On the Robustness of Bandit Multiple Testing

Zhengyu Zhou, Weiwei Liu*

School of Computer Science,
Wuhan University, China
{zzysince1999, liuweiwei863}@gmail.com

Abstract

Bandit multiple hypothesis testing has broad applications in biological sciences, clinical testing for drug discovery, and online A/B/n testing. The framework utilizes an adaptive sampling strategy for multiple testing which aims to maximize statistical power while ensuring anytime false discovery rate control. This paper proposes a robust approach for bandit multiple testing, allowing for (at most) ε fraction of arbitrary distribution corruption, as in Huber's contamination model. Specifically, we introduce two adaptive sampling strategies designed to minimize the number of samples required to exceed a target true positive rate, while providing anytime control over the false discovery rate. We analyze the sample complexity of our proposed methods and perform numerical simulations to demonstrate their efficiency and robustness. Furthermore, we extend our methods to address scenarios where distributions have infinite variance and situations involving multiple agents collaborating on the same bandit task.

1 Introduction

Consider a scenario with K potential treatments, such as drugs in a clinical trial. Each treatment either has a positive expected effect relative to a baseline (an actual positive) or no effect (null). The goal is to identify as many actual positive treatments as possible within a total measurement budget of B. A common approach is to allocate B/K measurements to each treatment on average and then identify predicted actual positives based on the measured effect sizes. To further enhance statistical power, the bandit multiple testing framework has been proposed (Yang et al. 2017; Jamieson and Jain 2018; Xu, Wang, and Ramdas 2021). This method adaptively allocates the measurement budget and achieves near-optimal statistical power while maintaining control of the false discovery rate (FDR).

In this paper, we investigate bandit multiple testing under the critical assumption that measurements may be subject to adversarial corruption. Specifically, consider K arms, where the reward from an arm $i \in [K] := \{1, \dots, K\}$ is drawn from a sub-Gaussian distribution $\mathbb{P}_i$. At each round t, the observed reward may be corrupted with probability ε. When a reward sample is corrupted at time t, it is replaced by a sample drawn from a contamination model with distribution $\mathbb{Q}_i$, which is distinct from $\mathbb{P}_i$. This setup has significant practical implications. For example, in the context of measuring drug responses, multiple compounds are tested to identify those with a positive effect on a disease. It is reasonable to assume that a fraction of the test results may be reported incorrectly or that some samples may be contaminated (Keogh-Brown et al. 2007).

1.1 Problem Statement

Consider the bandit model of K arms. The reward of arm $i \in [K]$ is generated from a probability measure $\mathbb{P}_i$ with mean $\mu_i \in \mathbb{R}$. At each round t, a player interacts with the bandit by pulling one of the arm $A_t \in [K]$, generating the random reward $X_{A_t,t} \in \mathbb{R}$. Then, the adversary flips a coin, whose outcome is denoted by the random variable D_t, where $D_t \sim \mathrm{Bern}(\varepsilon)$[1]. If $D_t = 0$, the adversary sends the true reward to the player; if $D_t = 1$, the adversary replaces the true reward with a corrupted sample drawn from an adversarial distribution $\mathbb{Q}_{A_t}$.

Hence, at each round t, the adversarial action can be modeled by a Bernoulli random variable $D_t \sim \mathrm{Bern}(\varepsilon)$, determining the form of the observed reward as follows:

$$R_{A_t,t} = \begin{cases} X_{A_t,t} \sim \mathbb{P}_{A_t}, & \text{if } D_t = 0 \\ X'_{A_t,t} \sim \mathbb{Q}_{A_t}, & \text{if } D_t = 1 \end{cases}. \tag{1}$$

The contamination model mentioned above is first introduced in Huber (1964) to study the theory of robust estimation. Accordingly, the contamination model for arm $i \in [K]$ in the bandit instance is described by a mixture distribution:

$$\tilde{\mathbb{P}}_i = (1-\varepsilon)\mathbb{P}_i + \varepsilon\mathbb{Q}_i. \tag{2}$$

For a known threshold μ_0, define the sets

$$\mathcal{H}_1 = \{i \in [K] : \mu_i > \mu_0\} \text{ and } \mathcal{H}_0 = \{i \in [K] : \mu_i \le \mu_0\}.$$

The means μ_i for $i \in [K]$ and the size of $\mathcal{H}_1$ are unknown. Arms (treatments) in $\mathcal{H}_1$ have means greater than μ_0 (indicating a positive effect), while those in $\mathcal{H}_0$ have means less than or equal to μ_0 (indicating no effect over the

*Corresponding author.

[1] A Bernoulli random variable $\mathrm{Bern}(\varepsilon)$ takes the value 1 with probability ε and the value 0 with probability $1-\varepsilon$.

baseline). Accordingly, each arm i is associated with a null hypothesis $H_{i,0} : \mu_i \leq \mu_0$ and an alternative hypothesis $H_{i,1} : \mu_i > \mu_0$. At each round t, after the player selects an arm, the player also outputs a set of indices $\mathcal{S}_t \subseteq [K]$ that are interpreted as discoveries or rejections of the null hypothesis. For a $\tau \in \mathbb{N}$, the objective is to maximize the number of true discoveries $|\mathcal{S}_t \cap \mathcal{H}_1|$, while keeping the number of false alarms $|\mathcal{S}_t \cap \mathcal{H}_0|$ small uniformly over all times $t \in \mathbb{N}$. We now formally define the concepts of false discovery rate (FDR) and true positive rate (TPR) as in (Jamieson and Jain 2018).

Definition 1 (False Discovery Rate, FDR-δ)**.** Fix some $\delta \in (0, 1)$. We say an algorithm is FDR-δ if, for all possible problem instances $(\{\mathbb{P}_i\}_{i=1}^K, \mu_0)$, it satisfies

$$\mathrm{FDR}(\mathcal{S}_t) := \mathbb{E}\left[\frac{|\mathcal{S}_t \cap \mathcal{H}_0|}{|\mathcal{S}_t| \vee 1}\right] \leq \delta,$$

for all $t \in \mathbb{N}$ simultaneously.

