Rotting Bandits

Nir Levine [1] **Koby Crammer** [1] **Shie Mannor** [1]

Abstract

The Multi-Armed Bandits (MAB) framework highlights the tension between acquiring new knowledge (Exploration) and leveraging available knowledge (Exploitation). In the classical MAB problem, a decision maker must choose an arm at each time step, upon which she receives a reward. The decision maker's objective is to maximize her cumulative expected reward over the time horizon. The MAB problem has been studied extensively, specifically under the assumption of the arms' rewards distributions being stationary, or quasi-stationary, over time. We consider a variant of the MAB framework, which we termed *Rotting Bandits*, where each arm's expected reward decays as a function of the number of times it has been pulled. We are motivated by many real-world scenarios such as online advertising, content recommendation, crowdsourcing, and more. We present algorithms, accompanied by simulations, and derive theoretical guarantees.

1. Introduction

One of the most fundamental trade-offs in stochastic decision theory is the well celebrated Exploration vs. Exploitation dilemma. Should one acquire new knowledge on the expanse of possible sacrifice in the immediate reward (Exploration), or leverage past knowledge in order to maximize instantaneous reward (Exploitation)? Solutions that have been demonstrated to perform well are those which succeed in balancing the two. First proposed by Thompson (1933) in the context of drug trials, and later formulated in a more general setting by Robbins (1985), MAB problems serve as a distilled framework for this dilemma. In the classical setting of the MAB, at each time step, the decision maker must choose (pull) between a fixed number of arms. After pulling an arm, she receives a reward which is a realization drawn from the arm's underlying reward distribution. The decision maker's objective is to maximize her cumulative expected reward over the time horizon. An equivalent, more typically studied, is the *regret*, which is defined as the difference between the optimal cumulative expected reward (under full information) and that of the policy deployed by the decision maker.

MAB formulation has been studied extensively, and was leveraged to formulate many real-world problems. Some examples for such modeling are, but not limited to, online advertising (Pandey et al., 2007), routing of packets (Awerbuch & Kleinberg, 2004), and online auctions (Kleinberg & Leighton, 2003).

Most past work (Section 6) on the MAB framework has been performed under the assumption that the underlying distributions are stationary, or possibly quasi-stationary. In many real-world scenarios, this assumption may seem simplistic. Specifically, we are motivated by real-world scenarios where the expected reward of an arm decreases over time instances that it has been pulled. We term this variant of decaying expected rewards *Rotting Bandits*. For motivational purposes, we present the following two examples.

- Consider an online advertising problem where an agent must choose which ad (arm) to present (pull) to a user. It seems reasonable that the effectiveness (reward), on a user, of a specific ad would deteriorate over exposures. Similarly, in the content recommendation context, Agarwal et al. (2009) showed that articles' CTR decay over amount of exposures.

- Consider the problem of assigning projects through crowdsourcing systems (Tran-Thanh et al., 2012). Given that the assignments primarily require human perception, subjects may fall into boredom and their performance would decay (e.g., license plate transcriptions (Du et al., 2013)).

As opposed to the stationary case, where the optimal policy is to always choose some specific arm, in the case of Rotting Bandits the optimal policy consists of choosing different arms. This results in the notion of *adversarial regret* vs. *policy regret* (Arora et al., 2012) (see Section 6). In this work we tackle the harder problem of minimizing the policy regret.

[1]Electrical Engineering Department, The Technion - Israel Institute of Technology, Haifa 32000, Israel. Correspondence to: Nir Levine <levin.nir1@gmail.com>.

The rest of the paper is organized as follows: in Section 2 we present the model and relevant preliminaries. In Section 3 we present our algorithms along with theoretical guarantees for the asymptotically vanishing case. In Section 4 we do the same for the asymptotically non-vanishing case, followed by simulations in Section 5. In Section 6 we review related work, and conclude with a discussion in Section 7.

2. Model and Preliminaries

We consider the problem of Rotting Bandits (RB); we are given K arms and at each step $t = 1, 2, ..$ one of the arms must be pulled. We denote the arm that is pulled at time step t as $i(t) \in [K] = \{1, .., K\}$. When arm i is pulled for the n^{th} time, we receive a time independent, $\sigma^2(n)$ sub-Gaussian random reward[1], r_t, with mean $\mu_i^S(n; \theta_i^*) = \mu_i^c + \mu(n; \theta_i^*)$, determined by a rotting parameter $\theta_i^* \in \Theta$ and an unknown constant μ_i^c. The functions $\sigma^2(n)$ and $\{\mu(n; \theta)\}_{\theta \in \Theta}$ are known in advance (but not the true underlying rotting models $\Theta^* = \{\theta_i^*\}_{i=1}^K$).

In this work we consider two different cases. The first is the asymptotically vanishing case (AV), i.e., $\forall i : \mu_i^c = 0$. The second is the asymptotically non-vanishing case (ANV), i.e., $\forall i : \mu_i^c \in \mathbb{R}$.

Let $N_i(t)$ be the number of pulls of arm i at time t not including ($N_i(1) = 0$), and Π the set of all (possibly randomized) mappings $(\{1, .., K\} \times \mathbb{R})^{t-1} \to \{1, .., K\}$ for any $t \in \mathbb{N}$. i.e., a mapping from action indices and observed rewards to an action index (also referred to as a policy). We denote the arm that is chosen by policy π at time t as $\pi(t)$. The objective of an agent is to maximize the expected total reward in time T, defined for policy $\pi \in \Pi$ by,

$$J(T; \pi) = \mathbb{E}\left[\sum_{t=1}^{T} \mu_{\pi(t)}^S\left(N_{\pi(t)}(t) + 1; \theta_{\pi(t)}^*\right)\right] \quad (1)$$

We consider the equal objective of minimizing the regret in time T defined by,

$$\mathcal{R}(T; \pi) = \max_{\tilde{\pi} \in \Pi}\{J(T; \tilde{\pi})\} - J(T; \pi). \quad (2)$$

2.1. Rotting Bandits Redemption

We demonstrate a significant difference between standard MAB and the RB problems in the following example.

Scenario I: Standard Multi-Armed Bandits (MAB) setup. Two Bernoulli arms with probabilities $\{0.5, 0.4\}$. For $T = 3$, clearly, as for any time horizon, the optimal policy would be to always pull the arm w.p 0.5. Each time the decision maker pulls the arm w.p 0.4 she increases the regret without a possibility to decrease it at a later point in time.

Scenario II: RB setup. Two Bernoulli arms with probabilities $\{0.5/n, 0.4/n\}$ where n is the number of times that arm was pulled, including current pull, if chosen. For $T = 3$ the optimal policy would be to pull twice the arm with $0.5/n$ and once the other. Note that at time T it does not matter what was the order of pulls. It is possible to increase the regret (e.g., pull the arm w.p $0.4/n$ at the first time step), and then decrease the regret (e.g., by then pulling twice consecutively the arm w.p $0.5/n$).

The above example demonstrates a crucial difference between the two setups. In the RB problem the decision maker can possibly "redeem" herself for past mistakes in future time. This observation naturally raises the following question: what regret can be achieved by an algorithm for the RB problem?

For the AV case, we will answer this question by presenting algorithms that, for long enough horizon, achieve zero regret(!) with high probability, or regret in $o(1)$ in expectation. As for the ANV case, we present an algorithm with regret in $O(\ln(T))$ with high probability.

2.2. Definitions

Definition 2.1 *Let $\mathcal{S}, \mathcal{D} : f \times \mathbb{N}^2 \to \mathbb{R}$ be the Interval Sum, and the Interval Difference operators (respectively) defined on a function $f : \mathbb{N} \to \mathbb{R}$ and two indices $m > n$ by,*

$$\mathcal{S}(f, n, m) = \sum_{k=n}^{m} f(k)$$

$$\mathcal{D}(f, n, m) = \sum_{k=n}^{n+\lfloor\frac{m-n+1}{2}\rfloor-1} f(k) - \sum_{n+\lfloor\frac{m-n+1}{2}\rfloor}^{m} f(k)$$

Definition 2.2 *For any $\theta_1 \neq \theta_2 \in \Theta^2$, define $det_{\theta_1,\theta_2}, Ddet_{\theta_1,\theta_2} : \mathbb{N} \to \mathbb{R}$ as,*

$$det_{\theta_1,\theta_2}(n) = \frac{\mathcal{S}(\sigma^2, 1, n)}{\left(\mathcal{S}(\mu(\cdot; \theta_1), 1, n) - \mathcal{S}(\mu(\cdot; \theta_2), 1, n)\right)^2}$$

$$Ddet_{\theta_1,\theta_2}(n) = \frac{\mathcal{S}(\sigma^2, 1, n)}{\left(\mathcal{D}(\mu(\cdot; \theta_1), 1, n) - \mathcal{D}(\mu(\cdot; \theta_2), 1, n)\right)^2}$$

Definition 2.3 *Let $\mu_{max}(n)$ and $\mu_{min}(n)$ be defined as $\max_{\theta \in \Theta}\{\mu(n; \theta)\}$ and $\min_{\theta \in \Theta}\{\mu(n; \theta)\}$, respectively.*

Definition 2.4 *Let $bal : \mathbb{N} \cup \infty \to \mathbb{N} \cup \infty$ be defined at each point $n \in \mathbb{N}$ as the solution for the following prob-*

[1] The results also hold for the case where the variance depends on the rotting model, with a slight modification in the assumptions. We present it this way for clearer analysis.

lem,

$$\min \alpha$$
$$s.t \begin{cases} \alpha \in \mathbb{N} \\ \mu_{max}(\alpha) \leq \mu_{min}(n) \end{cases}$$

We define $bal(\infty) = \infty$.

