

PROBLEM SETTING: HTPS BANDITS

Multi-Armed Bandit: At each round $t \in [T]$, the learner selects an arm $I_t \in [K]$ and receives reward $X_{I_t,t}$

Piecewise-Stationarity: an instance ν is a sequence of Υ sets of K probability distributions with means $\{\mu_i^{(j)}\}_{i \in [K]}$. During epoch $j \in [\Upsilon]$ the mean reward of action $i \in [K]$ is $\mu_i^{(j)}$

Heavy-tailed Rewards: for every action $i \in [K]$ and time $t \in [T]$, $\mathbb{E}[|X - \mathbb{E}[X]|^{1+\epsilon}] \leq v$ for some $\epsilon \in (0, 1]$

Dynamic Regret of Policy π

$$\mathbb{E}[R^\pi(T)] = \sum_{j \in [\Upsilon]} \sum_{i \in [K]} \Delta_i^{(j)} \mathbb{E}[N_{i,j}^\pi(T)].$$

Relevant Quantities

- $\nu \in \mathcal{B}(v, \epsilon, \Upsilon)$: HTPS instance, composed of a sequence of Υ reward distribution with means $\{\mu_i^{(j)}\}_{i \in [K], j \in [\Upsilon]}$
- K : number of arms
- Υ : number of changes
- v, ϵ : heaviness of the tails

Lower Bound on the Regret for every Policy π

$$\sup_{\nu \in \mathcal{B}(v, \epsilon, \Upsilon)} \mathbb{E}_\nu[R^\pi(T)] \geq \frac{1}{25} (K\Upsilon)^{\frac{\epsilon}{1+\epsilon}} (vT)^{\frac{1}{1+\epsilon}}.$$

CATONI-FCS-DETECTOR

Catoni Estimator $\sum_{i=1}^n \phi_\epsilon(\lambda_i(X_i - \hat{\mu}_c)) = 0$, where $\phi_\epsilon(x) = \log\left(1 + |x| + \frac{|x|^{1+\epsilon}}{1+\epsilon}\right)$, and $\{\lambda_i\}_{i=1}^n$ is predictable

Catoni Confidence Sequence $\left\{ m \in \mathbb{R} \mid \sum_{i=1}^t \phi_\epsilon(\lambda_i(X_i - m)) \in \left[-\frac{v \sum_{i=1}^t \lambda_i^{1+\epsilon}}{2} - \log \frac{2}{\gamma}, \frac{v \sum_{i=1}^t \lambda_i^{1+\epsilon}}{2} + \log \frac{2}{\gamma} \right] \right\}$

Catoni-FCS-detector (based on repeated-FCS-detector from Shekhar and Ramdas, 2023)

At every round t , we receive a new sample X_t and initialize a new Catoni $(1 - \gamma)$ -CS for the mean. Moreover, we update all previously initialized CS $\{CI^{(i)}\}_{i < t}$ using X_t . We define the stopping time, τ , as the first time at which the intersection of all initialized CS becomes empty

Guarantees of Catoni-FCS-detector

Let τ be the stopping time returned by Catoni-FCS-detector with $\gamma = 2T^{-3}$, t_c the true change-point, and δ the magnitude of the mean change. Then for a stream of **heavy-tailed random variables**:

- Bound on the Detection Delay:** $\mathbb{P}_{t_c} \left((\tau - t_c)^+ \leq \mathcal{O} \left(v^{\frac{1}{\epsilon}} \frac{\log(T)}{\delta^{\frac{1+\epsilon}{\epsilon}}} \right) \right) \geq 1 - \frac{14}{T}$
- Bound on the False Alarm Ratio:** $\mathbb{P}_{t_c} (\tau < t_c) \leq \frac{14}{T}$

ROBUST-CPD-UCB

High-level Recipe of Robust-CPD-UCB

- Run a **stationary heavy-tailed bandit algorithm** π_s (i.e., Robust-UCB using a median-of-means estimator)
- Inject **cyclic uniform exploration** so every arm is sampled often enough for reliable change detection
- For each arm, run an instance of **Catoni-FCS-detector** to track shifts
- When a detector fires, **reset** the learner and restart the stationary policy

