Bandit Algorithm for Multiple Hypothesis Testing

1 Problem Formulation

We consider an agent who has a set $\mathcal{M} = \{1, 2, \cdots, m\}$ of different options (arms) with the unknown mean values $\{\mu_1, \mu_2, \cdots, \mu_m\}$. At any time step $t \in \mathbb{N}$, the agent observes a new data point X_i^t for any arm i which is generated according to

$$X_i^t = \mu_i + n_i^t, \tag{1}$$

where n_i^t is zero mean noise components associated to arm i at time t. The noise components are independent across different arms $i = 1, 2, \cdots, m$ and iid distributed across different time steps $t \in \mathbb{N}$ for any arm i. The agent's goal is to identify which arms are *close* in performance to a base arm– called arm 0– and which arms are *different* from the base arm in a sequential manner. The base arm has a mean value of μ_0 and for a given sensitivity level $\epsilon > 0$, arm i is considered close to the base arm if $|\mu_0 - \mu_i| < \epsilon$ and different from the base arm if $|\mu_0 - \mu_i| > \epsilon$. Specifically, the agent wants to develop an algorithm which, at any time $t \in \mathbb{N}$ and based on the observed data $\{X_i^1, X_i^2, \cdots, X_i^t\}$ for any arm $i \in \mathcal{M}$, makes one of the following decisions for arm i:

1. confidently accepts the null hypothesis $H_0^i : |\mu_i - \mu_0| < \epsilon$ against the alternative hypothesis $H_1^i : |\mu_i - \mu_0| > \epsilon$,

2. confidently rejects the null hypothesis $H_0^i : |\mu_i - \mu_0| < \epsilon$ against the alternative hypothesis $H_1^i : |\mu_i - \mu_0| > \epsilon$,

3. waits for more observations of arm i.

The agent stops running the algorithm once all the null hypotheses are either accepted or rejected.

For such an algorithm, let $\mathcal{R}$ and $\mathcal{A}$ be the set of rejected and accepted null hypotheses returned at the stopping time, respectively. Also, let $\mathcal{M}_r = \{i \in \mathcal{M} : |\mu_i - \mu_0| > \epsilon\}$ and $\mathcal{M}_a = \{i \in \mathcal{M} : |\mu_i - \mu_0| < \epsilon\}$ be the set of ϵ-different and ϵ-equal arms to the base arm, respectively. For the problem to be well-defined, we assume that there is no arm with a mean value $\mu_0 + \epsilon$ or $\mu_0 - \epsilon$. Furthermore, let R, A, m_r and $m_a = m - m_r$, denote the size of the sets $\mathcal{R}, \mathcal{A}, \mathcal{M}_r$ and $\mathcal{M}_a$, respectively. Now, define $V = |\mathcal{M}_a \cap \mathcal{R}|$ as the number of false negatives (number of incorrectly rejected arms) and $W = |\mathcal{M}_r \cap \mathcal{A}|$ be as the number

	accepted	rejected	total
H_0 true	$m_a - V$	V	m_a
H_0 false	W	$m_r - W$	m_r
total	A	R	m

Table 1: Various Parameters in Multiple Hypothesis Testing

of false positives (number of incorrectly accepted arms) by the the algorithm. The type I and type II Family-Wise Error Rate (FWER) of the algorithm are then defined as

$$E_1 = \mathbb{P}[V \geq 1], \quad E_2 = \mathbb{P}[W \geq 1],$$

respectively. Table 1 summarizes the relations among different parameters defined above.

The agent's goal is to design a sequential hypotheses testing algorithm which guarantees the following FWER bounds

$$E_1 \leq \alpha, \quad E_2 \leq \beta,$$

for any given value of $\alpha, \beta \in (0, 1)$.

2 The Bandit Algorithm

We now propose a bandit algorithm for the above hypothesis testing problem. First, we focus on the scenario in which the performance of the base arm (i.e., μ_0) is known to the agent. Later, we describe how we can modify our proposed algorithm to relax this assumption. Furthermore, we make the following assumption on the tail of the noise components in (1).

Assumption 1. *There exists $\sigma > 0$ such that the noise component n_i^t is $\frac{\sigma^2}{4}$-subgaussian for all i, t.*

For example, if X_i^t is a Binary random variables for all i, t, then we meet Assumption 1 by taking $\sigma^2 = 1$. Furthermore, if there exists, a_i, b_i for each i such that $X_i^t \in [a_i, b_i]$ almost surely for all $t \in \mathbb{N}$, then n_i^t is $(b_i - a_i)^2$-subgaussian for any t and hence, taking $\sigma^2 = \max_i \ (b_i - a_i)^2$ will satisfy Assumption 1.

