Rotting Infinitely Many-armed Bandits beyond the Worst-case Rotting: An Adaptive Approach

Jung-hun Kim
Seoul National University
junghunkim@snu.ac.kr

Milan Vojnović
London School of Economics
m.vojnovic@lse.ac.uk

Se-Young Yun
KAIST
yunseyoung@kaist.ac.kr

Abstract

In this study, we consider the infinitely many armed bandit problems in rotting environments, where the mean reward of an arm may decrease with each pull, while otherwise, it remains unchanged. We explore two scenarios capturing problem-dependent characteristics regarding the decay of rewards: one in which the cumulative amount of rotting is bounded by V_T, referred to as the slow-rotting scenario, and the other in which the number of rotting instances is bounded by S_T, referred to as the abrupt-rotting scenario. To address the challenge posed by rotting rewards, we introduce an algorithm that utilizes UCB with an adaptive sliding window, designed to manage the bias and variance trade-off arising due to rotting rewards. Our proposed algorithm achieves tight regret bounds for both slow and abrupt rotting scenarios. Lastly, we demonstrate the performance of our algorithms using synthetic datasets.

1 Introduction

We consider multi-armed bandit problems, which are fundamental sequential learning problems where an agent plays an action, referred to as an 'arm,' at each time and receives a corresponding reward. The core of the problems is in dealing with the exploration-exploitation trade-off. The significance of bandit problems extends across diverse real-world applications, such as recommendation systems (Li et al., 2010) and clinical trials (Villar et al., 2015). In a recommendation system, an arm could represent an item, and the goal is to maximize the click-through rate by making effective recommendations.

Real-world observations reveal that the mean rewards associated with arms may decrease over repeated interactions. For instance, in content recommendation systems, the click rates for each action (item) may diminish due to user boredom with repeated exposure to the same content. Another example is evident in clinical trials, where the efficacy of a medication can decline over time due to drug tolerance induced by repeated administration. The decline in mean rewards associated with selected arms, referred to as *(rested) rotting bandits*, has been studied by Levine et al. (2017); Seznec et al. (2019, 2020). The previous work focuses on finite K arms, in which Seznec et al. (2019) proposed algorithms achieving $\tilde{O}(\sqrt{KT})$ regret. This suggests that rotting bandits with a finite number of arms are no harder than the stationary case. However, in real-world scenarios like recommendation systems, where the content items such as movies or articles are numerous, the prior methods encounter limitations as the parameter K

Table 1: Summarized results of our regret analysis.

Type	Regret upper bounds for $\beta \geq 1$	Regret upper bounds for $0 < \beta < 1$	Regret lower bounds for $\beta > 0$
Slow rotting (V_T)	$\tilde{O}\left(\max\{V_T^{\frac{1}{\beta+2}} T^{\frac{\beta+1}{\beta+2}}, T^{\frac{\beta}{\beta+1}}\}\right)$	$\tilde{O}\left(\max\{V_T^{\frac{1}{3}} T^{\frac{2}{3}}, \sqrt{T}\}\right)$	$\Omega\left(\max\{V_T^{\frac{1}{\beta+2}} T^{\frac{\beta+1}{\beta+2}}, T^{\frac{\beta}{\beta+1}}\}\right)$
Abrupt rotting (S_T)	$\tilde{O}\left(\max\left\{S_T^{\frac{1}{\beta+1}} T^{\frac{\beta}{\beta+1}}, V_T\right\}\right)$	$\tilde{O}\left(\max\left\{\sqrt{S_T T}, V_T\right\}\right)$	$\Omega\left(\max\left\{S_T^{\frac{1}{\beta+1}} T^{\frac{\beta}{\beta+1}}, V_T\right\}\right)$

increases, resulting in a trivial regret. This emphasizes the necessity for studying rotting scenarios with *infinitely* many arms, particularly when there is a lack of information about the features of each item. The consideration of infinitely many arms for rested rotting bandits fundamentally distinguishes these problems from those with a finite number of arms, which we will explain later.

The study of multi-armed bandit problems with an infinite number of arms has been extensively conducted in the context of *stationary* reward distributions Berry et al. (1997); Wang et al. (2009); Bonald and Proutière (2013); Carpentier and Valko (2015); Bayati et al. (2020). Initially, the distribution of the mean rewards for the arms was assumed to be uniform over the interval $[0, 1]$ Berry et al. (1997); Bonald and Proutière (2013). This assumption was expanded to include a much wider range of distributions with $\beta > 0$ satisfying $\mathbb{P}(\mu(a) > \mu^* - x) = \Theta(x^\beta)$, where $\mu(a)$ represents the mean reward of arm a, μ^* is the mean reward of the best-performing arm (Wang et al., 2009; Carpentier and Valko, 2015; Bayati et al., 2020). While Kim et al. (2022) explores the concept of diminishing rewards in the context of bandits with infinitely many arms, their focus is limited to the basic case where the initial mean rewards are uniformly distributed, and the decay rate of mean reward per play is bounded by a maximum decay rate $\rho(= o(1))$. Crucially, their focus is on the regret for the worst-case rather than problem-dependent cases with respect to the decay rates. We also note that feature information for each arm is not required for multi-armed bandit problems with infinitely many arms, which differs from linear bandits Abbasi-Yadkori et al. (2011) or continuum-armed bandits Auer et al. (2007); Kleinberg (2004), where feature information for each arm, either for the Lipschitz or linear structure, is involved.

In this work, we explore rotting bandits with infinitely many arms, subject to moderate constraints on the initial mean reward distribution and the rate at which the mean reward of an arm declines. We adopt the assumption that the initial mean reward distribution satisfies the aforementioned condition with parameter $\beta > 0$. Our examination of the diminishing, or 'rotting,' rewards covers two cases: one with the total amount of rotting bounded by V_T, and the other with the total number of rotting instances bounded by S_T. This allows us to capture the *problem-dependency* regarding rotting rates. Similar quantifications have been studied in the context of nonstationary finite K-armed bandit problems Besbes et al. (2014); Auer et al. (2019); Russac et al. (2019), in which the reward distribution changes over time independently of the agent. Adhering to established terminology for nonstationary bandits, we refer to the environment for a bounded total amount of rotting with V_T as the *slow rotting* scenario and the one for a bounded total number of rotting instances with S_T as the *abrupt rotting* scenario.

Here we provide why (rested) rotting for infinitely many arms are fundamentally different from that for finite arms. In the case of finite arms, rested rotting is known to be no harder than stationary case (Seznec et al., 2019, 2020).

This result comes from the confinement of mean rewards of optimal arms and selected arms within confidence bounds (Lemma 1 in Seznec et al. (2019, 2020)). However, in the case of infinite arms, which allows for having an infinite number of near-optimal arms, there always exist near-optimal arms outside of explored arms. Therefore, the mean reward gap may not be confined within confidence bounds. This makes our setting fundamentally different for finite armed bandits, introducing additional challenges.

For the infinite arm settings, there exists an additional cost for exploring new arms to find near-optimal arms while eliminating explored suboptimal arms. If the total rotting effect on explored arms is significant, then the frequency at which new near-optimal arms must be sought increases substantially, resulting in a large regret. This is why the rested rotting significantly affects the exploration cost regarding V_T or S_T in our setting.

To solve our problem, we introduce algorithms that employ an adaptive sliding window mechanism, effectively managing the tradeoff between bias and variance stemming from rotting rewards. Notably, our paper is the first to consider the slow and abrupt rotting scenarios, capturing problem-dependent characteristics, specifically for initial mean reward distributions that satisfy the condition parameterized with β, in the context of infinitely many armed bandits.

Summary of our Contributions. The key contributions of this study are summarized in the following points. Refer to Table 1 for a summary of our regret bounds.

- To address the slow and abrupt rotting scenarios capturing the problem-dependency of rotting rates, we propose a UCB-based algorithm using an adaptive sliding window and a threshold parameter. This algorithm allows for effectively managing the bias and variance trade-off arising from rotting rewards.
- In the context of both slow rotting with V_T and abrupt rotting with S_T, for any $\beta > 0$, we present regret upper bounds achieved by our algorithm with an appropriately tuned threshold parameter.
- We establish regret lower bounds for both slow rotting and abrupt rotting scenarios. These regret lower bounds imply tightness of our upper bounds for the case when $\beta \geq 1$. In the other case, when $0 < \beta < 1$, there is a gap between our upper bounds and the corresponding lower bounds, similar to what can be found in related literature, which is discussed in the paper.
- Lastly, we demonstrate the performance of our algorithm through experiments using synthetic datasets, validating our theoretical results.

2 Problem Statement

We consider rotting bandits with infinitely many arms where the mean reward of an arm may decrease when the agent pulls the arm. Let $\mathcal{A}$ be the set of infinitely many arms and let $\mu_t(a)$ denote the unknown mean reward of arm $a \in \mathcal{A}$ at time t. At each time t, an agent pulls arm $a_t^\pi \in \mathcal{A}$ according to policy π and observes stochastic reward r_t given by $r_t = \mu_t(a_t^\pi) + \eta_t$, where η_t is a noise term following a 1-sub-Gaussian distribution. To simplify, we use a_t for a_t^π when there is no confusion about the policy. We assume that initial mean rewards $\{\mu_1(a)\}_{a\in\mathcal{A}}$ are i.i.d. random variables on $[0, 1]$, which is widely considered in the context of infinitely many-armed bandits (Bonald and

Proutière, 2013; Berry et al., 1997; Wang et al., 2009; Carpentier and Valko, 2015; Bayati et al., 2020; Kim et al., 2022).

As in Wang et al. (2009); Carpentier and Valko (2015); Bayati et al. (2020), we consider, to our best knowledge, the most generalized distribution of the initial mean reward $\mu_1(a)$, for every $a \in \mathcal{A}$, satisfying the following condition: there exists a *constant* $\beta > 0$ such that for every $a \in \mathcal{A}$ and all $x \in [0, 1]$,

$$\mathbb{P}(\mu_1(a) > 1 - x) = \mathbb{P}(\Delta_1(a) < x) = \Theta(x^\beta), \tag{1}$$

where $\Delta_1(a) = 1 - \mu_1(a)$ is the initial sub-optimality gap.

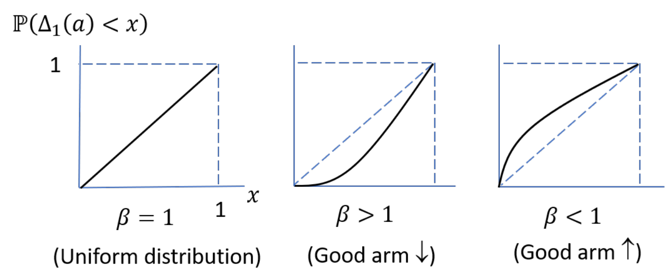

Figure 1: $\mathbb{P}(\Delta_1(a) < x) = x^\beta$ for different values of β.

To discuss the effect of β on the distribution of $\Delta_1(a)$ and the probability of sampling a good arm (having small $\Delta_1(a)$), we consider the case when $\mathbb{P}(\Delta_1(a) < x) = x^\beta$, which is shown in Figure 1 for some values of β. It is noteworthy that the uniform distribution is a special case when $\beta = 1$. Importantly, the larger the value of β, the smaller the probability of sampling a good arm.

The rotting of arms is defined in a similar manner to Kim et al. (2022); Levine et al. (2017); Seznec et al. (2019, 2020). The mean reward of the played arm a_t at time $t > 0$ undergoes the following change:

$$\mu_{t+1}(a_t) = \mu_t(a_t) - \varrho_t(a_t),$$

where rotting rate $\varrho_t(a_t) \geq 0$ is *arbitrarily* determined after pulling a_t and it is unknown to the agent. The mean rewards of other arms remain unchanged as $\mu_{t+1}(a) = \mu_t(a)$ with $\rho_t(a) = 0$ for $a \in \mathcal{A}/\{a_t\}$, which are also determined after pulling a_t. We can express $\mu_t(a) = \mu_1(a) - \sum_{s=1}^{t-1} \varrho_s(a)$, allowing $\mu_t(a)$ to take negative values. For simplicity, we use ρ_t for $\rho_t(a_t)$ when there is no confusion. We consider two distinct measures for quantifying the amount of rotting; the total cumulative rotting over T is bounded by V_T as $\sum_{t=1}^{T-1} \rho_t \leq V_T$, and the number of rotting instances (plus one) is bounded by S_T as $1 + \sum_{t=1}^{T-1} \mathbb{1}(\rho_t \neq 0) \leq S_T (\leq T)$. While we have $S_T \leq T$ because the number of rotting instances is at most $T - 1$, the upper bound for V_T may not exist due to the lack of constraints on ρ_t's. We will shortly introduce and discuss an assumption for V_T.

The goal of this problem is to find a policy that minimizes the expected cumulative regret over a time horizon of T time steps, which, for a given policy π, is defined as $\mathbb{E}[R^\pi(T)] = \mathbb{E}[\sum_{t=1}^{T}(1 - \mu_t(a_t))]$. The use of 1 in the regret definition for the optimal mean reward is justified because there exists an arm whose mean reward is sufficiently close to 1 among the infinite arms with initial mean rewards following the distribution specified in (1).[1]

[1] For any $\epsilon > 0$, there exists $a \in \mathcal{A}$ s.t. $\Delta_1(a) < \epsilon$ with probability 1 because $\lim_{n\to\infty} 1 - \mathbb{P}(\Delta_1(a) \geq \epsilon)^n = 1$.

Here we discuss an assumption for the cumulative amount of rotting (V_T). In the case of $V_T > T$, the problem becomes trivial as shown in the following proposition.

Proposition 2.1. *In the case of $V_T > T$, a simple policy that samples a new arm every round achieves the optimal regret of $\Theta(T)$.*

Proof. The proof is provided in Appendix A.1 □

From the above proposition, when $V_T > T$, the regret lower bound of this problem is $\Omega(T)$, which can be achieved from a simple policy. Therefore, we consider the following assumption for the region of non-trivial problems.

Assumption 2.2. $V_T \leq T$.

The above assumption is not strong, as it frequently arises in real-world scenarios and is more general than the assumption made in prior work, as described in the following remarks.

Remark 2.3. *The assumption $V_T \leq T$ is satisfied when mean rewards consistently remain positive, ensuring $0 \leq \mu_t(a_t) \leq 1$ for all $t \in [T]$, because this condition implies $\rho_t \leq 1$. Such a scenario is frequently encountered in real-world applications, where reward is represented by metrics like click rates or ratings in content recommendation systems.*

Remark 2.4. *Our rotting scenario with $V_T \leq T$ is more general in scope than the one with a maximum rotting rate constraint where $\rho_t \leq \rho = o(1)$ for all $t \in [T-1]$, which was explored in Kim et al. (2022). This is because for our setting, ρ_t is not necessarily bounded by $o(1)$, and for the maximum rotting constraint setting with $\rho_t \leq \rho = o(1)$, $V_T \leq T$ is always satisfied. Additionally, our work delves into the realm of a generalized initial mean reward distribution with $\beta > 0$, offering a broader perspective than the uniform distribution case ($\beta = 1$) considered in Kim et al. (2022).*

3 Algorithms and Regret Analysis

Previous Work. Before moving on to our algorithm, we review the previously suggested algorithm, `UCB-TP` (Kim et al., 2022), which is designed under the case of maximum rotting rate constraint as $\rho_t \leq \rho$ for all $t > 0$. The mean estimator in `UCB-TP` considers the worst-case scenario with the maximum rotting rate ρ as $\tilde{\mu}_t^o(a) - \rho n_t(a)$ where $\tilde{\mu}_t^o$ is an estimator for the initial mean reward and $n_t(a)$ is the number of pulling arm a until $t-1$, which leads to achieve $\tilde{O}(\max\{\rho^{1/3}T, \sqrt{T}\})$. The estimator is not appropriate to deal with ρ_t delicately because it aims to attain the worst-case regret bound regarding rotting rates.

Therefore, we propose using *an adaptive sliding window* for delicately controlling bias and variance tradeoff of the mean reward estimator. This is why our algorithm can adapt to varying rotting rates ρ_t and achieve tight regret bounds with respect to V_T or even S_T. Furthermore, our algorithm accommodates the general mean reward distribution with β by employing a carefully optimized threshold parameter.

Here we describe our proposed algorithm (Algorithm 1) in detail. We define $\widehat{\mu}_{[t_1,t_2]}(a) = \sum_{t=t_1}^{t_2} r_t \mathbb{1}(a_t = a)/n_{[t_1,t_2]}(a)$ where $n_{[t_1,t_2]}(a) = \sum_{t=t_1}^{t_2} \mathbb{1}(a_t = a)$ for $t_1 \leq t_2$. Then for window-UCB index of the algorithm,

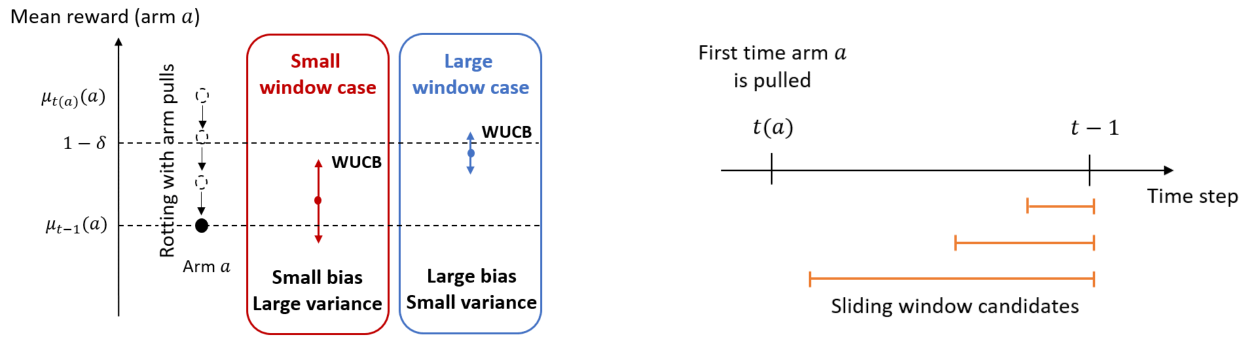

Figure 2: Illustrations for the adaptive sliding window used by our algorithm: (left) the effect of the sliding window length on the mean reward estimation, (right) sliding window candidates with doubling lengths.

we define $WUCB(a, t_1, t_2, T) = \widehat{\mu}_{[t_1,t_2]}(a) + \sqrt{12\log(T)/n_{[t_1,t_2]}(a)}$. In Algorithm 1, we first select an arbitrary arm $a \in \mathcal{A}'$ without prior knowledge regarding the arms in $\mathcal{A}'$, denoting the corresponding time as $t(a)$. We define $\mathcal{T}_t(a)$ as the set of starting times for sliding windows of doubling lengths, defined as $\mathcal{T}_t(a) = \{s \in [T] : t(a) \leq s \leq t-1 \text{ and } s = t - 2^{i-1} \text{ for some } i \in \mathbb{N}\}$. Then the algorithm pulls the arm consecutively until the following threshold condition is satisfied:

$$\min_{s \in \mathcal{T}_t(a)} WUCB(a, s, t-1, T) < 1 - \delta,$$

in which the sliding window having minimized window-UCB is utilized for adapting nonstationarity. If the threshold condition holds, then the algorithm considers the arm to be a sub-optimal (bad) arm and withdraws the arm. Then it samples a new arm and repeats this procedure.

Algorithm 1 UCB-Threshold with Adaptive Sliding Window

Given: $T, \delta, \mathcal{A}$; **Initialize:** $\mathcal{A}' \leftarrow \mathcal{A}$
Sample an arm $a \in \mathcal{A}'$
Pull arm a and get reward r_1
$t(a) \leftarrow 1$
for $t = 2, \dots, T$ **do**
  if $\min_{s \in \mathcal{T}_t(a)} WUCB(a, s, t-1, T) < 1 - \delta$ **then**
    $\mathcal{A}' \leftarrow \mathcal{A}'/\{a\}$
    Select an arm $a \in \mathcal{A}'$
    Pull arm a and get reward r_t
    $t(a) \leftarrow t$
  else
    Pull arm a and get reward r_t
  end if
end for

Utilizing the adaptive sliding window having minimized window UCB index enhances the algorithm's ability to dynamically identify poorly-performing arms across varying rotting rates. This adaptability is achieved by managing the tradeoff between bias and variance. The concept is depicted in Figure 2 (left), where an arm a is rotted several times, and WUCB with a smaller window exhibits minimal bias with the arm's most recent mean reward but introduces higher variance. Conversely, WUCB with a larger window displays increased bias but reduced variance. In this visual representation, the value of WUCB with a small window attains a minimum, enabling the algorithm to compare this value with $1-\delta$ to identify the suboptimal arm. Moreover, as illustrated in Figure 2 (right), by taking into account the constraint of $s = t - 2^{i-1}$ for the size of the adaptive windows, we can reduce the computation time for determining the appropriate window from $O(T^2)$ to $O(T \log T)$ across T time steps, and reduce the required memory from $O(t)$ to $O(\log t)$ for each time t.

Slow Rotting with V_T. Here we consider slowly rotting reward distribution with given $V_T(\leq T)$. We analyze the regret of Algorithm 1 with tuned δ using β and V_T. We define $\delta_V(\beta) = \max\{(V_T/T)^{1/(\beta+2)}, 1/T^{1/(\beta+1)}\}$ when $\beta \geq 1$ and $\delta_V(\beta) = \max\{(V_T/T)^{1/3}, 1/\sqrt{T}\}$ when $0 < \beta < 1$. The algorithm with $\delta_V(\beta)$ achieves a regret bound in the following theorem.

Theorem 3.1. *The policy π of Algorithm 1 with $\delta = \delta_V(\beta)$ achieves:*

$$\mathbb{E}[R^\pi(T)] = \begin{cases} \tilde{O}(\max\{V_T^{\frac{1}{\beta+2}} T^{\frac{\beta+1}{\beta+2}}, T^{\frac{\beta}{\beta+1}}\}) & for \;\; \beta \geq 1, \\ \tilde{O}(\max\{V_T^{\frac{1}{3}} T^{\frac{2}{3}}, \sqrt{T}\}) & for \;\; 0 < \beta < 1. \end{cases}$$

We observe that when β increases above 1, the regret bound becomes worse because the likelihood of sampling a good arm decreases. However, when β decreases below 1, the regret bound remains the same due to the inability to avoid a certain level of regret arising from estimating the mean reward. Further discussion will be provided later. Also, we observe that when $V_T = O(\max\{1/T^{1/(\beta+1)}, 1/\sqrt{T}\})$ where the problem becomes near-stationary, the regret bound in Theorem 3.1 matches the previously known regret bound $\tilde{O}(\max\{T^{\beta/(\beta+1)}, \sqrt{T}\})$ for the stationary infinitely many-armed bandits in Wang et al. (2009); Bayati et al. (2020).

Proof sketch. The full proof is provided in Appendix A.2. Here we outline the main ideas of the proof. The main difficulties in the proof come from analyzing regret delicately for varying ρ_t over time horizon T with an adaptive sliding window and for the generalized initial mean reward distribution with parameter β, which do not appear in Kim et al. (2022).

We initially separate the regret into two components: one associated with pulling initially good arms and another with pulling initially bad arms. The good arms indicate arms a which satisfy initial mean reward $\mu_1(a) \geq 1 - 2\delta$ and otherwise, the arms are recognized as bad arms. The reason why the separation is required is that our adaptive algorithm exhibits distinct behaviors depending on the category of arms. In short, good arms may be pulled continuously by our algorithm when rotting rates are sufficiently small but bad arms are not. This separation for regret is denoted as

$$R^\pi(T) = R^{\mathcal{G}}(T) + R^{\mathcal{B}}(T),$$

where $R^{\mathcal{G}}(T)$ is regret from good arms and $R^{\mathcal{B}}(T)$ is regret from bad arms.

We first provide a bound for $\mathbb{E}[R^{\mathcal{G}}(T)]$. For analyzing regret from good arms, we analyze the cumulative amount of rotting while pulling a sampled good arm before withdrawing the arm by the algorithm. Let $\mathcal{A}_T^{\mathcal{G}}$ be a set of good arms sampled until T, $t_1(a)$ be the initial time step at which arm a is pulled, and $t_2(a)$ be the final time step at which the arm is pulled by the algorithm so that the threshold condition holds when $t = t_2(a)+1$. For simplicity, we use t_1 and t_2 for $t_1(a)$ and $t_2(a)$, when there is no confusion. For any time steps $n \le m$, we define $V_{[n,m]}(a) = \sum_{t=n}^{m} \rho_t(a)$ and $\overline{\rho}_{[n,m]}(a) = V_{[n,m]}(a)/n_{[n,m]}(a)$. We show that the regret is decomposed as

$$R^{\mathcal{G}}(T) = \sum_{a\in\mathcal{A}_T^{\mathcal{G}}} \left(\Delta_1(a) n_{[t_1,t_2]}(a) + \sum_{t=t_1+1}^{t_2} V_{[t_1,t-1]}(a) \right), \tag{2}$$

which consists of regret from the initial mean reward and the cumulative amount of rotting for each arm. For the first term of $\sum_{a\in\mathcal{A}_T^{\mathcal{G}}} \Delta_1(a) n_{[t_1,t_2]}(a)$ in (2), since $\Delta_1(a) = O(\delta)$ from the definition of good arms $a \in \mathcal{A}_T^{\mathcal{G}}$, we have

$$\mathbb{E}\left[\sum_{a\in\mathcal{A}_T^{\mathcal{G}}} \Delta_1(a) n_{[t_1,t_2]}(a) \right] = O(\delta T). \tag{3}$$

The main difficulty lies in dealing with the second term of $\sum_{a\in\mathcal{A}_T^{\mathcal{G}}} \sum_{t=t_1+1}^{t_2} V_{[t_1,t-1]}(a)$ in (2), where we need to analyze the amount of cumulative rotting until the arm is eliminated from the adaptive threshold condition. Let $w(a) = \lceil \log(T)^{1/3} / \overline{\rho}_{[t_1,t_2-2]}(a)^{2/3} \rceil$, which denotes the optimized window size to control variance-bias tradeoff. We define an arm set $\overline{\mathcal{A}}_T^{\mathcal{G}} = \{a \in \mathcal{A}_T^{\mathcal{G}}; n_{[t_1,t_2-1]}(a) > w(a)\}$ which includes arms pulled more than $w(a)$ times. Then we analyze the second term in (2) by separating it as

$$\sum_{a\in\mathcal{A}_T^{\mathcal{G}}} \sum_{t=t_1+1}^{t_2} V_{[t_1,t-1]}(a) = \sum_{a\in\mathcal{A}_T^{\mathcal{G}}/\overline{\mathcal{A}}_T^{\mathcal{G}}} \sum_{t=t_1+1}^{t_2} V_{[t_1,t-1]}(a) + \sum_{a\in\overline{\mathcal{A}}_T^{\mathcal{G}}} \sum_{t=t_1+1}^{t_2} V_{[t_1,t-1]}(a). \tag{4}$$

From the restricted number of pulls of $w(a)$ times for each arm $a \in \mathcal{A}_T^{\mathcal{G}}/\overline{\mathcal{A}}_T^{\mathcal{G}}$, by omitting details here, we show that

$$\sum_{a\in\mathcal{A}_T^{\mathcal{G}}/\overline{\mathcal{A}}_T^{\mathcal{G}}} \sum_{t=t_1+1}^{t_2} V_{[t_1,t-1]}(a) = \tilde{O}\left(V_T + \sum_{a\in\mathcal{A}_T^{\mathcal{G}}/\overline{\mathcal{A}}_T^{\mathcal{G}}} V_{[t_1,t_2-2]}(a)^{\frac{1}{3}} n_{[t_1,t_2-2]}(a)^{\frac{2}{3}} \right), \tag{5}$$

where the first term comes from considering the worst case of rotting.