Definition 2 (True Positive Rate, TPR-δ, τ)**.** Fix some $\delta \in (0, 1)$. We say an algorithm is TPR-δ, τ on instance $(\{\mathbb{P}_i\}_{i=1}^K, \mu_0)$ if

$$\mathrm{TPR}(\mathcal{S}_t) := \mathbb{E}\left[\frac{|\mathcal{S}_t \cap \mathcal{H}_1|}{|\mathcal{H}_1|}\right] \geq 1 - \delta,$$

for all $t \geq \tau$.

Definition 3 (Sample Complexity)**.** Fix some $\delta \in (0, 1)$ and an algorithm $\mathcal{A}$ that is FDR-δ over all possible problem instances. Fix a particular problem instance $(\{\mathbb{P}_i\}_{i=1}^K, \mu_0)$. At each time $t \in \mathbb{N}$, $\mathcal{A}$ chooses an arm $i \in [K]$ to obtain an observation, and before proceeding to the next round outputs a set $\mathcal{S}_t \subseteq [K]$. The sample complexity of $\mathcal{A}$ on $(\{\mathbb{P}_i\}_{i=1}^K, \mu_0)$ is the smallest time $\tau \in \mathbb{N}$ such that $\mathcal{A}$ is TPR-δ, τ.

The main contributions of this paper are as follows:

- We address the bandit multiple testing problem under the critical assumption of adversarial corruption in observations. We propose two algorithms that adaptively select arms to minimize the number of samples required to achieve the desired TPR, while ensuring anytime control of the FDR.
- We provide a detailed analysis of the sample complexity of the proposed methods and validate their efficiency and robustness through numerical simulations.
- We extend our methods to tackle scenarios where the underlying distributions exhibit infinite variance and situations involving multiple agents collaborating on the same bandit task.

1.2 Related Work

Framework for Bandit Multiple Testing. Bandit Multiple Testing integrates the challenges of multiple hypothesis testing with multi-armed bandits (MAB). In a doubly sequential style, Yang et al. (2017) propose replacing a sequence of A/B tests with a sequence of best-arm MAB instances that can be continuously monitored by data scientists. In this setup, each MAB instance is used to compare a single placebo arm with several treatment arms. If at least one treatment arm is found to outperform the placebo, the goal is to identify and return the best treatment arm. Consequently, each MAB instance functions as a single adaptive sequential hypothesis test. The primary goal in Yang et al. (2017) is to control FDR throughout the entire sequence of MAB instances.

In contrast, Jamieson and Jain (2018) focus on a single MAB instance, where the objective is to adaptively identify treatments that are positive relative to a baseline. They propose an algorithm that improves the sample complexity to exceed a target TPR while guaranteeing anytime control of FDR. The improvement in sample efficiency is driven by a carefully designed adaptive sampling strategy.

Xu, Wang, and Ramdas (2021) propose a unified framework that separates each algorithm into two components: an exploration component and an evidence component. Additionally, the proposed meta-algorithm leverages "e-processes," which are introduced as an alternative to p-processes by Ramdas et al. (2022) for addressing various testing problems. Wang and Ramdas (2022) highlight numerous advantages of e-values over p-values, emphasizing their flexibility. This flexibility opens the door to exploring how bandit multiple testing can be applied to other settings.

Robust Statistics. In recent years, the concept of statistical robustness has primarily referred to the framework introduced by Huber (1964), which addresses scenarios where data are subject to a certain level of contamination. This contamination can occur either at the level of the underlying distribution (Huber 1964; Maronna et al. 2019; Diakonikolas et al. 2019) or directly on the observed data (Lecué and Lerasle 2020; Lugosi and Mendelson 2021; Minsker and Ndaoud 2021).

Prominent approaches in robust estimation include M-estimators and trimming. M-estimators, introduced by (Huber 1964), enhance robustness by curbing and limiting the influence of individual data points on statistical estimates. In contrast, trimming refers to the direct removal of outliers (Anscombe 1960), and trimmed means have been widely recognized as robust estimators for decades (Bickel 1965).

2 Algorithm

2.1 False Discovery Rate Control

We introduce our main technical tool for ensuring FDR control at stopping times: *e-processes* (Ramdas et al. 2023, 2022; Fan, Jiao, and Wang 2024; Zhou and Liu 2024). An *e-variable*, E, is a nonnegative random variable that $\mathbb{E}[E] \leq 1$ under the null hypothesis. The term *"e-value"* refers to the realized values of an e-variable E. A null hypothesis can be rejected when observing a large e-value. For instance, to control the FDR at 0.05 for a single hypothesis, we reject the null when $e \geq 20$. This follows from Markov's inequality, which ensures $\mathbb{P}[E \geq 20] \leq 0.05$. E-processes extend e-variables to sequential settings. A stochastic process $(E_t)_{t\geq 1}$ is called an e-process if it satisfies $\sup_{\tau \in \mathcal{T}} \mathbb{E}[E_\tau] \leq 1$, where $\mathcal{T}$ denotes the sets of all stopping times.

e-Benjamini-Hochberg (e-BH) controls FDR. The e-BH procedure proposed by Wang and Ramdas (2022) uses e-values to control FDR. In this case, let $e_1, \ldots, e_K$ be the

realized e-values for a set of e-variables $E_1, \ldots, E_K$. Define $e_{[i]}$ to be the ith largest e-value for $i \in [K]$. The e-BH procedure rejects all hypotheses with the largest k^* e-values, where

$$k^* := \max\left\{k \in [K] : \frac{ke_{[k]}}{K} \geq \frac{1}{\delta}\right\}. \tag{3}$$

The resulting rejection set is given by:

$$\text{eBH}[\delta](e_1, \ldots, e_K) := \{i \in [K] : e_i \geq e_{[k^*]}\}. \tag{4}$$

Proposition 1 (e-BH controls FDR. (Wang and Ramdas 2022))**.** $\text{FDR}(\text{eBH}[\delta](e_1, \ldots, e_K)) \leq \delta$ *regardless of the dependence structure among the e-values.*

Proposition 1 holds regardless of the dependence structure of the e-variables, which is crucial in our setting. Bandit algorithms often induce complex dependencies between reward statistics (Nie et al. 2018; Shin, Ramdas, and Rinaldo 2019, 2020). These dependencies are further compounded when using both adaptive sampling and stopping rules. Nevertheless, e-variable-based algorithms allow us to guarantee FDR control without assumptions on the sampling method.