Our proposed algorithm –described in Algorithm 1– takes as input the set of arms $\mathcal{M}$, mean value of the base arm μ_0, ϵ, desired errors α, β and a sequence of concentration functions $\Big\{\Delta_t : (0,1) \to \mathbb{R}\Big\}_{t \in \mathbb{N}}$ which for any $x \in (0, 1)$ satisfies

$$\mathbb{P}\big[|\mu_i - \hat{\mu}_i^t| \leq \Delta_t(x), \ \forall \ i \in \mathcal{M}, t \in \mathbb{N}\big] \geq 1 - x, \tag{2}$$

where $\hat{\mu}_i^t$ is the empirical mean of arm i at time t. One standard choice for Δ is the error bound of Hoeffding inequality combined with a union bound over all arms and all time steps; i.e.,

$$\Delta_t(x) = \sqrt{\frac{\sigma^2}{2t} \log\Big(\frac{4mt^2}{x}\Big)}. \tag{3}$$

The following proposition shows that this sequence of Δ_t satisfies (2).

Algorithm 1 Bandit Multiple Hypothesis Testing

Input: $\mathcal{M}, \mu_0, \epsilon, \alpha, \beta, \{\Delta_t\}_{t\in\mathbb{N}}$
Initialize: $t = 1$ and $M_t = \mathcal{M}$
while M_t is non-empty **do**
  pull each arm $i \in M_t$ once and observe the data
  Compute empirical means $\hat{\mu}_i^t,\ \forall\ i \in M_t$
  for $i \in M_t$ **do**
    if $|\hat{\mu}_i^t - \mu_0| > \big(\epsilon + \Delta_t(\alpha)\big)$ **then**
      Reject H_0^i
      Remove i from M_t and add it to $\mathcal{R}$
    end if
    if $|\hat{\mu}_i^t - \mu_0| < \big(\epsilon - \Delta_t(\beta)\big)$ **then**
      Accept H_0^i
      Remove i from M_t and add it to $\mathcal{A}$
    end if
  end for
  $t \leftarrow t + 1$
end while

Proposition 1. *With Δ_t given in* (3)*, we have*

$$\mathbb{P}\big[|\mu_i - \hat{\mu}_i^t| \le \Delta_t(x),\ \forall\ i \in \mathcal{M}, t \in \mathbb{N}\big] \ge 1 - x,$$

for any $x \in (0, 1)$.

Proof. First note that for any $x \in (0, 1)$, we have

$$\begin{aligned}\mathbb{P}\big[\exists i \in \mathcal{M}, t \in \mathbb{N} : |\mu_i - \hat{\mu}_i^t| > \Delta_t(x)\big] &\overset{\text{(a)}}{\le} \sum_{i=1}^{m}\sum_{t=1}^{\infty}\mathbb{P}\big[|\mu_i - \hat{\mu}_i^t| > \Delta_t(x)\big] \\ &\overset{\text{(b)}}{\le} \sum_{i=1}^{m}\sum_{t=1}^{\infty} 2\exp\bigg(-\frac{2t\Delta_t(x)^2}{\sigma^2}\bigg),\end{aligned}$$

where **(a)** follows from a union bound and **(b)** is an application of Hoeffding bound for $\frac{\sigma^2}{4}$-subgaussian random variables. Substituting from (3) for $\Delta_t(x)$ in the above equation

Algorithm 2 Bernstein Multiple Hypothesis Testing

Input: $\mathcal{M}, \mu_0, \epsilon, \alpha, \beta, R$
Initialize: $t = 1$ and $M_t = \mathcal{M}$
while M_t is non-empty **do**
  pull each arm $i \in M_t$ once and observe the data
  Compute empirical means $\hat{\mu}_i^t,\ \forall\ i \in M_t$
  Compute empirical standard deviations , $\hat{\nu}_i^t,\ \forall\ i \in M_t$
  Compute $Delta_t^i,\ \forall\ i \in M_t$ according to (4)
  for $i \in M_t$ **do**
    if $|\hat{\mu}_i^t - \hat{\mu}_0^t| > \big(\epsilon + \Delta_t^i(\alpha)\big)$ **then**
      Reject H_0^i
      Remove i from M_t and add it to $\mathcal{R}$
    end if
    if $|\hat{\mu}_i^t - \hat{\mu}_0^t| < \Big(\epsilon - \Delta_t^i(\beta)\Big)$ **then**
      Accept H_0^i
      Remove i from M_t and add it to $\mathcal{A}$
    end if
  end for
  $t \leftarrow t + 1$
end while

gives

$$\begin{aligned}
\mathbb{P}\big[|\mu_i - \hat{\mu}_i^t| \le \Delta_t(x),\ \forall\ i \in \mathcal{M}, t \in \mathbb{N}\big] &= 1 - \mathbb{P}\big[\exists i \in \mathcal{M}, t \in \mathbb{N} : |\mu_i - \hat{\mu}_i^t| > \Delta_t(x)\big] \\
&\ge 1 - \sum_{i=1}^{m}\sum_{t=1}^{\infty} 2\exp\Big(-\frac{2t}{\sigma^2}\frac{\sigma^2}{2t}\log\Big(\frac{4mt^2}{x}\Big)\Big) \\
&= 1 - \sum_{i=1}^{m}\sum_{t=1}^{\infty} 2\frac{x}{4mt^2} \\
&= 1 - \frac{x}{2}\sum_{t=1}^{\infty}\frac{1}{t^2} \\
&= 1 - \frac{x}{2}\frac{\pi^2}{6} \\
&\ge 1 - x.
\end{aligned}$$

□

Figure 1 summarizes the nature of the proposed bandit algorithm. At any time t, three different regions ("rejection", "acceptance" and "no decision") are recognized and a decision is made for each arm based on the region its empirical mean lies in.

