2 \leq v \ \ \forall i \in \{1,\dots,K\}.$$

Then, the regret of the modified robust UCB policy satisfies

$$R_n \leq \sum_{i:\Delta_i>0}\left(\frac{8v\log n}{\Delta_i} + 8\Delta_i\log n + 5\Delta_i\right).$$

The regret bound has better numerical constants than its analog based on the median-of-means estimator.

III. Discussion and Conclusion

In this paper, we have extended the UCB algorithm to heavy-tailed stochastic multiarmed bandit problems in which the reward distributions have only moments of order $1+\varepsilon$ for some $\varepsilon \in (0,1]$. In this setting, we have compared three estimators for the mean reward of the arms: median-of-means, truncated mean, and Catoni's M-estimator. The median-of-means estimator gives a regret bound that depends on the central $(1+\varepsilon)$-moments of the reward distributions, without need of knowing bounds on these moments. The truncated mean estimator, instead, delivers a regret bound that depends on the raw $(1+\varepsilon)$-moments, and requires the knowledge of a bound u on these moments. Finally, Catoni's estimator depends on the central moments like the median-of-means, but it requires the knowledge of a bound v on the central moments, and only works in the special case $\varepsilon = 1$ (where it gives the best leading constants on the regret). A tradeoff in the choice of the estimator appears if we take into account the computational costs involved in the update of each estimator as new rewards are observed. Indeed, while the truncated mean requires constant time and space per update, the median-of-means is slightly more difficult to update, requiring $O(\log\delta^{-1})$ space and $O(\log\log\delta^{-1})$ time per update. Finally, Catoni's M-estimator requires linear space per update, which is an unfortunate feature in this sequential setting.

It is an interesting question whether there exists an estimator with the same good concentration properties as the median-of-means, but requiring only constant time and space per update. The truncated mean has good computational properties but the knowledge of raw moment bounds is required. So it is natural to ask whether we may drop this requirement for the truncated mean or some variants of it. Finally, our proof techniques heavily rely on the independence of rewards for each arm. It is unclear whether similar results could be obtained for heavy-tailed bandits with dependent reward processes.

While we focused our attention on bandit problems, the concentration results presented in this paper may be naturally applied to other related sequential decision settings. Such examples include the racing algorithms of [16], and more generally nonparametric Monte Carlo estimation, see [17] and [18]. These techniques are based on mean estimators, and current results are limited to the application of the empirical mean to bounded reward distributions.

Appendix
Empirical Mean

In this Appendix, we discuss the behavior of the standard empirical mean when only moments of order $1+\varepsilon$ are available. We focus on finite-sample guarantees (i.e., nonasymptotic results), as this is the key property to obtain finite-time results for the multiarmed bandit problem.

Let $X, X_1,\dots,X_n$ be a real i.i.d. sequence with finite mean μ. We assume that for some $\varepsilon \in (0,1]$ and $v \geq 0$, one has $\mathbb{E}|X-\mu|^{1+\varepsilon} \leq v$. We also denote by u an upper bound on the raw moment of order $1+\varepsilon$, that is $\mathbb{E}|X|^{1+\varepsilon} \leq u$.

Lemma 3: Let $\widehat{\mu}$ be the empirical mean

$$\widehat{\mu} = \frac{1}{n}\sum_{t=1}^{n} X_t.$$

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, one has

$$\widehat{\mu} \leq \mu + \left(\frac{3v}{\delta n^{\varepsilon}}\right)^{\frac{1}{1+\varepsilon}}.$$

Proof: Let $\eta, a > 0$,

$$\mathbb{P}(\widehat{\mu}-\mu > \eta) \leq \mathbb{P}\big(\exists t \in \{1,\dots,n\} : |X_t-\mu| > a\big) + \mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n}(X_t-\mu)\mathbb{1}_{|X_t-\mu|\leq a} > \eta\right).$$

The first term on the right-hand side can be bounded by using a union bound followed by Chebyshev's inequality (for moments of order $1+\varepsilon$):

$$\mathbb{P}\big(\exists t \in \{1,\dots,n\} : |X_t-\mu| > a\big) \leq n\frac{\mathbb{E}|X-\mu|^{1+\varepsilon}}{a^{1+\varepsilon}} \leq \frac{nv}{a^{1+\varepsilon}}.$$

On the other hand, Chebyshev's inequality together with the fact that $\mathbb{E}(X-\mu)\mathbb{1}_{|X-\mu|\leq a} = -\mathbb{E}(X-\mu)\mathbb{1}_{|X-\mu|>a}$ give for the second term