Robust Test Martingale. Nonnegative martingales play a central role in characterizing admissible e-processes—every e-process is upper bounded by a nonnegative martingale (Ramdas et al. 2020, Corollary 24). A stochastic process $\{M_t\}_{t\geq 1}$ adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 1}$ is called a martingale if $\mathbb{E}[M_t|\mathcal{F}_{t-1}] = M_{t-1}$ and a supermartingale if $\mathbb{E}[M_t|\mathcal{F}_{t-1}] \leq M_{t-1}$. In the following, we demonstrate how to construct a test martingale, given a sequence of adversarially corrupted data.

Consider testing the null hypothesis $H_{i,0} : \mu_i \leq \mu_0$ against the alternative $H_{i,1} : \mu_i > \mu_0$, given a sequence of corrupted data $R_{i,t_i(s)} \overset{\text{i.i.d.}}{\sim} \tilde{\mathbb{P}}_i = (1-\varepsilon)\mathbb{P}_i + \varepsilon\mathbb{Q}_i$, $s \geq 1$. We construct a nonnegative test supermartingale as follows:

$$E_{i,t} = \prod_{s=1}^{T_i(t)} \frac{\exp(\phi(\lambda(R_{i,t_i(s)} - \mu_0)))}{1 + \lambda^2\sigma^2/2 + 1.5\varepsilon}, \tag{5}$$

where λ is a tunable parameter, $T_i(t) := \sum_{s=1}^{t} \mathbb{I}\{A_s = i\}$ represents the number of times arm i has been pulled up to time t, $t_i(s) := \inf\{t \geq 1 : T_i(t) \geq s\}$ denotes the time at which arm i has been pulled s times. The function ϕ is the influence function, defined in (Catoni 2012) as:

$$\phi(x) = \begin{cases} -\log 2, & x < -1, \\ \log(1 + x + x^2/2), & -1 \leq x < 0, \\ -\log(1 - x + x^2/2), & 0 \leq x < 1, \\ \log 2, & x \geq 1. \end{cases} \tag{6}$$

It is straightforward to verify that ϕ satisfies:

$$-\log(1 - x + x^2/2) \leq \phi(x) \leq \log(1 + x + x^2/2),$$

and $|\phi(x)| \leq \log 2$. The constant 1.5 in Eq (5) is not arbitrary; it stems from the fact that the maximum and minimum values of $\exp(\phi(x))$ are 2 and $1/2$, respectively, with a difference of $3/2$. In fact, the proof of Proposition 2 relies on the boundedness of $\exp(\phi(x))$, as detailed in Appendix A.

Proposition 2. *For any σ-sub-Gaussian $\mathbb{P}_i$, arbitrary $\mathbb{Q}_i$ and $\lambda > 0$, let $R_{i,t_i(1)}, R_{i,t_i(2)}, \ldots \overset{\text{i.i.d.}}{\sim} \tilde{\mathbb{P}}_i = (1-\varepsilon)\mathbb{P}_i + \varepsilon\mathbb{Q}_i$, $\{E_{i,t}\}_{t\geq 1}$, with $E_{i,0} = 1$, is a nonnegative supermartingale.*

Since $\{E_{i,t}\}_{t\geq 1}$ is a nonnegative supermartingale, it is an e-process by optional stopping theorem (Grimmett and Stirzaker 2001).

2.2 Sampling Strategies

To minimize the time required to identify the positive arms, it is intuitive to repeatedly pull the arm currently believed to be the best until it is rejected. However, since the true mean is unknown, we must gather sufficient evidence to determine which arm is likely the best. This section presents two sampling strategies: one based on the trimmed mean and the other on e-processes.

Mean-based Sampling Strategy. While the sample mean is widely used as an estimator for the true mean, it is not robust to adversarial corruption. Instead, we adopt the α-trimmed mean, a well-known robust estimator for the mean (Bickel 1965).

For each arm $i \in [K]$, let $\mathbf{R}_i^t := \{R_{i,t_i(s)} : 1 \leq s \leq T_i(t)\}$ denote the sequence of samples. We denote $W_1, W_2, \ldots, W_{T_i(t)}$ as the order statistics of $\mathbf{R}_i^t$. The α-trimmed mean for $\mathbf{R}_i^t$ is defined as in (Bickel 1965):

$$\hat{\mu}_i(t) = \frac{1}{T_i(t) - 2r} \sum_{i=r+1}^{T_i(t)-r} W_i, \tag{7}$$

where $r = \lfloor T_i(t)\alpha \rfloor$ is the integer part of $T_i(t)\alpha$, with α representing the trimming proportion. For the trimmed mean, we have the following concentration results.

Lemma 3 (Lemma 4.1 in Mukherjee et al. (2021))**.** *In the presence of adversary, for the α-trimmed mean estimator with $\alpha = \varepsilon/2$, there exist $T(\alpha, \delta) := \frac{2}{\alpha^2}\log\frac{1}{\delta}$ such that for all $t > T(\alpha, \delta)$, we have with probability at least $1 - \delta$,*

$$|\hat{\mu}_t - \mu| \leq U + \frac{\sigma}{1-\varepsilon}\sqrt{\frac{2}{t}\log\frac{2}{\delta}}, \tag{8}$$

where $U = \mathcal{O}\left(\sigma\varepsilon\sqrt{\log\frac{1}{\varepsilon}}\right)$ is an uncertainty term and $\hat{\mu}_t$ is the α-trimmed mean based on t samples, and μ is the true mean.