Regarding the arms in $\overline{\mathcal{A}}_T^{\mathcal{G}}$, where each arm is played sufficiently to have the optimized window, a careful analysis of the adaptive threshold policy is required to limit the total variation in rotting. By examining the estimation errors arising from variance and bias due to the threshold condition, we can establish an upper bound for the cumulative amount of rotting, given by:

$$\sum_{a\in\overline{\mathcal{A}}_T^{\mathcal{G}}} \sum_{t=t_1+1}^{t_2} V_{[t_1,t-1]}(a) = \tilde{O}\left(T\delta + V_T + \sum_{a\in\overline{\mathcal{A}}_T^{\mathcal{G}}} V_{[t_1,t_2-2]}(a)^{\frac{1}{3}} n_{[t_1,t_2-2]}(a)^{\frac{2}{3}} \right), \tag{6}$$

where the first term comes from the threshold parameter for determining when to withdraw arms. Therefore, from $\delta = \delta_V(\beta)$, $V_T \le T$, and equations (2)-(6), using Hölder's inequality, we have

$$\mathbb{E}[R^{\mathcal{G}}(T)] = \begin{cases} \tilde{O}(\max\{V_T^{\frac{1}{\beta+2}} T^{\frac{\beta+1}{\beta+2}}, T^{\frac{\beta}{\beta+1}}\}) & for \ \ \beta \ge 1, \\ \tilde{O}(\max\{V_T^{\frac{1}{3}} T^{\frac{2}{3}}, \sqrt{T}\}) & for \ \ 0 < \beta < 1. \end{cases} \tag{7}$$

Next, we provide a bound for $\mathbb{E}[R^{\mathcal{B}}(T)]$. We employ episodic regret analysis, defining an episode as the time steps between consecutively sampled distinct good arms by the algorithm. By analyzing bad arms within each episode, we can derive an upper bound for the overall regret stemming from bad arms. We define the regret from bad arms over $m^{\mathcal{G}}$ episodes as $R^{\mathcal{B}}_{m^{\mathcal{G}}}$. We first consider the case of $V_T = \omega(\max\{1/\sqrt{T}, 1/T^{1/(\beta+1)}\})$. In this case, by setting $m^{\mathcal{G}} = \lceil 2V_T/\delta \rceil$, we can show that $R^{\mathcal{B}}(T) \le R^{\mathcal{B}}_{m^{\mathcal{G}}}$. By analyzing $R^{\mathcal{B}}_{m^{\mathcal{G}}}$ with the episodic analysis, we can show that

$$\mathbb{E}[R^{\mathcal{B}}(T)] \le \mathbb{E}[R^{\mathcal{B}}_{m^{\mathcal{G}}}] = \tilde{O}\left(\max\left\{T^{\frac{\beta+1}{\beta+2}} V_T^{\frac{1}{\beta+2}}, T^{\frac{2}{3}} V_T^{\frac{1}{3}}\right\}\right). \tag{8}$$

As in the similar manner, when $V_T = O(\max\{1/\sqrt{T}, 1/T^{1/(\beta+1)}\})$, by setting $m^{\mathcal{G}} = C_3$ for some constant $C_3 > 0$, we can show that

$$\mathbb{E}[R^{\mathcal{B}}(T)] \le \mathbb{E}[R^{\mathcal{B}}_{m^{\mathcal{G}}}] = \tilde{O}\left(\max\left\{T^{\frac{\beta}{\beta+1}}, \sqrt{T}\right\}\right). \tag{9}$$

From (8) and (9), we have

$$\mathbb{E}[R^{\mathcal{B}}(T)] = \begin{cases} \tilde{O}(\max\{V_T^{\frac{1}{\beta+2}} T^{\frac{\beta+1}{\beta+2}}, T^{\frac{\beta}{\beta+1}}\}) & for \;\; \beta \ge 1, \\ \tilde{O}(\max\{V_T^{\frac{1}{3}} T^{\frac{2}{3}}, \sqrt{T}\}) & for \;\; 0 < \beta < 1. \end{cases} \tag{10}$$

Finally, using (7) and (10), we can conclude the proof from $\mathbb{E}[R^{\pi}(T)] = \mathbb{E}[R^{\mathcal{G}}(T)] + \mathbb{E}[R^{\mathcal{B}}(T)]$. □

Remark 3.2. *We compare our result in Theorem 3.1 with that in Kim et al. (2022) which, recall, is under the maximum rotting constraint $\rho_t \le \rho = o(1)$ for all t and uniform distribution of initial mean rewards ($\beta = 1$). Under the maximum rotting constraint, we have $V_T \le \rho T$. Then with $\beta = 1$, we can observe that the regret bound of Algorithm 1 is tighter than that of* `UCB-TP` *(Kim et al., 2022) as*

$$\tilde{O}\left(\max\left\{V_T^{\frac{1}{3}} T^{\frac{2}{3}}, \sqrt{T}\right\}\right) \le \tilde{O}\left(\max\left\{\rho^{\frac{1}{3}} T, \sqrt{T}\right\}\right), \tag{11}$$

where the right term is the regret bound of `UCB-TP`*. Here we provide an example to clarify the above inequality. We consider a case of $\rho_t = 1/(t\log(T))$ for all t so that $\rho = \rho_1 = 1/\log(T)$ and $V_T = \sum_{t=1}^{T-1} \rho_t = O(1)$. In this case, the regret bound of Algorithm 1, $\tilde{O}(\max\{V_T^{1/3} T^{2/3}, \sqrt{T}\}) = \tilde{O}(T^{2/3})$, is sublinear while that of* `UCB-TP`*, $\tilde{O}(\max\{\rho^{1/3} T, \sqrt{T}\}) = \tilde{O}(T/\log(T))$, is nearly linear. This is because* `UCB-TP` *is highly contingent on the maximum rotting rate rather than the cumulative rotting rates. We will demonstrate this further in our numerical results.*

Abrupt Rotting with S_T. Here we consider abruptly rotting reward distribution with given $S_T (\le T)$. We consider Algorithm 1 with δ newly tuned by S_T and β. We define $\delta_S(\beta) = (S_T/T)^{1/(\beta+1)}$ for $\beta \ge 1$ and $\delta_S(\beta) = (S_T/T)^{1/2}$ for $0 < \beta \le 1$. Then in the following theorem, we analyze the regret of Algorithm 1 with $\delta_S(\beta)$.

Theorem 3.3. *The policy π of Algorithm 1 with $\delta = \delta_S(\beta)$ achieves:*

$$\mathbb{E}[R^{\pi}(T)] = \begin{cases} \tilde{O}(\max\{S_T^{\frac{1}{\beta+1}} T^{\frac{\beta}{\beta+1}}, V_T\}) & for \;\; \beta \ge 1, \\ \tilde{O}(\max\{\sqrt{S_T T}, V_T\}) & for \;\; 0 < \beta < 1. \end{cases}$$

As in the slow rotting case, for the abrupt rotting scenario, we observe that when β increases above 1, the regret bound worsens as the likelihood of sampling a good arm decreases. However, when β decreases below 1, the regret bound remains the same because we cannot avoid a certain level of regret arising from estimating the mean reward. Also, we observe that for the case of $S_T = 1$, where the problem becomes stationary (implying $V_T = 0$), the regret bound in Theorem 3.1 matches the previously known regret bound of $\tilde{O}(\max\{T^{\beta/(\beta+1)}, \sqrt{T}\})$ for the stationary infinitely many-armed bandits Wang et al. (2009); Bayati et al. (2020). Additionally, we observe that the regret bound is linearly bounded by V_T, which is attributed to the algorithm's necessity to pull a rotted arm at least once to determine its status as bad. Later, in the analysis of regret lower bounds, we will establish the impossibility of avoiding V_T regret in the worst-case.

Proof sketch. The full proof is provided in Appendix A.3. Here we provide a proof outline. We follow the proof framework of Theorem 3.1 but the main difference lies in carefully dealing with substantially rotted arms. For the ease of presentation, we consider each arm that experiences abrupt rotting as if it were newly sampled by the algorithm, treating the arm before and after abrupt rotting as distinct arms. The definition of a good arm and a bad arm is based on the mean reward at the time when it is newly sampled. Then we divide the regret into regret from good and bad arms as $R^\pi(T) = R^{\mathcal{G}}(T) + R^{\mathcal{B}}(T)$. From the definition of good arms, we can easily show that

$$\mathbb{E}[R^{\mathcal{G}}(T)] = O(\delta_S(\beta)T) = \begin{cases} \tilde{O}(S_T^{\frac{1}{\beta+1}} T^{\frac{\beta}{\beta+1}}) & for \;\; \beta \geq 1, \\ \tilde{O}(\sqrt{S_T T}) & for \;\; 0 < \beta < 1. \end{cases}$$

For dealing with $R^{\mathcal{B}}(T)$, we partition the regret into two scenarios: one where the bad arm is initially bad sampled from the distribution of (1) and another where it becomes bad after rotting. This can be expressed as

$$R^{\mathcal{B}}(T) = R^{\mathcal{B},1}(T) + R^{\mathcal{B},2}(T).$$

Then for the former regret, $R^{\mathcal{B},1}(T)$, as in the proof of Theorem 3.1, by using the episodic analysis with $m^{\mathcal{G}} = S_T$, we can show that

$$\mathbb{E}[R^{\mathcal{B},1}(T)] \leq \mathbb{E}[R^{\mathcal{B}}_{m^{\mathcal{G}}}] = \begin{cases} \tilde{O}(S_T^{\frac{1}{\beta+1}} T^{\frac{\beta}{\beta+1}}) & for \;\; \beta \geq 1, \\ \tilde{O}(\sqrt{S_T T}) & for \;\; 0 < \beta < 1. \end{cases}$$

For the regret from rotted bad arms, $R^{\mathcal{B},2}(T)$, it is critical to analyze significant rotting instances to obtain a tight bound with respect to S_T. We analyze that when there exists significant rotting, then the algorithm can detect it as a bad arm and eliminate it by pulling it at once. From this analysis, we can show that

$$\mathbb{E}[R^{\mathcal{B},2}(T)] = \begin{cases} \tilde{O}(\max\{S_T^{\frac{\beta}{\beta+1}} T^{\frac{1}{\beta+1}}, V_T\}) & for \;\; \beta \geq 1, \\ \tilde{O}(\max\{\sqrt{S_T T}, V_T\}) & for \;\; 0 < \beta < 1. \end{cases}$$

Putting all the results together with $\mathbb{E}[R^\pi(T)] = \mathbb{E}[R^{\mathcal{G}}(T)] + \mathbb{E}[R^{\mathcal{B},1}(T)] + \mathbb{E}[R^{\mathcal{B},2}(T)]$ and $S_T \leq T$, we can conclude the proof. $\square$

Remarkably, our proposed method utilizing an adaptive sliding window yields a tight bound (lower bounds will be presented later) not only for slow rotting but also for abrupt rotting scenarios characterized by a limited number of

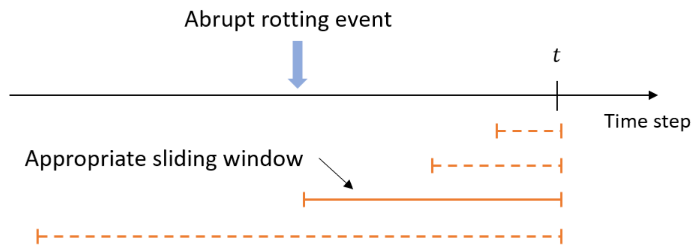

Figure 3: Adaptive sliding window for abrupt rotting.

rotting instances. The rationale behind the effectiveness of the adaptive sliding window in controlling the bias and variance tradeoff with respect to abrupt rotting is as follows. It can be observed that the adaptive threshold condition of $\min_{s\in\mathcal{T}_t(a)} WUCB(a,s,t-1,T) < 1-\delta$ is equivalent to the condition of $WUCB(a,s,t-1,T) < 1-\delta$ for some s such that $t_1(a) \le s \le t-1$ (ignoring the computational reduction trick). The latter expression represents the threshold condition tested for every time step before t, encompassing the time step immediately following an abrupt rotting event. Consequently, as illustrated in Figure 3, this adaptive threshold condition can identify substantially rotted arms by mitigating bias and variance using the window starting from the time step following the occurrence of rotting.

In cases where abrupt rotting occasionally occurs with $S_T = O(\min\{V_T^{(\beta+1)/(\beta+2)}T^{1/(\beta+2)}, V_T^{2/3}T^{1/3}\})$, we can observe that Algorithm 1 with $\delta_S(\beta)$ achieves a tighter regret bound compared to the one with $\delta_V(\beta)$ from Theorems 3.1 and 3.3 as follows. For simplicity, when $\beta = 1$, we have

$$\tilde{O}\left(\max\left\{\sqrt{S_T T}, V_T\right\}\right) \le \tilde{O}\left(\max\left\{V_T^{\frac{1}{3}}T^{\frac{2}{3}}, \sqrt{T}\right\}\right). \tag{12}$$

We will demonstrate this later from our numerical results.

Rotting with V_T and S_T. In what follows, we study the case of rotting with given both V_T and S_T. Then Algorithm 1 with $\delta = \min\{\delta_V(\beta), \delta_S(\beta)\}$ can achieve a tighter regret bound as noted in the following corollary, which can be obtained from Theorems 3.1 and 3.3.

Corollary 3.4. *Let R_V and R_S be defined as*

$$R_V := \begin{cases} \max\{V_T^{\frac{1}{\beta+2}}T^{\frac{\beta+1}{\beta+2}}, T^{\frac{\beta}{\beta+1}}\} & for \ \ \beta \ge 1, \\ \max\{V_T^{1/3}T^{2/3}, \sqrt{T}\} & for \ \ 0<\beta<1 \end{cases} \quad and \ R_S := \begin{cases} \max\{S_T^{\frac{1}{\beta+1}}T^{\frac{\beta}{\beta+1}}, V_T\} & for \ \ \beta \ge 1, \\ \max\{\sqrt{S_T T}, V_T\} & for \ \ 0<\beta<1. \end{cases}$$

The policy π of Algorithm 1 with $\delta = \min\{\delta_V(\beta), \delta_S(\beta)\}$ achieves the following regret bound: $\mathbb{E}[R^\pi(T)] = \tilde{O}\left(\min\left\{R_V, R_S\right\}\right)$.

Case without Prior Knowledge of V_T, S_T, and β. Here we study the case when the algorithm does not have prior information about the values of V_T, S_T, and β. These parameters play a crucial role in determining the optimal threshold parameter δ in Algorithm 1. In cases where these parameter values are unknown, an additional step is necessary to estimate the optimal δ. To address this, we employ the Bandit-over-Bandit (BoB) approach Cheung et al. (2019). This approach involves a master and several bases with a time block size of H, in which each base algorithm represents Algorithm 1 with a candidate value of δ, and the role of the master is to find the best base having near optimal δ. Further details of the algorithm (Algorithm 2) can be found in Appendix A.4.

With a time block size of H (where $H = \lceil \sqrt{T} \rceil$), the algorithm operates over $\lceil T/H \rceil$ blocks. The regret is composed of two factors from the master and bases. Let $V_{H,i}$ represent the cumulative amount of rotting in the time steps within the i-th block. It is possible to encounter $V_{H,i} = V_T$ for some i, potentially resulting in an arm's mean reward having a significantly low negative value, leading to suboptimal behavior by the master. To address this, we introduce the assumption of equally distributed cumulative rotting $V_{H,i}$'s, stated as follows:

Assumption 3.5. *$V_{H,i} \leq H$ for all $i \in [\lceil T/H \rceil]$*

Remark 3.6. *As highlighted in Remark 2.3, this assumption is satisfied when mean rewards consistently stay positive, guaranteeing $0 \leq \mu_t(a_t) \leq 1$ for all $t \in [T]$, which is frequently encountered in real-world applications. Additionally, it is noteworthy that Assumption 3.5 implies Assumption 2.2.*

We require additional assumptions for the regret from the bases. Regret regarding V_T or S_T differs from each base because the behavior of each base policy varies with different candidate values of δ, and the rotting rates are arbitrarily determined. To ensure that the regret bound concerning V_T and S_T remains guaranteed irrespective of the bases chosen, we consider either of the following assumptions.

Assumption 3.7. *Rotting rates ρ_t for all $t > 0$ are determined from an oblivious adversary regardless of the actions. In other words, $\{\rho_t\}_{t \in [T]}$ are determined arbitrarily before starting the game and each ρ_t is independent of the selected arm a_t.*

Assumption 3.8. *Let Π be the set of all feasible policies. Then we consider that V_T and S_T satisfies the worst-case policy bounds regarding rotting rates such that $\max_{\pi \in \Pi} \sum_{t \in [T-1]} \rho_t(a_t^\pi) \leq V_T$ and $1 + \max_{\pi \in \Pi} \sum_{t \in [T-1]} \mathbb{1}(\rho_t(a_t^\pi) \neq 0) \leq S_T$, respectively.*

We provide a regret bound of Algorithm 2 under Assumption 3.5 and either of Assumption 3.7 or Assumption 3.8 in the following.

Theorem 3.9. *Let R'_V and R'_S be defined as*

$$R'_V := \begin{cases} V_T^{\frac{1}{\beta+2}} T^{\frac{\beta+1}{\beta+2}} + T^{\frac{2\beta+1}{2\beta+2}} & for \ \ \beta \geq 1, \\ V_T^{\frac{1}{3}} T^{\frac{2}{3}} + T^{\frac{3}{4}} & for \ \ 0 < \beta < 1 \end{cases} \text{and } R'_S := \begin{cases} \max\{S_T^{\frac{1}{\beta+1}} T^{\frac{\beta}{\beta+1}} + T^{\frac{2\beta+1}{2\beta+2}}, V_T\} & for \ \ \beta \geq 1, \\ \max\{\sqrt{S_T T} + T^{\frac{3}{4}}, V_T\} & for \ \ 0 < \beta < 1. \end{cases}$$

Then, the policy π of Algorithm 2 with $H = \lceil \sqrt{T} \rceil$ achieves the following regret bound:

$$\mathbb{E}[R^\pi(T)] = \tilde{O}(\min\{R'_V, R'_S\})$$

Proof. The proof is provided in Appendix A.5. □

We can observe that there is the additional regret cost of $T^{(2\beta+1)/(2\beta+2)}$ for $\beta \geq 1$ or $T^{3/4}$ for $0 < \beta < 1$ compared to Algorithm 1. This additional cost originates from the additional procedure to learn the optimal value of δ in

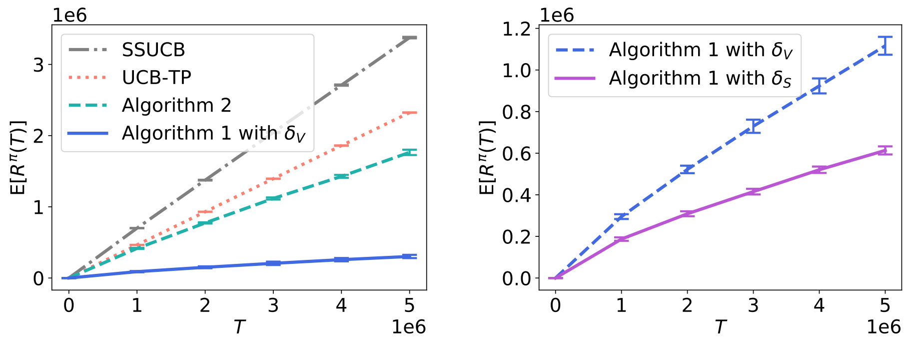

Figure 4: Performance of algorithms demonstrating (left) Eq.(11) and (right) Eq.(12).

Algorithm 2, which is negligible when V_T and S_T are large enough. We also note that the well-known black-box framework proposed for addressing nonstationarity (Wei and Luo, 2021) is not applicable to this problem because the UCB of the chosen arm in this context does not consistently surpass the optimal mean reward, violating the necessary assumption for the framework. Attaining the optimal regret bound under a parameter-free algorithm remains an unresolved issue.

4 Regret Lower Bounds

In this section, we analyze regret lower bounds of our problem under Assumption 2.2 to provide guidance on the tightness of our achieved upper bounds. First, we provide a regret lower bound for slow rotting with V_T in the following.

Theorem 4.1. *Given $V_T \leq T$ and $\beta > 0$, for any policy π, there always exist $\rho_t \geq 0$ for $t \in [T-1]$ such that the regret of π satisfies*

$$\mathbb{E}[R^\pi(T)] = \Omega\left(\max\left\{V_T^{\frac{1}{\beta+2}} T^{\frac{\beta+1}{\beta+2}}, T^{\frac{\beta}{\beta+1}}\right\}\right).$$

Proof. The proof is provided in Appendix A.6. □

Next, we provide a regret lower bound for abrupt rotting with S_T in the following.

Theorem 4.2. *Given S_T and $\beta > 0$, for any policy π, there always exist $\rho_t \geq 0$ for $t \in [T-1]$ with $V_T \leq T$ such that the regret of π satisfies*

$$\mathbb{E}[R^\pi(T)] = \Omega\left(\max\left\{S_T^{\frac{1}{\beta+1}} T^{\frac{\beta}{\beta+1}}, V_T\right\}\right).$$

Proof. The proof is provided in Appendix A.7. □

We note that even for abrupt rotting with S_T, it is unavoidable to incur $\Omega(V_T)$ regret in the worst case because an arm may be rotted V_T amount all at once and an algorithm pulls this rotted arm at least once.

Discussion on Optimal Regret. As we summarize our results in Table 1, Algorithm 1 achieves near-optimal regret only when $\beta \geq 1$. Here, we discuss the discrepancies between lower and upper bounds when $0 < \beta < 1$. From (1), we can observe that as β decreases below 1, the probability to sample good arms may increase, which appears to be beneficial with respect to regret. However, the regret upper bounds for $0 < \beta < 1$ in Theorems 3.1 and 3.3 remain the same as the case when $\beta = 1$ while the regret lower bounds in Theorems 4.1 and 4.2 decrease as β decreases, resulting in a gap between the regret upper and lower bounds. The phenomenon that the regret upper bound remains the same when β decreases has also been observed in previous literature on infinitely many-armed bandits (Carpentier and Valko, 2015; Bayati et al., 2020; Wang et al., 2009). As mentioned in Carpentier and Valko (2015), although there are likely to be many good arms when β is small, it is not possible to avoid a certain amount of regret from estimating mean rewards to distinguish arms under sub-Gaussian reward noise. Therefore, we believe that our regret upper bounds are near-optimal across the entire range of β, and achieving tighter regret lower bounds when $\beta < 1$ is left for future research.

5 Experiments

In this section, we provide empirical validation of our theoretical results using synthetic datasets under uniform distribution for initial mean rewards ($\beta = 1$).

We first compare the performance of our algorithms with previously proposed ones to demonstrate (11). We note that `UCB-TP` Kim et al. (2022) is the state-of-the-art algorithm for the rotting setting and `SSUCB` Bayati et al. (2020) is known to achieve near-optimal regret in stationary infinitely many armed bandits. We set $\rho_t = 1/(t\log(T))$ for all t so that $\rho = \rho_1 = 1/\log(T) = o(1)$ and $V_T = \sum_{t=1}^{T-1}\rho_t = O(1)$. In Figure 4 (left), we can observe that Algorithms 1 and 2 perform better than `UCB-TP` and `SSUCB` (and Algorithm 1 outperforms Algorithm 2), which is in agreement with the insights from our regret analysis for the case of $\beta = 1$. In this case, the regret bounds of Algorithms 1 and 2 are $\tilde{O}(T^{2/3})$ and $\tilde{O}(T^{3/4})$ from Theorems 3.1 and 3.9, respectively, which are tighter than the regret bound of `UCB-TP`, $\tilde{O}(T/\log(T)^{1/3})$.

Now we compare the performance of our algorithms to demonstrate (12). Let $t(s)$ be the time step when the s-th abrupt rotting occurs. Then for the case of abruptly rotting rewards, we set $S_T = T^{1/4}$ and $\rho_{t(s)} = T^{1/2}$ for all $s \in [S_T]$, which implies $V_T = \sum_{s=1}^{S_T}\rho_{t(s)} = T^{3/4}$. We consider that the S_T abrupt rotting events are equally distributed over T. In Figure 4 (right), we can observe that Algorithm 1 with $\delta_S(\beta)$ has better performance than Algorithm 1 with $\delta_V(\beta)$, which observation is consistent with the insight from our regret analysis for the case $\beta = 1$, in which the regret bound of Algorithm 1 with $\delta_S(\beta)$, $\tilde{O}(T^{3/4})$, is tighter than that of one with $\delta_V(\beta)$, $\tilde{O}(T^{11/12})$.

6 Conclusion

We explore the challenges posed by infinitely many-armed bandit problems with rotting rewards, examining slow rotting (V_T) and abrupt rotting (S_T), under the generalized initial mean reward distribution with $\beta > 0$. To address the challenges, we propose an algorithm incorporating an adaptive sliding window, which achieves tight regret bounds.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

Auer, P., Gajane, P., and Ortner, R. (2019). Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, pages 138–158. PMLR.

Auer, P., Ortner, R., and Szepesvári, C. (2007). Improved rates for the stochastic continuum-armed bandit problem. In *International Conference on Computational Learning Theory*, pages 454–468. Springer.