Definition 2.5 *For a function* $f : \mathbb{N} \to \mathbb{R}$*, we define the function* $f^{\star\downarrow} : \mathbb{R} \to \mathbb{N} \cup \{\infty\}$ *by the following rule: given* $\zeta \in \mathbb{R}$*,* $f^{\star\downarrow}(\zeta)$ *returns the smallest* $N \in \mathbb{N}$ *such that* $\forall n \geq N : f(n) \leq \zeta$*, or* ∞ *if such* N *does not exist.*

2.3. Assumptions

Assumption 2.1 ***(Discreteness)*** *The rotting models,* Θ*, compose a discrete*[2] *(possibly infinite) known set.*

Assumption 2.2 ***(Rotting)*** $\mu(n;\theta)$ *is positive, monotonically decreasing in* n*, and* $\mu(n;\theta) \in o(1)$*,* $\quad \forall \theta \in \Theta$.

Example 2.1 *Sub-Gaussian random rewards with means* $\mu_i^S(n;\theta_i^*) = n^{-\theta_i^*}$ *and constant variances* σ^2*. Where* $\theta_i^* \in \Theta = \{\theta_1, \theta_2, ..\}$*, and* $\forall \theta \in \Theta : 0.01 \leq \theta \leq 0.49$.

Even though the expected rewards are decaying, even in the AV case, the expected total reward can still diverge (e.g., Example 2.1), keeping this problem relevant and meaningful.

2.4. Optimal Policy

Let π^{max} be a policy defined by,

$$\pi^{max}(t) \in \underset{i\in[K]}{\operatorname{argmax}}\{\mu(N_i(t)+1;\theta_i^*)\} \tag{3}$$

where, in a case of tie, break it randomly.

Lemma 2.1 π^{max} *is an optimal policy for the RB problem.* ***Proof:*** *See Appendix B.*

3. Closest To Origin (AV)

The Closest To Origin (CTO) approach for RB is a heuristic that simply states that we hypothesize that the true underlying model for an arm is the one that best fits the past rewards. The fitting criterion is proximity to the origin of the sum of expected rewards shifted by the observed rewards.

We remember that in the AV case, the expected rewards are given by, $\mu_i^S(n;\theta i^*) = 0 + \mu(n;\theta_i^*)$. Let $r_1^i, r_2^i, .., r_{N_i(t)}^i$

[2] The discreteness assumption enables us to convey this setting's potential in the clearest manner.

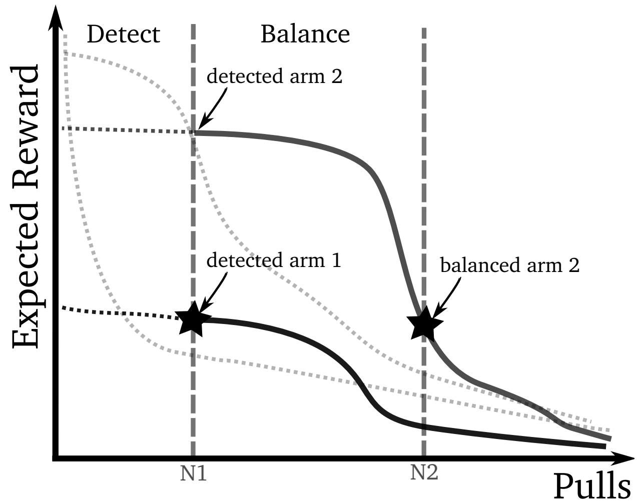

Figure 1. An example with 4 rotting models and 2 arms. After pulling each of the arms $N1$ times we detect the true underlying models, given by solid lines (Detect stage). We then balance the arms' expected rewards by pulling arm 2 a total of $N2$ pulls (Balance stage). After $N1 + N2$ time steps ($N1$ for arm 1 and $N2$ for arm 2), the regret vanishes. (The figure is best seen on screen)

be the sequence of rewards observed from arm i up until time t. Define a set of shifted sum of rewards per arm,

$$Y(i,t;\Theta) = \left\{ \sum_{j=1}^{N_i(t)} r_j^i - \sum_{j=1}^{N_i(t)} \mu(j;\theta) \right\}_{\theta\in\Theta}. \tag{4}$$

The CTO approach dictates that at each decision point, we assume that the true underlying rotting model corresponds to the following proximity to origin rule (hence the name),

$$\hat{\theta}_i(t) = \underset{\theta\in\Theta}{\operatorname{argmin}}\{|Y(i,t;\theta)|\}. \tag{5}$$

Next, we provide two versions of the CTO approach: CTO$_{DB}$ with one decision point after a finite time, and CTO$_{SIM}$ with a decision at each time step.

3.1. CTO$_{DB}$

The CTO$_{DB}$ version tackles the RB problem by two consecutive stages. Detect: pulling each arm a finite number of times, such that it ensures that the true rotting models are correctly detected with high probability. Balance: since the expected rewards of all the arms are decaying, the algorithm "redeems" past mistakes, following by puling arms while keeping their expected rewards balanced. Combining these two steps results in that from some finite time step, the regret vanishes. An illustration of the process is given by Figure 1.

3.1.1. HIGH PROBABILITY BOUND

The goal is to ensure with high probability that the regret vanishes from some point in time.

We define the following optimization problem, indicating the number of samples required for ensuring correct detection of the rotting models with high probability. For some arm i with (unknown) rotting model θ_i^*,

$$\min m \quad \text{s.t} \begin{cases} m, l \in \mathbb{N} \\ P\left(\hat{\theta}_i(l) \neq \theta_i^*\right) \leq p, \quad \forall l \geq m \\ \text{while pulling only arm } i. \end{cases} \tag{6}$$

We denote the solution to the above problem, when we use proximity rule (5), by $m^*(p;\theta_i^*)$, and define $m^*(p) \triangleq \max_{\theta\in\Theta}\{m^*(p;\theta)\}$. Pseudo algorithm for CTO$_{DB}$ is given by Algorithm 1. In case of a tie in the Balance stage, it may be arbitrarily broken. The following theorem demonstrates the strength of the algorithm for the AV RB problem.

Assumption 3.1 ***(Detection ability)***

$$\max_{\theta_1\neq\theta_2\in\Theta^2}\left\{det^{\star\downarrow}_{\theta_1,\theta_2}(\epsilon)\right\} \leq M(\epsilon) < \infty, \quad \forall\epsilon > 0$$

The above assumption ensures that the underlying models could be distinguished from the others, for any given probability, by their sums of expected rewards. This holds, for instance, for Example 2.1 (Appendix F.1).

Theorem 3.1 *Suppose Assumptions 2.1, 2.2, and 3.1 hold. For $\delta \in (0,1)$, CTO$_{DB}$ algorithm achieves **zero regret**, with probability of at least $1-\delta$, starting from (finite) time step $T^*(\delta)$ which is the solution for the following optimization problem,*

$$\min \|t\|_1 \quad \textit{s.t} \begin{cases} t \in \mathbb{N}^K \\ t_i \geq m^*(\delta/K), \quad \forall i \in [K] \\ \tilde{i} \triangleq \operatorname{argmin}_{i\in[K]}\left\{\mu\left(m^*(\delta/K);\theta_i^*\right)\right\} \\ \mu(t_i+1;\theta_i^*) \leq \mu\left(m^*(\delta/K);\theta_{\tilde{i}}^*\right), \quad \forall i \in [K] \end{cases} \tag{7}$$

Proof*: See Appendix C.1.*

Remark 3.1 *Instead of calculating $m^*(\delta/K)$, it is possible to use any upper bound of it (e.g., as shown in Appendix C.1, $\max_{\theta_1\neq\theta_2\in\Theta^2}\left\{det^{\star\downarrow}_{\theta_1,\theta_2}\left(\frac{1}{8}\ln^{-1}\left(\frac{2K}{\delta}\right)\right)\right\}$).*

Algorithm 1 CTO$_{DB}$

Input : K, Θ, δ
Initialization : $N_i = 0, \ \forall i \in [K]$
Samples : $m^*(\delta/K) \leftarrow$ sol. to (6) with $p = \delta/K$
for $t = 1, 2, .., K \times m^*(\delta/K)$ **do**
  Explore : $i(t)$ by Round Robin, and update $N_{i(t)}$
end for
Detect : determine $\{\hat{\theta}_i\}$ by Eq. (5)
for $t = K \times m^*(\delta/K) + 1, ...,$ **do**
  Balance : $i(t) \in \operatorname{argmax}_{i\in[K]}\left\{\mu\left(N_i+1;\hat{\theta}_i\right)\right\}$
  Update : $N_{i(t)} \leftarrow N_{i(t)} + 1$
end for

3.1.2. IN EXPECTATION BOUND

Here, the goal is to ensure that, in expectation, the regret is upper bounded by an expression that decays in T (and is always upper bounded by $\mu_{max}(1)$).