The concentration bound for the α-trimmed mean estimator indicates that it holds when the sample size is at least $T(\alpha, \delta)$. This threshold ensures that the estimator effectively removes outliers with high probability. Notably, the uncertainty associated with the α-trimmed mean estimator, $U = \mathcal{O}\left(\sigma\varepsilon\sqrt{\log(1/\varepsilon)}\right)$, is consistent with the fundamental lower bound on uncertainty that any estimator can achieve under Huber's contamination model Eq (2), assuming a sub-Gaussian distribution (Diakonikolas and Kane 2019). Based on the trimmed mean, we propose the mean-based method, outlined in Algorithm 1.

Algorithm 1: A mean-based algorithm for bandit multiple testing under adversarial contamination.

Input: Threshold μ_0, desired level of FDR control δ, contamination level ε.
Initialize: $\mathcal{S}_0 = \emptyset$, $e_{i,0} = 1$ for all $i \in [K]$, $N_i = 1$ for all $i \in [K]$
Define $t(s) := \left\lceil \frac{4s}{\varepsilon^2} \log\left(\frac{K\pi^2 s^2}{3\delta}\right) \right\rceil, s = 1, 2, \ldots$
Sample each arm in $[K]$ for $t(1)$ times.
for $l = 1, 2, \ldots$ **do**
  $t = \sum_{i=1}^{K} t(N_i)$
  $A_t = \operatorname{argmin}_{i \in [K] \setminus \mathcal{S}_t} \hat{\mu}_{i,t} + \frac{\sigma\varepsilon}{(1-\varepsilon)\sqrt{2N_i}}$
  Sample arm A_t for $t(N_{A_t} + 1) - t(N_{A_t})$ times.
  $N_{A_t} \leftarrow N_{A_t} + 1$
  Update e-process for each queried arm not in $\mathcal{S}_{t-1}$ as in Eq (5).
  $\mathcal{S}_t = \mathrm{eBH}[\delta](e_{1,t}, \ldots, e_{K,t})$.
end for

The proposed mean-based method involves a sorting routine to compute the trimmed mean, with a computational complexity of $\mathcal{O}\left(\sum_{i=1}^{K} T_i(t) \log T_i(t)\right)$. Additionally, the e-BH subroutine requires sorting, contributing $\mathcal{O}(K \log K)$. Consequently, the total computational cost for the mean-based method up to time T is $\mathcal{O}\left(\sum_{t=1}^{T} \sum_{i=1}^{K} T_i(t) \log T_i(t) + TK \log K\right)$.

This method incurs a per-update cost of $\mathcal{O}(\sum_{i=1}^{K} T_i(t) \log T_i(t) + K \log K)$, which scales with the number of samples. Such scaling can result in significant computational overhead when the sample size is large. To mitigate this issue, we propose the e-process-based method as a more computationally efficient alternative.

E-processes-based Sampling Strategy. The e-process defined in Eq (5) exhibits faster growth for arms with larger true means, reflecting their stronger potential to provide evidence against the null hypothesis. Leveraging this property, we propose an adaptive sampling strategy that dynamically prioritizes arms with higher e-process values, thereby efficiently allocating resources to the most promising arms. This strategy, detailed in Algorithm 2, ensures that sampling is continuously guided by the evolving evidence, optimizing the detection of positive arms.

Updating the e-process for each arm requires only constant time, resulting in a computational cost of $\mathcal{O}(K)$ for this step. Selecting the arm with the largest e-process value involves a sorting routine with complexity $\mathcal{O}(K \log K)$, and the e-BH subroutine also operates with complexity $\mathcal{O}(K \log K)$. Thus, the total computational cost for the e-process-based method up to time T is $\mathcal{O}(TK \log K)$.

Importantly, the e-process-based method incurs a fixed computational cost of $\mathcal{O}(K \log K)$ per update, making it computationally efficient and constant with respect to the number of updates when the number of arms K is fixed.

Algorithm 2: An e-process-based algorithm for bandit multiple testing under adversarial contamination.

Input: Threshold μ_0, desired level of FDR control δ, contamination level ε.
Initialize: $\mathcal{S}_0 = \emptyset$, $e_{i,0} = 1$ for all $i \in [K]$, $B(\delta, \varepsilon) := \left\lceil 4\varepsilon^{-1} \log\left(\frac{4K}{\delta^2}\right) \right\rceil$
for $t \geq 1$ **do**
  if $t \leq KB(\delta, \varepsilon)$ and $t \mod K \neq 0$ **then**
    $A_t = t \mod K$.
  end if
  if $t \mod K = 0$ **then**
    $A_t = K$
  end if
  if $t > KB(\delta, \varepsilon)$ **then**
    $A_t = \operatorname{argmax}_{i \in [K] \setminus \mathcal{S}_t} e_{i,t-1}$.
    Update e-process for each queried arm not in $\mathcal{S}_{t-1}$ as in Eq (5).
    $\mathcal{S}_t = \mathrm{eBH}[\delta](e_{1,t}, \ldots, e_{K,t})$.
  end if
end for

3 Theoretical Guarantees

In this section, we assume that the distribution of each arm in $[K]$ is σ-sub-Gaussian.

Assumption 4. For every $i \in [K]$, there exists $\sigma > 0$ such that $\mathbb{E}_{X \sim \mathbb{P}_i}[\exp(t(X - \mu_i))] \leq \exp(\sigma^2 t^2/2), \forall t$, where $\mu_i = \mathbb{E}_{X \sim \mathbb{P}_i}[X]$ is the mean reward of arm i and $\sigma > 0$ is a known constant.

From Lemma 3, we note the presence of an uncertainty term that persists even with an infinite number of samples. This inherent uncertainty is unavoidable when estimating the mean in a contamination setting (Diakonikolas and Kane 2019). To address this challenge, we make the following assumption to ensure the feasibility of identifying the true positive arms.

Assumption 5. The indexes of the positive arms remain unaffected by contamination, meaning that for every $i \in \mathcal{H}_0$, the following condition holds:

$$\Delta_i := \min_{j \in \mathcal{H}_1} (\mu_j - U) - (\mu_i + U) > 0, \tag{9}$$

where U represents the uncertainty induced by the contamination defined in Lemma 3.

We provide a theoretical guarantee for the mean-based sampling strategy, with the proof deferred to Appendix B.