Bayati, M., Hamidi, N., Johari, R., and Khosravi, K. (2020). The unreasonable effectiveness of greedy algorithms in multi-armed bandit with many arms. *arXiv preprint arXiv:2002.10121*.

Berry, D. A., Chen, R. W., Zame, A., Heath, D. C., and Shepp, L. A. (1997). Bandit problems with infinitely many arms. *The Annals of Statistics*, 25(5):2103 – 2116.

Besbes, O., Gur, Y., and Zeevi, A. (2014). Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27.

Bonald, T. and Proutière, A. (2013). Two-target algorithms for infinite-armed bandits with bernoulli rewards. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 2184–2192, Red Hook, NY, USA. Curran Associates Inc.

Brown, D. G. (2011). How I wasted too long finding a concentration inequality for sums of geometric variables. *Found at https://cs. uwaterloo. ca/˜ browndg/negbin. pdf*, 8(4).

Carpentier, A. and Valko, M. (2015). Simple regret for infinitely many armed bandits. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1133–1141. JMLR.org.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. (2019). Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1079–1087. PMLR.

Dylan J. Foster, A. R. (2022). Statistical reinforcement learning and decision making: Course notes. *MIT Lecture notes for course 9.S915 Fall 2022*.

Kim, J.-h., Vojnovic, M., and Yun, S.-Y. (2022). Rotting infinitely many-armed bandits. In *International Conference on Machine Learning*, pages 11229–11254. PMLR.

Kleinberg, R. (2004). Nearly tight bounds for the continuum-armed bandit problem. *Advances in Neural Information Processing Systems*, 17.

Levine, N., Crammer, K., and Mannor, S. (2017). Rotting bandits. *Advances in neural information processing systems*, 30.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670.

Rigollet, P. and Hütter, J.-C. (2015). High dimensional statistics. *Lecture notes for course 18S997*, 813:814.

Russac, Y., Vernade, C., and Cappé, O. (2019). Weighted linear bandits for non-stationary environments. *arXiv preprint arXiv:1909.09146*.

Seznec, J., Locatelli, A., Carpentier, A., Lazaric, A., and Valko, M. (2019). Rotting bandits are no harder than stochastic ones. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2564–2572. PMLR.

Seznec, J., Menard, P., Lazaric, A., and Valko, M. (2020). A single algorithm for both restless and rested rotting bandits. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3784–3794. PMLR.

Tsun, A. (2020). Probability & statistics with applications to computing.

Villar, S. S., Bowden, J., and Wason, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199.

Wang, Y., Audibert, J.-y., and Munos, R. (2009). Algorithms for infinitely many-armed bandits. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Wei, C.-Y. and Luo, H. (2021). Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Conference on learning theory*, pages 4300–4354. PMLR.

A Appendix

A.1 Proof of Proposition 2.1

Recall $\Delta_1(a) = 1 - \mu_1(a)$. We first show that $\mathbb{E}[\mu_1(a)] = \Theta(1)$. For any randomly sampled $a \in \mathcal{A}$, we have $\mathbb{E}[\mu_1(a)] \geq y\mathbb{P}(\mu_1(a) \geq y) = y\mathbb{P}(\Delta_1(a) < 1 - y)$ for $y \in [0, 1]$. With $y = 1/2$, we have $\mathbb{E}[\mu_1(a)] \geq (1/2)\mathbb{P}(\Delta_1(a) < (1/2)) = \Theta(1)$ from constant $\beta > 0$ and (1). Then with $\mathbb{E}[\mu_1(a)] \leq 1$, we can conclude $\mathbb{E}[\mu_1(a)] = \Theta(1)$ (Especially when $\mathbb{P}(\Delta(a) < x) = x^\beta$, we have $\mathbb{E}[\Delta_1(a)] = \int_0^1 \mathbb{P}(\Delta_1(a) \geq x)dx = 1 - \int_0^1 \mathbb{P}(\Delta_1(a) < x)dx = 1 - \int_0^1 x^\beta dx = 1 - \frac{1}{\beta+1}$, which implies $\mathbb{E}[\mu_1(a)] = \Theta(1)$ with constant $\beta > 0$). We then think of a policy π' that randomly samples a new arm and pulls it only once every round. Since $\mathbb{E}[\mu_1(a)] = \Theta(1)$ for any randomly sampled a, we have $\mathbb{E}[R^{\pi'}(T)] = \Theta(T)$.

Next we show that the policy π' is optimal for the worst case of $V_T > T$. We think of any policy π'' except π'. For any policy π'', there always exists an arm a such that the policy must pull arm a at least twice. Let t' and t'' be the rounds when the policy pulls arm a. If we consider $\rho_{t'} = V_T$ then such policy has $\Omega(V_T)$ regret bound. Since

$V_T > T$, any algorithm except π' has $\Omega(T)$ in the worst case. Therefore we can conclude that π' is the optimal algorithm for achieving the optimal regret of $\Theta(T)$.

A.2 Proof of Theorem 3.1: Regret Upper Bound of Algorithm 1 for Slow Rotting with V_T

Let $\Delta_t(a) = 1 - \mu_t(a)$. Using a threshold parameter δ, we classify an arm a as *good* if $\Delta_1(a) \le \delta/2$, *near-good* if $\delta/2 < \Delta_1(a) \le 2\delta$, and otherwise, we classify a as a *bad* arm. In $\mathcal{A}$, let $\bar{a}_1, \bar{a}_2, \dots,$ be a sequence of arms, which have i.i.d. mean rewards with uniform distribution on $[0,1]$. Without loss of generality, we assume that the policy samples arms, which are pulled at least once, according to the sequence of $\bar{a}_1, \bar{a}_2, \dots,$. Let $\mathcal{A}_T$ be the set of sampled arms over the horizon of T time steps, which satisfies $|\mathcal{A}_T| \le T$. Let $\mathcal{A}_T^{\mathcal{G}}$ be a set of good or near good arms in $\mathcal{A}_T$.

Let $\overline{\mu}_{[s_1,s_2]}(a) = \sum_{t=s_1}^{s_2} \mu_t(a)/n_{[s_1,s_2]}(a)$ for the time steps $0 < s_1 \le s_2$. We define event $E_1 = \{|\widehat{\mu}_{[s_1,s_2]}(a) - \overline{\mu}_{[s_1,s_2]}(a)| \le \sqrt{12\log(T)/n_{[s_1,s_2]}(a)}$ for all $1 \le s_1 \le s_2 \le T, a \in \mathcal{A}_T\}$. By following the proof of Lemma 35 in Dylan J. Foster (2022), from Lemma A.21 we have

$$
\begin{aligned}
&P\left(\left|\widehat{\mu}_{[s_1,s_2]}(a) - \overline{\mu}_{[s_1,s_2]}(a)\right| \le \sqrt{\frac{12\log T}{n_{[s_1,s_2]}(a)}}\right) \\
&\le \sum_{n=1}^{T} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \le \sqrt{12\log(T)/n}\right) \\
&\le \frac{2}{T^5}, \qquad (13)
\end{aligned}
$$

where $X_i = r_{\tau_i} - \mu_{\tau_i}(a)$ and τ_i is the i-th time that the policy pulls arm a starting from s_1. We note that even though X_i's seem to depend on each other from τ_i's, each value of X_i is independent of each other. Then using union bound for s_1, s_2, and $a \in \mathcal{A}_T$, we have $\mathbb{P}(E_1^c) \le 2/T^2$. From the cumulative amount of rotting V_T, we note that $\Delta_t(a) = O(V_T + 1)$ for any a and t, which implies $\mathbb{E}[R^\pi(T)|E_1^c] = o(T^2)$ from $V_T \le T$. For the case where E_1 does not hold, the regret is $\mathbb{E}[R^\pi(T)|E_1^c]\mathbb{P}(E_1^c) = O(1)$, which is negligible compared to the regret when E_1 holds, which we show later. Therefore, for the rest of the proof, we assume that E_1 holds.

For regret analysis, we divide $R^\pi(T)$ into two parts, $R^{\mathcal{G}}(T)$ and $R^{\mathcal{B}}(T)$ corresponding to regret of good or near-good arms, and bad arms over time T, respectively, such that $R^\pi(T) = R^{\mathcal{G}}(T) + R^{\mathcal{B}}(T)$. We first provide a bound of $R^{\mathcal{G}}(T)$ in the following lemma.

Lemma A.1. *Under E_1 and policy π, we have*

$$
\mathbb{E}[R^{\mathcal{G}}(T)] = \tilde{O}\left(T\delta + T^{2/3}V_T^{1/3}\right).
$$

Proof. Here we consider arms $a \in \mathcal{A}_T^{\mathcal{G}}$. Let $V_{[n,m]}(a) = \sum_{l=n}^{m} \rho_l(a)$ and $\overline{\rho}_{[n,m]}(a) = \sum_{l=n}^{m} \rho_l(a)/n_{[n,m]}(a)$ for time steps $n \le m$. For ease of presentation, for time steps $r > q$, we define $V_{[r,q]}(a) = n_{[r,q]}(a) = \overline{\rho}_{[r,q]}(a) =$

$\sum_{t=r}^{q} x(t) = 0$ for $x(t) \in \mathbb{R}$ and $1/0 = \infty$. Then, for any s such that $n \le s \le m$, under E_1 we have

$$\begin{aligned}
\widehat{\mu}_{[s,m]}(a) &\le \bar{\mu}_{[s,m]}(a) + \sqrt{12\log(T)/n_{[s,m]}(a)} \\
&\le \mu_m(a) + \sum_{l=s}^{m-1} \rho_l \mathbb{1}(a_l = a) + \sqrt{12\log(T)/n_{[s,m]}(a)} \\
&= \mu_n(a) - \sum_{l=n}^{m-1} \rho_l \mathbb{1}(a_l = a) + \sum_{l=s}^{m-1} \rho_l \mathbb{1}(a_l = a) + \sqrt{12\log(T)/n_{[s,m]}(a)} \\
&\le \mu_n(a) - V_{[n,m-1]}(a) + \overline{\rho}_{[s,m-1]}(a) n_{[s,m]}(a) + \sqrt{12\log(T)/n_{[s,m]}(a)}.
\end{aligned}$$

Therefore, from $\mu_n(a) \le 1$ we obtain

$$\begin{aligned}
&\widehat{\mu}_{[s,m]}(a) + \sqrt{12\log(T)/n_{[s,m]}(a)} \\
&\le 1 - V_{[n,m-1]}(a) + \overline{\rho}_{[s,m-1]}(a) n_{[s,m]}(a) + 2\sqrt{12\log(T)/n_{[s,m]}(a)}.
\end{aligned} \tag{14}$$

Let $t_1(a)$ be the initial time when the arm a is sampled and pulled and $t_2(a)$ be the final time when the policy pulls the arm. For simplicity, we use t_1 and t_2 instead of $t_1(a)$ and $t_2(a)$, respectively, when there is no confusion. We define $\mathcal{A}^0$ as a set of arms $a \in \mathcal{A}_T^{\mathcal{G}}$ such that $t_2(a) = t_1(a)$ and define $\mathcal{A}^1$ as a set of arms $a \in \mathcal{A}_T^{\mathcal{G}}$ such that $t_2(a) = t_1(a)+1$. We also define a set of arms $\overline{\mathcal{A}}_T^{\mathcal{G}} = \{a \in \mathcal{A}_T^{\mathcal{G}}/\{\mathcal{A}^0 \cup \mathcal{A}^1\} : n_{[t_1,t_2-1]}(a) > \lceil (\log T)^{1/3}/\overline{\rho}_{[t_1,t_2-2]}(a)^{2/3} \rceil\}$. Let $w(a) = \lceil (\log T)^{1/3}/\overline{\rho}_{[t_1,t_2-2]}(a)^{2/3} \rceil$. For simplicity, we use w for $w(a)$ when there is no confusion. Then with the fact that $\mu_t(a) = \mu_{t_1}(a) - \sum_{t=t_1(a)}^{t-1} \rho_t(a) = \mu_{t_1}(a) - V_{[t_1,t-1]}(a)$ for $t_1(a) \le t \le t_2(a)$, we have

$$\begin{aligned}
\mathbb{E}[R^{\mathcal{G}}(T)] = \mathbb{E}\left[\sum_{a \in \mathcal{A}_T^{\mathcal{G}}} \sum_{t=t_1(a)}^{t_2(a)} (1 - \mu_t(a))\right] &= \mathbb{E}\left[\sum_{a \in \mathcal{A}_T^{\mathcal{G}}} \left(\Delta_1(a) n_{[t_1,t_2]}(a) + \sum_{t=t_1(a)+1}^{t_2(a)} V_{[t_1,t-1]}(a)\right)\right] \\
&\le \mathbb{E}\left[2T\delta + \sum_{a \in \mathcal{A}^1} \rho_{t_1(a)} + \sum_{a \in \mathcal{A}_T^{\mathcal{G}}/\{\overline{\mathcal{A}}_T^{\mathcal{G}} \cup \mathcal{A}^0 \cup \mathcal{A}^1\}} \sum_{t=t_1(a)+1}^{t_2(a)} V_{[t_1,t-1]}(a)\right. \\
&\quad \left. + \sum_{a \in \overline{\mathcal{A}}_T^{\mathcal{G}}} \left(\sum_{t=t_1(a)+1}^{t_1(a)+w(a)} V_{[t_1,t-1]}(a) + \sum_{t=t_1(a)+w(a)+1}^{t_2(a)} V_{[t_1,t-1]}(a)\right)\right],
\end{aligned} \tag{15}$$

where the first inequality comes from $\Delta_1(a) \le 2\delta$ for any $a \in \mathcal{A}_T^{\mathcal{G}}$. For the second term in the right hand side of the last inequality (15),

$$\sum_{a \in \mathcal{A}^1} \rho_{t_1(a)} \le V_T. \tag{16}$$

For the third term in (15), from the fact that $n_{[t_1+1,t_2]}(a) = n_{[t_1,t_2-1]}(a) < w(a)$ for any $a \in \mathcal{A}_T^{\mathcal{G}}/\overline{\mathcal{A}}_T^{\mathcal{G}}$ from the definition of $\overline{\mathcal{A}}_T^{\mathcal{G}}$, we have

$$\begin{aligned}
&\sum_{a\in\mathcal{A}_T^{\mathcal{G}}/\{\overline{\mathcal{A}}_T^{\mathcal{G}}\cup\mathcal{A}^0\cup\mathcal{A}^1\}}\sum_{t=t_1(a)+1}^{t_2(a)} V_{[t_1,t-1]}(a)\\
&\leq \sum_{a\in\mathcal{A}_T^{\mathcal{G}}/\{\overline{\mathcal{A}}_T^{\mathcal{G}}\cup\mathcal{A}^0\cup\mathcal{A}^1\}} n_{[t_1+1,t_2]}(a)V_{[t_1,t_2-2]}(a)+\rho_{t_2(a)-1}\\
&=O\left(V_T+\sum_{a\in\mathcal{A}_T^{\mathcal{G}}/\{\overline{\mathcal{A}}_T^{\mathcal{G}}\cup\mathcal{A}^0\cup\mathcal{A}^1\}} w(a)V_{[t_1,t_2-2]}(a)\right)\\
&=\tilde{O}\left(V_T+\sum_{a\in\mathcal{A}_T^{\mathcal{G}}/\{\overline{\mathcal{A}}_T^{\mathcal{G}}\cup\mathcal{A}^0\cup\mathcal{A}^1\}} n_{[t_1,t_2-2]}(a)^{2/3}V_{[t_1,t_2-2]}(a)^{1/3}\right).
\end{aligned} \tag{17}$$

Now, we focus on the fourth term in (15). From $t_1(a)+w(a)+1\leq t_2(a)$ for $a\in\overline{\mathcal{A}}_T^{\mathcal{G}}$ from the definition of $\overline{\mathcal{A}}_T^{\mathcal{G}}$ and (17), we first have

$$\begin{aligned}
\sum_{a\in\overline{\mathcal{A}}_T^{\mathcal{G}}}\sum_{t=t_1(a)+1}^{t_1(a)+w(a)} V_{[t_1,t-1]}(a) &= \sum_{a\in\overline{\mathcal{A}}_T^{\mathcal{G}}}\sum_{t=t_1(a)+1}^{t_1(a)+w(a)}\sum_{s=t_1}^{t-1}\rho_s\\
&\leq \sum_{a\in\overline{\mathcal{A}}_T^{\mathcal{G}}}\sum_{t=t_1(a)+1}^{t_1(a)+w(a)}\sum_{s=t_1(a)}^{t_2(a)-2}\rho_s\\
&\leq \sum_{a\in\overline{\mathcal{A}}_T^{\mathcal{G}}} w(a)V_{[t_1,t_2-2]}(a)\\
&=\tilde{O}\left(\sum_{a\in\overline{\mathcal{A}}_T^{\mathcal{G}}} n_{[t_1,t_2-2]}(a)^{2/3}V_{[t_1,t_2-2]}(a)^{1/3}\right).
\end{aligned} \tag{18}$$

Now we focus on $\sum_{a\in\overline{\mathcal{A}}_T^{\mathcal{G}}}\sum_{t=t_1(a)+w(a)+1}^{t_2(a)} V_{[t_1,t-1]}(a)$ in (15). From the definition of t_2 and the threshold condition in the algorithm with (14), for any $t_1\leq t\leq t_2$ and any $t_1\leq s\leq t-1$ s.t. $s=t-2^{l-1}$ for $l\in\mathbb{Z}^+$, we have

$$1-V_{[t_1,t-2]}(a)+n_{[s,t-1]}(a)\overline{\rho}_{[s,t-2]}(a)+2\sqrt{12\log(T)/n_{[s,t-1]}(a)}\geq 1-\delta. \tag{19}$$

For $t\geq t_1+w(a)+1$, there always exists $t_1\leq s(t)\leq t-1$ such that $w(a)/2\leq n_{[s(t),t-1]}(a)\leq w(a)$ and $s(t)=t-2^{l-1}$ for $l\in\mathbb{Z}^+$. Then from (19) with $s=s(t)$, we have

$$V_{[t_1,t-2]}(a)=\tilde{O}\left(\delta+\overline{\rho}_{[s(t),t-2]}(a)/\overline{\rho}_{[t_1,t_2-2]}(a)^{2/3}+\overline{\rho}_{[t_1,t_2-2]}(a)^{1/3}\right). \tag{20}$$

Using the facts that $n_{[s(t),t-2]}(a)\geq n_{[s(t),t-1]}(a)/2\geq w(a)/4$ and $t-s(t)\leq w(a)$ from $n_{[s(t),t-1]}(a)\leq w(a)$, we can obtain that

$$\sum_{t=t_1(a)+1+w(a)}^{t_2(a)} \overline{\rho}_{[s(t),t-2]}(a) \leq \sum_{t=t_1(a)+1+w(a)}^{t_2(a)} \frac{\sum_{k=t-w(a)}^{t-2} \rho_k}{n_{[s(t),t-2]}(a)}$$
$$\leq \sum_{t=t_1(a)+1}^{t_2(a)-2} \frac{w(a)\rho_t}{n_{[s(t),t-2]}(a)}$$
$$\leq 4 \sum_{t=t_1(a)}^{t_2(a)-2} \rho_t, \tag{21}$$

where the second inequality is obtained from the fact that the number of times that ρ_t is duplicated for each $t \in [t_1(a)+1, t_2(a)-2]$ in the expression $\sum_{t=t_1(a)+1+w(a)}^{t_2(a)} \sum_{k=t-w(a)}^{t-2} \rho_k$ is at most $w(a)$. Then with (20) and (21), using the fact that

$$\sum_{t_1(a)+1+w(a)}^{t_2(a)} \overline{\rho}_{[t_1,t_2-2]}(a)^{1/3} \leq n_{[t_1,t_2-2]}(a)\overline{\rho}_{[t_1,t_2-2]}(a)^{1/3} = O(n_{[t_1,t_2-2]}(a)^{2/3} V_{[t_1,t_2-2]}(a)^{1/3}),$$

we have

$$\begin{aligned}
&\sum_{a \in \overline{\mathcal{A}}_T^{\mathcal{G}}} \sum_{t=t_1(a)+1+w(a)}^{t_2(a)} V_{[t_1,t-1]}(a) \\
&\leq \sum_{a \in \overline{\mathcal{A}}_T^{\mathcal{G}}} \sum_{t=t_1(a)+1+w(a)}^{t_2(a)} V_{[t_1,t-2]}(a) + \rho_{t_2(a)-1} \\
&= \tilde{O}\left(\delta T + V_T + \sum_{a \in \overline{\mathcal{A}}_T^{\mathcal{G}}} \sum_{t=t_1(a)+1+w(a)}^{t_2(a)} \overline{\rho}_{[s(t),t-2]}(a)/\overline{\rho}_{[t_1,t_2-2]}(a)^{2/3} + \sum_{a \in \overline{\mathcal{A}}_T^{\mathcal{G}}} \sum_{t=t_1(a)+1+w(a)}^{t_2(a)} \overline{\rho}_{[t_1,t_2-2]}(a)^{1/3} \right) \\
&= \tilde{O}\left(\delta T + V_T + \sum_{a \in \overline{\mathcal{A}}_T^{\mathcal{G}}} \sum_{t=t_1(a)+1+w(a)}^{t_2(a)} \overline{\rho}_{[s(t),t-2]}(a)/\overline{\rho}_{[t_1,t_2-2]}(a)^{2/3} + \sum_{a \in \overline{\mathcal{A}}_T^{\mathcal{G}}} n_{[t_1,t_2-2]}(a)^{2/3} V_{[t_1,t_2-2]}(a)^{1/3} \right) \\
&= \tilde{O}\left(\delta T + V_T + \sum_{a \in \overline{\mathcal{A}}_T^{\mathcal{G}}} \sum_{t=t_1(a)}^{t_2(a)-2} \rho_t/\overline{\rho}_{[t_1,t_2-2]}(a)^{2/3} + \sum_{a \in \overline{\mathcal{A}}_T^{\mathcal{G}}} n_{[t_1,t_2-2]}(a)^{2/3} V_{[t_1,t_2-2]}(a)^{1/3} \right) \\
&= \tilde{O}\left(\delta T + V_T + \sum_{a \in \overline{\mathcal{A}}_T^{\mathcal{G}}} n_{[t_1,t_2-2]}(a)^{2/3} V_{[t_1,t_2-2]}(a)^{1/3} \right).
\end{aligned} \tag{22}$$

Then putting the results from (15),(17),(18), and (22) altogether, we have

$$
\begin{aligned}
&\mathbb{E}[R^{\mathcal{G}}(T)] \\
&\le \mathbb{E}\left[\sum_{a\in\mathcal{A}_T^{\mathcal{G}}}\left(\Delta_1(a)n_{[t_1,t_2]}(a)+\sum_{t=t_1(a)+1}^{t_2(a)}V_{[t_1,t-1]}(a)\right)\right] \\
&= \tilde{O}\left(T\delta+V_T+\sum_{a\in\mathcal{A}_T^{\mathcal{G}}/\{\mathcal{A}^0\cup\mathcal{A}^1\}}V_{[t_1,t_2-2]}(a)^{1/3}n_{[t_1,t_2-2]}(a)^{2/3}\right) \\
&= \tilde{O}\left(T\delta+V_T^{1/3}T^{2/3}\right),
\end{aligned}
\tag{23}
$$

where the last equality comes from Hölder's inequality and $V_T \le T$. This concludes the proof. $\square$

Now, we provide a bound for $R^{\mathcal{B}}(T)$. We note that the initially bad arms can be defined only when $2\delta < 1$. Otherwise when $2\delta \ge 1$, we have $R(T) = R^{\mathcal{G}}(T)$, which completes the proof. Therefore, for the regret from bad arms, we consider the case of $2\delta < 1$. We adopt the episodic approach in Kim et al. (2022) for the remaining regret analysis. The episodic approach is reformulated using the cumulative amount of rotting instead of the maximum rotting rate. In the following, we define some notation.

Given a policy sampling arms in the sequence order, let $m^{\mathcal{G}}$ be the number of samples of distinct good arms and $m_i^{\mathcal{B}}$ be the number of consecutive samples of distinct bad arms between the $i-1$-st and i-th sample of a good arm among $m^{\mathcal{G}}$ good arms. We refer to the period starting from sampling the $i-1$-st good arm before sampling the i-th good arm as the i-th *episode*. Observe that $m_1^{\mathcal{B}}, \dots, m_{m^{\mathcal{G}}}^{\mathcal{B}}$ are i.i.d. random variables with geometric distribution with parameter 2δ, given a fixed value of $m^{\mathcal{G}}$. Therefore, for non-negative integer k we have $\mathbb{P}(m_i^{\mathcal{B}} = k) = (1-2\delta)^k 2\delta$, for $i = 1, \dots, m^{\mathcal{G}}$. Define $\tilde{m}_T$ to be the number of episodes from the policy π over the horizon T, $\tilde{m}_T^{\mathcal{G}}$ to be the total number of samples of a good arm by the policy π over the horizon T such that $\tilde{m}_T^{\mathcal{G}} = \tilde{m}_T$ or $\tilde{m}_T^{\mathcal{G}} = \tilde{m} - 1$, and $\tilde{m}_{i,T}^{\mathcal{B}}$ to be the number of samples of a bad arm in the i-th episode by the policy π over the horizon T.

Under a policy π, let $R_{i,j}^{\mathcal{B}}$ be the regret (summation of mean reward gaps) contributed by pulling the j-th bad arm in the i-th episode. Then let $R_{m^{\mathcal{G}}}^{\mathcal{B}} = \sum_{i=1}^{m^{\mathcal{G}}}\sum_{j\in[m_i^{\mathcal{B}}]} R_{i,j}^{\mathcal{B}}$, which is the regret from initially bad arms over the period of $m^{\mathcal{G}}$ episodes.

Let $a(i)$ be a good arm in the i-th episode and $a(i,j)$ be a j-th bad arm in the i-th episode. We define $V_T(a) = \sum_{t=1}^T \rho_t \mathbb{1}(a_t = a)$. Then excluding the last episode $\tilde{m}_T$ over T, we provide lower bounds of the total rotting variation over T for $a(i)$, denoted by $V_T(a(i))$, in the following lemma.