The Hoeffding-based sequence Δ_t defined in (3) does not take the variance of each arm into account. In the scenarios where the variance of an arm is very small compared to the

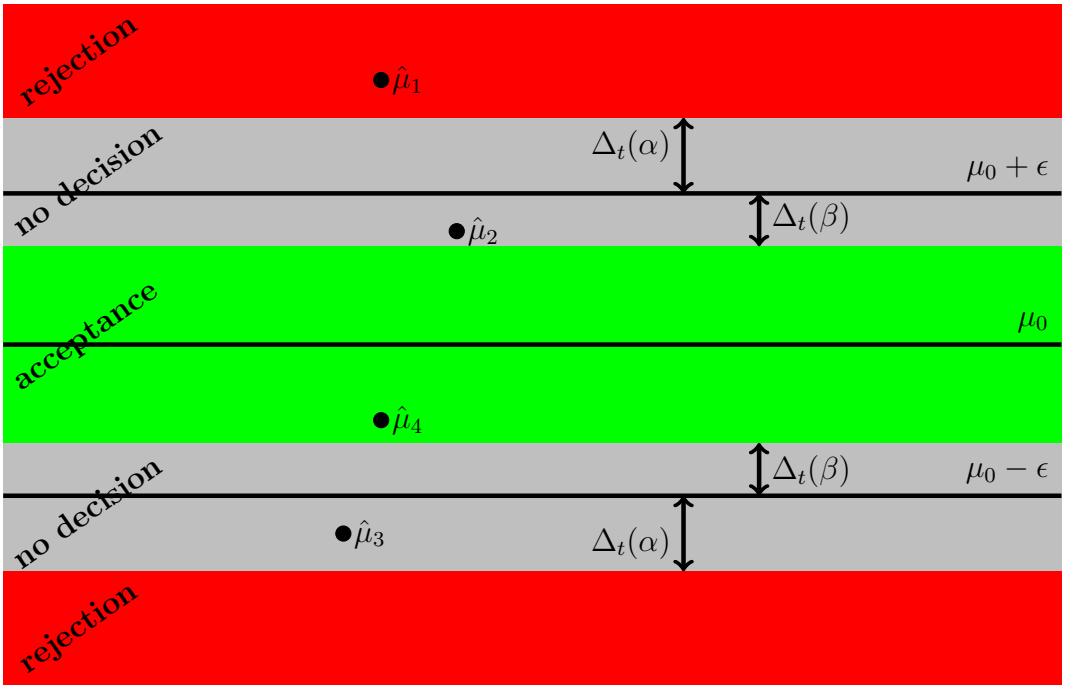

Figure 1: Different regions recognized by the bandit algorithm at time t. At this time, H_0^1 is rejected, H_0^4 is accepted and no decision has been made for H_0^2 and H_0^3.

range of the possible observations, it is very helpful to use a variance dependent concentration bound, such as the Bernstein inequality. This includes for example, a binary multiple hypothesis testing where we know a priori that the conversion rates are very smaller than 1. Therefore in the scenarios where $X_i^t \in [0, R]$ for some $R \geq 0$ and $Var(X_i^t) \ll R$, we suggest running the proposed algorithm with the following arm-dependent sequence:

$$\Delta_t^i(x) = \hat{\nu}_i^t \sqrt{\frac{2}{t} \log\left(\frac{6mt^2}{x}\right)} + \frac{3R}{t} \log\left(\frac{6mt^2}{x}\right), \tag{4}$$

where $\hat{\nu}_i^t$ is the empirical standard deviation of arm i at time t computed as

$$\hat{\nu}_i^t = \sqrt{\frac{1}{t} \sum_{\tau=1}^{t} \left(X_i^t - \hat{\mu}_i^t\right)^2}. \tag{5}$$

Successive steps and details of the algorithm with the above choice of Δ are summarized in Algorithm 2. While a similar analysis of our proposed algorithm can be carried out with this choice of concentration function, we specialize our analysis in the following section to the Hoeffding-based sequence $\{\Delta_t\}$ defined in (3).

3 Analysis

In this section, we analyze the performance of the proposed bandit algorithm and show that it can guarantee the desired type I and type II errors. Furthermore, we derive an upper bound on the run time of the algorithm.

