# Non-stationary Bandits with Heavy Tail

Weici Pan, Zhenhua Liu
Stony Brook University
weici.pan@stonybrook.edu, zhenhua.liu@stonybrook.edu

## ABSTRACT

In this study, we investigate the performance of multi-armed bandit algorithms in environments characterized by heavy-tailed and non-stationary reward distributions, a setting that deviates from the conventional risk-neutral and sub-Gaussian assumptions. We specifically focus on extending Upper-Confidence Bound (UCB)-based policies to accommodate heavy-tailed reward distributions while preserving their performance guarantees in non-stationary contexts by change-point detection. We provide a rigorous analysis of the proposed algorithm by establishing a regret bound in the style of $\sqrt{T} + \log T$ in the time horizon $T$, in stochastic cases. Our results contribute to the understanding of multi-armed bandits in more complex and realistic environments, with potential implications for various applications in machine learning and decision-making under uncertainty.

## 1. INTRODUCTION

Decision-making in uncertain environments is a crucial challenge in fields like engineering, economics, social science, and ecology. Two significant obstacles in these problems are nonstationarity, where the environment changes over time, and heavy-tailedness, which involves extreme events that are rare but impactful. An effective strategy in such environments needs to balance exploration, or gathering new information, and exploitation, or using existing knowledge to make optimal decisions.

The Multi-armed Bandit (MAB) framework introduced by [8] is a well-known approach for sequential decision-making that navigates the exploration-exploitation trade-off. It is described as follows: an agent facing $K$ actions (or bandit arms) selects one arm at every time step. With each arm $i \in \{1, \cdots, K\}$ there is an associated probability distribution $\nu_i$ with finite mean $\mu_i$. These distributions are unknown to the agent. At each round $t = 1, \cdots, T$, the agent chooses an arm $I_t$, and observes a reward drawn from $\nu_{I_t}$ independently from the past given $I_t$. The goal of the agent is to minimize the regret

$$R_T = T \max_{i=1,\cdots,K} \mu_i - \sum_{t=1}^{T} \mathbb{E}\nu_{I_t}.$$

The simplicity and versatility of the MAB framework have led to its application in various fields. For example, MAB is used in online advertising [10], recommendation systems [12], network routing [5], portfolio optimization [6], and so on.

However, the traditional assumptions of stationarity and sub-Gaussian reward distributions in classic MAB problems are often too restrictive for real-world situations, where environments can change, and rewards can have heavy tails due to extreme events. In this work, we extend the MAB framework to address both nonstationarity and heavy-tailedness, aiming to create more robust and adaptable algorithms for sequential decision-making in complex environments. Non-stationarity and heavy-tailedness are challenging but necessary since. Even though the two problems are explored since a long time ago, little work has been seen fully combining the two challenges together except for [11, 2]. The former relies on the knowledge of changing frequency and the budget while the latter majorly focused on the pseudo-regret according to some kind of risk.

To the best of our knowledge, there has been no prior work addressing the topic of regret minimization on the heavy-tailed non-stationary bandits. With this in mind, we would like to introduce our main contributions, which are as follows:

- We design a novel regret minimization algorithm in the novel heavy-tailed piecewise-stationary bandit setting, with common assumptions. The algorithm is based on the change-point detection, and repeated restarting the optimal bandit algorithm.
- We prove our algorithm achieves a regret of $O(\sqrt{KYT} + \sigma \sum_{i \neq i^*} \Delta_i^{-1} \log T)$, $Y$ being the number of change points. Its $O(\sqrt{T})$ dependency on time horizon $T$ matches the previous work where it is sub-Gaussian but non-stationary, as shown in Table 1. The other item in the regret bound shows a exponential relation, that is, $O(\sigma)$, in the heavy-tailedness parameters $\sigma$, showing that the heavier the tails are, the larger would the regret grow. In the instance-independent meaning, the algorithm achieves regret of $O(\sqrt{KT}(\sqrt{\sigma} \log T + \sqrt{Y}))$.

| Regret | Stationary | Non-stationary |
|---|---|---|
| Sub-Gaussain | $\log T$[1] | $\sqrt{T}$[4] |
| Heavy-tailed | $\log T$-style [3] | $\sqrt{T} \log T$ |

**Table 1: Best regret bound achieved by different bandits under different stochastic settings**

## 2. PRELIMINARIES

We assume that there exists a unique optimal arm. It is a common assumption in MAB and RL literature. Denote the optimal arm in hindsight to be the $i^*$-th arm, and define $\Delta_i$ for other arms to be the suboptimally gap between it and the optimal arm, that is,

$$\Delta_i = \mu_i - \mu_{i^*}.$$

### 2.1 Piecewise stationary

To model the non-stationarity, we define the piecewise-stationary reward process similar to [13]. It changes its distribution arbitrarily and at arbitrary rounds, but otherwise remains stationary. Let $r_t$ represent the reward vector at time $t$, with the $i$-th element $r_t(i)$ representing the reward associated with the $i$-th arm. The reward sequence $r_1, r_2, \cdots$ is an independent sequence of random variables that undergoes abrupt changes in distribution at unknown rounds $y_1, y_2, \cdots \in Y$ called change-points. Let $\nu(t) = \{\nu_1(t), \cdots, \nu_K(t)\}$ denote the distribution of $r_t$. Hence, $r_{y_j}$, $r_{y_j+1}, \cdots, r_{y_{j+1}-1}$ have common distribution $\nu(y_j)$. And we can rewrite the number of the piecewise-stationary regimes $Y$ divided by the change points as follows:

$$Y = 1 + \sum_{t=1}^{T-1} \mathbb{1}\{\nu_i(t) \neq \nu_i(t+1) \text{ for some } i \in \{1, \cdots, K\}\}.$$

Throughout the paper we also make the following non-parametric assumption on the distributions families.