Assumption 3.2 ***(Balance after Detection ability)***

$$bal\left(\max_{\theta_1\neq\theta_2\in\Theta^2}\left\{det^{\star\downarrow}_{\theta_1,\theta_2}\left(\frac{1}{8}\ln^{-1}(\zeta)\right)\right\}\right) \in o(\zeta)$$

The above assumption ensures that, starting from some horizon T, the underlying models could be distinguished from the others, w.p $1-1/T$, by their sums of expected rewards, and the arms could then be balanced, all within the horizon. This holds, for instance, for Example 2.1 (Appendix F.2).

Corollary 3.1.1 *Suppose Assumptions 2.1, 2.2, and 3.2 hold. By choosing $\delta = 1/T$, there exists a (finite) step T^*_{DB}, such that for all $T \geq T^*_{DB}$, CTO$_{DB}$ achieves regret upper bounded by $\max_{\theta\in\Theta^*}\left\{\mu\left(m^*\left(\frac{1}{KT};\theta\right)\right)\right\}$.*
Proof*: See Appendix C.2.*

Remark 3.2 *This result requires the knowledge of T in advance. $\max_{\theta\in\Theta^*}\left\{\mu\left(m^*\left(\frac{1}{KT};\theta\right)\right)\right\}$ is upper bounded by a constant, $\mu_{max}(1)$. But, in addition, by assumption (2.2), it is monotonically decreasing in $m^*(\cdot)$, and as a result in T.*

Remark 3.3 *Given a function $L(\epsilon)$ that satisfies, $\forall n \geq L(\epsilon)$,*

$$bal\left(\max_{\theta_1\neq\theta_2\in\Theta^2}\left\{det^{\star\downarrow}_{\theta_1,\theta_2}\left(\frac{1}{8}\ln^{-1}(n)\right)\right\}\right) \leq \epsilon n$$

we have,

$$T^*_{DB} \leq \frac{L\left(\left(2K^2\right)^{-1}\right)}{2K}$$

as seen in Appendix C.2.

3.2. CTO$_{SIM}$

The CTO$_{SIM}$ version tackles the RB problem by simultaneously detecting the true rotting models and balancing between the expected rewards. In this approach, every time step, each arm's rotting model is hypothesized according to the proximity rule (5). Then the algorithm simply follows an argmax rule, where least number of pulls is used for tie breaking (randomly between an equal number of pulls). Pseudo algorithm for CTO$_{SIM}$ is given by Algorithm 2.

Assumption 3.3 ***(Simultaneous Balance and Detection ability)***

$$bal\left(\max_{\theta_1\neq\theta_2\in\Theta^2}\left\{det^{\star\downarrow}_{\theta_1,\theta_2}\left(\frac{1}{16}\ln^{-1}(\zeta)\right)\right\}\right)\in o(\zeta)$$

The above assumption is very similar to Assumption 3.2, except that it ensures distinction w.p $1-1/T^2$. This holds, for instance, for Example 2.1 (Appendix F.2).

Theorem 3.2 *Suppose Assumptions 2.1 ,2.2, and 3.3 hold. There exist a finite step T^*_{SIM}, such that for all $T\geq T^*_{SIM}$, CTO$_{SIM}$ achieves regret upper bounded by $\mathbf{o}(\mathbf{1})$ (which is upper bounded by $\mu_{max}(1)$). Furthermore, T^*_{SIM} is upper bounded by the solution for the following optimization problem,*

$$\min T$$

$$s.t\begin{cases}T,b\in\mathbb{N}\cup\{0\},t\in\mathbb{N}^K\\ \forall b,\exists t:\begin{cases}\|t\|_1\leq T+b\\ t_i\geq\max_{\theta\in\Theta^*}\left\{m^*\left(\frac{1}{K(T+b)^2};\theta\right)\right\}\\ \mu\left(t_i+1;\theta_i^*\right)\leq\\ \quad\mu_{min}\left(\max_{\theta\in\Theta^*}\left\{m^*\left(\frac{1}{K(T+b)^2};\theta\right)\right\}\right)\end{cases}\end{cases}\tag{8}$$

***Proof**: See Appendix D.1.*

Remark 3.4 *Regret upper bounded by $o(1)$ is achieved by proving that w.p of at least $1-1/T$ the regret vanishes, and in case it does not, it is still bounded by a decaying term.*

Remark 3.5 *Given a function $U(\epsilon)$ that satisfies, $\forall n\geq U(\epsilon)$,*

$$bal\left(\max_{\theta_1\neq\theta_2\in\Theta^2}\left\{det^{\star\downarrow}_{\theta_1,\theta_2}\left(\frac{1}{16}\ln^{-1}(n)\right)\right\}\right)\leq\epsilon n$$

we have,

$$T^*_{SIM}\leq\frac{U\left(\left(K\sqrt{2K}\right)^{-1}\right)}{\sqrt{2K}}$$

as seen in Appendix D.1.

Algorithm 2 CTO$_{SIM}$

Input : K,Θ
Initialization : $N_i=0,\ \forall i\in[K]$
for $t=1,2,..,K$ **do**
  Ramp up : $i(t)=t$,and update $N_{i(t)}$
end for
for $t=K+1,...,$ **do**
  Detect : determine $\{\hat{\theta}_i\}$ by Eq. (5)
  Balance : $i(t)\in\mathrm{argmax}_{i\in[K]}\left\{\mu\left(N_i+1;\hat{\theta}_i\right)\right\}$
  Update : $N_{i(t)}\leftarrow N_{i(t)}+1$
end for

4. Differences Closest To Origin (ANV)

In the ANV case, the expected rewards are given by, $\mu_i^S(n;\theta_i^*)=\mu_i^c+\mu(n;\theta_i^*)$. Thereby, as opposed to the AV case, we need to estimate both the rotting model and the arm's constant term. The Differences Closest To Origin (D-CTO) for RB takes a two stage approach. First, detecting the underlying rotting models, then estimating and controlling the pulls due to the constant terms. We denote $a^*=\mathrm{argmax}_{i\in[K]}\{\mu_i^c\}$, and $\Delta_i=\mu_{a^*}^c-\mu_i^c$.

4.1. Models Detection

In order to detect the underlying rotting models, we cancel the influence of the constant terms. Once we do this, we can take a similar approach as for the AV case. Specifically, we define a criterion of proximity to the origin as follows: define the following set per arm,

$$Z(i,t;\Theta)=\left\{\left(\sum_{j=1}^{\left\lfloor\frac{N_i(t)}{2}\right\rfloor}r_j^i-\sum_{\left\lfloor\frac{N_i(t)}{2}\right\rfloor+1}^{n}r_j^i\right)-\left(\sum_{j=1}^{\left\lfloor\frac{N_i(t)}{2}\right\rfloor}\mu(j;\theta)-\sum_{\left\lfloor\frac{N_i(t)}{2}\right\rfloor+1}^{n}\mu(j;\theta)\right)\right\}_{\theta\in\Theta}.\tag{9}$$

The D-CTO approach is that in each decision point, we assume that the true underlying model corresponds to the following rule,

$$\hat{\theta}_i(t)=\underset{\theta\in\Theta}{\mathrm{argmin}}\{|Z(i,t;\theta)|\}\tag{10}$$

We refer to the optimization problem (6) of model detection, described for the AV case. We denote the solution for the ANV case (i.e., when we use proximity rule (10)) by $m^*_{\text{diff}}(p;\theta_i^*)$, and define $m^*_{\text{diff}}(p)\triangleq\max_{\theta\in\Theta}\{m^*_{\text{diff}}(p;\theta)\}$.