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n}(X_t-\mu)\mathbb{1}_{|X_t-\mu|\leq a} > \eta\right) \\
&\quad\leq \frac{1}{\eta^2}\mathbb{E}\left(\frac{1}{n}\sum_{t=1}^{n}(X_t-\mu)\mathbb{1}_{|X_t-\mu|\leq a}\right)^2 \\
&\quad\leq \frac{\mathbb{E}(X-\mu)^2\mathbb{1}_{|X-\mu|\leq a}}{n\eta^2} + \frac{\big(\mathbb{E}(X-\mu)\mathbb{1}_{|X-\mu|\leq a}\big)^2}{\eta^2} \\
&\quad= \frac{\mathbb{E}(X-\mu)^2\mathbb{1}_{|X-\mu|\leq a}}{n\eta^2} + \frac{\big(\mathbb{E}(X-\mu)\mathbb{1}_{|X-\mu|> a}\big)^2}{\eta^2}.
\end{aligned}$$

By applying a trivial manipulation on the first term, and using Hölder's inequality with exponents $p = 1+\varepsilon$ and $q = 1+1/\varepsilon$ for the second term, we obtain that the last expression above is upper bounded by

$$\frac{\mathbb{E}|X-\mu|^{1+\varepsilon}a^{1-\varepsilon}}{n\eta^2} + \frac{\left(\mathbb{E}|X-\mu|^{1+\varepsilon}\right)^{\frac{2}{1+\varepsilon}}\left(\mathbb{P}(|X-\mu|>a)\right)^{\frac{2\varepsilon}{1+\varepsilon}}}{\eta^2}$$
$$\leq \frac{va^{1-\varepsilon}}{n\eta^2} + \frac{v^{\frac{2}{1+\varepsilon}}v^{\frac{2\varepsilon}{1+\varepsilon}}}{\eta^2 a^{2\varepsilon}}.$$

Thus, we proved that

$$\mathbb{P}(\widehat{\mu}-\mu>\eta) \leq \frac{nv}{a^{1+\varepsilon}} + \frac{va^{1-\varepsilon}}{n\eta^2} + \frac{v^2}{\eta^2 a^{2\varepsilon}}.$$

Taking $a = n\eta$ entails

$$\mathbb{P}(\widehat{\mu}-\mu>\eta) \leq \frac{2v}{n^{\varepsilon}\eta^{1+\varepsilon}} + \left(\frac{v}{n^{\varepsilon}\eta^{1+\varepsilon}}\right)^2.$$

Note that if $\frac{v}{n^{\varepsilon}\eta^{1+\varepsilon}} > 1$ then the bound is trivial, and thus we always have

$$\mathbb{P}(\widehat{\mu}-\mu>\eta) \leq \frac{3v}{n^{\varepsilon}\eta^{1+\varepsilon}}. \tag{12}$$

The proof now follows by straightforward computations. ■

It is easy to see that the order of magnitude of (12) is tight up to a constant factor. Indeed, let $\gamma \in (0,1)$ and consider the distribution $(1-\gamma^{1+\varepsilon})\delta_0 + \gamma^{1+\varepsilon}\delta_{1/\gamma}$ (with δ_x being the Dirac distribution on x). Clearly, for this distribution we have $\mathbb{E}|X-\mu|^{1+\varepsilon} \leq 1$, so (12) shows that for an i.i.d. sequence drawn from this distribution, one has

$$\mathbb{P}(\widehat{\mu}-\mu>\eta) \leq \frac{3}{n^{\varepsilon}\eta^{1+\varepsilon}}.$$

We can restrict our attention to the case where $\eta > n^{-\frac{\varepsilon}{1+\varepsilon}}$, for otherwise the above upper bound is trivial. Now consider $\gamma = \frac{1}{2n\eta}$. Note that we have $\mu = \gamma^{\varepsilon} = \frac{1}{(2n\eta)^{\varepsilon}} < \eta$ and in particular this implies $1/\gamma = 2n\eta > n(\eta+\mu)$. From this last inequality and basic computations, we obtain