Lemma A.2. *Under E_1, given $\tilde{m}_T$, for any $i \in [\tilde{m}_T^{\mathcal{G}}]/\{\tilde{m}_T\}$ we have*

$$V_T(a(i)) \ge \delta/2.$$

Proof. Suppose that $V_T(a(i))=\delta/2$, then we have

$$\begin{aligned}
&\min_{t_1(a(i))\le s\le t_2(a(i))}\left\{\widehat{\mu}_{[s,t_2(a(i))]}(a(i))+\sqrt{12\log(T)/n_{[s,t_2(a(i))]}(a(i))}\right\}\\
&\ge \min_{t_1(a(i))\le s\le t_2(a(i))}\{\overline{\mu}_{[s,t_2(a(i))]}(a(i))\}\\
&\ge \mu_{t_2(a(i))}(a(i))\\
&\ge \mu_1(a(i))-V_T(a(i))\\
&\ge 1-\delta,
\end{aligned}$$

where the first inequality is obtained from E_1, and the last inequality is from $V_T(a(i))=\delta/2$ and $\mu_1(a(i))\ge 1-\delta/2$. Therefore, policy π must pull arm $a(i)$ until its total rotting amount is greater than $\delta/2$, which implies $V_T(a(i))\ge\delta/2$. □

In the following, we consider two different cases with respect to V_T; large and small V_T.

Case 1: We consider $V_T=\omega(\max\{1/\sqrt{T},1/T^{1/(\beta+1)}\})$ in the following.

In this case, we have $\delta=\delta_V(\beta)=\max\{(V_T/T)^{1/(\beta+2)},(V_T/T)^{1/3}\}$. Here, we define the policy π after time T such that it pulls a good arm until its total rotting variation is equal to or greater than $\delta/2$ and does not pull a sampled bad arm. We note that defining how π works after T is only for the proof to get a regret bound over time horizon T. For the last arm $\tilde{a}$ over the horizon T, it pulls the arm until its total variation becomes $\max\{\delta/2,V_T(\tilde{a})\}$ if $\tilde{a}$ is a good arm. For $i\in[m^{\mathcal{G}}]$, $j\in[m_i^{\mathcal{B}}]$ let $V_i^{\mathcal{G}}$ and $V_{i,j}^{\mathcal{B}}$ be the total rotting variation of pulling the good arm in i-th episode and j-th bad arm in i-th episode from the policy, respectively. Here we define $V_i^{\mathcal{G}}$'s and $V_{i,j}^{\mathcal{B}}$'s as follows:

If $\tilde{a}$ is a good arm,

$$V_i^{\mathcal{G}}=\begin{cases}V_T(a(i)) & \text{for } i\in[\tilde{m}_T^{\mathcal{G}}-1]\\ \max\{\delta/2,V_T(a(i))\} & \text{for } i\in[m^{\mathcal{G}}]/[\tilde{m}_T^{\mathcal{G}}-1]\end{cases},\ V_{i,j}^{\mathcal{B}}=\begin{cases}V_T(a(i,j)) & \text{for } i\in[\tilde{m}_T^{\mathcal{G}}], j\in[\tilde{m}_{i,T}^{\mathcal{B}}]\\ 0 & \text{for } i\in[m^{\mathcal{G}}]/[\tilde{m}_T^{\mathcal{G}}], j\in[m_i^{\mathcal{B}}].\end{cases}$$

Otherwise,

$$V_i^{\mathcal{G}}=\begin{cases}V_T(a(i)) & \text{for } i\in[\tilde{m}_T^{\mathcal{G}}]\\ \delta/2 & \text{for } i\in[m^{\mathcal{G}}]/[\tilde{m}_T^{\mathcal{G}}]\end{cases},\ V_{i,j}^{\mathcal{B}}=\begin{cases}V_T(a(i,j)) & \text{for } i\in[\tilde{m}_T^{\mathcal{G}}], j\in[\tilde{m}_{i,T}^{\mathcal{B}}]\\ 0 & \text{for } i\in[m^{\mathcal{G}}]/[\tilde{m}_T^{\mathcal{G}}-1], j\in[m_i^{\mathcal{B}}]/[\tilde{m}_{i,T}^{\mathcal{B}}].\end{cases}$$

For $i\in[m^{\mathcal{G}}]$, $j\in[m_i^{\mathcal{B}}]$ let $n_{i,j}^{\mathcal{B}}$ be the number of pulling the j-th bad arm in i-th episode from the policy. We define $n_T(a)$ be the total amount of pulling arm a over T. Here we define $n_{i,j}^{\mathcal{B}}$'s as follows:

$$n_{i,j}^{\mathcal{B}}=\begin{cases}n_T(a(i,j)) & \text{for } i\in[\tilde{m}_T^{\mathcal{G}}], j\in[\tilde{m}_{i,T}^{\mathcal{B}}]\\ 0 & \text{for } i\in[m^{\mathcal{G}}]/[\tilde{m}_T^{\mathcal{G}}], j\in[m_i^{\mathcal{B}}].\end{cases}$$

Then we provide $m^{\mathcal{G}}$ such that $R^{\mathcal{B}}(T)\le R_{m^{\mathcal{G}}}^{\mathcal{B}}$ in the following lemma.

Lemma A.3. *Under E_1, when $m^{\mathcal{G}} = \lceil 2V_T/\delta \rceil$ we have*

$$R^{\mathcal{B}}(T) \leq R^{\mathcal{B}}_{m^{\mathcal{G}}}.$$

Proof. From Lemma A.2, we have

$$\sum_{i \in [m^{\mathcal{G}}]} V_i^{\mathcal{G}} \geq m^{\mathcal{G}} \frac{\delta}{2} \geq V_T,$$

which implies that $R^{\mathcal{B}}(T) \leq R^{\mathcal{B}}_{m^{\mathcal{G}}}$. □

From the result of Lemma A.3, we set $m^{\mathcal{G}} = \lceil 2V_T/\delta \rceil$. We analyze $R^{\mathcal{B}}_{m^{\mathcal{G}}}$ for obtaining a bound for $R^{\mathcal{B}}(T)$ in the following.

Lemma A.4. *Under E_1 and policy π, we have*

$$\mathbb{E}[R^{\mathcal{B}}_{m^{\mathcal{G}}}] = \tilde{O}\left(\max\{T^{(\beta+1)/(\beta+2)} V_T^{1/(\beta+2)}, T^{2/3} V_T^{1/3}\}\right).$$

Proof. Let $a(i,j)$ be a sampled arm for j-th bad arm in the i-th episode and $\tilde{m}_T$ be the number of episodes from the policy π over the horizon T. Suppose that the algorithm samples arm $a(i,j)$ at time $t_1(a(i,j))$. Then the algorithm stops pulling arm $a(i,j)$ at time $t_2(a(i,j))+1$ if $\widehat{\mu}_{[s,t_2(a(i,j))]}(a) + \sqrt{12\log(T)/n_{[s,t_2(a(i,j))]}(a)} < 1-\delta$ for some s such that $t_1(a(i,j)) \leq s \leq t_2(a(i,j))$ and $s = t_2(a(i,j)) + 1 - 2^{l-1}$ for $l \in \mathbb{Z}^+$. For simplicity, we use t_1 and t_2 instead of $t_1(a(i,j))$ and $t_2(a(i,j))$ when there is no confusion. We first consider the case where the algorithm stops pulling arm $a(i,j)$ because the threshold condition is satisfied. For the regret analysis, we consider that for $t > t_2$, arm a is virtually pulled. We note that under E_1, we have

$$\begin{aligned}\widehat{\mu}_{[s,t_2]}(a(i,j)) + \sqrt{12\log(T)/n_{[s,t_2]}(a(i,j))} &\leq \overline{\mu}_{[s,t_2]}(a(i,j)) + 2\sqrt{12\log(T)/n_{[s,t_2]}(a(i,j))} \\ &\leq \mu_1(a(i,j)) + 2\sqrt{12\log(T)/n_{[s,t_2]}(a(i,j))}.\end{aligned}$$

Then we assume that $\tilde{t}_2 (\geq t_2)$ is the smallest time that there exists $t_1 \leq s \leq \tilde{t}_2$ with $s = \tilde{t}_2 + 1 - 2^{l-1}$ for $l \in \mathbb{Z}^+$ such that the following threshold condition is met:

$$\mu_1(a(i,j)) + 2\sqrt{12\log(T)/n_{[s,\tilde{t}_2]}(a(i,j))} < 1-\delta. \tag{24}$$

From the definition of $\tilde{t}_2$, we observe that for given $\tilde{t}_2$, the time step $s = s'$ which satisfying (24) equals to t_1 (i.e. $s' = t_1$). Then, we can observe that $n_{[s',\tilde{t}_2]}(a(i,j)) = n_{[t_1,\tilde{t}_2]}(a(i,j)) = \lceil C_2 \log(T)/(\Delta_{t_1}(a(i,j)) - \delta)^2 \rceil$ for some constant $C_2 > 0$, which satisfies (24). Then from $n_{[t_1,t_2]}(a(i,j)) \leq n_{[t_1,\tilde{t}_2]}(a(i,j))$, for all $i \in [\tilde{m}_T], j \in [\tilde{m}^{\mathcal{B}}_{i,T}]$ we have $n^{\mathcal{B}}_{i,j} = \tilde{O}(1/(\Delta_1(a(i,j)) - \delta)^2)$. Then with the facts that $n^{\mathcal{B}}_{i,j} = 0$ for $i \in [m^{\mathcal{G}}]/[\tilde{m}^{\mathcal{G}}_T]$, $j \in [m^{\mathcal{B}}_i]/[\tilde{m}^{\mathcal{B}}_{i,T}]$, we have, for any $i \in [m^{\mathcal{G}}]$ and $j \in [m^{\mathcal{B}}_i]$,

$$n^{\mathcal{B}}_{i,j} = \tilde{O}(1/(\Delta_{t_1}(a(i,j)) - \delta)^2).$$

For $2\delta < x \leq 1$, let $b(x) = \mathbb{P}(\Delta_1(a) = x | a \text{ is a bad arm})$. Then we have

$$\begin{aligned}b(x) &= \mathbb{P}(\Delta_1(a) = x | \Delta_1(a) > 2\delta) \\ &= \mathbb{P}(\Delta_1(a) = x)/\mathbb{P}(\Delta_1(a) > 2\delta) \\ &= \mathbb{P}(\Delta_1(a) = x)/(1 - C(2\delta)^\beta).\end{aligned}$$

We note that $2\delta < \Delta_{t_1}(a(i,j)) = \Delta_1(a(i,j)) \le 1$. Since $n^{\mathcal{B}}_{i,j} = \tilde{O}(1/(\Delta_{t_1}(a(i,j)) - \delta)^2) = \tilde{O}(1/\delta^2)$, we have

$$
\begin{aligned}
\mathbb{E}[R^{\mathcal{B}}_{i,j}] &= \mathbb{E}\left[\sum_{t=t_1(a(i,j))}^{t_2(a(i,j))} \Delta_{t_1}(a(i,j)) + \sum_{t=t_1(a(i,j))}^{t_2(a(i,j))-1} \sum_{s=t_1(a(i,j))}^{t} \rho_s\right] \\
&\le \mathbb{E}[\Delta_1(a(i,j)) n^{\mathcal{B}}_{i,j} + V^{\mathcal{B}}_{i,j} n^{\mathcal{B}}_{i,j}] \\
&\le \mathbb{E}[\Delta_1(a(i,j)) n^{\mathcal{B}}_{i,j} + V^{\mathcal{B}}_{i,j}(1/\delta^2)] \\
&= \tilde{O}\left(\int_{2\delta}^{1} \frac{1}{(x-\delta)^2} x b(x) dx + \mathbb{E}[V^{\mathcal{B}}_{i,j}(1/\delta^2)]\right).
\end{aligned} \tag{25}
$$

Recall that we consider $2\delta < 1$ for regret from bad arms. We adopt some techniques introduced in Appendix D of Bayati et al. (2020) to deal with the generalized mean reward distribution with β. Let $K = (1-2\delta)/\delta$, $a_j = \frac{2}{j\delta}$, and $p_j = \int_{j\delta}^{(j+1)\delta} b(t+\delta)dt$. Then for obtaining a bound of the last equality in (25) we have

$$
\begin{aligned}
\int_{2\delta}^{1} \left(\frac{1}{(x-\delta)^2} x\right) b(x) dx &= \int_{\delta}^{1-\delta} (\frac{1}{t} + \frac{\delta}{t^2}) b(t+\delta) dt \\
&= \sum_{j=1}^{K} \int_{j\delta}^{(j+1)\delta} (\frac{1}{t} + \frac{\delta}{t^2}) b(t+\delta) dt \\
&\le \sum_{j=1}^{K} \frac{2}{j\delta} \int_{j\delta}^{(j+1)\delta} b(t+\delta) dt \\
&= \sum_{j=1}^{K} a_j p_j
\end{aligned} \tag{26}
$$

We note that $\sum_{i=1}^{j} p_i \le C_0 (j\delta)^\beta$ for all $j \in [K]$ for some constant $C_0 > 0$. Then for getting a bound of the last equality in (26), we have

$$
\begin{aligned}
\sum_{j=1}^{K} a_j p_j &= \sum_{j=1}^{K-1} (a_j - a_{j+1}) \left(\sum_{i=1}^{j} p_i\right) + a_K \sum_{i=1}^{K} p_i \\
&\le \sum_{j=1}^{K-1} (a_j - a_{j+1}) C_0 (j\delta)^\beta + a_K C_0 (K\delta)^\beta \\
&= C_0 \delta^\beta a_1 + \sum_{j=2}^{K} C_0 (j^\beta - (j-1)^\beta) \delta^\beta a_j \\
&= O\left(\left(\frac{1}{\delta}\right)\delta^\beta + \sum_{j=2}^{K} \left(\frac{1}{j\delta}\right)\left((j\delta)^\beta - ((j-1)\delta)^\beta\right)\right) \\
&= O\left(\delta^{\beta-1} + \sum_{j=2}^{K} \left(\frac{1}{j}\delta^{\beta-1}\right)\left(j^\beta - (j-1)^\beta\right)\right).
\end{aligned} \tag{27}
$$

Now we analyze the term in the last equality in (27) according to the criteria for β. For $\beta = 1$, we can obtain

$$
O\left(\delta^{\beta-1} + \sum_{j=2}^{K} \left(\frac{1}{j}\delta^{\beta-1}\right)\left(j^\beta - (j-1)^\beta\right)\right) = \tilde{O}(1). \tag{28}
$$

For $\beta > 1$, we have $j^\beta - (j-1)^\beta \le \beta j^{\beta-1}$ using the mean value theorem. Therefore, we obtain the following.

$$\begin{aligned}
&O\left(\delta^{\beta-1} + \sum_{j=2}^{K}\left(\frac{1}{j}\delta^{\beta-1}\right)\left(j^\beta - (j-1)^\beta\right)\right) = O\left(\sum_{j=1}^{K}\left(\frac{1}{j}\delta^{\beta-1}\right)j^{\beta-1}\right)\\
&= O\left(\sum_{j=2}^{K}\delta^{\beta-1}j^{\beta-2}\right)\\
&= O\left(\delta^{\beta-1}\frac{1}{\beta-1}\left((K+1)^{\beta-1}-1\right)\right)\\
&= O(1).
\end{aligned} \tag{29}$$

For $\beta < 1$, when $j > 1$ we have $j^\beta - (j-1)^\beta \le \beta(j-1)^{\beta-1}$ using the mean value theorem. Therefore, we obtain

$$\begin{aligned}
&O\left(\delta^{\beta-1} + \sum_{j=1}^{K}\left(\frac{1}{j}\delta^{\beta-1}\right)\left(j^\beta - (j-1)^\beta\right)\right) = O\left(\delta^{\beta-1} + \sum_{j=2}^{K}\left(\frac{1}{j}\delta^{\beta-1}\right)(j-1)^{\beta-1}\right)\\
&= O\left(\delta^{\beta-1} + \sum_{j=2}^{K}\delta^{\beta-1}(j-1)^{\beta-2}\right)\\
&= O\left(\delta^{\beta-1} + \delta^{\beta-1}\frac{1}{\beta-1}\left((K+1)^{\beta-1}-1\right)\right)\\
&= O\left(\delta^{\beta-1} + \delta^{\beta-1}\frac{1-((1-\delta)/\delta)^{\beta-1}}{1-\beta}\right) = O(\delta^{\beta-1}).
\end{aligned} \tag{30}$$

From (26),(27),(28),(29), and (30), we have

$$\int_{2\delta}^{1}\left(\frac{1}{(x-\delta)^2}x\right)b(x)dx = \tilde{O}(\max\{1, \delta^{\beta-1}\}).$$

Then for any $i \in [m^{\mathcal{G}}]$, $j \in [m_i^{\mathcal{B}}]$, we have

$$\begin{aligned}
\mathbb{E}[R_{i,j}^{\mathcal{B}}] &\le \mathbb{E}\left[\Delta(a(i,j))n_{i,j}^{\mathcal{B}} + V_{i,j}^{\mathcal{B}}n_{i,j}^{\mathcal{B}}\right]\\
&= \tilde{O}\left(\max\{1, \delta^{\beta-1}\} + \mathbb{E}[V_{i,j}^{\mathcal{B}}]/\delta^2\right).
\end{aligned} \tag{31}$$

Recall that $R_{m^{\mathcal{G}}}^{\mathcal{B}} = \sum_{i=1}^{m^{\mathcal{G}}}\sum_{j\in[m_i^{\mathcal{B}}]} R_{i,j}^{\mathcal{B}}$. With $\delta = \max\{(V_T/T)^{1/(\beta+2)}, (V_T/T)^{1/3}\}$ and $m^{\mathcal{G}} = \lceil 2V_T/\delta\rceil$, from the fact that $m_i^{\mathcal{B}}$'s are i.i.d. random variables with geometric distribution with $\mathbb{E}[m_i^{\mathcal{B}}] = 1/(2\delta)^\beta - 1$, we have

$$\begin{aligned}
\mathbb{E}[R_{m^{\mathcal{G}}}^{\mathcal{B}}] &= O\left(\mathbb{E}\left[\sum_{i=1}^{m^{\mathcal{G}}}\sum_{j\in[m_i^{\mathcal{B}}]} R_{i,j}^{\mathcal{B}}\right]\right)\\
&= \tilde{O}\left((V_T/\delta)\frac{1}{\delta^\beta}\max\{1, \delta^{\beta-1}\} + V_T/\delta^2\right)\\
&= \tilde{O}\left(\max\{T^{(\beta+1)/(\beta+2)}V_T^{1/(\beta+2)}, T^{2/3}V_T^{1/3}\}\right).
\end{aligned} \tag{32}$$

□

From $R^{\pi}(T) = R^{\mathcal{G}}(T) + R^{\mathcal{B}}(T)$ and Lemmas A.1, A.3, A.4, with $\delta = \max\{(V_T/T)^{1/(\beta+2)}, (V_T/T)^{1/3}\}$ we have

$$\mathbb{E}[R^{\pi}(T)] = \tilde{O}\left(\max\{T^{(\beta+1)/(\beta+2)}V_T^{1/(\beta+2)}, T^{2/3}V_T^{1/3}\}\right). \tag{33}$$

Case 2: Now we consider $V_T = O(\max\{1/\sqrt{T}, 1/T^{1/(\beta+1)}\})$ in the following. In this case, we have $\delta = \max\{1/T^{\frac{1}{\beta+1}}, 1/\sqrt{T}\}$. For getting $R^{\mathcal{B}}_{m^{\mathcal{G}}}$, here we define the policy π after time T such that it pulls V_T amount of rotting variation for a good arm and 0 for a bad arm. We note that defining how π works after T is only for the proof to get a regret bound over time horizon T. For the last arm $\tilde{a}$ over the horizon T, it pulls the arm up to V_T amount of rotting variation if $\tilde{a}$ is a good arm. For $i \in [m^{\mathcal{G}}]$, $j \in [m_i^{\mathcal{B}}]$ let $V_i^{\mathcal{G}}$ and $V_{i,j}^{\mathcal{B}}$ be the amount of rotting variation from pulling the good arm in i-th episode and j-th bad arm in i-th episode from the policy, respectively. Here we define $V_i^{\mathcal{G}}$'s and $V_{i,j}^{\mathcal{B}}$'s as follows:

If $\tilde{a}$ is a good arm,

$$V_i^{\mathcal{G}} = \begin{cases} V_T(a(i)) & \text{for } i \in [\tilde{m}_T^{\mathcal{G}} - 1] \\ V_T & \text{for } i \in [m^{\mathcal{G}}]/[\tilde{m}_T^{\mathcal{G}} - 1] \end{cases}, V_{i,j}^{\mathcal{B}} = \begin{cases} V_T(a(i,j)) & \text{for } i \in [\tilde{m}_T^{\mathcal{G}}], j \in [\tilde{m}_{i,T}^{\mathcal{B}}] \\ 0 & \text{for } i \in [m^{\mathcal{G}}]/[\tilde{m}_T^{\mathcal{G}}], j \in [m_i^{\mathcal{B}}]. \end{cases}$$

Otherwise,

$$V_i^{\mathcal{G}} = \begin{cases} V_T(a(i)) & \text{for } i \in [\tilde{m}_T^{\mathcal{G}}] \\ V_T & \text{for } i \in [m^{\mathcal{G}}]/[\tilde{m}_T^{\mathcal{G}}] \end{cases}, V_{i,j}^{\mathcal{B}} = \begin{cases} V_T(a(i,j)) & \text{for } i \in [\tilde{m}_T^{\mathcal{G}}], j \in [\tilde{m}_{i,T}^{\mathcal{B}}] \\ 0 & \text{for } i \in [m^{\mathcal{G}}]/[\tilde{m}_T^{\mathcal{G}} - 1], j \in [m_i^{\mathcal{B}}]/[\tilde{m}_{i,T}^{\mathcal{B}}]. \end{cases}$$

For $i \in [m^{\mathcal{G}}]$, $j \in [m_i^{\mathcal{B}}]$ let $n_{i,j}^{\mathcal{B}}$ be the number of pulling the j-th bad arm in i-th episode from the policy. We define $n_T(a)$ be the total amount of pulling arm a over T. Here we define $n_{i,j}^{\mathcal{B}}$'s as follows:

$$n_{i,j}^{\mathcal{B}} = \begin{cases} n_T(a(i,j)) & \text{for } i \in [\tilde{m}_T^{\mathcal{G}}], j \in [\tilde{m}_{i,T}^{\mathcal{B}}] \\ 0 & \text{for } i \in [m^{\mathcal{G}}]/[\tilde{m}_T^{\mathcal{G}}], j \in [m_i^{\mathcal{B}}]. \end{cases}$$

Then we provide $m^{\mathcal{G}}$ such that $R^{\mathcal{B}}(T) \le R^{\mathcal{B}}_{m^{\mathcal{G}}}$ in the following lemma.

Lemma A.5. *Under E_1, when $m^{\mathcal{G}} = C_3$ for some large enough constant $C_3 > 0$, we have*

$$R^{\mathcal{B}}(T) \le R^{\mathcal{B}}_{m^{\mathcal{G}}}.$$

Proof. From Lemma A.2, under E_1 we can find that $V_i^{\mathcal{G}} \ge \min\{\delta/2, V_T\}$ for $i \in [m^{\mathcal{G}}]$. Then if $m^{\mathcal{G}} = C_3$ with some large enough constant $C_3 > 0$, then with $\delta = \Theta(\max\{1/T^{1/(\beta+1)}, 1/T^{1/2}\})$ and $V_T = O(\max\{1/T^{1/(\beta+1)}, 1/\sqrt{T}\})$, we have

$$\sum_{i \in [m^{\mathcal{G}}]} V_i^{\mathcal{G}} \ge C_3 \min\{\delta/2, V_T\} > V_T,$$

which implies $R^{\mathcal{B}}(T) \le R^{\mathcal{B}}_{m^{\mathcal{G}}}$.

□

We analyze $R^{\mathcal{B}}_{m^{\mathcal{G}}}$ for obtaining a bound for $R^{\mathcal{B}}(T)$ in the following.

Lemma A.6. *Under E_1 and policy π, we have*

$$\mathbb{E}[R^{\mathcal{B}}_{m^{\mathcal{G}}}] = \tilde{O}\left(\max\{T^{\beta/(\beta+1)}, \sqrt{T}\}\right).$$

Proof. From (31), for any $i \in [m^{\mathcal{G}}]$, $j \in [m^{\mathcal{B}}_i]$, we have

$$\begin{aligned}\mathbb{E}[R^{\mathcal{B}}_{i,j}] &\le \mathbb{E}\left[\Delta(a(i,j))n^{\mathcal{B}}_{i,j} + V^{\mathcal{B}}_{i,j}n^{\mathcal{B}}_{i,j}\right] \\ &= \tilde{O}\left(\max\{1, \delta^{\beta-1}\} + \mathbb{E}[V^{\mathcal{B}}_{i,j}]/\delta^2\right).\end{aligned}$$

Recall that $R^{\mathcal{B}}_{m^{\mathcal{G}}} = \sum_{i=1}^{m^{\mathcal{G}}}\sum_{j\in[m^{\mathcal{B}}_i]} R^{\mathcal{B}}_{i,j}$. With $\delta = \max\{(1/T)^{1/(\beta+1)}, 1/T^{1/2}\}$ and $m^{\mathcal{G}} = C_3$, from the fact that $m^{\mathcal{B}}_i$'s are i.i.d. random variables with geometric distribution with $\mathbb{E}[m^{\mathcal{B}}_i] = 1/(2\delta)^\beta - 1$, we have

$$\begin{aligned}\mathbb{E}[R^{\mathcal{B}}_{m^{\mathcal{G}}}] &= O\left(\mathbb{E}\left[\sum_{i=1}^{m^{\mathcal{G}}}\sum_{j\in[m^{\mathcal{B}}_i]} R^{\mathcal{B}}_{i,j}\right]\right) \\ &= \tilde{O}\left(\frac{1}{\delta^\beta}\max\{1, \delta^{\beta-1}\} + V_T/\delta^2\right) \\ &= \tilde{O}\left(\max\{T^{\beta/(\beta+1)}, \sqrt{T}\}\right).\end{aligned}$$

□

From Lemma A.1, with $\delta = \max\{1/T^{\frac{1}{\beta+1}}, 1/\sqrt{T}\}$ we have

$$\mathbb{E}[R^{\mathcal{G}}(T)] = \tilde{O}\left(\max\{T^{\beta/(\beta+1)}, \sqrt{T}\}\right).$$

From $R^\pi(T) = R^{\mathcal{G}}(T) + R^{\mathcal{B}}(T)$ and Lemmas A.1, A.5, A.6 with $\delta = \max\{1/T^{\frac{1}{\beta+1}}, 1/\sqrt{T}\}$ we have

$$\mathbb{E}[R^\pi(T)] = \tilde{O}\left(\max\{T^{\beta/(\beta+1)}, \sqrt{T}\}\right). \tag{34}$$

Conclusion: Overall, from (33) and (34), we have

$$\mathbb{E}[R^\pi(T)] = \tilde{O}\left(\max\{V_T^{1/(\beta+2)}T^{(\beta+1)/(\beta+2)}, V_T^{1/3}T^{2/3}, T^{\beta/(\beta+1)}, \sqrt{T}\}\right).$$

A.3 Proof of Theorem 3.3: Regret Upper Bound of Algorithm 1 for Abrupt Rotting with S_T

Using the threshold parameter δ in the algorithm, we define an arm a as a *good* arm if $\Delta_t(a) \le \delta/2$, a *near-good* arm if $\delta/2 < \Delta_t(a) \le 2\delta$, and otherwise, a is a *bad* arm at time t. For analysis, *we consider abrupt change as sampling a new arm.* In other words, if a sudden change occurs to an arm a by pulling the arm a, then the arm is considered to be two different arms; before and after the change. The type of abruptly rotted arms (good, near-good, or bad) after the change is determined by the current value of rotted mean reward. Without loss of generality, we assume that the policy samples arms, which are pulled at least once, in the sequence of $\bar{a}_1, \bar{a}_2, \dots,$. Let $\mathcal{A}_T$ be the set of sampled arms, which are pulled at least once, over the horizon of T time steps, which satisfies $|\mathcal{A}_T| \le T$. We also define $\mathcal{A}_S$ as a set of arms that have been rotted and pulled at least once, which satisfies $|\mathcal{A}_S| \le S_T$. To better understand the definitions, we provide an example. If an arm a suffers abrupt rotting at first, then the arm a is considered to be a different arm a' after the rotting. If the arm a' again suffers abrupt rotting, then it is considered to be a'' after the rotting. If arms a, a', a'' are pulled at least once, then $\{a, a', a''\} \in \mathcal{A}_T$ and $\{a', a''\} \in \mathcal{A}_S$ but $a \notin \mathcal{A}_S$. If arm a'' is not pulled at least once but a and a' are pulled at least once, then $\{a, a'\} \in \mathcal{A}_T$ and $a' \in \mathcal{A}_S$ but $a'' \notin \mathcal{A}_S$.