4.2. D-CTO$_{UCB}$

We next describe an approach with one decision point, and later on remark on the possibility of having a decision point at each time step. As explained above, after detecting the rotting models, we move to tackle the constant terms aspect of the expected rewards. This is done in a UCB1-like approach (Auer et al., 2002a). Given a sequence of rewards from arm i, $\{r_k^i\}_{k=1}^{N_i(t)}$, we modify them using the estimated rotting model $\hat{\theta}_i$, then estimate the arm's constant term, and finally choose the arm with the highest estimated expected reward, plus an upper confident term. i.e., at time t, we pull arm $i\left(t\right)$, according to the rule,

$$i\left(t\right) \in \underset{i\in[K]}{\operatorname{argmax}} \left\{ \hat{\mu}_i^c\left(t\right) + \mu\left(N_i\left(t\right)+1;\hat{\theta}_i\right) + c_{t,N_i(t)} \right\} \tag{11}$$

where $\hat{\theta}_i$ is the estimated rotting model (obtained in the first stage), and,

$$\begin{cases} \hat{\mu}_i^c\left(t\right) = \frac{\sum_{j=1}^{N_i(t)}\left(r_j^i - \mu\left(j;\hat{\theta}_i\right)\right)}{N_i(t)} \\ c_{t,s} = \frac{\left(8\ln(t)\mathcal{S}\left(\sigma^2,1,s\right)\right)^{1/2}}{s} \end{cases}$$

In a case of a tie in the UCB step, it may be arbitrarily broken. Pseudo algorithm for D-CTO$_{UCB}$ is given by Algorithm 3, accompanied with the following theorem.

Assumption 4.1 ***(D-Detection ability)***

$$\max_{\theta_1\neq\theta_2\in\Theta^2}\left\{ Ddet^{\star\downarrow}_{\theta_1,\theta_2}\left(\epsilon\right) \right\} \le B\left(\epsilon\right) < \infty, \quad \forall\epsilon > 0$$

The above assumption is similar in nature to Assumption 3.1, aside from using a different manner to distinguish. Specifically, the differences (in pulls) between the first and second halves of the models' sums of expected rewards. This holds, for instance, for Example 2.1 (Appendix F.3).

Theorem 4.1 *Suppose Assumptions 2.1, 2.2, and 4.1 hold. For $\delta \in (0,1)$, with probability of at least $1-\delta$, D-CTO$_{UCB}$ algorithm achieves regret bounded at time T by,*

$$\sum_{\substack{i\in[K]\\ i\neq a^*}} \left[\max\left\{ m^*_{diff}\left(\delta/K\right), \mu^{\star\downarrow}\left(\epsilon_i;\theta_i^*\right), m_{sep}\left(\ln\left(T\right)\right) \right\} \times \left(\Delta_i + \mu\left(1;\theta^*_{a^*}\right)\right) \right] + C\left(\Theta^*,\{\mu_i^c\}\right) \tag{12}$$

for any sequence $\epsilon_i \in (0,\Delta_i), \forall i \neq a^*$. Where $C\left(\Theta^*,\{\mu_i^c\}\right)$ is time independent, and $m_{sep}\left(\ln\left(T\right)\right)$ is the solution for,

$$\min m \quad s.t \begin{cases} m \in \mathbb{N} \\ \frac{m^2}{\mathcal{S}(\sigma^2,1,m)} \ge \frac{32\ln(T)}{(\Delta_i-\epsilon_i)^2} \end{cases} \tag{13}$$

Proof*: See Appendix E.1.*

Algorithm 3 D-CTO$_{UCB}$

Input : K, Θ, δ
Initialization : $N_i = 0,\ \forall i \in [K]$
Samples : $m^*_{\text{diff}}\left(\delta/K\right) \leftarrow$ sol. to (6) with $p = \delta/K$
for $t = 1,2,..,K\times m^*_{\text{diff}}\left(\delta/K\right)$ **do**
  Explore : $i\left(t\right)$ by Round Robin, and update $N_{i(t)}$
end for
Detect : determine $\{\hat{\theta}_i\}$ by Eq. (10)
for $t = K\times m^*_{\text{diff}}\left(\delta/K\right)+1, ...,$ **do**
  UCB : $i\left(t\right)$ according to Eq. (11)
  Update : $N_{i(t)} \leftarrow N_{i(t)}+1$
end for

Remark 4.1 *Instead of calculating $m^*_{diff}\left(\delta/K\right)$, it is possible to use any upper bound of it (e.g., as shown in Appendix E.1, $\max_{\theta_1\neq\theta_2\in\Theta^2}\left\{ Ddet^{\star\downarrow}_{\theta_1,\theta_2}\left(\frac{1}{8}\ln^{-1}\left(\frac{2K}{\delta}\right)\right)\right\}$ rounded to the nearest higher even number).*

Remark 4.2 *We note that both $m^*_{diff}\left(\delta/K\right)$ and $\mu^{\star\downarrow}\left(\epsilon_i;\theta_i^*\right)$ are independent of the horizon, and only depend on the structure and $\{\epsilon_i\}$. The only time horizon dependent term is $m_{sep}\left(\ln\left(T\right)\right)$. Moreover, for constant variance σ^2, we have $m_{sep}\left(\ln\left(T\right)\right) = \lceil 32\sigma^2\ln\left(T\right)/\left(\Delta_i-\epsilon_i\right)^2\rceil$ (hence the notation of $m_{sep}\left(\ln\left(T\right)\right)$, which reduces to the sub-Gaussian UCB1 horizon dependent bound, for $\{\epsilon_i\}\to 0$, as shown by Liu & Zhao (2011). Note that we cannot hope for a better rate than $\ln\left(T\right)$ as stochastic MAB is a special case of the ANV RB.*

Remark 4.3 *We can convert the D-CTO$_{UCB}$ algorithm to have a decision point in each step. Simply, at each time step, determine the rotting models according to proximity rule (10), followed by pulling an arm according to Eq. (11). We term this version D-CTO$_{SIM-UCB}$.*

5. Simulations

In this section we compare the performance of the CTO approach with benchmark algorithms in both the AV and the ANV cases.

Setups We set the number of arms to be $K = 10$, and chose the set of possible rotting models to follow the rule,

$$\mu\left(j;\theta\right) = \left(\textbf{int}\left(\frac{j}{100}\right)+1\right)^{-\theta}$$

where **int**$(\cdot)$ is the lower rounded integer, and $\Theta = \{0.1, 0.15, .., 0.4\}$ (i.e., plateaus of length 100, with decay between plateaus according to θ).

AV Case: Reward functions were chosen to be Normal distributions with randomly assigned rotting models as means and $\sigma^2(n) \equiv 0.2$. i.e., $r_j^i \sim \mathcal{N}(\mu(j; \theta_i^*), 0.2)$, where $\theta_i^* \in \Theta$ were randomly sampled with replacement from Θ, independently across arms and across trajectories. The Horizon was chosen to be $T = 20,000$.
ANV Case: Similar to the AV case, but only differs by the means being sums of randomly assigned rotting models and random constants drawn from $[0, 1]$. i.e., $r_j^i \sim \mathcal{N}(\mu_i^c + \mu(j; \theta_i^*), 0.2)$, where μ_i^c were randomly assigned, independently across arms and across trajectories. The Horizon was chosen to be $T = 40,000$.

Algorithms We implemented *Random*; which simply randomly chooses actions, *UCB1* by Auer et al. (2002a), *Discounted UCB* (DUCB) and *Sliding-Window UCB* (SWUCB) by Garivier & Moulines (2008), and EXP3S by Auer et al. (2002b). Finally, for the AV case we implemented CTO$_{SIM}$, and for the ANV case we implemented D-CTO$_{SIM-UCB}$.

Grid Searches were performed in order to determine the different algorithms' parameters. For *DUCB*, following Kocsis & Szepesvári (2006), the discount factor was chosen from $\gamma \in \{0.9, 0.99, .., 0.999999\}$. For *SWUCB* window size was chosen from $\tau \in \{1K, 2K, .., 16K\}$. For EXP3S, following Auer et al. (2002b), we chose $\gamma \in \{0.05, 0.1, .., 0.95\}$ and $\alpha \in \{1e-5, 2e-5, .., 1e-4\}$.

Performance For each of the cases, we present a plot of the averaged regret over 100 trajectories, specify the number of 'wins' of each algorithm over the others, and report the p-value of a paired T-test between the (end of trajectories) regrets of each pair of algorithms. For each trajectory and two algorithms, the *'winner'* is defined as the algorithm with the lesser regret at the end of the horizon.

5.1. AV Case Results

Parameters that were chosen by the grid search are as follows: DUCB's $\gamma = 0.999999$, SWUCB's $\tau = 8K$, EXP3S's $\gamma = 0.1$ and $\alpha = 7e-5$. We note that DUCB's performance improved as $\gamma \to 1$.

Results The averaged regret for the different algorithms is given by Figure 2. The upper half of Table 1 shows the number of 'wins' and p-values of the paired T-tests. The table is to be read as the following: the entries under the diagonal are the number of times the algorithms from the left column 'won' against the algorithms from the top row, and the entries above the diagonal are the p-values between the two (e.g., DUCB 'won' against UCB1 48 times, and their corresponding p-value is 0.085). We note that CTO$_{SIM}$ 'won' consistently (in every trajectory) against all other algorithms with very high significance.