$$\begin{aligned}
\mathbb{P}(\widehat{\mu}-\mu>\eta) &\geq \mathbb{P}\big(\exists\, i \in \{1,\dots,n\} : X_i \geq n(\eta+\mu)\big) \\
&\geq \mathbb{P}\big(\exists\, i \in \{1,\dots,n\} : X_i = 1/\gamma\big) \\
&= 1-(1-\gamma^{1+\varepsilon})^n \\
&= 1-\exp\left(n\ln\left(1-\frac{1}{(2n\eta)^{1+\varepsilon}}\right)\right) \\
&\geq 1-\exp\left(-\frac{1}{n^{\varepsilon}(2\eta)^{1+\varepsilon}}\right) \\
&= \frac{1}{n^{\varepsilon}(2\eta)^{1+\varepsilon}} + o\left(\frac{1}{n^{\varepsilon}(2\eta)^{1+\varepsilon}}\right)
\end{aligned}$$

which shows that (12) is tight up to a constant factor for this distribution.

Clearly, the concentration properties of the empirical mean are much weaker than for the truncated empirical mean or the median-of-means. Indeed, while the dependency on n in the confidence term is similar for the three estimators, the dependency on $1/\delta$ is polynomial for the empirical mean and polylogarithmic for the truncated empirical mean and the median-of-means. As we just showed, this is not an artifact of the proof method, and the empirical mean indeed has polynomial deviations (as opposed to the exponential deviations of the two other estimators). This remark is at the basis of the theory of robust statistics and many approaches to fix the above issue have been proposed, see for example [19], [20].

REFERENCES

[1] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, vol. 58, pp. 527–535, 1952.

[2] S. Bubeck and N. Cesa-Bianchi, Regret analysis of stochastic and nonstochastic multi-armed bandit problems 2012 [Online]. Available: arXiv:1204.5721, to be published

[3] R. Agrawal, "Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem," *Adv. Appl. Math.*, vol. 27, pp. 1054–1078, 1995.

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learning J.*, vol. 47, no. 2-3, pp. 235–256, 2002.

[5] J. Liebeherr, A. Burchard, and F. Ciucu, "Delay bounds in communication networks with heavy-tailed and self-similar traffic," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 1010–1024, Feb. 2012.

[6] C. P. Gomes, B. Selman, N. Crato, and H. Kautz, "Heavy-tailed phenomena in satisfiability and constraint satisfaction problems," *J. Autom. Reasoning*, vol. 24, no. 1-2, pp. 67–100, 2000.

[7] M. Gagliolo and J. Schmidhuber, "Algorithm portfolio selection as a bandit problem with unbounded losses," *Annu. Math. Artif. Intell.*, vol. 61, no. 2, pp. 49–86, 2011.

[8] K. Liu and Q. Zhao, "Deterministic sequencing of exploration and exploitation for multi-armed bandit problems," *ArXiv e-Prints*, 2011.

[9] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, pp. 4–22, 1985.

[10] P. J. Bickel, "On some robust estimates of location," *Annu. Math. Statist.*, vol. 36, pp. 847–858, 1965.

[11] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," presented at the 24th Annu. Conf. Learning Theory, 2011.

[12] S. Bubeck, "Bandits Games and Clustering Foundations," Ph.D. dissertation, Université Lille 1, Lille, France, 2010.

[13] A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*. New York, NY, USA: Wiley, 1983.

[14] N. Alon, Y. Matias, and M. Szegedy, "The space complexity of approximating the frequency moments," *J. Comput. Syst. Sci.*, vol. 58, pp. 137–147, 2002.

[15] O. Catoni, Challenging the empirical mean and empirical variance: A deviation study 2010 [Online]. Available: arXiv:1009.2048, to be published

[16] O. Maron and A. Moore, "The racing algorithm: Model selection for lazy learners," *Artif. Intell. Rev.*, vol. 11, no. 1, pp. 193–225, 1997.

[17] P. Dagum, R. Karp, M. Luby, and S. Ross, "An optimal algorithm for Monte Carlo estimation," *SIAM J. Comput.*, vol. 29, no. 5, pp. 1484–1496, 2000.

[18] C. Domingo, R. Gavaldà, and O. Watanabe, "Adaptive sampling methods for scaling up knowledge discovery algorithms," *Data Mining Knowl. Discov.*, vol. 6, no. 2, pp. 131–152, 2002.

[19] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, pp. 73–101, 1964.

[20] P. J. Huber, *Robust Statistics*. New York, NY, USA: Wiley Interscience, 1981.

Sébastien Bubeck, biography not available at the time of publication.

Nicolò Cesa-Bianchi, biography not available at the time of publication.

Gábor Lugosi, biography not available at the time of publication.