Let $\overline{\mu}_{[t_1,t_2]}(a) = \sum_{t=t_1}^{t_2} \mu_t(a)\mathbb{1}(a_t = a)/n_{[t_1,t_2]}(a)$. We define the event $E_1 = \{|\widehat{\mu}_{[s_1,s_2]}(a) - \overline{\mu}_{[s_1,s_2]}(a)| \le \sqrt{12\log(T)/n_{[s_1,s_2]}(a)}$ for all $1 \le s_1 \le s_2 \le T, a \in \mathcal{A}_T\}$. By following the proof of Lemma 35 in Dylan J. Foster (2022), from Lemma A.21 we have

$$\begin{aligned}
&P\left(\left|\widehat{\mu}_{[s_1,s_2]}(a) - \overline{\mu}_{[s_1,s_2]}(a)\right| \le \sqrt{\frac{12\log T}{n_{[s_1,s_2]}(a)}}\right) \\
&\le \sum_{n=1}^{T} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \le \sqrt{12\log(T)/n}\right) \\
&\le \frac{2}{T^5},
\end{aligned} \tag{35}$$

where $X_i = r_{\tau_i} - \mu_{\tau_i}(a)$ and τ_i is the i-th time that the policy pulls arm a starting from s_1. We note that even though X_i's seem to depend on each other from τ_i's, each value of X_i is independent of each other. Then using union bound for s_1, s_2, and $a \in \mathcal{A}_T$, we have

$$\mathbb{P}(E_1^c) \le 2/T^2.$$

Let $t(s)$ be the time when s-th abrupt rotting occurs with $\rho_{t(s)}$ for $s \in [S_T]$. Then we have $\Delta_t(a) = O(1 + \sum_{s=1}^{S_T} \rho_{t(s)}) = O(1 + V_T)$ for any a and t, which implies $\mathbb{E}[R^\pi(T)|E_1^c] = O(T + TV_T)$. For the case that E_1 does not hold, the regret is $\mathbb{E}[R^\pi(T)|E_1^c]\mathbb{P}(E_1^c) = O((1 + V_T)/T)$, which is negligible comparing with the regret when E_1 holds true which we show later. Therefore, in the rest of the proof we assume that E_1 holds true.

Recall that $R^\pi(T) = \sum_{t=1}^{T}(1 - \mu_t(a_t))$. For regret analysis, we divide $R^\pi(T)$ into two parts, $R^{\mathcal{G}}(T)$ and $R^{\mathcal{B}}(T)$ corresponding to regret of good or near-good arms, and bad arms over time T, respectively, such that $R^\pi(T) = R^{\mathcal{G}}(T) + R^{\mathcal{B}}(T)$. Recall that we consider abrupt change as sampling a new arm in this analysis. Then, from $\Delta_t(a) \le 2\delta$ for any good or near-good arms a at time t, we can easily obtain that

$$\mathbb{E}[R^{\mathcal{G}}(T)] = O(\delta T) = O(\max\{S_T^{1/(\beta+1)} T^{\beta/(\beta+1)}, \sqrt{S_T T}\}). \tag{36}$$

Now we analyze $R^{\mathcal{B}}(T)$. We divide regret $R^{\mathcal{B}}(T)$ into two regret from bad arms in $\mathcal{A}_T/\mathcal{A}_S$, denoted by $R^{\mathcal{B},1}(T)$, and regret from bad arms in $\mathcal{A}_S$, denoted by $R^{\mathcal{B},2}(T)$ such that $R^{\mathcal{B}}(T) = R^{\mathcal{B},1}(T) + R^{\mathcal{B},2}(T)$. We denote bad arms in $\mathcal{A}_S$ by $\mathcal{A}_S^{\mathcal{B}}$. We first analyze $R^{\mathcal{B},1}(T)$ in the following. For regret analysis, we adopt the episodic approach suggested in Kim et al. (2022). The main difference lies in analyzing our adaptive window UCB and a more generalized mean-reward distribution with β. In the following, we introduce some notation. *Here we only consider* *arms in* $\mathcal{A}_T/\mathcal{A}_S$ so that the following notation is defined without considering (rotted) arms in $\mathcal{A}_S$. We note that from the definition of $\mathcal{A}_T$, arms a before having undergone rotting are contained in $\mathcal{A}_T/\mathcal{A}_S$. Here we consider the case of $2\delta_S(\beta) < 1$ since otherwise when $2\delta_S(\beta) \geq 1$, bad arms are not defined in $\mathcal{A}_T/\mathcal{A}_S$.

Given a policy sampling arms in the sequence order, let $m^{\mathcal{G}}$ be the number of samples of distinct good arms and $m_i^{\mathcal{B}}$ be the number of consecutive samples of distinct bad arms between the $i-1$-st and i-th sample of a good arm among $m^{\mathcal{G}}$ good arms. We refer to the period starting from sampling the $i-1$-st good arm before sampling the i-th good arm as the i-th *episode*. Observe that $m_1^{\mathcal{B}}, \dots, m_{m^{\mathcal{G}}}^{\mathcal{B}}$ are i.i.d. random variables with geometric distribution with parameter 2δ, given a fixed value of $m^{\mathcal{G}}$. Therefore, for non-negative integer k we have $\mathbb{P}(m_i^{\mathcal{B}} = k) = (1-2\delta)^k 2\delta$, for $i = 1, \dots, m^{\mathcal{G}}$.

Define $\tilde{m}_T^{\mathcal{G}}$ to be the total number of samples of a good arm by the policy π over the horizon T and $\tilde{m}_{i,T}^{\mathcal{B}}$ to be the number of samples of a bad arm in the i-th episode by the policy π over the horizon T. For $i \in [\tilde{m}_T^{\mathcal{G}}]$, $j \in [\tilde{m}_{i,T}^{\mathcal{B}}]$, let $\tilde{n}_i^{\mathcal{G}}$ be the number of pulls of the good arm in the i-th episode and $\tilde{n}_{i,j}^{\mathcal{B}}$ be the number of pulls of the j-th bad arm in the i-th episode by the policy π over the horizon T. Let $\tilde{a}$ be the last sampled arm over time horizon T by π.

With a slight abuse of notation, we use π for a modified strategy after T. Under a policy π, let $R_{i,j}^{\mathcal{B}}$ be the regret (summation of mean reward gaps) contributed by pulling the j-th bad arm in the i-th episode. Then let $R_{m^{\mathcal{G}}}^{\mathcal{B}} = \sum_{i=1}^{m^{\mathcal{G}}} \sum_{j \in [m_i^{\mathcal{B}}]} R_{i,j}^{\mathcal{B}}$, which is the regret from initially bad arms over the period of $m^{\mathcal{G}}$ episodes. For getting $R_{m^{\mathcal{G}}}^{\mathcal{B}}$, here we define the policy π after T such that it pulls T amounts for a good arm and zero for a bad arm. After T we can assume that there are no abrupt changes. For the last arm $\tilde{a}$ over the horizon T, it pulls the arm up to T amounts if $\tilde{a}$ is a good arm and $\tilde{n}_{\tilde{m}_T^{\mathcal{G}}}^{\mathcal{G}} < T$. For $i \in [m^{\mathcal{G}}]$, $j \in [m_i^{\mathcal{B}}]$ let $n_i^{\mathcal{G}}$ and $n_{i,j}^{\mathcal{B}}$ be the number of pulling the good arm in i-th episode and j-th bad arm in i-th episode under π, respectively. Here we define $n_i^{\mathcal{G}}$'s and $n_{i,j}^{\mathcal{B}}$'s as follows:

If $\tilde{a}$ is a good arm,

$$n_i^{\mathcal{G}} = \begin{cases} \tilde{n}_i^{\mathcal{G}} & \text{for } i \in [\tilde{m}_T^{\mathcal{G}} - 1] \\ T & \text{for } i = \tilde{m}_T^{\mathcal{G}} \\ 0 & \text{for } i \in [m^{\mathcal{G}}]/[\tilde{m}_T^{\mathcal{G}}] \end{cases}, n_{i,j}^{\mathcal{B}} = \begin{cases} \tilde{n}_{i,j}^{\mathcal{B}} & \text{for } i \in [\tilde{m}_T^{\mathcal{G}}], j \in [\tilde{m}_{i,T}^{\mathcal{B}}] \\ 0 & \text{for } i \in [m^{\mathcal{G}}]/[\tilde{m}_T^{\mathcal{G}}], j \in [m_i^{\mathcal{B}}]/[\tilde{m}_{i,T}^{\mathcal{B}}]. \end{cases}$$

Otherwise,

$$n_i^{\mathcal{G}} = \begin{cases} \tilde{n}_i^{\mathcal{G}} & \text{for } i \in [\tilde{m}_T^{\mathcal{G}}] \\ T & \text{for } i = \tilde{m}_T^{\mathcal{G}} + 1 \\ 0 & \text{for } i \in [m^{\mathcal{G}}]/[\tilde{m}_T^{\mathcal{G}} + 1] \end{cases}, n_{i,j}^{\mathcal{B}} = \begin{cases} \tilde{n}_{i,j}^{\mathcal{B}} & \text{for } i \in [\tilde{m}_T^{\mathcal{G}}], j \in [\tilde{m}_{i,T}^{\mathcal{B}}] \\ 0 & \text{for } i \in [m^{\mathcal{G}}]/[\tilde{m}_T^{\mathcal{G}} - 1], j \in [m_i^{\mathcal{B}}]/[\tilde{m}_{i,T}^{\mathcal{B}}]. \end{cases}$$

Using the above notation and newly defined π after T, we show that if $m^{\mathcal{G}} = S_T + 1$, then $R^{\mathcal{B}}(T) \leq R_{m^{\mathcal{G}}}^{\mathcal{B}}$ in the following.

Lemma A.7. *Under E_1, when $m^{\mathcal{G}} = S_T$ we have*

$$R^{\mathcal{B},1}(T) \leq R_{m^{\mathcal{G}}}^{\mathcal{B}}.$$

Proof. There are $S_T - 1$ number of abrupt changes over T. We consider two cases; there are S_T abrupt changes before sampling S_T-th good arm or there are not. For the former case, if π samples the S_T-th good arm and there are $S_T - 1$ number of abrupt changes before sampling the good arm, then it continues to pull the good arm until T. This is because when the algorithm samples a good arm a at time t', from E_1 and the stationary period, we have

$$\widehat{\mu}_{[t',t]}(a) + \sqrt{12\log(T)/n_{[t',t]}(a)} \geq \mu_{t'}(a) \geq 1 - \delta.$$

This implies that from the threshold condition, the algorithm does not stop pulling the good arm a. After T, from the definition of π for the case when $\tilde{a}$ is a good arm, $n^{\mathcal{G}}_{\tilde{m}^{\mathcal{G}}_T} = T$. Therefore, the algorithm pulls the good arm for T rounds.

Now we consider the latter case, such that π samples the S_T-th good arm before the $S_T - 1$-st abrupt change over T. Before sampling the S_T-th good arm, there must exist two consecutive good arms such that there is no abrupt change between the two sampled good arms. This is a contraction because π must pull the first good arm among the two up to T under E_1 and $S_T - 1$-st abrupt change must occur after T.

Therefore, it is enough to consider the former case. When $m^{\mathcal{G}} = S_T$, we have

$$\sum_{i\in[m^{\mathcal{G}}]} n^{\mathcal{G}}_i \geq T,$$

which implies $R^{\mathcal{B},1}(T) \leq R^{\mathcal{B}}_{m^{\mathcal{G}}}$. □

From the above lemma, we set $m^{\mathcal{G}} = S_T$. We analyze $R^{\mathcal{B}}_{m^{\mathcal{G}}}$ to get a bound for $R^{\mathcal{B},1}(T)$ in the following lemma.

Lemma A.8. *Under E_1 and policy π, we have*

$$\mathbb{E}\left[R^{\mathcal{B}}_{m^{\mathcal{G}}}\right] = \tilde{O}\left(\max\{S_T^{1/(\beta+1)}T^{\beta/(\beta+1)}, \sqrt{S_T T}\}\right).$$

Proof. Recall that we consider arms in $\mathcal{A}_T/\mathcal{A}_S$. Let $a(i,j)$ be a sampled arm for j-th bad arm in the i-th episode and $\tilde{m}_T$ be the number of episodes from the policy π over the horizon T. Suppose that the algorithm samples arm $a(i,j)$ at time $t_1(a(i,j))$. Then the algorithm stops pulling arm $a(i,j)$ at time $t_2(a(i,j)) + 1$ if $\widehat{\mu}_{[s,t_2(a(i,j))]}(a) + \sqrt{12\log(T)/n_{[s,t_2(a(i,j))]}(a)} < 1 - \delta$ for some s such that $t_1(a(i,j)) \leq s \leq t_2(a(i,j))$ and $s = t_2(a(i,j)) + 1 - 2^{l-1}$ for $l \in \mathbb{Z}^+$. For simplicity, we use t_1 and t_2 instead of $t_1(a(i,j))$ and $t_2(a(i,j))$ when there is no confusion. For the regret analysis, we consider that for $t > t_2$, arm a is virtually pulled. With E_1, we assume that $\tilde{t}_2(\geq t_2)$ is the smallest time that there exists $t_1 \leq s \leq \tilde{t}_2$ with $s = \tilde{t}_2 + 1 - 2^{l-1}$ for $l \in \mathbb{Z}^+$ such that the following condition is met:

$$\mu_{t_1}(a(i,j)) + 2\sqrt{12\log(T)/n_{[s,\tilde{t}_2]}(a(i,j))} < 1 - \delta. \tag{37}$$

From the definition of $\tilde{t}_2$, we observe that for given $\tilde{t}_2$, the time step $s = s'$ satisfying (37) equals to t_1 (i.e. $s' = t_1$). Then, we can observe that $n_{[s',\tilde{t}_2]}(a(i,j)) = n_{[t_1,\tilde{t}_2]}(a(i,j)) = \lceil C_2\log(T)/(\Delta_{t_1}(a(i,j)) - \delta)^2\rceil$ for some constant $C_2 > 0$, which satisfies (37). Then from $n_{[t_1,t_2]}(a(i,j)) \leq n_{[t_1,\tilde{t}_2]}(a(i,j))$, for all $i \in [\tilde{m}_T], j \in [\tilde{m}^{\mathcal{B}}_{i,T}]$ we have $n^{\mathcal{B}}_{i,j} = \tilde{O}(1/(\Delta_{t_1}(a(i,j)) - \delta)^2)$. We note that this bound for the number of pulling an arm holds for not only the case where the arm stops being pulled from the threshold condition but also the case where the arm stops being pulled from meeting an abrupt change (recall that abrupt changes are considered as sampling a new arm) or T. Then with the facts that $n^{\mathcal{B}}_{i,j} = 0$ for $i \in [m^{\mathcal{G}}]/[\tilde{m}_T]$, $j \in [m^{\mathcal{B}}_i]/[\tilde{m}^{\mathcal{B}}_{i,T}]$, we have, for any $i \in [m^{\mathcal{G}}]$ and $j \in [m^{\mathcal{B}}_i]$,

$$n^{\mathcal{B}}_{i,j} = \tilde{O}(1/(\Delta_{t_1}(a(i,j)) - \delta)^2).$$

For $2\delta < x \leq 1$, let $b(x) = \mathbb{P}(\Delta_{t_1}(a) = x | a \text{ is a bad arm})$. Then we have $\mathbb{P}(\Delta_{t_1}(a) = x | a \text{ is a bad arm}) = \mathbb{P}(\Delta_{t_1}(a) = x | \Delta_{t_1}(a) > 2\delta) = \mathbb{P}(\Delta_{t_1}(a) = x)/\mathbb{P}(\Delta_1(a) > 2\delta) = \mathbb{P}(\Delta_{t_1}(a) = x)/(1 - C(2\delta)^{\beta}) = O(\mathbb{P}(\Delta_{t_1}(a) = x))$, where the last equality comes from small δ with $S_T = o(T)$. For any $i \in [m^{\mathcal{G}}]$, $j \in [m^{\mathcal{B}}_i]$, we have

$$\begin{aligned}\mathbb{E}[R^{\mathcal{B}}_{i,j}] &\leq \mathbb{E}\left[\Delta_{t_1(a(i,j))}(a(i,j)) n^{\mathcal{B}}_{i,j}\right] \\ &= \tilde{O}\left(\int_{2\delta}^{1} \frac{1}{(x-\delta)^2} x b(x) dx\right). \end{aligned} \tag{38}$$

From the above results in (38),(26),(27),(28),(29),(30), for $\beta > 0$ we have

$$\mathbb{E}[R^{\mathcal{B}}_{i,j}] = \tilde{O}(\max\{1, \delta^{\beta-1}\}).$$

Recall that $R^{\mathcal{B}}_{m^{\mathcal{G}}} = \sum_{i=1}^{m^{\mathcal{G}}} \sum_{j \in [m^{\mathcal{B}}_i]} R^{\mathcal{B}}_{i,j}$. With $\delta = \max\{(S_T/T)^{1/(\beta+1)}, (S_T/T)^{1/2}\}$ and $m^{\mathcal{G}} = S_T$, from Lemma A.7 and the fact that $m^{\mathcal{B}}_i$'s are i.i.d. random variables following geometric distribution with $\mathbb{E}[m^{\mathcal{B}}_i] = 1/(2\delta)^{\beta} - 1$, we have

$$\begin{aligned}\mathbb{E}[R^{\mathcal{B}}_{m^{\mathcal{G}}}] &= O\left(\mathbb{E}\left[\sum_{i=1}^{m^{\mathcal{G}}} \sum_{j \in [m^{\mathcal{B}}_i]} R^{\mathcal{B}}_{i,j}\right]\right) \\ &= \tilde{O}\left(S_T \frac{1}{\delta^{\beta}} \max\{1, \delta^{\beta-1}\}\right) \\ &= \tilde{O}\left(\max\{S_T^{1/(\beta+1)} T^{\beta/(\beta+1)}, \sqrt{S_T T}\}\right). \end{aligned}$$

□

From Lemma A.8, we have $\mathbb{E}[R^{\mathcal{B},1}(T)] = \mathbb{E}[R^{\mathcal{B}}_{m^{\mathcal{G}}}] = \tilde{O}\left(\max\{S_T^{1/(\beta+1)} T^{\beta/(\beta+1)}, \sqrt{S_T T}\}\right)$.

Now we analyze $R^{\mathcal{B},2}(T)$ in the following lemma. Here, we consider arms in $\mathcal{A}^{\mathcal{B}}_S$, which is allowed to have negative mean rewards.

Lemma A.9. *Under E_1 and policy π, we have*

$$\mathbb{E}\left[R^{\mathcal{B},2}(T)\right] = \tilde{O}\left(\max\{S_T/\delta, V_T\}\right).$$

Proof. Recall that we consider arms $a \in \mathcal{A}^{\mathcal{B}}_S$ so that $\Delta_{t_1}(a) > 2\delta$ from definition. Suppose that the arm a is sampled and pulled for the first time at time $t_1(a)$. Then the algorithm stops pulling arm a at time $t_2(a) + 1$ if $\widehat{\mu}_{[s,t_2(a)]}(a) + \sqrt{12\log(T)/n_{[s,t_2(a)]}(a)} < 1 - \delta$ for some s such that $s \leq t_2(a)$ and $s = t_2(a) + 1 - 2^{l-1}$ for $l \in \mathbb{Z}^+$. For simplicity, we use t_1 and t_2 instead of $t_1(a)$ and $t_2(a)$ when there is no confusion. For regret analysis, we consider that for $t > t_2$, arm a is virtually pulled. With E_1, we assume that $\tilde{t}_2(\geq t_2)$ is the smallest time that there exists $t_1 \leq s \leq \tilde{t}_2$ with $s = \tilde{t}_2 + 1 - 2^{l-1}$ for $l \in \mathbb{Z}^+$ such that the following condition is met:

$$\mu_{t_1}(a) + 2\sqrt{12\log(T)/n_{[s,\tilde{t}_2]}(a)} < 1 - \delta. \tag{39}$$

From the definition of $\tilde{t}_2$, we observe that for given $\tilde{t}_2$, the time step s, which satisfies (39), equals to t_1. Then, we can observe that $n_{[t_1,\tilde{t}_2]}(a) = \max\{\lceil C_2 \log(T)/(\Delta_{t_1}(a) - \delta)^2\rceil, 1\}$ for some constant $C_2 > 0$, which satisfies (39). From the above, for any $a \in \mathcal{A}_S^{\mathcal{B}}$ satifying $\Delta_{t_1}(a) \geq \sqrt{C_2 \log(T)} + \delta$, we have $n_{[t_1,\tilde{t}_2]}(a) = 1$. This implies that after pulling the arm a once, the arm is eliminated and after that, the arm is not pulled anymore. Therefore, for any arm a' which was rotted to a, we have $\Delta_{t_1(a')}(a') < \sqrt{C_2 \log(T)} + \delta$. This is because otherwise such that $\Delta_{t_1(a')}(a') \geq \sqrt{C_2 \log(T)} + \delta$, the arm a' is eliminated and a cannot be pulled which means $a \notin \mathcal{A}_S^{\mathcal{B}}$, which is a contradiction. Then for any arm $a \in \mathcal{A}_S^{\mathcal{B}}$, we have $\Delta_{t_1}(a) \leq \sqrt{C_2 \log(T)} + \delta + \rho_{t_1(a)-1}$. Recall that we consider abrupt rotting of an arm as sampling a new arm. Let $t(s)$ be the time step when the s-th abrupt rotting occurs. Then we note that $\rho_{t_1(a)-1} = \rho_{t(s)}$ when arm a is a sampled arm from s-th abrupt rotting for $s \in [S_T]$.

From $n_{[t_1,t_2]}(a) \leq n_{[t_1,\tilde{t}_2]}(a)$, we have $n_{[t_1,t_2]}(a) = \tilde{O}(\max\{1/(\Delta_{t_1}(a) - \delta)^2, 1\})$. We note that this bound for number of pulling an arm holds for not only the case where the arm stops to be pulled from the threshold condition, but also the case where the arm stops to be pulled from meeting an abrupt change (recall that abrupt changes are considered as sampling a new arm) or T. From the definition of bad arms, we have $\Delta_{t_1}(a) \geq 2\delta$. Then the regret from arm a, denoted by $R(a)$, is bounded as follows: $R(a) = \Delta_{t_1}(a) n_{[t_1,t_2]}(a) = \tilde{O}(\max\{\Delta_{t_1}(a)/(\Delta_{t_1}(a) - \delta)^2, \Delta_{t_1}(a)\})$. Since $x/(x-\delta)^2 \leq 2/\delta$ for any $x \geq 2\delta$, we have $R(a) = \tilde{O}(\max\{1/\delta, \Delta_{t_1}(a)\})$. Therefore, with the fact that $\Delta_{t_1}(a) \leq \sqrt{C_2 \log(T)} + \delta + \rho_{t(s)}$ for the corresponding $s \in [S_T]$ such that $\rho_{t_1(a)-1} = \rho_{t(s)}$, we have

$$\begin{aligned}
\sum_{a \in \mathcal{A}_S^{\mathcal{B}}} R(a) &= \tilde{O}\left(\max\left\{S_T/\delta, \sum_{a \in \mathcal{A}_S^{\mathcal{B}}} \Delta_{t_1}(a)\right\}\right) \\
&= \tilde{O}(\max\{S_T/\delta, S_T + \sum_{s=1}^{S_T} \rho_{t(s)}\}) \\
&= \tilde{O}(\max\{S_T/\delta, \sum_{s=1}^{S_T} \rho_{t(s)}\}) \\
&= \tilde{O}(\max\{S_T/\delta, V_T\}),
\end{aligned}$$

where the second last equality comes from $S_T/\delta \geq S_T$. $\square$

Finally, from $R^\pi(T) = R^{\mathcal{G}}(T) + R^{\mathcal{B}}(T)$, (36), and Lemmas A.8, A.9, we have

$$\mathbb{E}[R^\pi(T)] = \tilde{O}\left(\max\{S_T^{1/(\beta+1)} T^{\beta/(\beta+1)}, \sqrt{S_T T}, V_T\}\right).$$

A.4 Pseudo-code of the Adaptive Algorithm for Unknown Parameters

The parameters of β, V_T, and S_T are used to set the optimal threshold parameter δ in Algorithm 1. Therefore, when the parameters are not given, the procedure to find the optimal value δ is required. We adopt the Bandit-over-Bandit (`BoB`) approach in Cheung et al. (2019); Kim et al. (2022) by additionally considering adaptive window. In

Algorithm 2 Adaptive UCB-Threshold with Adaptive Sliding Window

Given: $T, H, \mathcal{B}, \mathcal{A}, \alpha, \kappa, C$

Initialize: $\mathcal{A}' \leftarrow \mathcal{A}, w(\delta') \leftarrow 1$ for $\delta' \in \mathcal{B}$

for $i = 1, 2, \ldots, \lceil T/H \rceil$ **do**

  $t' \leftarrow (i-1)H + 1$

  Select an arm $a \in \mathcal{A}'$

  Pull arm a and get reward $r_{(i-1)H+1}$

  $p(\delta') \leftarrow (1-\alpha)\frac{w(\delta')}{\sum_{k \in \mathcal{B}} w(k)} + \alpha\frac{1}{B}$ for $\delta' \in \mathcal{B}$

  Select $\delta \leftarrow \delta'$ with probability $p(\delta')$ for $\delta' \in \mathcal{B}$

  for $t = (i-1)H + 2, \ldots, i \cdot H \wedge T$ **do**

    if $\min_{s \in \mathcal{T}_t(a)} WUCB(a, s, t-1, H) < 1 - \delta$ **then**

      $\mathcal{A}' \leftarrow \mathcal{A}'/\{a\}$

      Select an arm $a \in \mathcal{A}'$

      Pull arm a and get reward r_t

      $t' \leftarrow t$

    else

      Pull arm a and get reward r_t

    end if

  end for

  $w(\delta) \leftarrow w(\delta)\exp\left(\frac{\alpha}{Bp(\delta)}\left(\frac{1}{2} + \frac{\sum_{t=(i-1)H}^{i \cdot H \wedge T} r_t}{CH\log(H) + 4\sqrt{H\log T}}\right)\right)$

end for

Algorithm 2, the algorithm consists of a master and several base algorithms with $\mathcal{B}$. For the master, we use EXP3 (Auer et al., 2002) to find a nearly best base in $\mathcal{B}$. Each base represents Algorithm 1 with a candidate threshold $\delta' \in \mathcal{B}$. The algorithm divides the time horizon into several blocks of length H. At each block, the algorithm samples a base in $\mathcal{B}$ from the EXP3 strategy and runs the base over the time steps of the block. Using the feedback from the block, the algorithm updates EXP3 and samples a new base for the next block. By block time passes, the master is likely to find an optimized δ in $\mathcal{B}$. Let $B = |\mathcal{B}|$. Then for Algorithm 2, we set $\alpha = \min\{1, \sqrt{B\log B/((e-1)\lceil T/H \rceil)}\}$ and $C > 0$ to be a large enough constant.