5.2. ANV Case Results

Parameters that were chosen by the grid search are as follows: DUCB's $\gamma = 0.999999$, SWUCB's $\tau = 16K$, EXP3S's $\gamma = 0.1$ and $\alpha = 2e-5$. We note that SWUCB's performance improved with increasing the windows size, such that it performed as UCB1 until the time step which equaled the window size, and from there on deteriorated. The remark for DUCB stands as in the AV case.

Results The averaged regret for the different algorithms is given by Figure 3. The lower half of Table 1 shows the number of 'wins' and p-values of the paired T-tests, and is to be read as explained in the AV case. We note that CTO$_{SIM-UCB}$ 'won' consistently (in every trajectory) against all other algorithms with very high significance.

6. Related Work

We turn to reviewing related work while emphasizing the differences from our problem.

Stochastic MAB In the stochastic MAB setting (Lai & Robbins, 1985), the underlying reward distributions are stationary over time. The notion of regret is the same as in our work, but the optimal policy in this setting is one that pulls a fixed arm throughout the trajectory. The two most common approaches for this problem are constructing Upper Confidence Bounds (e.g., Auer et al. (2002a); Garivier & Cappé (2011); Maillard et al. (2011)), and Bayesian heuristics such as Thompson Sampling (e.g., Kaufmann et al. (2012); Agrawal & Goyal (2013); Gopalan et al. (2014)).

Adversarial MAB In the Adversarial MAB setting (also referred to as the Experts Problem, see the book of Cesa-Bianchi & Lugosi (2006) for a review), the sequence of rewards are selected by an adversary (i.e., can be arbitrary). In this setting the notion of *adversarial regret* is adopted (Auer et al., 2002b; Hazan & Kale, 2011), where the regret is measured against the best possible fixed action that could have been taken in hindsight. This is as opposed to the *policy regret* we adopt, where the regret is measured against the best sequence of actions in hindsight.

Hybrid models Some past work consider settings between the Stochastic and the Adversarial settings. Garivier & Moulines (2008) consider the case where the reward distributions remain constant over epochs and change arbitrarily at unknown time instants, similarly to Yu & Mannor (2009) who consider the same setting, only with the availability of side observations. Chakrabarti et al. (2009) consider the case where arms can expire and be replaced with new arms with arbitrary expected reward, but as long as an arm does not expire its statistics remain the same.

Table 1. Number of 'wins' and p-values between the different algorithms

		UCB1	DUCB	SWUCB	EXP3S	CTO
AV	UCB1		0.085	<1e-5	<1e-5	<1e-5
	DUCB	48		<1e-5	<1e-5	<1e-5
	SWUCB	75	79		<1e-5	<1e-5
	EXP3S	85	86	80		<1e-5
	CTO	**100**	**100**	**100**	**100**	
ANV	UCB1		0.055	<1e-5	<1e-5	<1e-5
	DUCB	40		<1e-5	< 1e-5	<1e-5
	SWUCB	6	3		<1e-5	<1e-5
	EXP3S	0	0	0		<1e-5
	CTO	**100**	**100**	**100**	**100**	

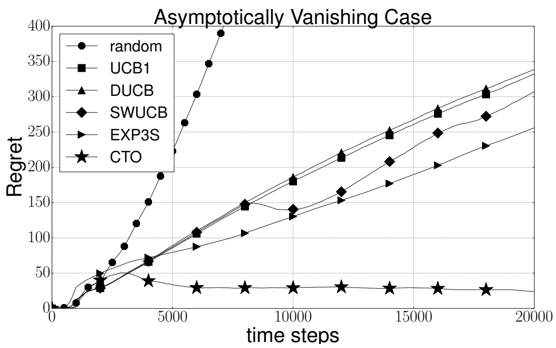

Figure 2. AV Case: Averaged regret over 100 trajectories

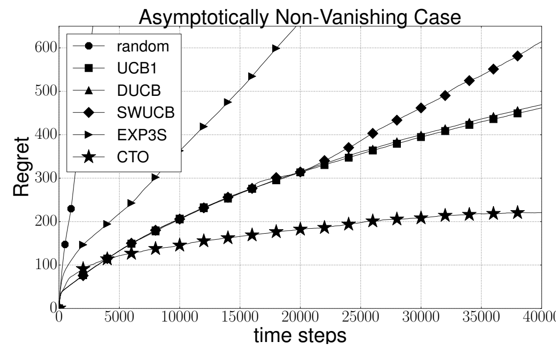

Figure 3. ANV Case: Averaged regret over 100 trajectories

Non-Stationary MAB Most related to our problem is the so-called Non-Stationary MAB. Originally proposed by Jones & Gittins (1972), who considered a case where the reward distribution of a chosen arm can change, and gave rise to a sequence of works (e.g., Whittle et al. (1981); Tekin & Liu (2012)) which were termed *Restless Bandits* and *Rested Bandits*. In the *Restless Bandits* setting, termed by Whittle (1988), the reward distributions change in each step according to a known stochastic process. Komiyama & Qin (2014) consider the case where each arm decays according to a linear combination of decaying basis functions. This is similar to our case in that the reward distributions decay, but differs fundamentally in that it belongs to the *Restless Bandits* setup (ours to the *Rested Bandits*), and in that there is a finite number of basis functions, together resulting in regret being bounded linearly by the number of basis functions times the square root of the horizon. More examples in this line of work are Slivkins & Upfal (2008) who consider evolution of rewards according to Brownian motion, and Besbes et al. (2014) who consider bounded total variation of expected rewards. In the *Rested Bandits* setting, only the reward distribution of a chosen arm changes, which is the case we consider. Heidari et al. (2016) consider the case where the reward decays (as we do), but with no statistical noise (deterministic rewards) which significantly simplifies the problem. Another somewhat closely related setting is suggested by Bouneffouf & Feraud (2016), in which statistical noise exists, but the expected reward shape is known up to a multiplicative factor.

7. Discussion

We introduced a novel variant of the *Rested Bandits* framework, which we termed *Rotting Bandits*. This setting deals with the case where the expected rewards generated by an arm decay as a function of pulls of that arm. This is motivated by many real-world scenarios.

We first tackled the asymptotically vanishing case, where the rewards decay to zero. We started by introducing an algorithm for ensuring, with high probability, zero regret. We then introduced algorithms for ensuring, in expectation, regret upper bounded by a term that decays to zero with the horizon.

We then tackled the asymptotically non-vanishing case, where each arm's rewards decay to some unknown constant. We introduced an algorithm for ensuring, with high probability, regret upper bounded by a horizon-dependent rate which is optimal for the stationary case.

We concluded with simulations that demonstrated our algorithms' superiority over benchmark algorithms, both for the AV and the ANV cases.

We finally note that it would be interesting to consider a non-parameterized Rotting Bandits setting, where there is no knowledge of the rotting models in advance.

References

Agarwal, Deepak, Chen, Bee-Chung, and Elango, Pradheep. Spatio-temporal models for estimating click-through rate. In *Proceedings of the 18th international conference on World wide web*, pp. 21–30. ACM, 2009.

Agrawal, Shipra and Goyal, Navin. Further optimal regret bounds for thompson sampling. In *Aistats*, pp. 99–107, 2013.

Arora, Raman, Dekel, Ofer, and Tewari, Ambuj. Online bandit learning against an adaptive adversary: from regret to policy regret. *arXiv preprint arXiv:1206.6400*, 2012.

Auer, Peter, Cesa-Bianchi, Nicolo, and Fischer, Paul. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a.

Auer, Peter, Cesa-Bianchi, Nicolo, Freund, Yoav, and Schapire, Robert E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

Awerbuch, Baruch and Kleinberg, Robert D. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pp. 45–53. ACM, 2004.

Besbes, Omar, Gur, Yonatan, and Zeevi, Assaf. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems*, pp. 199–207, 2014.

Bouneffouf, Djallel and Feraud, Raphael. Multi-armed bandit problem with known trend. *Neurocomputing*, 205:16–21, 2016.

Cesa-Bianchi, Nicolo and Lugosi, Gábor. *Prediction, learning, and games*. Cambridge university press, 2006.

Chakrabarti, Deepayan, Kumar, Ravi, Radlinski, Filip, and Upfal, Eli. Mortal multi-armed bandits. In *Advances in Neural Information Processing Systems*, pp. 273–280, 2009.

Du, Shan, Ibrahim, Mahmoud, Shehata, Mohamed, and Badawy, Wael. Automatic license plate recognition (alpr): A state-of-the-art review. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(2):311–325, 2013.

Garivier, Aurélien and Cappé, Olivier. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *COLT*, pp. 359–376, 2011.

Garivier, Aurélien and Moulines, Eric. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*, 2008.

Gopalan, Aditya, Mannor, Shie, and Mansour, Yishay. Thompson sampling for complex online problems. In *ICML*, volume 14, pp. 100–108, 2014.

Hazan, Elad and Kale, Satyen. Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12 (Apr):1287–1311, 2011.