Assumption 1. *We assume that for all $t \in \{1, \cdots, T\}$ and for all $i \in \{1, \cdots, K\}$:*

$$\mathbb{E}_{r \sim \nu_i(t)} \|r - \mu_i(t)\|^2 \leq \sigma,$$

*for some known $\sigma < \infty$.*

This assumption encompass a wide range of families such as heavy-tailed distributions that do not have finite higher moments.

## 3. ALGORITHMS

Our framework consists of two main components: an optimal bandit algorithm for heavy-tailed distribution, and the restarted Bayesian online change-point detector. At each round $t$ and based on the past observations, the bandit outputs a decision. By playing action, the environment reveals a reward which is observed by both the bandit algorithm and the detector instance. The sequential change-point detector which monitors the distribution of each arm either sends a positive signal to restart the estimated parameters related to the played arm when a change point is detected, or sends a negative signal when no change is observed.

The detector algorithm is designed to solve the problem of sequential change-point detection in a setting where both the change points and the distributions before and after the change are assumed to be unknown. This setting corresponds exactly to the situation of an agent facing a multi-armed bandit whose distributions are unknown and may change abruptly at some unknown instants.

The framework is as follows:

By forced exploration, we try to ensure each arm is sampled enough and changes can also be detected on arms currently under-sampled by the bandit algorithm. In the majority of cases where the environment is described by several change-points, these change-point can affect sub-sampled arms. Thus, for local changes, it is not enough to combine even an optimal bandit algorithm with an optimal online change point detector strategy. In this way, the bandit will play the arm whose current index is maximal with high probability or sample uniformly the arm set with low probability with force exploration.

**Algorithm 1** Change-point Detection for Heavy-tailed Bandit

```
Input: Arms 1 to $K$, $\alpha \in (0,1)$: forced exploration rate, $T$: time horizon.
for $t = 1, \cdots, T$ do
    For all arm $i = 1, \cdots, K$, with probability $\alpha/K$, choose arm $i$    ▷ Forced exploration
    Otherwise with probability $1-\alpha$, choose the decision of the bandit subroutine, Algorithm 2
    Take the index of the chosen arm, say, $j$
    Take the reward sequence of rounds pulling $j$ since last restart, say, $r_{j_1}, r_{j_2}, \cdots$
    Run the detector subroutine, Algorithm 3 on the reward sequence of arm $j$ pulled
    if Positive answer from DETECT then
        Clear the history reward sequences and restart ESTIMATE
    end if
end for
```

### 3.1 Optimal Bandit

During each piecewise-stationary regime, that is, between every consecutive two restarts of Algorithm 1, we can run the bandit algorithm as stationary. Therefore, we can treat the mean of each arm $i$ as a time-invariant $\mu_i$. Here we elaborate on the optimal bandit subroutine algorithm used in Algorithm 1:

**Algorithm 2** ESTIMATE

```
Input: number of arms $K$, horizon $T$
Define the number of times arm $i$ being pulled at time $t$ to be $T_i(t)$
For all arm $i = 1, \cdots, K$, define $\hat{\mu}_{i,s,t}$ as the estimate of $\mu_i$
Define the index to be $B_{i,s,t} = \hat{\mu}_{i,s,t} + \left(\frac{8v \log t}{s}\right)^1 /2$ for $s \geq 8 \log t$, $s, t \geq 1$
Otherwise define $B_{i,s,t} = +\infty$
for $t = 1, \cdots, T$ do
    Draw the arm maximizing $B_{i,T_i(t-1),t}$
    Observe the reward
    Update the estimate $\hat{\mu}$ according to equation 1
    Update the pulling counts $T_i$ and indices $B$
end for
```

Here we define the Catoni's estimator as follows: Define function $\psi : \mathbb{R} \to \mathbb{R}$ to 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 $T \geq -4 \log \delta$ and the history reward sequence to be $x_1, \cdots, x_s$, the we have the Catoni's estimator as the unique value $\hat{\mu}$ such that

$$\sum_{j=1}^{s}\psi\left(\sqrt{\frac{-2\log\delta}{T(v+\frac{-2v\log\delta}{T+2\log\delta})}}(x_j-\hat{\mu})\right)=0 \qquad (1)$$