We define $\delta_V^\dagger = \max\{(V_T/T)^{1/(\beta+2)}, (V_T/T)^{1/3}, 1/H^{1/(\beta+1)}, 1/\sqrt{H}\}$ and $\delta_S^\dagger = \max\{(S_T/T)^{1/(\beta+1)}, (S_T/T)^{1/2}, 1/H^{1/(\beta+1)}, 1/\sqrt{H}\}$. Then the optimized threshold parameter is $\delta_{VS}^\dagger = \min\{\delta_S^\dagger, \delta_V^\dagger\}$. The optimized threshold parameter can be derived from the theoretical analysis in Appendix A.5. The target of the master is to find the parameter. From the above, we can observe that $1/\sqrt{H} \leq \delta_{VS}^\dagger \leq 1$. Therefore, we set $\mathcal{B} = \{1/2, \ldots, 1/2^{\log_2 \sqrt{H}}\}$ which is the candidate values for unknown $\delta^\dagger$.

A.5 Proof of Theorem 3.9: Regret Upper Bound of Algorithm 2

In the following, we deal with the cases of (a) $\delta_V^\dagger \leq \delta_S^\dagger$ so that $\delta_{VS}^\dagger = \delta_V^\dagger$ and (b) $\delta_V^\dagger > \delta_S^\dagger$ so that $\delta_{VS}^\dagger = \delta_S^\dagger$, separately.

A.5.1 Case of $\delta_V^\dagger \leq \delta_S^\dagger$

Let $\pi_i(\delta')$ for $\delta' \in \mathcal{B}$ denote the base policy for time steps between $(i-1)H+1$ and $i \cdot H \wedge T$ in Algorithm 2 using $1-\delta'$ as a threshold. Denote by $a_t^{\pi_i(\delta')}$ the pulled arm at time step t by policy $\pi_i(\delta')$. Then, for $\delta^\dagger \in \mathcal{B}$, which is set later for a near-optimal policy, we have

$$\mathbb{E}[R^\pi(T)] = \mathbb{E}\left[\sum_{t=1}^{T} 1 - \sum_{i=1}^{\lceil T/H \rceil} \sum_{t=(i-1)H+1}^{i \cdot H \wedge T} \mu_t(a_t^\pi)\right] = \mathbb{E}[R_1^\pi(T)] + \mathbb{E}[R_2^\pi(T)]. \tag{40}$$

where

$$R_1^\pi(T) = \sum_{t=1}^{T} 1 - \sum_{i=1}^{\lceil T/H \rceil} \sum_{t=(i-1)H+1}^{i \cdot H \wedge T} \mu_t(a_t^{\pi_i(\delta^\dagger)})$$

and

$$R_2^\pi(T) = \sum_{i=1}^{\lceil T/H \rceil} \sum_{t=(i-1)H+1}^{i \cdot H \wedge T} \mu_t(a_t^{\pi_i(\delta^\dagger)}) - \sum_{i=1}^{\lceil T/H \rceil} \sum_{t=(i-1)H+1}^{i \cdot H \wedge T} \mu_t(a_t^\pi).$$

Note that $R_1^\pi(T)$ accounts for the regret caused by the near-optimal base algorithm $\pi_i(\delta^\dagger)$'s against the optimal mean reward and $R_2^\pi(T)$ accounts for the regret caused by the master algorithm by selecting a base with $\delta \in \mathcal{B}$ at every block against the base with $\delta^\dagger$. In what follows, we provide upper bounds for each regret component. We first provide an upper bound for $\mathbb{E}[R_1^\pi(T)]$ by following the proof steps in Theorem 3.1. Then we provide an upper bound for $\mathbb{E}[R_2^\pi(T)]$. We set $H = \lceil T^{1/2} \rceil$ and $\delta^\dagger$ to be a smallest value in $\mathcal{B}$ which is larger than $\delta_V^\dagger = \max\{(V_T/T)^{1/(\beta+2)}, (V_T/T)^{1/3}, 1/H^{1/(\beta+1)}, 1/H^{1/2}\}$.

Upper Bounding $\mathbb{E}[R_1^\pi(T)]$. We refer to the period starting from time step $(i-1)H+1$ to time step $i \cdot H \wedge T$ as the i-th *block*. For any $i \in \lceil (T/H) - 1 \rceil$, policy $\pi_i(\delta^\dagger)$ runs over H time steps independent to other blocks so that each block has the same expected regret and the last block has a smaller or equal expected regret than other blocks. Therefore, we focus on finding a bound on the regret from the first block equal to $\sum_{t=1}^{H} 1 - \mu_t(a_t^{\pi_1(\delta^\dagger)})$. We define an arm a as a *good* arm if $\Delta(a) \leq \delta^\dagger/2$, a *near-good* arm if $\delta^\dagger/2 < \Delta(a) \leq 2\delta^\dagger$, and otherwise, a is a *bad* arm. In $\mathcal{A}$, let $\bar{a}_1, \bar{a}_2, \dots,$ be a sequence of arms, which have i.i.d. mean rewards following (1). Without loss of generality, we assume that the policy samples arms in the sequence of $\bar{a}_1, \bar{a}_2, \dots,$.

Denote by $\mathcal{A}(i)$ the set of sampled arms in the i-th block, which satisfies $|\mathcal{A}(i)| \leq H$. Let $\overline{\mu}_{[t_1,t_2]}(a) = \sum_{t=t_1}^{t_2} \mu_t(a)/n_{[t_1,t_2]}(a)$. We define the event $E_1 = \{|\widehat{\mu}_{[s_1,s_2]}(a) - \overline{\mu}_{[s_1,s_2]}(a)| \leq \sqrt{12\log(H)/n_{[s_1,s_2]}(a)}$ for all $1 \leq s_1 \leq s_2 \leq H, a \in \mathcal{A}(i)\}$. As in (13), we have

$$\mathbb{P}(E_1^c) \leq 2/H^2.$$

We denote by $V_{H,i}$ the cumulative amount of rotting in the time steps in the i-th block. From the cumulative amount of rotting, we note that $\Delta_t(a) = O(V_{H,i}+1)$ for any a and t in i-th block, which implies $\mathbb{E}[R^\pi(T)|E_1^c] = O(H^2)$ from $V_{H,i} \leq H$ under Assumption 3.5. For the case where E_1 does not hold, the regret is $\mathbb{E}[R^\pi(T)|E_1^c]\mathbb{P}(E_1^c) = O(1)$, which is negligible compared to the regret when E_1 holds, which we show later. For the case that E_1 does not hold, the regret is $\mathbb{E}[R^\pi(H)|E_1^c]\mathbb{P}(E_1^c) = O(1)$, which is negligible compared with the regret when E_1 holds true which we show later. Therefore, in the rest of the proof we assume that E_1 holds true.

In the following, we first provide a regret bound over the first block.

For regret analysis, we divide $R^{\pi_1(\delta^\dagger)}(H)$ into two parts, $R^{\mathcal{G}}(H)$ and $R^{\mathcal{B}}(H)$ corresponding to regret of good or near-good arms, and bad arms over time H, respectively, such that $R^{\pi_1(\delta^\dagger)}(H) = R^{\mathcal{G}}(H) + R^{\mathcal{B}}(H)$. We denote by $V_{H,i}$ the cumulative amount of rotting in the time steps in the i-th block. We first provide a bound of $R^{\mathcal{G}}(H)$ in the following lemma.

Lemma A.10. *Under E_1 and policy π, we have*

$$\mathbb{E}[R^{\mathcal{G}}(H)] = \tilde{O}\left(H\delta^\dagger + H^{2/3}V_{H,1}^{1/3}\right).$$

Proof. We can easily prove the theorem by following the proof steps in Lemma A.1 □

Now, we provide a regret bound for $R^{\mathcal{B}}(H)$. We note that the initially bad arms can be defined only when $2\delta^\dagger < 1$. Otherwise when $2\delta^\dagger \geq 1$, we have $R(T) = R^{\mathcal{G}}(T)$, which completes the proof. Therefore, for the regret from bad arms, we consider the case of $2\delta^\dagger < 1$. For the proof, we adopt the episodic approach in Kim et al. (2022) for regret analysis.

Given a policy sampling arms in the sequence order, let $m^{\mathcal{G}}$ be the number of samples of distinct good arms and $m_i^{\mathcal{B}}$ be the number of consecutive samples of distinct bad arms between the $i-1$-st and i-th sample of a good arm among $m^{\mathcal{G}}$ good arms. We refer to the period starting from sampling the $i-1$-st good arm before sampling the i-th good arm as the i-th *episode*. Observe that $m_1^{\mathcal{B}}, \dots, m_{m^{\mathcal{G}}}^{\mathcal{B}}$ are i.i.d. random variables with geometric distribution with parameter 2δ, given a fixed value of $m^{\mathcal{G}}$. Therefore, for non-negative integer k we have $\mathbb{P}(m_i^{\mathcal{B}} = k) = (1-2\delta^\dagger)^k 2\delta^\dagger$, for $i = 1, \dots, m^{\mathcal{G}}$. Define $\tilde{m}_H$ to be the number of episodes from the policy π over the horizon H, $\tilde{m}_H^{\mathcal{G}}$ to be the total number of samples of a good arm by the policy π over the horizon H such that $\tilde{m}_H^{\mathcal{G}} = \tilde{m}_H$ or $\tilde{m}_H^{\mathcal{G}} = \tilde{m} - 1$, and $\tilde{m}_{i,H}^{\mathcal{B}}$ to be the number of samples of a bad arm in the i-th episode by the policy $\pi_1(\delta^\dagger)$ over the horizon H.

Under a policy $\pi_1(\delta^\dagger)$, let $R_{i,j}^{\mathcal{B}}$ be the regret (summation of mean reward gaps) contributed by pulling the j-th bad arm in the i-th episode. Then let $R_{m^{\mathcal{G}}}^{\mathcal{B}} = \sum_{i=1}^{m^{\mathcal{G}}} \sum_{j \in [m_i^{\mathcal{B}}]} R_{i,j}^{\mathcal{B}}$, which is the regret from initially bad arms over the period of $m^{\mathcal{G}}$ episodes.

For obtaining a regret bound, we first focus on finding a required number of episodes, $m^{\mathcal{G}}$, such that $R^{\mathcal{B}}(T) \leq R_{m^{\mathcal{G}}}^{\mathcal{B}}$. Then we provide regret bounds for each bad arm and good arm in an episode. Lastly, we obtain a regret bound for $\mathbb{E}[R^{\mathcal{B}}(T)]$ using the episodic regret bound.

Let $a(i)$ be a good arm in the i-th episode and $a(i,j)$ be a j-th bad arm in the i-th episode. We define $V_H(a) = \sum_{t=1}^{H} \rho_t \mathbb{1}(a_t = a)$. Then excluding the last episode $\tilde{m}_H$ over H, we provide lower bounds of the total rotting variation over H for $a(i)$, denoted by $V_H(a(i))$, in the following lemma.

Lemma A.11. *Under E_1, given $\tilde{m}_H$, for any $i \in [\tilde{m}_H^{\mathcal{G}}]/\{\tilde{m}_H\}$ we have*

$$V_H(a(i)) \geq \delta^\dagger/2.$$

Proof. We can easily prove the theorem by following the proof steps in Lemma A.2 □

We first consider the case where $V_T = \omega(\max\{T/H^{3/2}, T/H^{(\beta+2)/(\beta+1)}\})$. In this case, we have $\delta^\dagger = \Theta(\max\{(V_T/T)^{1/(\beta+2)}, (V_T/T)^{1/3}\})$. Here, we define the policy π after time H such that it pulls a good arm until its total rotting variation is equal to or greater than $\delta^\dagger/2$ and does not pull a sampled bad arm. We note that defining how π works after H is only for the proof to get a regret bound over time horizon H. For the last arm $\tilde{a}$ over the horizon H, it pulls the arm until its total variation becomes $\max\{\delta^\dagger/2, V_H(\tilde{a})\}$ if $\tilde{a}$ is a good arm. For $i \in [m^{\mathcal{G}}]$, $j \in [m_i^{\mathcal{B}}]$ let $V_i^{\mathcal{G}}$ and $V_{i,j}^{\mathcal{B}}$ be the total rotting variation of pulling the good arm in i-th episode and j-th bad arm in i-th episode from the policy, respectively. Here we define $V_i^{\mathcal{G}}$'s and $V_{i,j}^{\mathcal{B}}$'s as follows:

If $\tilde{a}$ is a good arm,

$$V_i^{\mathcal{G}} = \begin{cases} V_H(a(i)) & \text{for } i \in [\tilde{m}_H^{\mathcal{G}} - 1] \\ \max\{\delta^\dagger/2, V_H(a(i))\} & \text{for } i \in [m^{\mathcal{G}}]/[\tilde{m}_H^{\mathcal{G}} - 1] \end{cases}, V_{i,j}^{\mathcal{B}} = \begin{cases} V_H(a(i,j)) & \text{for } i \in [\tilde{m}_H^{\mathcal{G}}], j \in [\tilde{m}_{i,H}^{\mathcal{B}}] \\ 0 & \text{for } i \in [m^{\mathcal{G}}]/[\tilde{m}_H^{\mathcal{G}}], j \in [m_i^{\mathcal{B}}]. \end{cases}$$

Otherwise,

$$V_i^{\mathcal{G}} = \begin{cases} V_H(a(i)) & \text{for } i \in [\tilde{m}_H^{\mathcal{G}}] \\ \delta^\dagger/2 & \text{for } i \in [m^{\mathcal{G}}]/[\tilde{m}_H^{\mathcal{G}}] \end{cases}, V_{i,j}^{\mathcal{B}} = \begin{cases} V_H(a(i,j)) & \text{for } i \in [\tilde{m}_H^{\mathcal{G}}], j \in [\tilde{m}_{i,H}^{\mathcal{B}}] \\ 0 & \text{for } i \in [m^{\mathcal{G}}]/[\tilde{m}_H^{\mathcal{G}} - 1], j \in [m_i^{\mathcal{B}}]/[\tilde{m}_{i,H}^{\mathcal{B}}]. \end{cases}$$

For $i \in [m^{\mathcal{G}}]$, $j \in [m_i^{\mathcal{B}}]$ let $n_{i,j}^{\mathcal{B}}$ be the number of pulling the good arm in i-th episode and j-th bad arm in i-th episode from the policy, respectively. We define $n_H(a)$ be the total amount of pulling arm a over H. Here we define $n_{i,j}^{\mathcal{B}}$'s as follows:

$$n_{i,j}^{\mathcal{B}} = \begin{cases} n_H(a(i,j)) & \text{for } i \in [\tilde{m}_H^{\mathcal{G}}], j \in [\tilde{m}_{i,H}^{\mathcal{B}}] \\ 0 & \text{for } i \in [m^{\mathcal{G}}]/[\tilde{m}_H^{\mathcal{G}}], j \in [m_i^{\mathcal{B}}]. \end{cases}$$

Then we provide $m^{\mathcal{G}}$ such that $R^{\mathcal{B}}(H) \le R_{m^{\mathcal{G}}}^{\mathcal{B}}$ in the following lemma.

Lemma A.12. *Under E_1, when $m^{\mathcal{G}} = \lceil 2V_{H,1}/\delta^\dagger \rceil$ we have*

$$R^{\mathcal{B}}(H) \le R_{m^{\mathcal{G}}}^{\mathcal{B}}.$$

Proof. We can easily show the theorem by following the proof steps of Lemma A.3 □

From the result of Lemma A.12, we set $m^{\mathcal{G}} = \lceil 2V_{H,1}/\delta^\dagger \rceil$. In the following, we anlayze $R_{m^{\mathcal{G}}}^{\mathcal{B}}$ for obtaining a regret bound for $R^{\mathcal{B}}(H)$.

Lemma A.13. *Under E_1 and policy π, we have*

$$\mathbb{E}[R_{m^{\mathcal{G}}}^{\mathcal{B}}] = \tilde{O}\left(\max\{V_{H,1}(T/V_T)^{(\beta+1)/(\beta+2)} + (T/V_T)^{\beta/(\beta+2)}, V_{H,1}(T/V_T)^{2/3} + (T/V_T)^{1/3}\}\right).$$

Proof. We can easily prove the theorem by following proof steps in Lemma A.4. From (31), for any $i \in [m^{\mathcal{G}}]$, $j \in [m_i^{\mathcal{B}}]$, we have

$$\begin{aligned}\mathbb{E}[R_{i,j}^{\mathcal{B}}] &\le \mathbb{E}\left[\Delta(a(i,j))n_{i,j}^{\mathcal{B}} + V_{i,j}^{\mathcal{B}} n_{i,j}^{\mathcal{B}}\right] \\ &= \tilde{O}\left(\max\{1, (\delta^\dagger)^{\beta-1}\} + \mathbb{E}[V_{i,j}^{\mathcal{B}}]/(\delta^\dagger)^2\right).\end{aligned}$$

Recall that $R^{\mathcal{B}}_{m^{\mathcal{G}}} = \sum_{i=1}^{m^{\mathcal{G}}}\sum_{j\in[m_i^{\mathcal{B}}]} R^{\mathcal{B}}_{i,j}$. With $\delta^\dagger = \max\{(V_T/T)^{1/(\beta+2)}, (V_T/T)^{1/3}\}$ and $m^{\mathcal{G}} = \lceil 2V_{H,1}/\delta^\dagger\rceil$, from the fact that $m_i^{\mathcal{B}}$'s are i.i.d. random variables with geometric distribution with $\mathbb{E}[m_i^{\mathcal{B}}] = 1/(2\delta^\dagger)^\beta - 1$, we have

$$\begin{aligned}
\mathbb{E}[R^{\mathcal{B}}_{m^{\mathcal{G}}}] &= O\left(\mathbb{E}\left[\sum_{i=1}^{m^{\mathcal{G}}}\sum_{j\in[m_i^{\mathcal{B}}]} R^{\mathcal{B}}_{i,j}\right]\right)\\
&= \tilde{O}\left((V_{H,1}/\delta^\dagger + 1)\frac{1}{(\delta^\dagger)^\beta}\max\{1, (\delta^\dagger)^{\beta-1}\} + V_{H,1}/(\delta^\dagger)^2\right)\\
&= \tilde{O}\left(\max\left\{\frac{V_{H,1}}{(\delta^\dagger)^{\beta+1}}, \frac{V_{H,1}}{(\delta^\dagger)^2}\right\} + \max\left\{\frac{1}{(\delta^\dagger)^\beta}, \frac{1}{\delta^\dagger}\right\}\right)\\
&= \tilde{O}\left(\max\{V_{H,1}(T/V_T)^{(\beta+1)/(\beta+2)} + (T/V_T)^{\beta/(\beta+2)}, V_{H,1}(T/V_T)^{2/3} + (T/V_T)^{1/3}\}\right).
\end{aligned}$$

□

From $R^{\pi_1(\delta^\dagger)}(H) = R^{\mathcal{G}}(H) + R^{\mathcal{B}}(H)$ and Lemmas A.10, A.12, A.13, with $\delta^\dagger = \max\{(V_T/T)^{1/(\beta+2)}, (V_T/T)^{1/3}\}$ we have

$$\begin{aligned}
&\mathbb{E}[R^{\pi_1(\delta^\dagger)}(H)]\\
&= \tilde{O}\Big(\max\Big\{V_{H,1}(T/V_T)^{(\beta+1)/(\beta+2)} + H(V_T/T)^{1/(\beta+2)} + (T/V_T)^{\beta/(\beta+2)},\\
&\qquad V_{H,1}(T/V_T)^{2/3} + H(V_T/T)^{1/3} + (T/V_T)^{1/3}\Big\} + H^{2/3}V_{H,1}^{1/3}\Big).
\end{aligned}$$

The above regret bound is for the first block. Therefore, by summing regrets from $\lceil T/H\rceil$ number of blocks, from $V_T = \omega(\max\{T/H^{(\beta+2)/(\beta+1)}, T/H^{3/2}\})$ and $H = \lceil T^{1/2}\rceil$, using Hölder's inequality we have shown that

$$\begin{aligned}
\mathbb{E}[R_1^\pi(T)] &= \tilde{O}\left(\max\{T^{(\beta+1)/(\beta+2)}V_T^{1/(\beta+2)}, T^{2/3}V_T^{1/3}\} + \frac{T}{H}\max\{(T/V_T)^{\beta/(\beta+2)}, (T/V_T)^{1/3}\}\right)\\
&= \tilde{O}\left(\max\{T^{(\beta+1)/(\beta+2)}V_T^{1/(\beta+2)}, T^{2/3}V_T^{1/3}\} + \max\{T^{(2\beta+1)/(2\beta+2)}, T^{3/4}\}\right). \qquad (41)
\end{aligned}$$

Now, we consider the case where $V_T = O(\max\{T/H^{3/2}, T/H^{(\beta+2)/(\beta+1)}\})$. In this case, we have $\delta^\dagger = \Theta(\max\{1/\sqrt{H}, 1/H^{\frac{1}{\beta+1}}\})$. From the result of Lemma A.12, by setting $m^{\mathcal{G}} = \lceil 2V_{H,1}/\delta^\dagger\rceil$ we have $R^{\mathcal{B}}(H) \le R^{\mathcal{B}}_{m^{\mathcal{G}}}$.

Lemma A.14. *Under E_1 and policy π, we have*

$$\mathbb{E}[R^{\mathcal{B}}_{m^{\mathcal{G}}}] = \tilde{O}\left(\max\{V_{H,1}(T/V_T)^{(\beta+1)/(\beta+2)} + (T/V_T)^{\beta/(\beta+2)}, V_{H,1}(T/V_T)^{2/3} + (T/V_T)^{1/3}\}\right).$$

Proof. We can easily prove the theorem by following proof steps in Lemma A.4. From (31), for any $i\in[m^{\mathcal{G}}]$, $j\in[m_i^{\mathcal{B}}]$, we have

$$\begin{aligned}
\mathbb{E}[R^{\mathcal{B}}_{i,j}] &\le \mathbb{E}\left[\Delta(a(i,j))n^{\mathcal{B}}_{i,j} + V^{\mathcal{B}}_{i,j}n^{\mathcal{B}}_{i,j}\right]\\
&= \tilde{O}\left(\max\{1, \delta^{\beta-1}\} + \mathbb{E}[V^{\mathcal{B}}_{i,j}]/\delta^2\right).
\end{aligned}$$

Recall that $R^{\mathcal{B}}_{m^{\mathcal{G}}} = \sum_{i=1}^{m^{\mathcal{G}}} \sum_{j\in[m_i^{\mathcal{B}}]} R^{\mathcal{B}}_{i,j}$. With $\delta^\dagger = \max\{1/H^{1/2}, 1/H^{1/(\beta+1)}\}$ and $m^{\mathcal{G}} = \lceil 2V_{H,1}/\delta^\dagger \rceil$, from the fact that $m_i^{\mathcal{B}}$'s are i.i.d. random variables with geometric distribution with $\mathbb{E}[m_i^{\mathcal{B}}] = 1/(2\delta^\dagger)^\beta - 1$, we have

$$
\begin{aligned}
\mathbb{E}[R^{\mathcal{B}}_{m^{\mathcal{G}}}] &= O\left(\mathbb{E}\left[\sum_{i=1}^{m^{\mathcal{G}}} \sum_{j\in[m_i^{\mathcal{B}}]} R^{\mathcal{B}}_{i,j}\right]\right) \\
&= \tilde{O}\left((V_{H,1}/\delta^\dagger + 1)\frac{1}{(\delta^\dagger)^\beta}\max\{1, (\delta^\dagger)^{\beta-1}\} + V_{H,1}/(\delta^\dagger)^2\right) \\
&= \tilde{O}\left(\max\left\{\frac{V_{H,1}}{(\delta^\dagger)^{\beta+1}}, \frac{V_{H,1}}{(\delta^\dagger)^2}\right\} + \max\left\{\frac{1}{(\delta^\dagger)^\beta}, \frac{1}{\delta^\dagger}\right\}\right) \\
&= \tilde{O}\left(V_{H,1}H + \max\{H^{\beta/(\beta+1)}, H^{1/2}\}\right).
\end{aligned}
$$

□

From $R^{\pi_1(\delta^\dagger)}(H) = R^{\mathcal{G}}(H) + R^{\mathcal{B}}(H)$ and Lemmas A.10, A.12, A.14, with $\delta^\dagger = \Theta(\max\{1/H^{1/2}, 1/H^{1/(\beta+1)}\})$ we have

$$
\mathbb{E}[R^{\pi_1(\delta^\dagger)}(H)] = \tilde{O}\left(\max\{H^{\beta/(\beta+1)}, H^{1/2}\} + H^{2/3}V_{H,1}^{1/3} + V_{H,1}H\right).
$$

Therefore, by summing regrets from $\lceil T/H \rceil$ number of blocks and from $V_T = O(\max\{T/H^{3/2}, T/H^{(\beta+2)/(\beta+1)}\})$, $H = \lceil T^{1/2} \rceil$, and the fact that length of time steps in each block is bounded by H, we have

$$
\begin{aligned}
\mathbb{E}[R_1^\pi(T)] &= \tilde{O}\left(\frac{T}{H}\max\{H^{\beta/(\beta+1)}, H^{1/2}\} + \sum_{i=1}^{\lceil T/H \rceil} H^{2/3}V_{H,i}^{1/3} + \sum_{i=1}^{\lceil T/H \rceil} V_{H,i}H\right) \\
&= \tilde{O}\left(\frac{T}{H}\max\{H^{\beta/(\beta+1)}, H^{1/2}\} + T^{2/3}V_T^{1/3} + V_T H\right) \\
&= \tilde{O}\left(\max\{T/H^{1/(\beta+1)}, T/H^{1/2}\}\right) \\
&= \tilde{O}\left(\max\{T^{(2\beta+1)/(2\beta+2)}, T^{3/4}\}\right), \qquad (42)
\end{aligned}
$$

where the second equality comes from Hölder's inequality.