3.1 FWER of the Algorithm

We start by driving an upper bound on the type I error of the bandit algorithm. We have

$$\begin{aligned}
E_1 &= \mathbb{P}[V \geq 1] \\
&= \mathbb{P}\big[\mathcal{R} \cap \mathcal{M}_a \text{ is non-empty}\big] \\
&= \mathbb{P}\Big[\exists\ i \in \mathcal{M}_a, t \in \mathbb{N}: \quad \big|\hat{\mu}_i^t - \mu_0\big| > \big(\epsilon + \Delta_t(\alpha)\big)\Big] \\
&\overset{\text{(a)}}{\leq} \mathbb{P}\Big[\exists\ i \in \mathcal{M}_a, t \in \mathbb{N}: \quad \big|\hat{\mu}_i^t - \mu_i\big| > \Delta_t(\alpha)\Big] \\
&= 1 - \mathbb{P}\Big[\big|\hat{\mu}_i^t - \mu_i\big| \leq \Delta_t(\alpha), \forall\ i \in \mathcal{M}_a, t \in \mathbb{N}\Big] \\
&\leq 1 - \mathbb{P}\Big[\big|\hat{\mu}_i^t - \mu_i\big| \leq \Delta_t(\alpha), \forall\ i \in \mathcal{M}, t \in \mathbb{N}\Big] \\
&\overset{\text{(b)}}{\leq} \alpha,
\end{aligned} \tag{6}$$

where **(a)** follows because the event $\Big[\exists\ i \in \mathcal{M}_a, t \in \mathbb{N}: \quad \big|\hat{\mu}_i^t - \mu_i\big| > \big(\epsilon + \Delta_t(\alpha)\big)\Big]$ implies the event $\Big[\exists\ i \in \mathcal{M}_a, t \in \mathbb{N}: \quad \big|\hat{\mu}_i^t - \mu_i\big| > \Delta_t(\alpha)\Big]$, and **(b)** follows from (2). Therefore, the proposed bandit algorithm guarantees the desired type I FWER.

To derive type II error rate of the bandit algorithm, we have

$$\begin{aligned}
E_2 &= \mathbb{P}[W \geq 1] \\
&= \mathbb{P}\big[\mathcal{A} \cap \mathcal{M}_r \text{ is non-empty}\big] \\
&= \mathbb{P}\Big[\exists\ i \in \mathcal{M}_r, t \in \mathbb{N}: \quad \big|\hat{\mu}_i^t - \mu_0\big| \leq \big(\epsilon - \Delta_t(\beta)\big)\Big] \\
&\overset{\text{(a)}}{\leq} \mathbb{P}\Big[\exists\ i \in \mathcal{M}_r, t \in \mathbb{N}: \quad \big|\hat{\mu}_i^t - \mu_i\big| > \Delta_t(\beta)\Big]\mathbb{I}[\Delta_t(\beta) < \epsilon] \\
&= 1 - \mathbb{P}\Big[\big|\hat{\mu}_i^t - \mu_i\big| \leq \Delta_t(\beta), \forall\ i \in \mathcal{M}_r, t \in \mathbb{N}\Big] \\
&\leq 1 - \mathbb{P}\Big[\big|\hat{\mu}_i^t - \mu_i\big| \leq \Delta_t(\beta), \forall\ i \in \mathcal{M}, t \in \mathbb{N}\Big] \\
&\overset{\text{(b)}}{\leq} \beta,
\end{aligned} \tag{7}$$

where **(a)** follows from the fact that given $\Delta_t(\beta) < \epsilon$, the event $\Big[\exists\ i \in \mathcal{M}_r, t \in \mathbb{N}: \quad \big|\hat{\mu}_i^t - \mu_i\big| \leq \big(\epsilon - \Delta_t(\beta)\big)\Big]$ implies the event $\Big[\exists\ i \in \mathcal{M}_r, t \in \mathbb{N}: \quad \big|\hat{\mu}_i^t - \mu_i\big| > \Delta_t(\beta)\Big]$ and **(b)** follows from (2). Therefore, the proposed bandit algorithm guarantees the desired type II FWER.

3.2 Run Time of the Algorithm

In this section, we derive a probabilistic bound for the run time of the algorithm as a measure of its efficiency. To derive such a bound, we focus on the case where the sequence Δ_t is given by an application of Hoeffding bound and union bound as in (3). A similar approach can be taken for any other types of concentration bound.

Indeed, the run time depends on the positions of the mean values of different arms w.r.t. the equality region $\{\mu : |\mu - \mu_0| < \epsilon\}$ (see Figure 1). For any $i \in \mathcal{M}$, let $d_i = \min\Big(|\mu_i - (\mu_0+\epsilon)|, |\mu_i - (\mu_0-\epsilon)|\Big)$ be the the distance of arm i's mean value to the boundaries of the equality region. Furthermore, define

$$\overline{d} = \min\big\{d_i : i \in \mathcal{M}_r\big\}, \quad \underline{d} = \min\big\{d_i : i \in \mathcal{M}_a\big\}, \tag{8}$$

and let

$$\overline{T} = \min\big\{t \geq 0 : \Delta_t(\alpha) = \overline{d}\big\}, \quad \underline{T} = \min\big\{t \geq 0 : \Delta_t(\beta) = \underline{d}\big\}. \tag{9}$$

We also let $T^* = \max(\underline{T}, \overline{T})$ and $d = \min(\overline{d}, \underline{d})$.