Theorem 6 (Sample Complexity for Mean-based Strategy)**.** *Under Assumptions 4 and 5, suppose* $0 < \varepsilon \leq 1/7$, $\lambda^2 = \frac{\varepsilon}{4\sigma^2}$ *and* $\min_{j \in \mathcal{H}_1} \mu_j > \mu_0 + 14\sigma\sqrt{\varepsilon}$, *for all* $t \in \mathbb{N}$, *the rejection set produced by Algorithm 1 satisfies:*

$$\mathbb{E}\left[\frac{|\mathcal{S}_t \cap \mathcal{H}_0|}{|\mathcal{S}_t| \vee 1}\right] \leq \delta.$$

Moreover, with probability at least $1 - 2\delta$, *there exists* T

such that:

$$T \leq \sum_{i \in \mathcal{H}_0} \frac{2\sigma^2}{\Delta_i^2(1-\varepsilon)^2}\left(\log\frac{K\pi^2}{12\delta} + 4\log\frac{\varepsilon\sigma}{\Delta_i(1-\varepsilon)}\right) + \frac{8\sqrt{3}|\mathcal{H}_1|}{\pi\varepsilon^2}\log\left(\frac{4|\mathcal{H}_1|}{\delta^2}\right) + \frac{8\sqrt{3}K}{\pi\varepsilon^2},$$

and for all $t \geq T$, $\mathbb{E}\left[\frac{|\mathcal{S}_t \cap \mathcal{H}_1|}{|\mathcal{H}_1|}\right] \geq 1-\delta$.

Remark 7. The sample complexity is decomposed into two parts: (i) the sampling cost associated with null arms $\mathcal{H}_0$, which scales inversely with the squared difference in mean Δ_i^2; and (ii) the cost associated with non-null arms, which depends on the number of true signals and the total number of arms K.

Theorem 8 (Sample Complexity for E-process-based Strategy)**.** *Under Assumption 4, suppose* $0 < \varepsilon \leq 1/7$, $\lambda^2 = \frac{\varepsilon}{4\sigma^2}$ *and* $\min_{j\in\mathcal{H}_1}\mu_j > \mu_0 + 14\sigma\sqrt{\varepsilon}$, *for all* $t \in \mathbb{N}$, *the rejection set produced by Algorithm 2 satisfies:*

$$\mathbb{E}\left[\frac{|\mathcal{S}_t \cap \mathcal{H}_0|}{|\mathcal{S}_t| \vee 1}\right] \leq \delta.$$

Moreover, with probability at least $1-\delta$, *there exists* T *such that:*

$$T \leq 4K\varepsilon^{-1}\log\left(\frac{4K}{\delta^2}\right) + 4|\mathcal{H}_1|\varepsilon^{-1}\log\left(\frac{4|\mathcal{H}_1|}{\delta^2}\right) + 4K\varepsilon^{-1},$$

and for all $t \geq T$, $\mathbb{E}\left[\frac{|\mathcal{S}_t \cap \mathcal{H}_1|}{|\mathcal{H}_1|}\right] \geq 1-\delta$.

Remark 9. The theorem guarantees that the e-process-based strategy controls the FDR within δ and achieves high TPR once sufficient samples are collected. The convergence rate depends on K, $|\mathcal{H}_1|$, and ε, ensuring statistically efficient identification of positive arms even under adversarial contamination. The proof is provided in Appendix B.

4 Numerical Simulations

We conduct simulations in the sub-Gaussian setting to illustrate the empirical efficiency of our proposed algorithm based on e-processes.

Simulation Setup. We set the true distribution to be $\mathbb{P}_i = \mathcal{N}(\mu_i, 1)$ where $\mu_i = 0$ if $i \in \mathcal{H}_0$ and $\mu_i = 1$ if $i \in \mathcal{H}_1$, with $1/10$ change of contamination from a Lévy stable distribution with location parameter 1000 and skewness parameter 0.5. We consider 3 setups, where we set the number of non-null hypotheses to be $|\mathcal{H}_1| = 2$, $\lfloor \log K \rfloor$ and $\lfloor \sqrt{K} \rfloor$, to see the effect of different magnitudes of non-null hypotheses on the sample complexity of each method. We set $\delta = 0.05$ and compare three different methods. The first exploration method is simply uniform sampling across each arm. The second is our proposed e-process-based method. The third is our proposed mean-based method. When using e-BH, we set our e-processes as in (5) with $\lambda = \sqrt{\varepsilon}/2\sigma$ as suggested in (Wang and Ramdas 2023).

Results. Figure 1 illustrates the empirical estimate of $\mathbb{E}\left[|\mathcal{S}_t \cap \mathcal{H}_1|/|\mathcal{H}_1|\right]$ at each time step t for each algorithm, averaged over 500 trials. The gray dashed line in Figure 1 represents the threshold level of 0.95. From Figure 1, we observe that our proposed e-process-based method achieves the desired true positive rate (TPR) in the shortest time. In contrast, the TPR of our proposed mean-based method begins to grow later than that of our proposed e-process-based method. This delay can primarily be attributed to the time required for the trimmed means to converge. However, once the trimmed means converge, the TPR rapidly reaches the desired level.

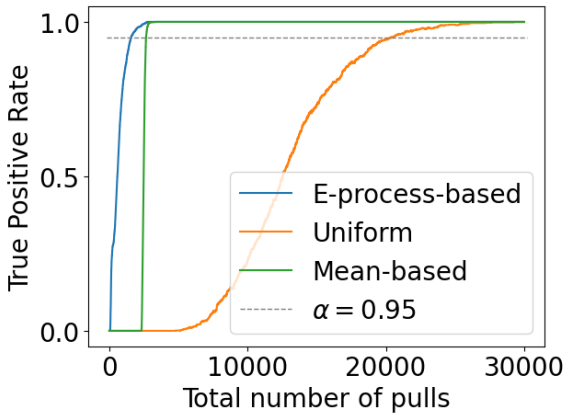


Figure 1: The performance of three methods. Each solid curve represents the average $|\mathcal{S}_t \cap \mathcal{H}_1|/|\mathcal{H}_1|$ over 500 trials for each method, evaluated at each time step. This comparison highlights the efficiency and robustness of our proposed approach.