Robust-CPD-UCB operates under the assumption that **change-points are sufficiently spaced**, a common assumption in the PS bandits literature

Guarantees of Robust-CPD-UCB

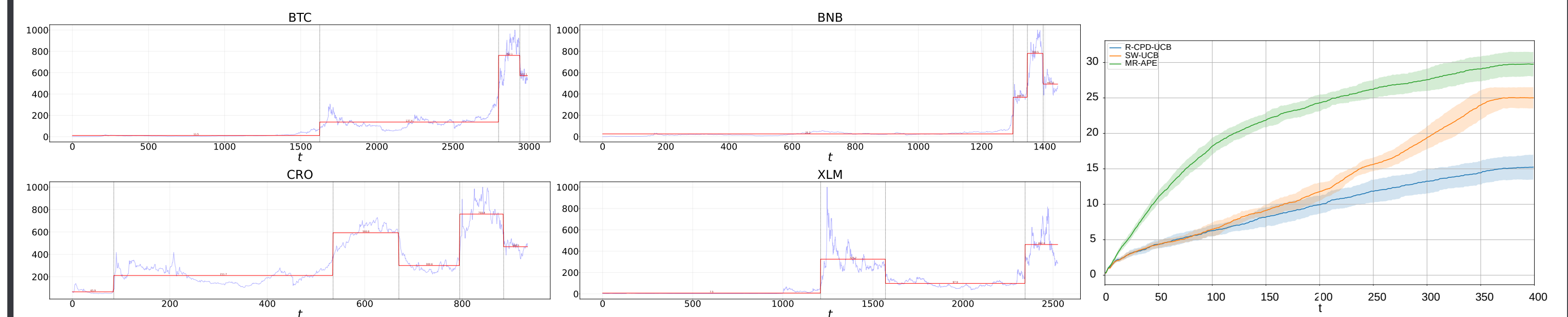
$$\mathbb{E}[R^\pi(T)] \leq \underbrace{\mathcal{O} \left(v^{\frac{1}{\epsilon}} \sum_{j=1}^{\Upsilon} \frac{\log(T)}{\left(\tilde{\delta}_{\min}^{(j)} \right)^{\frac{1+\epsilon}{\epsilon}}} \left\lceil \frac{K}{\eta} \right\rceil \Delta_{\max}^{(j)} \right)}_{\text{Detection Delay Contribution}} + \underbrace{\sum_{j=1}^{\Upsilon} \mathbb{E}[R^{\pi_s}(|E_j|)]}_{\text{Stationary Policy Regret}} + \underbrace{\eta \sum_{j=1}^{\Upsilon} |E_j| \Delta_{\max}^{(j)}}_{\text{Uniform Exploration}}$$

where, for every epoch j :

- $\tilde{\delta}_{\min}^{(j)} := \min\{\delta_{\min}^{(j-1)}, \delta_{\min}^{(j)}\}$ is the minimum change in the reward means of the epoch
- $|E_j|$ is the length of the epoch
- $\mathbb{E}[R^{\pi_s}(|E_j|)]$ is the expected cumulative regret of the stationary policy π_s
- $\Delta_{\max}^{(j)} := \max_{i \in [K]} \Delta_i^{(j)}$ is the maximum sub-optimality

NUMERICAL VALIDATION ON CRYPTOCURRENCY PRICES

We conducted numerical experiments by **fitting an HTPS instance on real-world financial data**:



Rescaled closing prices of four selected cryptocurrencies (blue) with a piecewise-constant approximation (red).

Cumulative regrets on HTPS built from cryptocurrency dataset. 20 trials, mean $\pm$ std.

REFERENCES

- Bubeck, S., Cesa-Bianchi, N., Lugosi, G. (2013). Bandits with heavy tail. IEEE Transactions on Information Theory, 59(11), 7711-7717.
- Shekhar, S., Ramdas, A. (2023). Reducing sequential change detection to sequential estimation. arXiv preprint arXiv:2309.09111.