Now, let S be the stopping time of the proposed algorithm. The following theorem provides a probabilistic upper bound on the run time of the proposed algorithm.

Theorem 2. *For any $T \geq T^*$, the following holds for the stopping time of the proposed bandit algorithm*

$$\mathbb{P}\big[S > T\big] \leq 2\Bigg(\sum_{i\in\mathcal{M}_r} \exp\bigg(-\frac{2T\big(d_i - \Delta_T(\alpha)\big)^2}{\sigma^2}\bigg) + \sum_{i\in\mathcal{M}_a} \exp\bigg(-\frac{2T\big(d_i - \Delta_T(\beta)\big)^2}{\sigma^2}\bigg)\Bigg) \tag{10}$$

Proof. We have

$$\begin{aligned}\mathbb{P}[S > T] &= \mathbb{P}\big[\exists i \in \mathcal{M} : i \in M_T\big] \\ &= \mathbb{P}\big[\exists i \in \mathcal{M}_r : i \in M_T\big] + \mathbb{P}\big[\exists i \in \mathcal{M}_a : i \in M_T\big].\end{aligned} \tag{11}$$

Now note that if $i \in M_T$ for some $i \in \mathcal{M}_r$, then $\hat{\mu}_i^T$ lies in one of the "no decision" regions at time T (see Fig. 1); i.e., $\hat{\mu}_i^T \in [\mu_0+\epsilon-\Delta_T(\beta), \mu_0+\epsilon+\Delta_T(\alpha)] \cup [\mu_0-\epsilon-\Delta_T(\alpha), \mu_0-\epsilon+\Delta_T(\beta)]$. Now note that since $|\mu_i - \mu_0| > \epsilon$ and $T \geq \overline{T}$, then $\Delta_T(\alpha) \leq d_i$ and hence

$$\begin{aligned}|\mu_i - \hat{\mu}_i^T| &\geq \inf\Big\{|\mu_i - \mu| : \mu \in [\mu_0 + \epsilon - \Delta_T(\beta), \mu_0 + \epsilon + \Delta_T(\alpha)] \cup [\mu_0 - \epsilon - \Delta_T(\alpha), \mu_0 - \epsilon + \Delta_T(\beta)]\Big\} \\ &= d_i - \Delta_T(\alpha).\end{aligned}$$

Therefore, we can write

$$\mathbb{P}\big[\exists i \in \mathcal{M}_r : i \in M_T\big] \leq \mathbb{P}\big[\exists i \in \mathcal{M}_r : |\mu_i - \hat{\mu}_i^T| \geq \big(d_i - \Delta_T(\alpha)\big)\big]. \tag{12}$$

Similarly, we can show that

$$\mathbb{P}\big[\exists i \in \mathcal{M}_a : i \in M_T\big] \leq \mathbb{P}\big[\exists i \in \mathcal{M}_a : |\mu_i - \hat{\mu}_i^T| \geq \big(d_i - \Delta_T(\beta)\big)\big]. \tag{13}$$

Combining (12) and (13) with (11) gives

$$\begin{aligned}
\mathbb{P}[S > T] &\leq \mathbb{P}\Big[\exists i \in \mathcal{M}_r : |\hat{\mu}_i^T - \mu_i| > \big(d_i - \Delta_T(\alpha)\big)\Big] + \mathbb{P}\Big[\exists i \in \mathcal{M}_a : |\hat{\mu}_i^T - \mu_i| > \big(d_i - \Delta_T(\beta)\big)\Big] \\
&\overset{\textbf{(a)}}{\leq} \sum_{i\in\mathcal{M}_r} \mathbb{P}\Big[|\hat{\mu}_i^T - \mu_i| > \big(d_i - \Delta_T(\alpha)\big)\Big] + \sum_{i\in\mathcal{M}_a} \mathbb{P}\Big[|\hat{\mu}_i^T - \mu_i| > \big(d_i - \Delta_T(\beta)\big)\Big] \\
&\overset{\textbf{(b)}}{\leq} \sum_{i\in\mathcal{M}_r} 2\exp\bigg(-\frac{2T\big(d_i - \Delta_T(\alpha)\big)^2}{\sigma^2}\bigg) + \sum_{i\in\mathcal{M}_a} 2\exp\bigg(-\frac{2T\big(d_i - \Delta_T(\beta)\big)^2}{\sigma^2}\bigg),
\end{aligned}$$

where **(a)** follow from the union bound and **(b)** is an application of the Hoeffding's inequality. □

Now, we can provide a more compact probabilistic bound on the run time of the bandit algorithm. In the simple case where $\overline{T}$ and $\underline{T}$ are small (i.e., $\overline{d}$ and $\underline{d}$ are large), the algorithm stops very soon as it is easy to distinguish the arms from the boundaries. Therefore, we are going to focus on the more interesting and complicated scenario in which $\overline{T}$ and $\underline{T}$ are not small.