From (41) and (42), we have

$$
\mathbb{E}[R_1^\pi(T)] = \tilde{O}(\max\{T^{(\beta+1)/(\beta+2)}V_T^{1/(\beta+2)} + T^{(2\beta+1)/(2\beta+2)}, T^{2/3}V_T^{1/3} + T^{3/4}\}). \qquad (43)
$$

Upper Bounding $\mathbb{E}[R_2^\pi(T)]$. We observe that the EXP3 is run for $\lceil T/H \rceil$ decision rounds and the number of policies (i.e. $\pi_i(\delta')$ for $\delta' \in \mathcal{B}$) is B. Denote the maximum absolute sum of rewards of any block with length H by a random variable Q'. We first provide a bound for Q' using concentration inequalities. For any block i, we have

$$
\left|\sum_{t=(i-1)H+1}^{i\cdot H\wedge T} \mu_t(a_t^\pi) + \eta_t\right| \le \left|\sum_{t=(i-1)H+1}^{i\cdot H\wedge T} \mu_t(a_t^\pi)\right| + \left|\sum_{t=(i-1)H+1}^{i\cdot H\wedge T} \eta_t\right|. \qquad (44)
$$

Denote by $\mathcal{T}_i$ the set of time steps in the i-th block. We define the event $E_2(i) = \{|\widehat{\mu}_{[s_1,s_2]}(a) - \overline{\mu}_{[s_1,s_2]}(a)| \le \sqrt{14\log(H)/n_{[s_1,s_2]}(a)},$ for all $s_1, s_2 \in \mathcal{T}_i, s_1 \le s_2, a \in \mathcal{A}(i)\}$ and $E_2 = \bigcap_{i\in[\lceil T/H \rceil]} E_2(i)$. From Lemma A.21, with $H = \lceil \sqrt{T} \rceil$ we have

$$\mathbb{P}(E_2^c) \leq \sum_{i \in [\lceil T/H \rceil]} \frac{2H^3}{H^6} \leq \frac{2}{T}.$$

By assuming that E_2 holds true, we can get a lower bound for $\mu_t(a_t^\pi)$, which may be a negative value from rotting, for getting an upper bound for $|\sum_{t=(i-1)H+1}^{i \cdot H \wedge T} \mu_t(a_t^\pi)|$. We can observe that $\sum_{t=(i-1)H+1}^{i \cdot H \wedge T} \mu_t(a_t^\pi) \leq H$. Therefore the remaining part is to get a lower bound for $\sum_{t=(i-1)H+1}^{i \cdot H \wedge T} \mu_t(a_t^\pi)$. For the proof simplicity, we consider that when an arm is rotted, then the arm is considered as a different arm after rotting. For instance, when arm a is rotted at time s, then arm a is considered as a different arm a' after s. Therefore, each arm can be considered to be stationary. The set of arms is denoted by $\mathcal{L}$. We denote by $\mathcal{L}^+$ the set of arms having $\mu_t(a) \geq 0$ for $a \in \mathcal{L}$. We first focus on the arms in $\mathcal{L}/\mathcal{L}^+$.

Let $\delta_{\max}$ denote the maximum value in $\mathcal{B}$ so that $\delta_{\max} = 1/2$. With E_2 and $a \in \mathcal{L}/\mathcal{L}^+$, we assume that $\tilde{t}_2 (\geq t_2)$ is the smallest time that there exists $t_1 \leq s \leq \tilde{t}_2$ with $s = \tilde{t}_2 + 1 - 2^{l-1}$ for $l \in \mathbb{Z}^+$ such that the following condition is met:

$$\mu_{t_1}(a) + \sqrt{12 \log(H)/n_{[s,\tilde{t}_2]}(a)} + \sqrt{14 \log(H)/n_{[s,\tilde{t}_2]}(a)} < 1 - \delta_{\max}. \tag{45}$$

From the definition of $\tilde{t}_2$, we observe that for given $\tilde{t}_2$, the time step s, which satisfies (45), equals to t_1. Then, we can observe that $n_{[t_1,\tilde{t}_2]}(a) = \max\{\lceil C_2 \log(H)/(\Delta_{t_1}(a) - \delta_{\max})^2 \rceil, 1\}$ for some constant $C_2 > 0$, which satisfies (45). From $n_{[t_1,t_2]}(a) \leq n_{[t_1,\tilde{t}_2]}(a)$, we have $n_{[t_1,t_2]}(a) \leq \max\{C_3 \log(H)/(\Delta_{t_1}(a) - \delta_{\max})^2, 1\}$ for some constant $C_3 > 0$. Then the regret from arm a, denoted by $R(a)$, is bounded as follows: $R(a) = \Delta_{t_1}(a) n_{[t_1,t_2]}(a) \leq \max\{C_3 \log(H) \Delta_{t_1}(a)/(\Delta_{t_1}(a) - \delta_{\max})^2, \Delta_{t_1}(a)\}$. Since $x/(x - \delta_{\max})^2 < 1/(1 - \delta_{\max})^2 = 4$ for any $x > 1$, we have $\Delta_{t_1}(a)/(\Delta_{t_1}(a) - \delta_{\max})^2 \leq 4$. Then we have $R(a) \leq \max\{C_4 \log(H), \Delta_{t_1}(a)\}$ for some constant $C_4 > 0$. Then from $|\mathcal{L}| \leq H$, we have $\sum_{a \in \mathcal{L}/\{\mathcal{L}^+\}} R(a) \leq \max\{C_4 H \log(H), H + V_{H,i}\}$.

Since $\sum_{a \in \mathcal{L}^+} R(a) \leq H$, we have $\sum_{a \in \mathcal{L}} R(a) \leq H + \max\{C_4 H \log(H), H + V_{H,i}\}$. Therefore from $R(a) = \sum_{t=t_1(a)}^{t_2(a)} (1 - \mu_t(a))$, we have

$$\sum_{t=(i-1)H+1}^{iH \wedge T} \mu_t(a_t) \geq -\max\{C_4 H \log(H), H + V_{H,i}\},$$

which implies that from $V_{H,i} \leq H$ under Assumption 3.5, for some $C_5 > 0$, we have

$$\left| \sum_{t=(i-1)H+1}^{i \cdot H \wedge T} \mu_t(a_t^\pi) \right| \leq \max\{C_4 H \log(H), H + V_{H,i}\} \leq C_5 H \log(H).$$

Next we provide a bound for $|\sum_{t=(i-1)H+1}^{i \cdot H \wedge T} \eta_t|$. We define the event $E_3(i) = \{|\sum_{t=(i-1)H+1}^{i \cdot H \wedge T} \eta_t| \leq 2\sqrt{H \log(T)}\}$ and $E_3 = \bigcap_{i \in [\lceil T/H \rceil]} E_3(i)$. From Lemma A.21, for any $i \in [\lceil T/H \rceil]$, we have

$$\mathbb{P}\left(E_3(i)^c\right) \leq \frac{2}{T^2}.$$

Then, under $E_2 \cap E_3$, with (44), we have

$$Q' \leq \max\{C_5 H \log H \varrho, H\} + 2\sqrt{H \log(T)} \leq C_5 H \log H + 2\sqrt{H \log(T)},$$

which implies $1/2 + \sum_{t=(i-1)H}^{i\cdot H\wedge T} r_t/(C_5 H\log H + 4\sqrt{H\log T}) \in [0,1]$ or some large enough $C > 0$. With the rescaling and translation of rewards in Algorithm 2, from Corollary 3.2. in Auer et al. (2002) with either of Assumption 3.7 or 3.8, we have

$$\mathbb{E}[R_2^\pi(T)|E_2\cap E_3] = \tilde{O}\left((C_5 H\log H + 2\sqrt{H\log T})\sqrt{BT/H}\right) = \tilde{O}\left(\sqrt{HBT}\right). \tag{46}$$

Regarding the utilization of the regret analysis of Corollary 3.2 in Auer et al. (2002), we note that the reward for each base can be defined independently of the master's action. Specifically, the reward for each base is defined by the rewards obtained when the corresponding base is selected, regardless of the master's actual action.

Note that the expected regret from EXP3 is trivially bounded by $o(H^2(T/H)) = o(TH)$ and $B = O(\log(T))$. Then, with (46), we have

$$\begin{aligned}\mathbb{E}[R_2^\pi(T)] &= \mathbb{E}[R_2^\pi(T)|E_2\cap E_3]\mathbb{P}(E_2\cap E_3) + \mathbb{E}[R_2^\pi(T)|E_2^c\cup E_3^c]\mathbb{P}(E_2^c\cup E_3^c)\\ &= \tilde{O}\left(\sqrt{HT}\right) + o\left(TH\right)\left(4/T^2\right)\\ &= \tilde{O}\left(\sqrt{HT}\right).\end{aligned} \tag{47}$$

Finally, from (40), (43), and (47), with $H = T^{1/2}$, we have

$$\mathbb{E}[R^\pi(T)] = \tilde{O}\left(\max\left\{V_T^{\frac{1}{\beta+2}}T^{\frac{\beta+1}{\beta+2}} + T^{\frac{2\beta+1}{2\beta+2}}, V_T^{\frac{1}{3}}T^{\frac{2}{3}} + T^{\frac{3}{4}}\right\}\right),$$

which concludes the proof.

A.5.2 Case of $\delta_V^\dagger > \delta_S^\dagger$

Let $\pi_i(\delta')$ for $\delta' \in \mathcal{B}$ denote the base policy for time steps between $(i-1)H+1$ and $i\cdot H\wedge T$ in Algorithm 2 using $1-\delta'$ as a threshold. Denote by $a_t^{\pi_i(\delta')}$ the pulled arm at time step t by policy $\pi_i(\delta')$. Then, for $\delta^\dagger \in \mathcal{B}$, which is set later for a near-optimal policy, we have

$$\mathbb{E}[R^\pi(T)] = \mathbb{E}\left[\sum_{t=1}^T 1 - \sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\mu_t(a_t^\pi)\right] = \mathbb{E}[R_1^\pi(T)] + \mathbb{E}[R_2^\pi(T)]. \tag{48}$$

where

$$R_1^\pi(T) = \sum_{t=1}^T 1 - \sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\mu_t(a_t^{\pi_i(\delta^\dagger)})$$

and

$$R_2^\pi(T) = \sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\mu_t(a_t^{\pi_i(\delta^\dagger)}) - \sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\mu_t(a_t^\pi).$$

Note that $R_1^\pi(T)$ accounts for the regret caused by the near-optimal base algorithm $\pi_i(\delta^\dagger)$'s against the optimal mean reward and $R_2^\pi(T)$ accounts for the regret caused by the master algorithm by selecting a base with $\delta \in \mathcal{B}$ at every block against the base with $\delta^\dagger$. In what follows, we provide upper bounds for each regret component. We first provide an upper bound for $\mathbb{E}[R_1^\pi(T)]$ by following the proof steps in Theorem 3.3. Then we provide an upper bound for $\mathbb{E}[R_2^\pi(T)]$. We set $\delta^\dagger$ to be a smallest value in $\mathcal{B}$ which is larger than

$\delta_S^\dagger = \max\{(S_T/T)^{1/(\beta+1)}, 1/H^{1/(\beta+1)}, (S_T/T)^{1/2}, 1/H^{1/2}\}$ such that we have $\delta^\dagger = \Theta(\max\{(S_T/T)^{1/(\beta+1)}, 1/H^{1/(\beta+1)}, (S_T/T)^{1/2}, 1/H^{1/2}\})$.

Upper Bounding $\mathbb{E}[R_1^\pi(T)]$. We refer to the period starting from time step $(i-1)H+1$ to time step $i \cdot H \wedge T$ as the i-th *block*. For any $i \in \lceil T/H - 1 \rceil$, policy $\pi_i(\delta^\dagger)$ runs over H time steps independent to other blocks so that each block has the same expected regret and the last block has a smaller or equal expected regret than other blocks. Therefore, we focus on finding a bound on the regret from the first block equal to $\sum_{t=1}^H 1 - \mu_t(a_t^{\pi_1(\delta^\dagger)})$. We define an arm a as a *good* arm if $\Delta_t(a) \le \delta^\dagger/2$, a *near-good* arm if $\delta^\dagger/2 < \Delta_t(a) \le 2\delta^\dagger$, and otherwise, a is a *bad* arm at time t. In $\mathcal{A}$, let $\bar{a}_1, \bar{a}_2, \dots,$ be a sequence of arms, which have i.i.d. mean rewards following (1). For analysis, *we consider abrupt change as sampling a new arm.* In other words, if a sudden change occurs to an arm a by pulling the arm a, then the arm is considered to be two different arms; before and after the change. The type of abruptly rotted arms (good, near-good, or bad) after the change is determined by the rotted mean reward. Without loss of generality, we assume that the policy samples arms, which are pulled at least once, in the sequence of $\bar{a}_1, \bar{a}_2, \dots,$.

Denote by $\mathcal{A}(i)$ the set of sampled arms, which are pulled at least once, in the i-th block, which satisfies $|\mathcal{A}(i)| \le H$. We also define $\mathcal{A}_S(i)$ as a set of arms that have been rotted and pulled at least once in the i-th block, which satisfies $|\mathcal{A}_S(i)| \le S_i$, where S_i is defined as the number of abrupt changes in the i-th block. Let $\overline{\mu}_{[t_1,t_2]}(a) = \sum_{t=t_1}^{t_2} \mu_t(a)/n_{[t_1,t_2]}(a)$. We define the event $E_1 = \{|\widehat{\mu}_{[s_1,s_2]}(a) - \overline{\mu}_{[s_1,s_2]}(a)| \le \sqrt{12\log(H)/n_{[s_1,s_2]}(a)}$ for all $1 \le s_1 \le s_2 \le H, a \in \mathcal{A}(i)\}$. From Lemma A.21, as in (13), we have

$$\mathbb{P}(E_1^c) \le 2/H^2.$$

For the case that E_1 does not hold, the regret is $\mathbb{E}[R^\pi(H)|E_1^c]\mathbb{P}(E_1^c) = O(1)$, which is negligible comparing with the regret when E_1 holds true which we show later. Therefore, in the rest of the proof we assume that E_1 holds true.

In the following, we first provide a regret bound over the first block.

For regret analysis, we divide $R_1^{\pi_1(\delta^\dagger)}(H)$ into two parts, $R^{\mathcal{G}}(H)$ and $R^{\mathcal{B}}(H)$ corresponding to regret of good or near-good arms, and bad arms over time T, respectively, such that $R_1^\pi(H) = R^{\mathcal{G}}(H) + R^{\mathcal{B}}(H)$. We can easily obtain that

$$\mathbb{E}[R^{\mathcal{G}}(H)] = O(\delta^\dagger H), \tag{49}$$

from $\Delta(a) \le 2\delta^\dagger$ for any good or near-good arms a.

Now we analyze $R^{\mathcal{B}}(H)$. We divide regret $R^{\mathcal{B}}(H)$ into two regret from bad arms in $\mathcal{A}(1)/\mathcal{A}_S(1)$, denoted by $R^{\mathcal{B},1}(H)$, and regret from bad arms in $\mathcal{A}_S(1)$, denoted by $R^{\mathcal{B},2}(H)$ such that $R^{\mathcal{B}}(H) = R^{\mathcal{B},1}(H) + R^{\mathcal{B},2}(H)$. We first analyze $R^{\mathcal{B},1}(H)$ in the following. We consider arms in $\mathcal{A}(1)/\mathcal{A}_S(1)$. For the proof, we adopt the episodic approach in Kim et al. (2022) for regret analysis. In the following, we introduce some notation. *Here we only consider arms in* $\mathcal{A}(1)/\mathcal{A}_S(1)$ so that the following notation is defined without considering (rotted) arms in $\mathcal{A}_S(1)$. Given a policy sampling arms in the sequence order, let $m^{\mathcal{G}}$ be the number of samples of distinct good arms and $m_i^{\mathcal{B}}$ be the number of consecutive samples of distinct bad arms between the $i-1$-st and i-th sample of a good arm among $m^{\mathcal{G}}$ good arms. We refer to the period starting from sampling the $i-1$-st good arm before sampling the i-th good arm as the i-th *episode*. Observe that $m_1^{\mathcal{B}},\dots,m_{m^{\mathcal{G}}}^{\mathcal{B}}$'s are i.i.d. random variables with geometric distribution with parameter $2\delta^\dagger$, conditional on the value of $m^{\mathcal{G}}$. Therefore, $\mathbb{P}(m_i^{\mathcal{B}}=k)=(1-2\delta^\dagger)^k2\delta^\dagger$, for $i=1,\dots,m^{\mathcal{G}}$.

Define $\tilde{m}_H^{\mathcal{G}}$ to be the total number of samples of a good arm by the policy $\pi_1(\delta^\dagger)$ over the horizon H and $\tilde{m}_{i,H}^{\mathcal{B}}$ to be the number of selections of a bad arm in the i-th episode by the policy π over the horizon H. For $i\in[\tilde{m}_H^{\mathcal{G}}]$, $j\in[\tilde{m}_{i,H}^{\mathcal{B}}]$, let $\tilde{n}_i^{\mathcal{G}}$ be the number of pulls of the good arm in the i-th episode and $\tilde{n}_{i,j}^{\mathcal{B}}$ be the number of pulls of the j-th bad arm in the i-th episode by the policy $\pi_1(\delta^\dagger)$ over the horizon H. Let $\tilde{a}$ be the last sampled arm over time horizon H by $\pi_1(\delta^\dagger)$.

With a slight abuse of notation, we use $\pi_1(\delta^\dagger)$ for a modified strategy after H. Under a policy $\pi_1(\delta^\dagger)$, let $R_{i,j}^{\mathcal{B}}$ be the regret (summation of mean reward gaps) contributed by pulling the j-th bad arm in the i-th episode. Then let $R_{m^{\mathcal{G}}}^{\mathcal{B}}=\sum_{i=1}^{m^{\mathcal{G}}}\sum_{j\in[m_i^{\mathcal{B}}]}R_{i,j}^{\mathcal{B}}$, which is the regret from initially bad arms over the period of $m^{\mathcal{G}}$ episodes. For getting $R_{m^{\mathcal{G}}}^{\mathcal{B}}$, here we define the policy $\pi_1(\delta^\dagger)$ after H such that it pulls H amounts for a good arm and zero for a bad arm. After H we can assume that there are no abrupt changes. For the last arm $\tilde{a}$ over the horizon H, it pulls the arm up to H amounts if $\tilde{a}$ is a good arm and $\tilde{n}_{\tilde{m}_H^{\mathcal{G}}}^{\mathcal{G}}<H$. For $i\in[m^{\mathcal{G}}]$, $j\in[m_i^{\mathcal{B}}]$ let $n_i^{\mathcal{G}}$ and $n_{i,j}^{\mathcal{B}}$ be the number of pulling the good arm in i-th episode and j-th bad arm in i-th episode under π, respectively. Here we define $n_i^{\mathcal{G}}$'s and $n_{i,j}^{\mathcal{B}}$'s as follows:

If $\tilde{a}$ is a good arm,

$$n_i^{\mathcal{G}}=\begin{cases}\tilde{n}_i^{\mathcal{G}} & \text{for } i\in[\tilde{m}_H^{\mathcal{G}}-1]\\ H & \text{for } i=\tilde{m}_H^{\mathcal{G}}\\ 0 & \text{for } i\in[m^{\mathcal{G}}]/[\tilde{m}_H^{\mathcal{G}}]\end{cases},\ n_{i,j}^{\mathcal{B}}=\begin{cases}\tilde{n}_{i,j}^{\mathcal{B}} & \text{for } i\in[\tilde{m}_H^{\mathcal{G}}], j\in[\tilde{m}_{i,H}^{\mathcal{B}}]\\ 0 & \text{for } i\in[m^{\mathcal{G}}]/[\tilde{m}_H^{\mathcal{G}}], j\in[m_i^{\mathcal{B}}]/[\tilde{m}_{i,H}^{\mathcal{B}}].\end{cases}$$

Otherwise,

$$n_i^{\mathcal{G}}=\begin{cases}\tilde{n}_i^{\mathcal{G}} & \text{for } i\in[\tilde{m}_H^{\mathcal{G}}]\\ H & \text{for } i=\tilde{m}_H^{\mathcal{G}}+1\\ 0 & \text{for } i\in[m^{\mathcal{G}}]/[\tilde{m}_H^{\mathcal{G}}+1]\end{cases},\ n_{i,j}^{\mathcal{B}}=\begin{cases}\tilde{n}_{i,j}^{\mathcal{B}} & \text{for } i\in[\tilde{m}_H^{\mathcal{G}}], j\in[\tilde{m}_{i,H}^{\mathcal{B}}]\\ 0 & \text{for } i\in[m^{\mathcal{G}}]/[\tilde{m}_H^{\mathcal{G}}-1], j\in[m_i^{\mathcal{B}}]/[\tilde{m}_{i,H}^{\mathcal{B}}].\end{cases}$$

With a slight abuse of notation, we define S_i to be the number of abrupt changes in i-th block. Then, we show that if $m^{\mathcal{G}}=S_1$, then $R^{\mathcal{B},1}(H)\le R_{m^{\mathcal{G}}}^{\mathcal{B}}$.

Lemma A.15. *Under E_1, when $m^{\mathcal{G}}=S_1$ we have*

$$R^{\mathcal{B},1}(H)\le R_{m^{\mathcal{G}}}^{\mathcal{B}}.$$

Proof. There are at most S_1-1 number of abrupt changes over the first block H. We consider two cases; there are S_1-1 abrupt changes before sampling S_1-th good arm or not. For the first case, if $\pi_1(\delta^\dagger)$ samples the S_1-th good arm and there are S_1-1 number of abrupt changes before sampling the good arm, then it continues to pull the good arm for H rounds from E_1 and the definition of $\pi_1(\delta^\dagger)$ after H.

Now we consider the second case. If $\pi_1(\delta^\dagger)$ samples the S_1-th good arm before T and there is at least one abrupt change after sampling the arm, then before sampling the S_1-th good arm, there must exist two consecutive good arms such that there is no abrupt change between sampling the two good arms. This is a contraction because $\pi_1(\delta^\dagger)$ must pull the first good arm up to H under E_1 and $S_1 - 1$-st abrupt change must occur after H.

Therefore, considering the first case, when $m^{\mathcal{G}} = S_1 + 1$, we have

$$\sum_{i \in [m^{\mathcal{G}}]} n_i^{\mathcal{G}} \geq H,$$

which implies $R^{\mathcal{B}}(H) \leq R^{\mathcal{B}}_{m^{\mathcal{G}}}$. □

From the above lemma, we set $m^{\mathcal{G}} = S_1$ and analyze $R^{\mathcal{B}}_{m^{\mathcal{G}}}$ to get a bound for $R^{\mathcal{B},1}(H)$ in the following lemma.

Lemma A.16. *Under E_1 and policy $\pi_1(\delta^\dagger)$, we have*

$$\mathbb{E}[R^{\mathcal{B}}_{m^{\mathcal{G}}}] = \tilde{O}\left(S_1 \max\{1/(\delta^\dagger)^\beta, 1/\delta^\dagger\}\right).$$

Proof. We can show this theorem by following the proof steps in Lemma A.8.

□

Now we analyze $R^{\mathcal{B},2}(H)$ in the following lemma. We denote by V_H a cumulative amount of rotting rates in the first block.