Heidari, Hoda, Kearns, Michael, and Roth, Aaron. Tight policy regret bounds for improving and decaying bandits.

Jones, David Morian and Gittins, John C. *A dynamic allocation index for the sequential design of experiments*. University of Cambridge, Department of Engineering, 1972.

Kaufmann, Emilie, Korda, Nathaniel, and Munos, Rémi. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pp. 199–213. Springer, 2012.

Kleinberg, Robert and Leighton, Tom. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on*, pp. 594–605. IEEE, 2003.

Kocsis, Levente and Szepesvári, Csaba. Discounted ucb. In *2nd PASCAL Challenges Workshop*, pp. 784–791, 2006.

Komiyama, Junpei and Qin, Tao. Time-decaying bandits for non-stationary systems. In *International Conference on Web and Internet Economics*, pp. 460–466. Springer, 2014.

Lai, Tze Leung and Robbins, Herbert. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Liu, Keqin and Zhao, Qing. Extended ucb policy for multi-armed bandit with light-tailed reward distributions. *arXiv preprint arXiv:1112.1768*, 2011.

Maillard, Odalric-Ambrym, Munos, Rémi, Stoltz, Gilles, et al. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *COLT*, pp. 497–514, 2011.

Pandey, Sandeep, Agarwal, Deepak, Chakrabarti, Deepayan, and Josifovski, Vanja. Bandits for taxonomies: A model-based approach. In *SDM*, pp. 216–227. SIAM, 2007.

Robbins, Herbert. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, pp. 169–177. Springer, 1985.

Slivkins, Aleksandrs and Upfal, Eli. Adapting to a changing environment: the brownian restless bandits. In *COLT*, pp. 343–354, 2008.

Tekin, Cem and Liu, Mingyan. Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58(8):5588–5611, 2012.

Thompson, William R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Tran-Thanh, Long, Stein, Sebastian, Rogers, Alex, and Jennings, Nicholas R. Efficient crowdsourcing of unknown experts using multi-armed bandits. In *European Conference on Artificial Intelligence*, pp. 768–773, 2012.

Whittle, Peter. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, pp. 287–298, 1988.

Whittle, Peter et al. Arm-acquiring bandits. *The Annals of Probability*, 9(2):284–292, 1981.

Yu, Jia Yuan and Mannor, Shie. Piecewise-stationary bandit problems with side observations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1177–1184. ACM, 2009.

A. Hoeffding's Inequality for Sub-Gaussian RVs

Let $X_1, .., X_n$ be independent, mean-zero, σ_i^2-sub-Gaussian random variables. Then for all $t \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left\{-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right\} \tag{14}$$

B. Optimal Policy

B.1. Proof of Lemma 2.1

In this section we show that π^{max}, defined by Eq. (3) is an optimal policy for the RB problem.
Assume on the contrary, that π^{max} is not an optimal policy. Thus, there exists a time horizon, T, for which there exists some other policy π^{cand} that satisfies $J\left(T; \pi^{cand}\right) > J\left(T; \pi^{max}\right)$.
Let m be the first time step in which π^{cand} deviates from π^{max}, since $J\left(T; \pi^{cand}\right) > J\left(T; \pi^{max}\right)$ we infer that $m \leq T$ (i.e., there is such time step). Let $\tilde{\pi}$ be a policy defined by,

$$\tilde{\pi}(t) = \begin{cases} \pi^{cand}(t), & \text{if } t < m \\ \text{argmax}_{i\in[K]}\{\mu(N_i(m)+1; \theta_i^*)\}, & \text{if } t = m \\ \pi^{cand}(t-1), & \text{if } t > m \end{cases}$$

where if there exist more than one member in $\text{argmax}_{i\in[K]}\{\mu(N_i(m)+1; \theta_i^*)\}$, $\tilde{\pi}$ chooses the same action as π^{max}. That is, $\tilde{\pi}$ mimics π^{cand} until time step m, then plays according to argmax rule, and then re-mimics π^{cand}. Let μ_m, μ_T be the expected rewards of the arms that $\tilde{\pi}$ chose at the m^{th} time step, and that π^{cand} chose at the T^{th} time step, respectively. It is easy to see that,

$$J(T; \tilde{\pi}) - J\left(T; \pi^{cand}\right) = \mu_m - \mu_T \geq 0$$

where the second transition holds by the argmax rule combined with the assumption that the expected rewards are monotonically decreasing (assumption 2.2). Thus, $J(T; \tilde{\pi}) \geq J\left(T; \pi^{cand}\right)$. If we apply the above logic steps recursively, we obtain a series of policies with non-decreasing values of expected total reward $J(T; \cdot)$, where the series ends when there is no time step which deviates from $\pi^{\max}$, i.e., $J\left(T; \pi^{max}\right) \geq J\left(T; \pi^{cand}\right)$, in contradiction to π^{max} being non-optimal. Thus, we infer that π^{max} is indeed an optimal policy.

C. CTO$_{DB}$

C.1. Proof of Thm. 3.1

Finiteness of $m^*(\delta/K)$
We first show that $m^*(\delta/K)$ is finite. Define,

$$\theta_i^{'}(\tilde{m}) \triangleq \underset{\theta \neq \theta_i^*}{\text{argmin}}\left\{\left|\sum_{j=1}^{\tilde{m}} \mu(j; \theta_i^*) - \sum_{j=1}^{\tilde{m}} \mu(j; \theta)\right|\right\}$$

Thus we have, when we sample only from arm i,

$$\begin{aligned} P\left(\hat{\theta}_i(\tilde{m}) \neq \theta_i^*\right) &= P\left(\exists \theta \neq \theta_i^* : |Y(i, \tilde{m}; \theta)| \leq |Y(i, \tilde{m}; \theta_i^*)|\right) \\ &\leq P\left(\left|\sum_{j=1}^{\tilde{m}} r_j^i - \sum_{j=1}^{\tilde{m}} \mu(j; \theta_i^*)\right| > \frac{1}{2}\left|\sum_{j=1}^{\tilde{m}} \mu(j; \theta_i^*) - \sum_{j=1}^{\tilde{m}} \mu\left(j; \theta_i^{'}(\tilde{m})\right)\right|\right) \\ &\leq 2\exp\left\{-\frac{1}{8 \times det_{\theta_i^*, \theta_i^{'}(\tilde{m})}(\tilde{m})}\right\} \end{aligned}$$

where the first inequality holds by inclusion of events, and the second inequality holds by Eq. (14) and the definition of $det_{\theta_i^*, \theta_i^{'}}$.

By assumption 3.1, there exists a finite $\tilde{m}$, for which,

$$\max_{\theta_1 \neq \theta_2 \in \Theta^2} \left\{ det_{\theta_1, \theta 2}\left(\tilde{m}\right) \right\} \leq \frac{1}{8} \ln^{-1}\left(\frac{2K}{\delta}\right)$$

Therefore, if we plug $\tilde{m}$ back in to the above equation we get,

$$2 \exp\left\{ -\frac{1}{8 \times det_{\theta_i^*, \theta_i'}\left(\tilde{m}\right)} \right\} \leq \frac{\delta}{K}$$

Thus, we have a finite $\tilde{m}$ that satisfies the constraints of prob. (6) for $p = \delta/K$, and by definition $m^*\left(\delta/K\right) \leq \tilde{m}$. i.e., $m^*\left(\delta/K\right)$ is finite.

Finiteness of $T^*\left(\delta\right)$ and optimization problem characterization

Detection The first stage of the CTO$_{DB}$ algorithm is to solve prob. (6) to obtain $m^*\left(\delta/K\right)$, which we already showed to be finite. After which, for the first $K \times m^*\left(\delta/K\right)$ time steps the algorithm samples each arm $m^*\left(\delta/K\right)$ times, then applying the proximity rule (5) in order to detect the rotting models. By definition of $m^*\left(\delta/K\right)$, and the union bound, we have that w.p of at least $1-\delta$ the detected models are the true models. From now on we omit the $1-\delta$ probability assertion, and assume that the true models were indeed detected, while bearing in mind that further discussion only holds by that probability.

Balancing Let $\tilde{i} \triangleq \mathrm{argmin}_{i\in[K]}\{\mu\left(m^*\left(\delta/K\right); \theta_i^*\right)\}$ (in general a subset of arms). For some arm $i \in [K]$, let $t_i \in \mathbb{N} \cup \infty$ be defined as the solution to the following problem,

$$\begin{aligned} &\min v \\ &\text{s.t} \begin{cases} v \in \mathbb{N} \cup \infty \\ v \geq m^*\left(\delta/K\right) \\ \mu\left(v+1; \theta_i^*\right) \leq \mu\left(m^*\left(\delta/K\right); \theta_{\tilde{i}}^*\right) \end{cases} \end{aligned}$$

By the limit assumption 2.2 we have that $t_i < \infty$ (i.e., finite). Thus, we can change the set constraint to $t \in \mathbb{N}$ and the solution to the above optimization problem is the same. We note that for $i \in \tilde{i}$ we have that $t_i = m^*\left(\delta/K\right)$.