Corollary 3. *Assume that* $\min(\overline{T}, \underline{T}) > 20$. *Then, for any* $T > 6T^*$ *the following holds:*

$$\mathbb{P}\big[S > T\big] \leq 2m\exp\Big(-\frac{0.023d^2T}{\sigma^2}\Big). \tag{14}$$

Proof. Since $\forall\ i \in \mathcal{M}_r :\ d_i \geq \overline{d}$ and by definition of $\overline{T}$ in (9), we have for any $T > \overline{T}$ and any $i \in \mathcal{M}_r$:

$$\begin{aligned}
\sqrt{T}\big(d_i - \Delta_T(\alpha)\big) &\geq \sqrt{T}\big(\overline{d} - \Delta_T(\alpha)\big) \\
&= \sqrt{T}\big(\Delta_{\overline{T}}(\alpha) - \Delta_T(\alpha)\big) \\
&= \sqrt{T}\left(\sqrt{\frac{\sigma^2}{\overline{T}}\log\left(2\sqrt{\frac{m}{\alpha}}\overline{T}\right)} - \sqrt{\frac{\sigma^2}{T}\log\left(2\sqrt{\frac{m}{\alpha}}T\right)}\right) \\
&= \sigma\left(\sqrt{\frac{T}{\overline{T}}\log\left(2\sqrt{\frac{m}{\alpha}}\overline{T}\right)} - \sqrt{\log\left(2\sqrt{\frac{m}{\alpha}}T\right)}\right).
\end{aligned} \tag{15}$$

On the other hand if $\overline{T} \geq 20$ and $T \geq 6\overline{T}$, we have

$$\begin{aligned}
& T \geq 6\overline{T} \\
\Rightarrow\ & T \geq 3.5\exp(0.53)\overline{T} \\
\Rightarrow\ & \log\left(2\sqrt{\frac{m}{\alpha}}T\right) \geq \log\left(2\sqrt{\frac{m}{\alpha}}\overline{T}\right) + \log(3.5) + 0.53 \\
\Rightarrow\ & \log\left(2\sqrt{\frac{m}{\alpha}}T\right) \geq \frac{\log\left(2\sqrt{\frac{m}{\alpha}}\,\overline{T}\right)^2 - \log(3.5)^2}{\log\left(2\sqrt{\frac{m}{\alpha}}\,\overline{T}\right) - \log(3.5)} + \frac{\log(3.5)^2}{\log\left(2\sqrt{\frac{m}{\alpha}}\,\overline{T}\right) - \log(3.5)}
\end{aligned}$$

where the last conclusion is based on the fact that $\sqrt{\frac{m}{\alpha}} \geq \sqrt{\frac{2}{0.5}} = 2$ and hence

$$\frac{\log(3.5)^2}{\log\left(2\sqrt{\frac{m}{\alpha}}\,\overline{T}\right) - \log(3.5)} \leq 0.53, \quad \text{for any } \overline{T} \geq 20.$$

Continuing from the above relations, it follows that

$$\begin{aligned}
&\Rightarrow \quad \log\left(2\sqrt{\frac{m}{\alpha}}T\right) \geq \frac{\log\left(2\sqrt{\frac{m}{\alpha}}\,\overline{T}\right)^2}{\log\left(2\sqrt{\frac{m}{\alpha}}\,\overline{T}\right) - \log(3.5)} \\
&\Rightarrow \quad \log\left(2\sqrt{\frac{m}{\alpha}}T\right)\log\left(2\sqrt{\frac{m}{\alpha}}\overline{T}\right) \geq \log\left(2\sqrt{\frac{m}{\alpha}}\overline{T}\right)^2 + \log(3.5)\log\left(2\sqrt{\frac{m}{\alpha}}T\right) \\
&\Rightarrow \quad \log\left(2\sqrt{\frac{m}{\alpha}}T\right) \geq \log\left(2\sqrt{\frac{m}{\alpha}}\overline{T}\right) + \log(3.5)\frac{\log\left(2\sqrt{\frac{m}{\alpha}}T\right)}{\log\left(2\sqrt{\frac{m}{\alpha}}\,\overline{T}\right)} \\
&\Rightarrow \quad \log\left(2\sqrt{\frac{m}{\alpha}}T\right) - \log\left(2\sqrt{\frac{m}{\alpha}}\overline{T}\right) \geq \log(3.5)\frac{\log\left(2\sqrt{\frac{m}{\alpha}}T\right)}{\log\left(2\sqrt{\frac{m}{\alpha}}\,\overline{T}\right)} \\
&\Rightarrow \quad \frac{T}{\overline{T}} \geq \log\left(\frac{T}{\overline{T}}\right) \geq \log(3.5)\frac{\log\left(2\sqrt{\frac{m}{\alpha}}T\right)}{\log\left(2\sqrt{\frac{m}{\alpha}}\,\overline{T}\right)} \\
&\Rightarrow \quad \sqrt{\frac{T}{\overline{T}}\log\left(2\sqrt{\frac{m}{\alpha}}\overline{T}\right)} \geq \sqrt{\log(3.5)}\sqrt{\log\left(2\sqrt{\frac{m}{\alpha}}T\right)},
\end{aligned}$$