Figure 2 shows the performance of three methods on different number of non-null hypotheses. From Figure 2, we can see that our proposed e-process-based method demonstrates superior performance among the three methods. Notably, the performance gap between the uniform sampling method and the e-process-based method widens as the number of arms increases. In contrast, the performance gap between the mean-based method and the e-process-based method remains largely unaffected by changes in the number of arms.

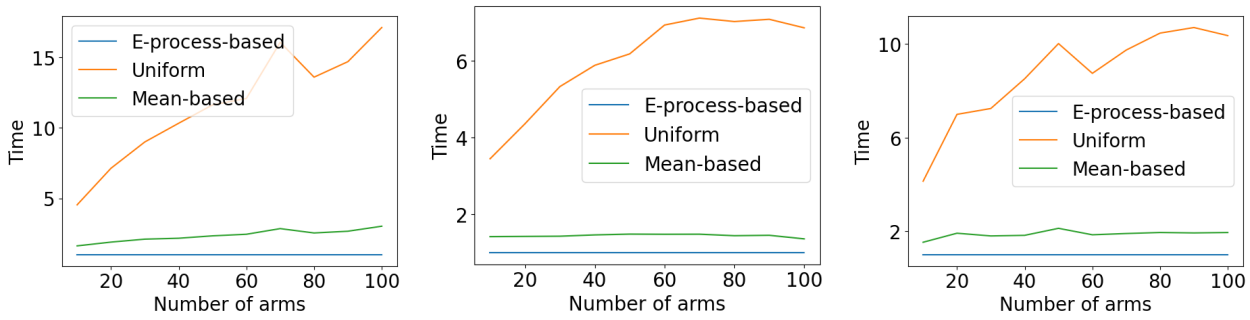
5 Extensions

5.1 Infinite Variance

We have demonstrated how to design a robust bandit multiple testing procedure in the sub-Gaussian setting, where the variance is finite. Next, we extend our analysis to the case where the true distribution only has finite p-th moments ($1 < p < 2$). Let $\mathcal{M}^1$ denote the set of all distributions with finite mean, and $\mu : \mathcal{M}^1 \to \mathbb{R}$ represent the mean functional:

$$\mu(\mathbb{P}) = \int x \mathrm{d}\mathbb{P}. \tag{10}$$

For $p > 1$, we define $\mathcal{M}^p$ as the subset of $\mathcal{M}^1$ consisting of distributions with both finite means and finite p-th moments. Let $v_p(\mathbb{P}) : \mathcal{M}^p \to \mathbb{R}$ denote the p-th absolute



(a) $|\mathcal{H}_1| = 2$ (b) $|\mathcal{H}_1| = \lfloor \sqrt{K} \rfloor$ (c) $|\mathcal{H}_1| = \lfloor \log K \rfloor$

Figure 2: The plot presents a comparison of the time t of the three methods required to obtain a rejection set $\mathcal{S}_t$ that satisfies $\mathrm{TPR}(\mathcal{S}_t) \geq 1 - \delta$ and $\mathrm{FDR}(\mathcal{S}_t) \leq \delta$, where $\delta = 0.05$. The comparison spans three different methods across varying numbers of arms (K) and densities of non-null hypotheses ($|\mathcal{H}_1|$). Time is reported as a ratio relative to the time taken by the e-process-based method. The results highlight that the e-process-based method consistently outperforms the other methods, requiring less time to achieve the desired performance metrics.

central moment functional:

$$v_p(\mathbb{P}) = \int |x - \mu(\mathbb{P})|^p \mathrm{d}\mathbb{P}. \tag{11}$$

Finally, we define $\mathcal{M}_\kappa^p = \{\mathbb{P} \in \mathcal{M}^p : v_p(\mathbb{P}) \leq \kappa\}$ as the set of distributions within $\mathcal{M}^p$ whose p-th absolute central moment is bounded by κ.

We now turn to the situation where distributions of each arm $\mathbb{P}_i$ belong to $\mathcal{M}_\kappa^p$ for $1 < p < 2$. Instead of (6), we define

$$\phi_p(x) = \begin{cases} -\log p, & x < -1, \\ \log(1 + x + |x|^p/p), & -1 \leq x < 0, \\ -\log(1 - x + x^p/p), & 0 \leq x < 1, \\ \log p, & x \geq 1. \end{cases} \tag{12}$$

We now construct e-processes as follows:

$$\begin{aligned} E_{i,0}^{(p)} &= 0, \\ E_{i,t}^{(p)} &= \prod_{s=1}^{T_i(t)} \frac{\exp(\phi_p(\lambda(R_{i,t_i(s)} - \mu_0)))}{1 + \lambda^p \kappa/p + (p - 1/p)\varepsilon}. \end{aligned} \tag{13}$$

The following proposition demonstrates that the stochastic process in (13) is a valid e-process. The proof is deferred to Appendix C.

Proposition 10. *For any $\mathbb{P}_i \in \mathcal{M}_\kappa^p$, arbitrary $\mathbb{Q}_i$ and $\lambda > 0$, let $R_{i,t_i(1)}, R_{i,t_i(2)}, \ldots \overset{\text{i.i.d.}}{\sim} \tilde{\mathbb{P}}_i = (1 - \varepsilon)\mathbb{P}_i + \varepsilon\mathbb{Q}_i$, $\{E_{i,t}^{(p)}\}_{t \geq 1}$, with $E_{i,0} = 1$, is a nonnegative supermartingale.*

Since $\{E_{i,t}^{(p)}\}_{t \geq 1}$ is a nonnegative supermartingale, it is an e-process by optional stopping theorem (Grimmett and Stirzaker 2001). Using $E_{i,t}^{(p)}$, we propose a bandit multiple testing procedure tailored for scenarios where the true distributions have infinite variance. This procedure is detailed in Algorithm 3 which is provided in Appendix D.