Lemma A.17. *Under E_1 and policy π, we have*

$$\mathbb{E}\left[R^{\mathcal{B},2}(H)\right] = \tilde{O}\left(\max\{S_1/\delta^\dagger, \sum_{s=1}^{S_1} \rho_{t(s)}\}\right).$$

Proof. We can show this theorem by following the proof steps in Lemma A.9. □

From Lemmas A.15, A.16, A.17, we have

$$\mathbb{E}[R^{\mathcal{B}}(H)] = \mathbb{E}[R^{\mathcal{B},1}(H)] + \mathbb{E}[R^{\mathcal{B},2}(H)] = \tilde{O}(S_1 \max\{1/(\delta^\dagger)^\beta, 1/\delta^\dagger\} + \sum_{s=1}^{S_1} \rho_{t(s)}) \tag{50}$$

From $R_1^\pi(H) = R^{\mathcal{G}}(H) + R^{\mathcal{B}}(H)$, (49), and (50), we have

$$\mathbb{E}[R^{\pi_1(\delta^\dagger)}_{m^{\mathcal{G}}}] = \tilde{O}\left(H\delta^\dagger + S_1 \max\{1/(\delta^\dagger)^\beta, 1/\delta^\dagger\} + \sum_{s=1}^{S_1} \rho_{t(s)}\right).$$

The above regret is for the first block. Therefore, by summing regrets over $\lceil T/H \rceil$ number of blocks, we have shown that

$$\mathbb{E}[R_1^\pi(T)] = \tilde{O}(T\delta^\dagger + (T/H + S_T) \max\{1/(\delta^\dagger)^\beta, 1/\delta^\dagger\} + \sum_{s=1}^{S_T} \rho_{t(s)}). \tag{51}$$

Upper bounding $\mathbb{E}[R_2^\pi(T)]$. By following the proof steps in Theorem 3.9, we have

$$\mathbb{E}[R_2^\pi(T)] = \tilde{O}\left(\sqrt{HT}\right). \tag{52}$$

Finally, from (48), (51), and (52), with $H = T^{1/2}$ and
$\delta^\dagger = \Theta(\max\{(S_T/T)^{1/(\beta+1)}, 1/H^{1/(\beta+1)}, (S_T/T)^{1/2}, 1/H^{1/2}\})$, we have

$$\begin{aligned}
\mathbb{E}[R^\pi(T)] &= \tilde{O}\left(T\delta^\dagger + (T/H + S_T)\max\{1/(\delta^\dagger)^\beta, 1/\delta^\dagger\} + \sqrt{HT} + \sum_{s=1}^{S_T}\rho_{t(s)}\right) \\
&= \tilde{O}\left(T\delta^\dagger + \max\{T/H, S_T\}\max\{1/(\delta^\dagger)^\beta, 1/\delta^\dagger\} + \sqrt{HT} + \sum_{s=1}^{S_T}\rho_{t(s)}\right) \\
&= \tilde{O}\left(2T\delta^\dagger + \sqrt{HT} + \sum_{s=1}^{S_T}\rho_{t(s)}\right) \\
&= \tilde{O}\left(\max\{S_T^{1/(\beta+1)}T^{\beta/(\beta+1)} + T^{(2\beta+1)/(2\beta+2)}, \sqrt{S_T T} + T^{3/4}, V_T\}\right),
\end{aligned}$$

which concludes the proof.

A.6 Proof of Theorem 4.1: Regret Lower Bound for Slowly Rotting Rewards

We first consider the case when $V_T = \Theta(T)$. Recall that $\Delta_1(a) = 1 - \mu_1(a)$. Then for any randomly sampled $a \in \mathcal{A}$, we have $\mathbb{E}[\mu_1(a)] \geq y\mathbb{P}(\mu_1(a) \geq y) = y\mathbb{P}(\Delta_1(a) < 1 - y)$ for $y \in [0,1]$. Then with $y = 1/2$, we have $\mathbb{E}[\mu_1(a)] \geq (1/2)\mathbb{P}(\Delta_1(a) < (1/2)) = \Theta(1)$ from constant $\beta > 0$ and (1). Then with $\mathbb{E}[\mu_1(a)] \leq 1$, we have $\mathbb{E}[\mu_1(a)] = \Theta(1)$. We then think of a policy π' that randomly samples a new arm and pulls it once every round. Since $\mathbb{E}[\mu_1(a)] = \Theta(1)$ for any randomly sampled a, we have $\mathbb{E}[R^{\pi'}(T)] = \Theta(T)$. Next, we think of any policy π'' except π'. Then any policy π'' must pull an arm a at least twice. Let t' and t'' be the rounds when the policy pulls arm a. If we consider $\rho_{t'} = V_T$ then such policy has $\Omega(V_T)$ regret bound. Since $V_T = \Theta(T)$, any algorithm has $\Omega(T)$ in the worst case. Therefore we can conclude that any algorithm including π' has a regret bound of $\Omega(T)$ in the worst case, which concludes the proof for $V_T = \Theta(T)$.

Now we think of the case where $V_T = o(T)$. For the lower bound, we adopt the proof methodology of Theorem 1 in Kim et al. (2022) by making necessary adjustments to accommodate V_T and β. We first categorize arms as either bad or good according to their initial mean reward values. For the categorization, we utilize two thresholds in the proof as follows. Consider $0 < \gamma < c < 1$ for γ, which will be specified, and a constant c. Then the value of $1 - \gamma$ represents a threshold value for identifying good arms, while $1 - c$ serves as the threshold for identifying bad arms. We refer to arms a satisfying $\mu_1(a) \leq 1 - c$ as 'bad' arms and arms a satisfying $\mu_1(a) > 1 - \gamma$ as 'good' arms. We also consider a sequence of arms in $\mathcal{A}$ denoted by $\bar{a}_1, \bar{a}_2, \dots$. Given a policy π, without loss of generality, we can assume that π selects arms according to the order of $\bar{a}_1, \bar{a}_2, \dots$. For the rotting rates, we define $\rho = V_T/(T-1)$. Then we consider $\rho_t = \rho$ for all $t \in [T-1]$ so that $\sum_{t=1}^{T-1}\rho_t = V_T$.

Case of $V_T = O(1/T^{1/(\beta+1)})$: When $V_T = O(1/T^{1/(\beta+1)})$, the lower bound of order $T^{\frac{\beta}{\beta+1}}$ for the stationary case, from Theorem 3 in Wang et al. (2009), is tight enough for the non-stationary case. From Theorem 3 in Wang et al. (2009), we have

$$\mathbb{E}[R^{\pi}(T)] = \Omega(T^{\frac{\beta}{\beta+1}}). \tag{53}$$

We note that even though the mean rewards are rotting in our setting, Theorem 3 in Wang et al. (2009) remains applicable without requiring any alterations in the proofs providing a tight regret bound for the near-stationary case. For the sake of completeness, we provide the proof of the theorem in the following. Let K_1 denote the number of bad arms a that satisfy $\mu_1(a) \leq 1-c$ before sampling the first good arm, which satisfies $\mu_1(a) > 1-\gamma$, in the sequence of arms $\bar{a}_1, \bar{a}_2, \dots$. Let $\overline{\mu}$ be the initial mean reward of the best arm among the sampled arms by π over time horizon T. Then for some $\kappa > 0$, we have

$$\begin{aligned} R^{\pi}(T) &= R^{\pi}(T)\mathbb{1}(\overline{\mu} \leq 1-\gamma) + R^{\pi}(T)\mathbb{1}(\overline{\mu} > 1-\gamma) \\ &\geq T\gamma\mathbb{1}(\overline{\mu} \leq 1-\gamma) + K_1 c\mathbb{1}(\overline{\mu} > 1-\gamma) \\ &\geq T\gamma\mathbb{1}(\overline{\mu} \leq 1-\gamma) + \kappa c\mathbb{1}(\overline{\mu} > 1-\gamma, K_1 \geq \kappa). \end{aligned} \tag{54}$$

By taking expectations on the both sides in (54) and setting $\kappa = T\gamma/c$, we have

$$\mathbb{E}[R^{\pi}(T)] \geq T\gamma\mathbb{P}(\overline{\mu} \leq 1-\gamma) + \kappa c(\mathbb{P}(\overline{\mu} > 1-\gamma) - \mathbb{P}(K_1 < \kappa)) = c\kappa\mathbb{P}(K_1 \geq \kappa).$$

We observe that K_1 follows a geometric distribution with success probability $\mathbb{P}(\mu_1(a) > 1-\gamma)/p(\mu_1(a) \notin (1-c, 1-\gamma]) = \overline{\gamma} \leq C_1\gamma^{\beta}/(1+C_2\gamma^{\beta}-C_3c^{\beta})$ for some constants $C_1, C_2, C_3 > 0$ from (1), in which the success probability is the probability of sampling a good arm given that the arm is either a good or bad arm. Here we set a constant $0 < c < 1$ satisfying $1 - C_3c^{\beta} > 0$. Then by setting $\gamma = 1/T^{\frac{1}{\beta+1}}$ with $\kappa = T^{\frac{\beta}{\beta+1}}/c$, for some constant $C > 0$ we have

$$\mathbb{E}[R^{\pi}(T)] \geq c\kappa(1-\overline{\gamma})^{\kappa} = \Omega\left(T^{\frac{\beta}{\beta+1}}(1-C\gamma^{\beta})^{T^{\frac{\beta}{\beta+1}}/c}\right) = \Omega(T^{\frac{\beta}{\beta+1}}),$$

where the last equality is obtained from $\log x \geq 1 - 1/x$ for all $x > 0$.

Case of $V_T = \omega(1/T^{1/(\beta+1)})$ and $V_T = o(T)$: When $V_T = \omega(1/T^{1/(\beta+1)})$, however, the lower bound of the stationary case is not tight enough. Here we provide the proof for the lower bound of $V_T^{1/(\beta+2)}T^{(\beta+1)/(\beta+2)}$ for the case of $V_T = \omega(1/T^{1/(\beta+1)})$. Let K_m denote the number of "bad" arms a that satisfy $\mu_1(a) \leq 1-c$ before sampling m-th "good" arm, which satisfies $\mu_1(a) > 1-\gamma$, in the sequence of arms $\bar{a}_1, \bar{a}_2, \dots$. Let N_T be the number of sampled good arms a such that $\mu_1(a) > 1-\gamma$ until T.

We can decompose $R^{\pi}(T)$ into two parts as follows:

$$R^{\pi}(T) = R^{\pi}(T)\mathbb{1}(N_T < m) + R^{\pi}(T)\mathbb{1}(N_T \geq m). \tag{55}$$

We set $m = \lceil (1/2)T^{1/(\beta+2)}V_T^{(\beta+1)/(\beta+2)} \rceil$ and $\gamma = (V_T/T)^{1/(\beta+2)}$ with $V_T = o(T)$. For the first term in (55), $R^{\pi}(T)\mathbb{1}(N_T < m)$, we consider the fact that the minimal regret is obtained from the situation where there are $m-1$ arms whose mean rewards are 1. In such a case, the optimal policy must sample the best $m-1$ arms until their mean rewards become below the threshold $1-\gamma$ (step 1) and then samples the best arm at each time for the remaining time steps (step 2). The number of times each arm needs to be pulled for the best $m-1$ arms until their mean reward falls below $1-\gamma$ is bounded from above by $\gamma/\varrho+1=\gamma((T-1)/V_T)+1$. Therefore, the regret from step 2 is $R=\Omega((T-m\gamma(T/V_T))\gamma)=\Omega(T^{(\beta+1)/(\beta+2)}V_T^{1/(\beta+2)})$ in which the optimal policy pulls arms which mean rewards are below $1-\gamma$ for the remaining time after step 1. Therefore, we have

$$R^\pi(T)\mathbb{1}(N_T<m)=\Omega(R\mathbb{1}(N_T<m))=\Omega(T^{(\beta+1)/(\beta+2)}V_T^{1/(\beta+2)}\mathbb{1}(N_T<m)). \tag{56}$$

For getting a lower bound of the second term in (55), $R^\pi(T)\mathbb{1}(N_T\geq m)$, we use the minimum number of sampled arms a that satisfy $\mu_1(a)\leq 1-c$. When $N_T\geq m$ and $K_m\geq\kappa$, the policy samples at least κ number of distinct arms a satisfying $\mu_1(a)\leq 1-c$ until T. Therefore, we have

$$R^\pi(T)\mathbb{1}(N_T\geq m)\geq c\kappa\mathbb{1}(N_T\geq m,K_m\geq\kappa). \tag{57}$$

We have $\overline{\gamma}=\Theta(\gamma^\beta)$ from (1) with constant $\beta>0$. By setting $\kappa=m/\overline{\gamma}-m-\sqrt{m}/\overline{\gamma}$, with $V_T=o(T)$ and constant $\beta>0$, we have

$$\kappa=\Theta(T^{(\beta+1)/(\beta+2)}V_T^{1/(\beta+2)}). \tag{58}$$

Then from (56), (57), and (58), we have

$$\begin{aligned}\mathbb{E}[R^\pi(T)]&=\Omega(T^{(\beta+1)/(\beta+2)}V_T^{1/(\beta+2)}\mathbb{P}(N_T<m)+T^{(\beta+1)/(\beta+2)}V_T^{1/(\beta+2)}\mathbb{P}(N_T\geq m,K_m\geq\kappa))\\&\geq\Omega(T^{(\beta+1)/(\beta+2)}V_T^{1/(\beta+2)}\mathbb{P}(K_m\geq\kappa)).\end{aligned} \tag{59}$$

Next we provide a lower bound for $\mathbb{P}(K_m\geq\kappa)$. Observe that K_m follows a negative binomial distribution with m successes and the success probability $\mathbb{P}(\mu_1(a)>1-\gamma)/\mathbb{P}(\mu_1(a)\notin(1-c,1-\gamma])=\overline{\gamma}$, in which the success probability is the probability of sampling a good arm given that the arm is either a good or bad arm. In the following lemma, we provide a concentration inequality for K_m.

Lemma A.18. *For any $1/2+\overline{\gamma}/m<\alpha<1$,*

$$\mathbb{P}(K_m\geq\alpha m(1/\overline{\gamma})-m)\geq 1-\exp(-(1/3)(1-1/\alpha)^2(\alpha m-\overline{\gamma})).$$

Proof. Let X_i for $i>0$ be i.i.d. Bernoulli random variables with success probability $\overline{\gamma}$. From Section 2 in Brown (2011), we have

$$\mathbb{P}\left(K_m\leq\left\lfloor\alpha m\frac{1}{\overline{\gamma}}\right\rfloor-m\right)=\mathbb{P}\left(\sum_{i=1}^{\lfloor\alpha m\frac{1}{\overline{\gamma}}\rfloor}X_i\geq m\right). \tag{60}$$

From (60) and Lemma A.20, for any $1/2 + \overline{\gamma}/m < \alpha < 1$ we have

$$\mathbb{P}\left(K_m \le \alpha m \frac{1}{\overline{\gamma}} - m\right) = \mathbb{P}\left(K_m \le \left\lfloor \alpha m \frac{1}{\overline{\gamma}} \right\rfloor - m\right)$$

$$= \mathbb{P}\left(\sum_{i=1}^{\left\lfloor \alpha m \frac{1}{\overline{\gamma}} \right\rfloor} X_i \ge m\right)$$
$$\le \exp\left(-\frac{(1-1/\alpha)^2}{3} \left\lfloor \alpha m \frac{1}{\overline{\gamma}} \right\rfloor \overline{\gamma}\right)$$
$$\le \exp\left(-\frac{(1-1/\alpha)^2}{3} (\alpha m - \overline{\gamma})\right),$$

in which the first inequality comes from Lemma A.20, which concludes the proof. □

From Lemma A.18 with $\alpha = 1 - 1/\sqrt{m}$ and large enough T, we have

$$\begin{aligned}
\mathbb{P}(K_m \ge \kappa) &\ge 1 - \exp\left(-\frac{1}{3}(m - \sqrt{m} - \overline{\gamma})\left(\frac{1}{\sqrt{m}-1}\right)^2\right) \\
&\ge 1 - \exp\left(-\frac{1}{6}(m - \sqrt{m})\left(\frac{1}{\sqrt{m}-1}\right)^2\right) \\
&= 1 - \exp\left(-\frac{1}{6}\frac{\sqrt{m}}{\sqrt{m}-1}\right) \\
&\ge 1 - \exp(-1/6).
\end{aligned} \tag{61}$$

Therefore, from (59) and (61), we have

$$\mathbb{E}[R^\pi(T)] = \Omega(T^{(\beta+1)/(\beta+2)} V_T^{1/(\beta+2)}). \tag{62}$$

Finally, from (53) and (62), we conclude that for any policy π, we have

$$\mathbb{E}[R^\pi(T)] = \Omega\left(\max\left\{T^{(\beta+1)/(\beta+2)} V_T^{1/(\beta+2)}, T^{\frac{\beta}{\beta+1}}\right\}\right).$$

A.7 Proof of Theorem 4.2: Regret Lower Bound for Abruptly Rotting Rewards

First, we deal with the case when $S_T = 1$ or $S_T = \Theta(T)$. When $S_T = 1$ (implying $V_T = 0$), from the definition, the problem becomes stationary without rotting instances, which implies $\mathbb{E}[R^\pi(T)] = \Omega(\sqrt{T})$ from Theorem 3 in Wang et al. (2009). When $S_T = \Theta(T)$, we consider that rotting occurs for the first $S_T - 1$ rounds with $\rho_t = 1$ for all $t \in [S_T - 1]$. Then it is always beneficial to pull new arms every round until $S_T - 1$ rounds because the mean rewards of rotted arms are below 0 and those of non-rotted arms lie in $[0, 1]$. This means that any ideal policy samples a new arm and pulls it every round until $S_T - 1$. Then for any randomly sampled $a \in \mathcal{A}$, we have $\mathbb{E}[\mu_1(a)] \ge y\mathbb{P}(\mu_1(a) \ge y) = y\mathbb{P}(\Delta_1(a) < 1 - y)$ for $y \in [0, 1]$. Then with $y = 1/2$, we have $\mathbb{E}[\mu_1(a)] \ge (1/2)\mathbb{P}(\Delta_1(a) < (1/2)) = \Theta(1)$ from constant $\beta > 0$ and (1). Then with $\mathbb{E}[\mu_1(a)] \le 1$, we have $\mathbb{E}[\mu_1(a)] = \Theta(1)$. Since $\mathbb{E}[\mu_1(a)] = \Theta(1)$ for any randomly sampled $a \in \mathcal{A}$, any ideal policy has $\mathbb{E}[R^\pi(T)] \geq \sum_{i=1}^{S_T} \mathbb{E}[\mu_1(a)] = \Omega(S_T) = \Omega(T)$, which concludes the proof for $S_T = \Theta(T)$.

Now we consider the case of $S_T = o(T)$ and $S_T \geq 2$. We first provide a regret bound with respect to the cumulative rotting amount of V_T. We first think of a policy π that randomly samples a new arm and pulls it once every round. Then for any randomly sampled $a \in \mathcal{A}$, we have $\mathbb{E}[\mu_1(a)] = \Theta(1)$. Then from constant $\beta > 0$, $\mathbb{E}[R^\pi(T)] = \Omega(T)$. Then there always exists ρ_t's satisfying $V_T = T$, which implies $\mathbb{E}[R^\pi(T)] = \Omega(V_T)$. Now we think of any nontrivial algorithm which must pull an arm a at least twice. Let t' and t'' be the rounds when the policy pulls arm a ($t' < t''$). If we consider $\rho_{t'} = V_T$ and $\rho_t = 0$ for $t \in [T-1]/\{t'\}$ in which $\sum_{t=1}^{T-1} \rho_t = V_T$ and $1 + \sum_{t=1}^{T-1} \rho_t \mathbb{1}(\rho_t \neq 0) \leq S_T$, then such policy has $\Omega(V_T)$ regret bound because it pulls the rotted arm a by $\rho_{t'}$ at time t''. Therefore, for any policy π, there always exist ρ_t's such that

$$\mathbb{E}[R^\pi(T)] = \Omega(V_T). \tag{63}$$

Next, for the regret bound with respect to S_T, we follow the proof steps in Theorem 4.1. We first categorize arms as either bad or good according to their initial mean reward values. For the categorization, we utilize two thresholds in the proof as follows. Consider $0 < \gamma < c < 1$ for γ, which will be specified, and a constant c. Then the value of $1 - \gamma$ represents a threshold value for identifying good arms, while $1 - c$ serves as the threshold for identifying bad arms. We refer to arms a satisfying $\mu_1(a) \leq 1 - c$ as 'bad' arms and arms a satisfying $\mu_1(a) > 1 - \gamma$ as 'good' arms. We also consider a sequence of arms in $\mathcal{A}$ denoted by $\bar{a}_1, \bar{a}_2, \dots$. Given a policy π, without loss of generality, we can assume that π selects arms according to the order of $\bar{a}_1, \bar{a}_2, \dots$.

Let K_m denote the number of bad arms a that satisfy $\mu_1(a) \leq 1 - c$ before sampling m-th good arm, which satisfies $\mu_1(a) > 1 - \gamma$, in the sequence of arms $\bar{a}_1, \bar{a}_2, \dots$. Let N_T be the number of sampled good arms a such that $\mu_1(a) > 1 - \gamma$ until T.

We can decompose $R^\pi(T)$ into two parts as follows:

$$R^\pi(T) = R^\pi(T)\mathbb{1}(N_T < m) + R^\pi(T)\mathbb{1}(N_T \geq m). \tag{64}$$

We set $m = S_T$ and $\gamma = (S_T/T)^{1/(\beta+1)}$ with $S_T = o(T)$. For getting a lower bound for the first term in (64), $R^\pi(T)\mathbb{1}(N_T < m)$, we consider the fact that the minimal regret is obtained from the situation where there are $m - 1$ arms whose mean rewards are 1. In such a case, the optimal policy must sample the best $m - 1$ arms until their mean rewards become equal to or below the threshold value of $1 - \gamma$ (step 1) and then samples the best arm at each time for the remaining time steps (step 2). In step 1, when the optimal policy pulls an optimal arm, we can think of the case when the mean reward of the arm is abruptly rotted to the value of $1 - \gamma$. This implies that the required number of rounds for step 1 is $m - 1$. The regret from step 2 is $R = \Omega((T - m + 1)\gamma) = \Omega(S_T^{1/(\beta+1)} T^{\beta/(\beta+1)})$, in which the optimal policy pulls arms which mean rewards are below or equal to $1 - \gamma$ for the remaining time after step 1. Therefore, we have

$$R^\pi(T)\mathbb{1}(N_T < m) = \Omega(R\mathbb{1}(N_T < m)) = \Omega(S_T^{1/(\beta+1)} T^{\beta/(\beta+1)} \mathbb{1}(N_T < m)). \tag{65}$$

For getting the above, we note that there always exists ρ_t's satisfying $V_T = O(\gamma m) = o(T)$ for the first step and $V_T = 0$ for the second step, which implies $V_T \leq T$. Such ρ_t's can be considered for the below. For getting a lower bound of the second term in (64), $R^\pi(T)\mathbb{1}(N_T \geq m)$, we use the minimum number of sampled arms a that satisfy $\mu_1(a) \leq 1-c$. When $N_T \geq m$ and $K_m \geq \kappa$, the policy samples at least κ number of distinct arms a satisfying $\mu_1(a) \leq 1-c$ until T. Therefore, we have

$$R^\pi(T)\mathbb{1}(N_T \geq m) \geq c\kappa\mathbb{1}(N_T \geq m, K_m \geq \kappa). \tag{66}$$

We set $\overline{\gamma} = \mathbb{P}(\mu_1(a) > 1-\gamma)/p(\mu_1(a) \notin (1-c, 1-\gamma])$. Then we have $\overline{\gamma} = \Theta(\gamma^\beta)$ from (1) with constant $\beta > 0$. By setting $\kappa = m/\overline{\gamma} - m - m/(\overline{\gamma}\sqrt{m+3})$, with $S_T = o(T)$ and constant $\beta > 0$, we have

$$\kappa = \Theta(S_T^{1/(\beta+1)}T^{\beta/(\beta+1)}). \tag{67}$$

Then from (65), (66), and (67), we have

$$\begin{aligned}\mathbb{E}[R^\pi(T)] &= \Omega(S_T^{1/(\beta+1)}T^{\beta/(\beta+1)}\mathbb{P}(N_T < m) + S_T^{1/(\beta+1)}T^{\beta/(\beta+1)}\mathbb{P}(N_T \geq m, K_m \geq \kappa)) \\ &\geq \Omega(S_T^{1/(\beta+1)}T^{\beta/(\beta+1)}\mathbb{P}(K_m \geq \kappa)).\end{aligned} \tag{68}$$

Next we provide a lower bound for $\mathbb{P}(K_m \geq \kappa)$. Observe that K_m follows a negative binomial distribution with m successes and the success probability $\mathbb{P}(\mu_1(a) > 1-\gamma)/\mathbb{P}(\mu_1(a) \notin (1-c, 1-\gamma]) = \overline{\gamma}$, in which the success probability is the probability of sampling a good arm given that the arm is either a good or bad arm. We recall Lemma A.18 for a concentration inequality for K_m in the following.

Lemma A.19. *For any $1/2 + \overline{\gamma}/m < \alpha < 1$,*

$$\mathbb{P}(K_m \geq \alpha m(1/\overline{\gamma}) - m) \geq 1 - \exp(-(1/3)(1-1/\alpha)^2(\alpha m - \overline{\gamma})).$$

From Lemma A.19 with $\alpha = 1 - 1/\sqrt{m+3}$ and large enough T, we have

$$\begin{aligned}\mathbb{P}(K_m \geq \kappa) &\geq 1 - \exp\left(-\frac{1}{3}(m - \frac{m}{\sqrt{m+3}} - \overline{\gamma})\left(\frac{1}{\sqrt{m+3}-1}\right)^2\right) \\ &\geq 1 - \exp\left(-\frac{1}{6}(m - \frac{m}{\sqrt{m+3}})\left(\frac{1}{\sqrt{m+3}-1}\right)^2\right) \\ &= 1 - \exp\left(-\frac{1}{6}\frac{m}{m+3}\frac{\sqrt{m+3}}{\sqrt{m+3}-1}\right) \\ &\geq 1 - \exp(-1/24),\end{aligned} \tag{69}$$

where the last inequality comes from $m/(m+3) = (S_T)/(S_T+3) \geq 1/4$ and $\sqrt{m+3}/(\sqrt{m+3}-1) \geq 1$. Therefore, from (68) and (69), we have

$$\mathbb{E}[R^\pi(T)] = \Omega(S_T^{1/(\beta+1)}T^{\beta/(\beta+1)}). \tag{70}$$

Overall from (63) and (70), for any π, there exist ρ_t's such that $\mathbb{E}[R^\pi(T)] = \Omega(\max\{S_T^{1/(\beta+1)}T^{\beta/(\beta+1)}, V_T\})$.

A.8 Lemmas for Concentration Inequalities

Lemma A.20 (Theorem 6.2.35 in Tsun (2020)). *Let $X_1, \ldots, X_n$ be identical independent Bernoulli random variables. Then, for $0 < \nu < 1$, we have*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq (1+\nu)\mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right) \leq \exp\left(-\frac{\nu^2 \mathbb{E}[\sum_{i=1}^{n} X_i]}{3}\right).$$

Lemma A.21 (Corollary 1.7 in Rigollet and Hütter (2015)). *Let $X_1, \ldots, X_n$ be independent random variables with σ-sub-Gaussian distributions. Then, for any $a = (a_1, \ldots, a_n)^\top \in \mathbb{R}^n$ and $t \geq 0$, we have*

$$\mathbb{P}\left(\sum_{i=1}^{n} a_i X_i > t\right) \leq \exp\left(-\frac{t^2}{2\sigma^2 \|a\|_2^2}\right) \text{ and } \mathbb{P}\left(\sum_{i=1}^{n} a_i X_i < -t\right) \leq \exp\left(-\frac{t^2}{2\sigma^2 \|a\|_2^2}\right).$$