After pulling each arm $m^*\left(\delta/K\right)$ times, CTO$_{DB}$ algorithm follows argmax policy, i.e., pulls an arm with the highest expected reward in each time step. Following this rule and the monotonicity of the expected rewards, we have that after $T = \|t\|_1$ time steps, the total expected reward is given by,

$$J\left(T; \pi^{CTO_{DB}}\right) = \sum_{i\in[K]} \sum_{j=1}^{t_i} \mu\left(j; \theta_i^*\right)$$

where $\pi^{CTO_{DB}}$ is the policy induced by the CTO$_{DB}$ algorithm. This is easily inferred by noting that each t_i simply balances each arm's reward w.r.t $\mu\left(m^*\left(\delta/K\right); \theta_{\tilde{i}}^*\right)$, i.e., we only keep pulling an arm (until the defined time, T) if it yields strictly greater reward than $\mu\left(m^*\left(\delta/K\right); \theta_{\tilde{i}}^*\right)$.

Assume on the contrary that, $J\left(T; \pi^{max}\right) \neq J\left(T; \pi^{CTO_{DB}}\right)$. On the one hand, by Lemma 2.1, we have, $J\left(T; \pi^{max}\right) \geq J\left(T; \pi^{CTO_{DB}}\right)$. On the other hand, Let $\{s_i\}_{i\in[K]}$ be the set of the arms' number of pulls at time T following π^{max}, i.e.,

$$J\left(T; \pi^{max}\right) = \sum_{i\in[K]} \sum_{j=1}^{s_i} \mu\left(j; \theta_i^*\right)$$

It is easily seen that $J\left(T; \pi^{CTO_{DB}}\right) - J\left(T; \pi^{max}\right)$ is a sum of pairs in the form of, $\mu\left(l; \theta_i^*\right) - \mu\left(h; \theta_j^*\right)$ where $l \leq t_i$, and $h > t_j$, for $i \neq j \in [K]$. By definition of $\{t_i\}$ and the monotonicity assumption 2.2, we have that $\mu\left(l; \theta_i^*\right) \geq \mu\left(m^*\left(\delta/K\right); \theta_{\tilde{i}}^*\right)$, and $\mu\left(m^*\left(\delta/K\right); \theta_{\tilde{i}}^*\right) \geq \mu\left(h; \theta_j^*\right)$, resulting in $J\left(T; \pi^{CTO_{DB}}\right) \geq J\left(T; \pi^{max}\right)$.

Thus, we infer that for T, which we also showed to be finite, as a finite sum of finite terms, $J\left(T; \pi^{max}\right) = J\left(T; \pi^{CTO_{DB}}\right)$, i.e., the regret vanishes. Finally, we note that since the solution for prob. (7) is separable in i, $T = \|t\|_1$ defined here is exactly the solution to the problem defined in the theorem, $T^*\left(\delta\right)$.

Optimality from $T^*(\delta)$ onward
We showed optimality for time step T defined above. We next show optimality for $T+1$. We examine the two possible cases.
Case 1: $\forall i \in [K] : t_i = s_i$. Since CTO$_{DB}$ follows the $\arg\max$ rule as π^{max} does, we infer that arms with equal expected reward would be chosen by both CTO$_{DB}$ and π^{max}. Thereby, holding $J(T+1;\pi^{max}) = J\left(T+1;\pi^{CTO_{DB}}\right)$. i.e., zero regret as stated.
Case 2: $\exists i : t_i \neq s_i$. Therefore, there must be an arm, denoted as i_{gap}, for which $t_{i_{gap}} < s_{i_{gap}}$. By the $\arg\max$ rule, CTO$_{DB}$ chooses an arm i_{T+1} such that, $\mu\left(t_{i_{T+1}}+1;\theta^*_{i_{T+1}}\right) \geq \mu\left(t_{i_{gap}}+1;\theta^*_{i_{gap}}\right)$. By the monotonicity assumption 2.2, and the definition of π^{max}, since $s_{i_{gap}} \geq t_{i_{gap}}+1$, we have $\mu\left(s_{j_{T+1}};\theta^*_{j_{T+1}}\right) \leq \mu\left(s_{i_{gap}};\theta^*_{i_{gap}}\right) \leq \mu\left(t_{i_{gap}}+1;\theta^*_{i_{gap}}\right)$, where j_{T+1} is the arm chosen by π^{max}. Thus, on the one hand we have $J(T+1;\pi^{max}) \leq J\left(T+1;\pi^{CTO_{DB}}\right)$. On the other hand, by Lemma 2.1, we have $J(T+1;\pi^{max}) \geq J\left(T+1;\pi^{CTO_{DB}}\right)$. Combining the two, we have $J(T+1;\pi^{max}) = J\left(T+1;\pi^{CTO_{DB}}\right)$. i.e., zero regret as stated.
The above argument can be applied recursively for any $t > T$, thus establishing the theorem result.

C.2. Proof of Corollary 3.1.1

Bounding number of steps
Following the same logic for the proof of Thm. 3.1, we need to have that w.p of at least $1-1/T$, CTO$_{DB}$ algorithm satisfies,

$$\{\text{\# of steps for Detection}\} + \{\text{\# of steps for Balance}\} \leq T$$

{# of steps for Detection}, per arm, is given by $m^*\left(\frac{1}{KT}\right)$ (where we use the union bound), where from the proof of Thm. 3.1, we have that,

$$m^*\left(\frac{1}{KT}\right) \leq \max_{\theta_1,\theta_2}\left\{det^{\star\downarrow}_{\theta_1,\theta_2}\left(\frac{1}{8}\ln^{-1}(2KT)\right)\right\}$$

After pulling each arm $m^*\left(\frac{1}{KT}\right)$ times, the smallest possible obtained expected reward is given by $\mu_{min}\left(m^*\left(\frac{1}{KT}\right)\right)$. Thus, the number of additional pulls of an arm in order to "balance" it, i.e. {# of steps for Balance}, per arm, is upper bounded, by definition, by $bal\left(m^*\left(\frac{1}{KT}\right)\right) - m^*\left(\frac{1}{KT}\right)$.
By the monotonicity assumption 2.2, it follows that $bal()$ is monotonically increasing, and by definition we have $bal(n) \geq n$. Combining the above observations we have,

$$\{\text{\# of steps for Detection}\} + \{\text{\# of steps for Balance}\} \leq K \times bal\left(\max_{\theta_1,\theta_2}\left\{det^{\star\downarrow}_{\theta_1,\theta_2}\left(\frac{1}{8}\ln^{-1}(2KT)\right)\right\}\right)$$

Feasibility
Let $\epsilon = \left(2K^2\right)^{-1}$. By assumption 3.2, we have that there exists a finite $\tilde{T}_{max}$ for which,

$$\forall \hat{T} \geq \tilde{T}_{max} : bal\left(\max_{\theta_1\neq\theta_2\in\Theta^2}\left\{det^{\star\downarrow}_{\theta_1,\theta_2}\left(\frac{1}{8}\ln^{-1}\left(\hat{T}\right)\right)\right\}\right) \leq \epsilon\hat{T}$$

By defining $T = (2K)^{-1}\hat{T}$ and multiply both sides by K, we get,

$$\forall T \geq \frac{\tilde{T}_{max}}{2K} : K \times bal\left(\max_{\theta_1\neq\theta_2\in\Theta^2}\left\{det^{\star\downarrow}_{\theta_1,\theta_2}\left(\frac{1}{8}\ln^{-1}(2KT)\right)\right\}\right) \leq T$$

which implies,

$$\{\text{\# of steps for Detection}\} + \{\text{\# of steps for Balance}\} \leq T$$

as required. Thus, we infer that $\forall T \geq \tilde{T}_{max}/2K$, CTO$_{DB}$ algorithm achieves zero regret, w.p of at least $1-1/T$.

Misdetection
As for the case where we misdetect any arm, we have $\forall T \geq \tilde{T}_{max}/2K$,

$$\begin{aligned}
J\left(T;\pi^{max}\right) - J\left(T;\pi^{CTO_{DB}}\right) &= \sum_{i=1}^{K}\sum_{j=1}^{N_i^{max}(T)} \mu\left(j;\theta_i^*\right) - \sum_{i=1}^{K}\sum_{j=1}^{N_i^{CTO_{DB}}(T)} \mu\left(j;\theta_i^*\right) \\
&\leq \sum_{i=1}^{K} I_{\{N_i^{max}(T) > N_i^{CTO_{DB}}(T)\}} \sum_{N_i^{CTO_{DB}}(T)+1}^{N_i^{max}(T)} \mu\left(j;\theta_i^*\right) \\
&\leq \sum_{i=1}^{K} I_{\{N_i^{max}(T) > N_i^{CTO_{DB}}(T)\}} \sum_{m^*\left(\frac{1}{KT}\right)}^{N_i^{max}(T)} \max_{\theta\in\Theta^*}\left\{\mu\left(m^*\left(\frac{1}{KT};\theta\right)\right)\right\} \\
&\leq T \max_{\theta\in\Theta^*}\left\{\mu\left(m^*\left(\frac{1}{KT};\theta\right)\right)\right\}
\end{aligned}$$

where the first inequality holds by only considering cases where $N_i^{max}(T) > N_i^{CTO_{DB}}(T)$, and not the other way around. The second inequality holds by the monotonicity assumption 2.2.