### 3.2 Change Detection

Then we introduce the restarted Bayesian change-point detector. More formally, for a sequence $X_1,\cdots,X_s$, we assume each $X_i$ is drawn from a distribution with mean $\theta_i$. Then the algorithm will take the sequence as input and return binary answer. If the answer is negative, it means that we have $\theta_t=\theta_1$ for all $t=2,\cdots,s$. Otherwise, when the algorithm outputs positive, it means that there exists some $i\leq s-1$ so that $\theta_t=\theta_1$ for $t=2,\cdots,i$ while $\theta_t=\theta_{i+1}$ for $t=i+1,\cdots,s$. Before we present the detector in Algorithm 3, we define $\prod_\theta$ to be the projection operator onto the set of $\theta$, and that

$$\text{clip}(x,\lambda)=x\min(1,\lambda/\|x\|).$$

**Algorithm 3** DETECTOR

```
Input: ${\eta_t}, \lambda > 0, \theta_0 \in \Theta, G$ the diameter of $\Theta, \chi \in (0,1)$
and the given sequence $X_1, \cdots, X_s$.
Initialize all estimation $\hat{\theta}_{t,t-1} \leftarrow \theta_0$
for $t = 1, \cdots, s$ do
    $\hat{\theta}_{i,t} \leftarrow \prod_\theta(\hat{\theta}_{i,t-1} - \eta_{t-i}\text{clip}(X_t - \hat{\theta}_{i,t-1}, \lambda))$ for every $i \leq t$
    if there exist $i$ satisfying the criteria in the building of [9] then
        Detect change at $t$ and output positive
    end if
end for
When no change ever detected, output negative
```

The detection delay is defined as the number of samples needed to detect a change. According to [7], the detection delay of our detector is asymptotically optimal in the sense that it reaches the existing lower bound. The false alarm rate corresponds to the probability of detecting a change at some instant where there is no change.

Note that we borrow the criteria from [9], which also gives an upper bound on the worst case detection delay. We will elaborate on this and the undetectable changes in future works.

## 4. MAIN RESULTS

In this section, we provide a mathematical analysis of the regret upper bound related to the application of the framework.

We state the regret bound on each piecewise-stationary regime, in other words, between each two consecutive restarts in Algorithm 1:

LEMMA 1. *Between each two consecutive restarts in Algorithm 1, denote the time span to be $T'$, we have the regret of Algorithm 2 in this interval to be:*

$$R'_{T'}\leq\sum_{i:\Delta_i>0}((8\sigma/\Delta_i)\log T+8\Delta_i\log T+5\Delta_i)$$

Finally, we combine the results to state the regret

THEOREM 2. *Algorithm 1 achieves regret that*

$$R_T\leq O(\sigma^{1/2}\sum_{i\neq i^*}\Delta_i^{-1/2}\log T+\sqrt{KY_TT})$$

*Instance-independently we have that*

$$R_T\leq O(\sqrt{\sigma KT\log T}+\sqrt{KY_TT})$$

This bound can be comprehended in two parts, that is, $O(\sigma^{1/2}\sum_{i\neq i^*}\Delta_i^{-1/2}\log T$ and $O(\sqrt{KY_TT})$. The former one corresponds to the regret bound in [3] where it is set to be stationary but heavy tailed. And the latter $O(\sqrt{T})$ dependency on time horizon $T$ matches most non-stationary work with sub-Gaussian distributions.

## 5. REFERENCES

[1] R. Agrawal. Sample mean based index policies by o (log n) regret for the multi-armed bandit problem. *Advances in applied probability*, 27(4):1054–1078, 1995.

[2] S. Bhatt, G. Fang, and P. Li. Piecewise stationary bandits under risk criteria. In *International Conference on Artificial Intelligence and Statistics*, pages 4313–4335. PMLR, 2023.

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

[4] A. Garivier and E. Moulines. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*, 2008.

[5] Y. Guo, Y. Wang, F. Khan, A. A. Al-Atawi, A. A. Abdulwahid, Y. Lee, and B. Marapelli. Traffic management in iot backbone networks using gnn and mab with sdn orchestration. *Sensors*, 23(16):7091, 2023.

[6] N. Gutowski, T. Amghar, O. Camp, and F. Chhel. Gorthaur-exp3: Bandit-based selection from a portfolio of recommendation algorithms balancing the accuracy-diversity dilemma. *Information Sciences*, 546:378–396, 2021.

[7] T. L. Lai and H. Xing. Sequential change-point detection when the pre-and post-change parameters are unknown. *Sequential analysis*, 29(2):162–175, 2010.

[8] H. Robbins. Some aspects of the sequential design of experiments. 1952.

[9] A. Sankararaman and B. Narayanaswamy. Online heavy-tailed change-point detection. In *Uncertainty in Artificial Intelligence*, pages 1815–1826. PMLR, 2023.

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

[11] L. Wei and V. Srivastava. Nonstationary stochastic multiarmed bandits: Ucb policies and minimax regret. *arXiv preprint arXiv:2101.08980*, 2021.

[12] K. Yang and L. Toni. Graph-based recommendation system. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 798–802. IEEE, 2018.

[13] J. Y. Yu and S. Mannor. Piecewise-stationary bandit problems with side observations. In *Proceedings of the 26th annual international conference on machine learning*, pages 1177–1184, 2009.