Example. Consider the following setup: the rewards for all arms are drawn from a heavy-tailed Student's t-distribution with 2 degrees of freedom[2], with a $1/10$ chance of contamination by a Lévy stable distribution with a location parameter of 1000 and a skewness parameter of 0.5. The shifted mean of each arm is determined as follows: $\mu_i = 0$ for $i \in \mathcal{H}_0$, and $\mu_i = 5$ for $i \in \mathcal{H}_1$. Note that the Student's t-distribution with 2 degrees of freedom has infinite variance. For $p = 1.85$, an upper bound for $v_p(\mathbb{P})$ is $\kappa = 50$.

We set the number of non-null hypotheses to $|\mathcal{H}_1| = 2$ and the total number of arms to $K = 50$. Our proposed Algorithm 3 is compared to uniform sampling under this setting. When using the e-BH procedure, the e-processes are defined as in (13), with $\lambda = (\varepsilon/\kappa)^{1/p}$.

Figure 3 demonstrates the superior performance of our proposed method compared to uniform sampling, highlighting its robustness and effectiveness in handling heavy-tailed rewards with infinite variance.

5.2 Multiple Agents

There are numerous scenarios where multiple agents interact with the same bandit and aim to cooperatively accumulate evidence. For instance, a research group might resume studying a hypothesis previously investigated by others and wish to combine existing evidence with new data collected from their own experiments. Another example is when multiple groups work on different subsets of a broader set of hypotheses, each contributing evidence that can be aggregated for a comprehensive analysis.

In such situations, the shared evidence—whether from prior studies or concurrent collaborators—might only be available in the form of e-values or p-values, with the raw data being obfuscated to maintain privacy. Consequently, these scenarios necessitate methods for merging multiple statistics into a single statistic that effectively represents the

[2]The probability density function of the Student's t-distribution is given by $f(t) \propto (1 + t^2/\nu)^{-(\nu+1)/2}$, where ν is the number of degrees of freedom.

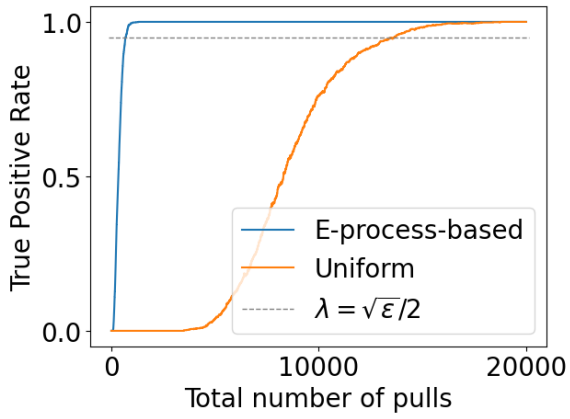


Figure 3: Performance of our proposed method compared to uniform sampling. Each solid curve represents the average $|\mathcal{S}_t \cap \mathcal{H}_1|/|\mathcal{H}_1|$ over 500 trials for each method, evaluated at each time step. This comparison highlights the efficiency and robustness of our proposed approach.

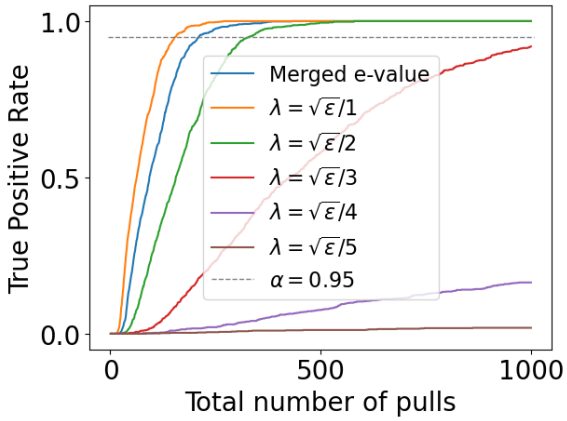


Figure 4: True Positive Rate (TPR) as a function of the total number of pulls for different values of λ and the merged e-value method. Each solid curve represents the average $|\mathcal{S}_t \cap \mathcal{H}_1|/|\mathcal{H}_1|$ over 500 trials for each value of λ, evaluated at each time step.

total evidence for rejecting a hypothesis. This allows for efficient collaboration and decision-making while respecting privacy constraints.

Suppose there are m agents, each contributing an e-value $E_1, \ldots, E_m$ that tests the same hypothesis. When these e-variables are independent, a natural way to combine them is by defining the ie-merging function f_{prod} as the product of the individual e-values:

$$f_{\text{prod}}(E_1, \ldots, E_m) := \prod_{i=1}^{m} E_i. \tag{14}$$

Proposition 11. *If $E_1, \ldots, E_m$ are independent e-variables, the combined value $f_{\text{prod}}(E_1, \ldots, E_m)$ remains a valid e-variable.*

This result follows from a basic property of independent random variables: the expectation of their product equals the product of their expectations. Hence, the validity of each individual e-variable ensures the combined value satisfies the e-variable property.

When the e-values $E_1, \ldots, E_m$ are dependent, however, the product may no longer preserve the e-variable property. In this case, we use an alternative approach:

$$f_{\text{mean}}(E_1, \ldots, E_m) := \frac{1}{m} \sum_{i=1}^{m} E_i. \tag{15}$$

Proposition 12. *If $E_1, \ldots, E_m$ are arbitrarily dependent, the function $f_{\text{mean}}(E_1, \ldots, E_m)$ is still a valid e-variable.*

This result holds because the mean of e-values, regardless of their dependence structure, satisfies the necessary properties of an e-variable. Averaging provides a robust way to merge evidence without assuming independence.

Vovk and Wang (2021) provide a theoretical foundation for symmetric e-merging functions. They show that the only admissible functions in this class are convex combinations of f_{mean} and the constant 1. Furthermore, they demonstrate a unique dominance property for f_{prod}: when all input e-values are at least 1, f_{prod} produces the largest combined e-value among all symmetric ie-merging functions.

Utilizing the e-merging function, we design an algorithm specifically suited for scenarios involving multiple agents collaborating on the same bandit task. The details of this algorithm are provided in Algorithm 4 which is deferred to Appendix D.