for any $T \geq 6\overline{T}$. Since $T^* \geq \overline{T}$, combining the above with (15) gives for any $T \geq 6T^*$, and $i \in \mathcal{M}_r$:

$$\begin{aligned}
\sqrt{T}\big(d_i - \Delta_T(\alpha)\big) &\geq \sigma\left(1 - \frac{1}{\sqrt{\log(3.5)}}\right)\sqrt{\frac{T}{\overline{T}}\log\left(2\sqrt{\frac{m}{\alpha}}\overline{T}\right)} \\
&\geq 0.106\overline{d}\sqrt{T}.
\end{aligned} \tag{16}$$

Similarly, we can show that for any $T \geq 6T^*$ and any $i \in \mathcal{M}_a$:

$$\sqrt{T}\big(d_i - \Delta_T(\beta)\big) \geq 0.106\underline{d}\sqrt{T}. \tag{17}$$

Combining (16) and (17) with Theorem 2 gives us that for any $T \geq 6T^*$, we have

$$\mathbb{P}[S > T] \leq 2\sum_{i\in\mathcal{M}_r}\exp\left(-\frac{0.023\overline{d}^2T}{\sigma^2}\right) + 2\sum_{i\in\mathcal{M}_a}\exp\left(-\frac{0.023\underline{d}^2T}{\sigma^2}\right) \leq 2m\exp\Big(-\frac{0.023d^2T}{\sigma^2}\Big).$$

□

The following Corollary gives the expected run time of the proposed bandit algorithm.

Corollary 4. *Assuming* $\min(\overline{T}, \underline{T}) \geq 20$*, the expected run time of the proposed algorithm is*

$$\mathbb{E}[S] = 6T^* + \frac{87mT^{*0.862}}{\log\left(2\sqrt{\frac{m}{\max(\alpha,\beta)}}T^*\right)} = O(T^*). \tag{18}$$

Proof. Using Corollary 3 , we have

$$\begin{aligned}
\mathbb{E}[S] &= \int_0^\infty \mathbb{P}[S > T]dT \\
&= \int_0^{6T^*} \mathbb{P}[S > T]dT + \int_{6T^*}^\infty \mathbb{P}[S > T]dT \\
&\leq 6T^* + 2m\int_{6T^*}^\infty \exp\left(-\frac{0.023d^2T}{\sigma^2}\right)dT \\
&= 6T^* + 87\frac{m\sigma^2}{d^2}\exp\left(-\frac{0.138d^2T^*}{\sigma^2}\right).
\end{aligned} \tag{19}$$

Note that by the definition of $\overline{T}$ in (9) and the fact that $\Delta_t(\alpha)$ is decreasing in t, we have

$$\begin{aligned}
\overline{d} &= \Delta_{\overline{T}}(\alpha) \\
&\geq \Delta_{T^*}(\alpha) \\
&= \sqrt{\frac{\sigma^2}{T^*}\log\left(2\sqrt{\frac{m}{\alpha}}T^*\right)} \\
&\geq \sqrt{\frac{\sigma^2}{T^*}\log\left(2\sqrt{\frac{m}{\max(\alpha,\beta)}}T^*\right)}.
\end{aligned} \tag{20}$$

Similarly, it can be shown that

$$\underline{d} \geq \sqrt{\frac{\sigma^2}{T^*}\log\left(2\sqrt{\frac{m}{\max(\alpha,\beta)}}T^*\right)}. \tag{21}$$

From (20), (21) and the fact that $d = \min(\overline{d}, \underline{d})$, it follows that

$$d \geq \sqrt{\frac{\sigma^2}{T^*}\log\left(2\sqrt{\frac{m}{\max(\alpha,\beta)}}T^*\right)}. \tag{22}$$

Applying the bound in (22) for d in (19) gives

Algorithm 3 Bandit Multiple Hypothesis Testing with Unknown μ_0

Input: $\mathcal{M}, \mu_0, \epsilon, \alpha, \beta, \{\Delta_t\}_{t\in\mathbb{N}}$
Initialize: $t = 1$ and $M_t = \mathcal{M}$
while M_t is non-empty **do**
  pull each arm $i \in M_t$ once and observe the data
  Compute empirical means $\hat{\mu}_0^t, \hat{\mu}_i^t, \ \forall\ i \in M_t$
  for $i \in M_t$ **do**
    if $|\hat{\mu}_i^t - \hat{\mu}_0^t| > \big(\epsilon + 2\Delta_t(\alpha)\big)$ **then**
      Reject H_0^i
      Remove i from M_t and add it to $\mathcal{R}$
    end if
    if $|\hat{\mu}_i^t - \hat{\mu}_0^t| < \big(\epsilon - 2\Delta_t(\beta)\big)$ **then**
      Accept H_0^i
      Remove i from M_t and add it to $\mathcal{A}$
    end if
  end for
  $t \leftarrow t + 1$
end while