Example. We set the true distribution as $\mathbb{P}_i = \mathcal{N}(\mu_i, 1)$, where $\mu_i = 0$ if $i \in \mathcal{H}_0$ and $\mu_i = 1$ if $i \in \mathcal{H}_1$, with a contamination probability $\varepsilon = 1/10$ from a Lévy stable distribution (location parameter 1000, skewness parameter 0.5). The total number of arms is $K = 10$, with $|\mathcal{H}_1| = 2$ non-nulls. For the tunable parameter λ in (5), we consider 5 agents, each selecting a different value: $\sqrt{\varepsilon}$, $\sqrt{\varepsilon}/2$, $\sqrt{\varepsilon}/3$, $\sqrt{\varepsilon}/4$, and $\sqrt{\varepsilon}/5$, resulting in 5 different e-processes per hypothesis. To account for dependency, we use the e-merging function f_{mean} to aggregate evidence across the e-processes. Figure 4 demonstrates that the largest λ achieves the fastest convergence to the target threshold ($\alpha = 0.95$), followed closely by the merged e-value method (blue curve), while smaller λ values show slower TPR growth.

6 Conclusion

In this paper, we present a robust framework for bandit multiple testing. By applying the e-BH procedure to the robust test supermartingale, we guarantee FDR control. We propose two adaptive sampling strategies, with the e-process-based method excelling in both statistical and computational efficiency. Comprehensive theoretical analyses are provided in Section , and numerical simulations in Section validate the efficiency and robustness of our approach. Moreover, we extend our methods to handle distributions with infinite variance and scenarios with multiple agents collaborating on the same bandit task.

Acknowledgments

This work is supported by the Key R&D Program of Hubei Province under Grant 2024BAB038, the National Key R&D Program of China under Grant 2023YFC3604702, the Fundamental Research Funds for the Central Universities under Grant 2042025kf0045.

References

Anscombe, F. J. 1960. Rejection of outliers. *Technometrics*.

Bickel, P. J. 1965. On some robust estimates of location. *The Annals of Mathematical Statistics*.

Catoni, O. 2012. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*.

Diakonikolas, I.; Kamath, G.; Kane, D.; Li, J.; Moitra, A.; and Stewart, A. 2019. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*.

Diakonikolas, I.; and Kane, D. M. 2019. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*.

Fan, Y.; Jiao, Z.; and Wang, R. 2024. Testing the mean and variance by e-processes. *Biometrika*.

Gong, X.; Yuan, D.; and Bao, W. 2021a. Discriminative metric learning for partial label learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8): 4428–4439.

Gong, X.; Yuan, D.; and Bao, W. 2021b. Understanding partial multi-label learning via mutual information. In *NeurIPS*.

Gong, X.; Yuan, D.; and Bao, W. 2022. Partial label learning via label influence function. In *ICML*.

Gong, X.; Yuan, D.; Bao, W.; and Luo, F. 2022. A unifying probabilistic framework for partially labeled data learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7): 8036–8048.

Grimmett, G.; and Stirzaker, D. 2001. *Probability and random processes*. Oxford University Press.

Huber, P. J. 1964. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*.

Jamieson, K. G.; and Jain, L. 2018. A Bandit Approach to Multiple Testing with False Discovery Control. In *NeurIPS*.

Keogh-Brown, M.; Bachmann, M.; Shepstone, L.; Hewitt, C.; Howe, A.; Ramsay, C. R.; Song, F.; Miles, J.; Torgerson, D.; Miles, S.; et al. 2007. Contamination in trials of educational interventions.

Lecué, G.; and Lerasle, M. 2020. Robust machine learning by median-of-means: theory and practice. *The Annals of Statistics*.

Lugosi, G.; and Mendelson, S. 2021. Robust multivariate mean estimation: the optimality of trimmed mean. *The Annals of Statistics*.

Maronna, R. A.; Martin, R. D.; Yohai, V. J.; and Salibián-Barrera, M. 2019. *Robust statistics: theory and methods (with R)*. John Wiley & Sons.

Minsker, S.; and Ndaoud, M. 2021. Robust and efficient mean estimation: an approach based on the properties of self-normalized sums. *Electronic Journal of Statistics*.

Mukherjee, A.; Tajer, A.; Chen, P.-Y.; and Das, P. 2021. Mean-based best arm identification in stochastic bandits under reward contamination. In *NeurIPS*.

Nie, X.; Tian, X.; Taylor, J.; and Zou, J. 2018. Why adaptively collected data have negative bias and how to correct for it. In *AISTATS*.

Ramdas, A.; Grünwald, P.; Vovk, V.; and Shafer, G. 2023. Game-Theoretic Statistics and Safe Anytime-Valid Inference. *Statistical Science*.

Ramdas, A.; Ruf, J.; Larsson, M.; and Koolen, W. 2020. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*.

Ramdas, A.; Ruf, J.; Larsson, M.; and Koolen, W. M. 2022. Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*.

Shin, J.; Ramdas, A.; and Rinaldo, A. 2019. Are sample means in multi-armed bandits positively or negatively biased? In *NeurIPS*.

Shin, J.; Ramdas, A.; and Rinaldo, A. 2020. On conditional versus marginal bias in multi-armed bandits. In *ICML*.

Vovk, V.; and Wang, R. 2021. E-values: Calibration, combination and applications. *The Annals of Statistics*.

Wang, H.; and Ramdas, A. 2023. Huber-robust confidence sequences. In *AISTATS*.

Wang, R.; and Ramdas, A. 2022. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*.

Xu, Z.; Wang, R.; and Ramdas, A. 2021. A unified framework for bandit multiple testing. In *NeurIPS*.

Yang, F.; Ramdas, A.; Jamieson, K. G.; and Wainwright, M. J. 2017. A framework for multi-a (rmed)/b (andit) testing with online fdr control. In *NeurIPS*.

Zheng, C.; Shi, Z.; Miao, R.; Liu, W.; Yang, T.; Cui, B.; and Uhlig, S. 2025. Answering Subset Query Over Multi-Attribute Data Streams Using Hyper-USS. *IEEE Transactions on Knowledge and Data Engineering*.

Zhou, Z.; and Liu, W. 2024. Sequential Kernel Goodness-of-fit Testing. In *ICML*.