$$
\begin{aligned}
\mathbb{E}[S] &\leq 6T^* + \frac{87mT^*}{\log\Big(2\sqrt{\frac{m}{\max(\alpha,\beta)}}T^*\Big)} \exp\bigg(-0.138\log\Big(2\sqrt{\frac{m}{\max(\alpha,\beta)}}T^*\Big)\bigg) \\
&= 6T^* + \frac{87mT^*}{\log\Big(2\sqrt{\frac{m}{\max(\alpha,\beta)}}T^*\Big)} \times \frac{\big(\max(\alpha,\beta)\big)^{0.138}}{(2\sqrt{m}T^*)^{0.138}} \qquad (23) \\
&\leq 6T^* + \frac{87mT^{*0.862}}{\log\Big(2\sqrt{\frac{m}{\max(\alpha,\beta)}}T^*\Big)}.
\end{aligned}
$$

□

4 Unknown μ_0

In this section, we consider the scenario in which the mean value of the base arm (μ_0) is unknown. Along with other arms, a new observation X_0^t is made for the base arm at time t which is generated as

$$X_0^t = \mu_0 + n_0^t, \qquad (24)$$

where the noise components $(n_0^1, n_0^2, \cdots)$ are iid and independent from the noise components associated with other arms. Also similar to Assumption 1, we assume that n_n^t is

$\frac{\sigma^2}{4}$-subgaussian. We show with a simple modification, our proposed algorithm can still classify the arms correctly within the desired error bounds.

Algorithm 3 describes the modification of our main algorithm to handle this scenario. In this case, we still build –at any time t– three regions and decide about each arm based on the region its empirical mean lies in. However, despite the previous case where these regions are centered at μ_0, here the regions are centered at the empirical mean of the base arm $\hat{\mu}_0^t$. Furthermore, in order to count for the additional uncertainty of not knowing μ_0, the length of the "no decision" region is doubled in this case. Along with other parameters, the algorithm takes as input a sequence of concentration functions $\{\Delta_t : (0,1) \to \mathbb{R}\}$ which for any $x \in (0,1)$ satisfies

$$\mathbb{P}\big[|\mu_i - \hat{\mu}_i^t| \leq \Delta_t(x),\ \forall\ i \in \mathcal{M} \cup \{0\}, t \in \mathbb{N}\big] \geq 1 - x.$$

Again, an standard choice is the Hoeffding error bound over all times and arms, given by

$$\Delta_t(x) = \sqrt{\frac{\sigma^2}{2t} \log\Big(\frac{4(m+1)t^2}{x}\Big)}. \tag{25}$$

Similar to the previous scenario where μ_0 was known, we can improve the performance of the Algorithm 3, by employing a Bernstein type concentration function. Specifically, in the scenarios where $X_i^t \in [0, R]$ and $Var(X_i^t) \ll R$ for any $t \in \mathbb{N}, i = 0, 1, \cdots, m$, we can use the following sequence

$$\Delta_t^i(x) = \hat{\nu}_i^t \sqrt{\frac{2}{t} \log\Big(\frac{6(m+1)t^2}{x}\Big)} + \frac{3R}{t} \log\Big(\frac{6(m+1)t^2}{x}\Big), \tag{26}$$

for each $i = 0, 1, \cdots, m$, where similar to (5), ν_0^t is the empirical standard deviation of arm 0 at time t. Successive steps and details of this version are summarized in Algorithm 4.

Algorithm 4 Bernstein Multiple Hypothesis Testing with Unknown μ_0

Input: $\mathcal{M}, \mu_0, \epsilon, \alpha, \beta, R$
Initialize: $t = 1$ and $M_t = \mathcal{M}$
while M_t is non-empty **do**
  pull each arm $i \in M_t$ once and observe the data
  Compute empirical means $\hat{\mu}_0^t, \hat{\mu}_i^t, \ \forall\ i \in M_t$
  Compute empirical standard deviations $\hat{\nu}_0^t, \hat{\nu}_i^t, \ \forall\ i \in M_t$
  Compute $\Delta_t^0, \Delta_t^i, \ \forall\ i \in M_t$ according to (26)
  for $i \in M_t$ **do**
    if $|\hat{\mu}_i^t - \hat{\mu}_0^t| > \left(\epsilon + \Delta_t^0(\alpha) + \Delta_t^i(\alpha)\right)$ **then**
      Reject H_0^i
      Remove i from M_t and add it to $\mathcal{R}$
    end if
    if $|\hat{\mu}_i^t - \hat{\mu}_0^t| < \left(\epsilon - \left(\Delta_t^0(\beta) + \Delta_t^i(\beta)\right)\right)$ **then**
      Accept H_0^i
      Remove i from M_t and add it to $\mathcal{A}$
    end if
  end for
  $t \leftarrow t + 1$
end while