Expectation
Combining the two above results we get that $\forall T \geq \tilde{T}_{max}/2K$,

$$\begin{aligned}
\mathcal{R}\left(T;\pi^{CTO_{DB}}\right) &= \mathcal{R}\left(T;\pi^{CTO_{DB}}|\text{true detection}\right) \times P\left(\text{true detection}\right) \\
&\quad + \mathcal{R}\left(T;\pi^{CTO_{DB}}|\text{misdetection}\right) \times P\left(\text{misdetection}\right) \\
&\leq \frac{1}{T} \times T \max_{\theta\in\Theta^*}\left\{\mu\left(m^*\left(\frac{1}{T};\theta\right)\right)\right\} = \max_{\theta\in\Theta^*}\left\{\mu\left(m^*\left(\frac{1}{T};\theta\right)\right)\right\}
\end{aligned}$$

We showed the result $\forall T \geq \tilde{T}_{max}/2K$, thus we infer that T^*_{DB}, for which the result holds, is upper bounded by $\tilde{T}_{max}/2K$, which is finite as stated.

D. CTO$_{SIM}$

D.1. Proof of Thm. 3.2

Bounding number of steps
We first characterize the bound, and later show feasibility (i.e., that the analysis we show here indeed holds within the horizon). In a similar manner to the proof for Thm. 3.1, we characterize an upper bound for T^*_{SIM} as an optimization problem, where we demand that each arm will be pulled enough times for truely detecting its rotting model w.h.p.
Let T be some *unknown* horizon. Given an hypothesize model, $\hat{\theta}_i$, for arm i, we term that arm 'saturated' if it has been pulled at least $m^*\left(\frac{1}{KT^2};\hat{\theta}_i\right)$ times (finite value by assumption 3.3, and the bound shown in the proof for Thm. 3.1). We assume that any 'saturated' arm is truely detected every time step, and omit this assertion from now on (we deal with the misdetection case later). i.e., we assume that once arm i hypothesize its rotting model to be $\hat{\theta}_i$ and also has been pulled at least $m^*\left(\frac{1}{KT^2};\hat{\theta}_i\right)$ times, then $\hat{\theta}_i = \theta_i^*$.
Since CTO$_{SIM}$ follows an argmax rule, combined with least # of pulls, we can bound the number of pulls of different arms, given the number of pulls of some other arm. Let s be the first time step for which $\min_{i[K]}\{N_i(s)\} = \max_{\theta\in\Theta^*}\{m^*\left(\frac{1}{KT^2};\theta\right)\}$ (this must occur at some point, noted later, as we next bound the number of pulls of the other arms). We have, by our assumption, that all the arms have been correctly detected at this point, and we have that for any arm j, $N_j(s)$ can be upper bounded by the following optimization problem,

$$\begin{aligned}
&\min t_j \\
&\text{s.t} \begin{cases} t_j \in \mathbb{N} \\ t_j \geq \max_{\theta\in\Theta^*}\{m^*\left(\frac{1}{KT^2};\theta\right)\} \\ \mu\left(t_j+1;\theta_i^*\right) \leq \mu_{min}\left(\max_{\theta\in\Theta^*}\left\{m^*\left(\frac{1}{KT^2};\theta\right)\right\}\right) \end{cases}
\end{aligned}$$

where the above optimization bound characterization holds trivially for any arm $j \in \text{argmin}_{i\in[K]}\{N_i(s)\}$. For any arm $j \notin \text{argmin}_{i\in[K]}\{N_i(s)\}$, this holds by that, according to the argmax rule and assumption 2.2, j would not be pulled such that $\mu\left(N_j(s);\theta_j^*\right) < \mu_{min}\left(\max_{\theta\in\Theta^*}\{m^*\left(\frac{1}{KT^2};\theta\right)\}\right)$, and by the tie breaking policy (least # of pulls), j would not be pulled such that its expected reward will be equal to $\mu_{min}\left(\max_{\theta\in\Theta^*}\{m^*\left(\frac{1}{KT^2};\theta\right)\}\right)$. Combining these last observations results in the stated upper bound. We note that by assumption 2.2, the above bound is finite, hence s is indeed a finite time step.

Since at time s all the arms are 'saturated', hence correctly detected, similarly to the proof of Thm. 3.1, a "balancing" stage takes place where $\mu_{min}\left(\max_{\theta\in\Theta^*}\left\{m^*\left(\frac{1}{KT^2};\theta\right)\right\}\right)$ serves as a lower bound that no arm's expected reward will cross (or even get to, if it wasn't its value after $\max_{\theta\in\Theta^*}\{m^*\left(\frac{1}{KT^2};\theta\right)\}$ pulls) before balancing out (thus achieving zero regret). The above statement holds since we showed that no arm crosses this value before all the arms are correctly detected, combined with the fact that the algorithm follows an argmax rule and least # of pulls for tie breaking, and additionally combined with assumption 2.2. From that point (zero regret), following the same arguments as in the proof of Thm 3.1, the algorithm maintains zero regret.

From all the observations above, we infer that the total number of steps before truely detecting and balancing the arms (achieve zero regret) is upper bounded by,

$$\min \|t\|_1$$
$$\text{s.t}\begin{cases} t\in\mathbb{N}^K \\ t_i \geq \max_{\theta\in\Theta^*}\{m^*\left(\frac{1}{KT^2};\theta\right)\}, \quad \forall i\in[K] \\ \mu\left(t_i+1;\theta_i^*\right) \leq \mu_{min}\left(\max_{\theta\in\Theta^*}\left\{m^*\left(\frac{1}{KT^2};\theta\right)\right\}\right), \quad \forall i\in[K]\end{cases}$$

If it happens to be that $\|t\|_1 \leq T$, then for that T, CTO$_{SIM}$ will achieve zero regret. Since we require that the result will hold from some T^*_{SIM} onward we need the above characterization to also hold for any $\tilde{T}\geq T$. We thereby infer that the smallest T such that for any $\tilde{T}\geq T$, there exists $\|t\|_1\leq\tilde{T}$ for which the above stated result holds (i.e., the solution to the optimization problem is indeed $\leq\tilde{T}$), can serve as an upper bound for T^*_{SIM}, achieving the stated optimization problem characterization.

Feasibility

In a similar manner to the proof of Corollary. 3.1.1, we wish to obtain,

$$\{\text{\# of steps for Detection}\} + \{\text{\# of steps for Balance}\} \leq T$$

where we require that the detection of each arm is w.p of at least $1-\frac{1}{KT^2}$. We note, that unlike the case in Corollary. 3.1.1, there is no distinction between the Detection and the Balance parts. Define $q(T) \triangleq \max_{\theta_1,\theta_2}\left\{det^{\star\downarrow}_{\theta_1,\theta_2}\left(\frac{1}{16}\ln^{-1}\left(\sqrt{2K}T\right)\right)\right\}$. In a similar manner to the proof for Thm. 3.1, we have that, after pulling an arm for $q(T)$ times, the probability of misdetection of its rotting model $\leq\frac{1}{KT^2}$. We refer to an arm that has been pulled at least $q(T)$ times as 'strongly saturated'. From now on we will assume that any 'strongly saturated' arm is truely detected at each decision point, and will discuss the other case later on.

On the one hand, by the definition of $bal()$, the monotonicity assumption 2.2, and the rule of tie breaking applied by CTO$_{SIM}$, we have that all arms become 'strongly saturated' after, at most, $q(T)+(K-1)\times bal(q(T))$ time steps.

On the other hand, from the definition of $bal()$, and CTO$_{SIM}$, we infer that no arm would be pulled $bal(q(T))+1$ times before all other arms would become 'strongly saturated'.

Combining the two above observations we have that, after at most, $q(T)+(K-1)\times bal(q(T))$ time steps, there exists a time step in which all arms have became 'strongly saturated', but were not pulled more than $bal(q(T))$ times. From that point, the number of pulls (in total, including former pulls) required in order to "balance" the arms, is bounded by $K\times bal(q(T))$. That is under the worst case scenario, where every arm that becomes 'strongly saturated' is detected to be an arm that requires $bal(q(T))$ pulls to "balance" itself w.r.t to another 'strongly saturated' arm. Thus, we infer that,

$$\{\text{\# of steps for Detection}\} + \{\text{\# of steps for Balance}\} \leq K\times bal(q(T))$$

Let $\epsilon = \left(K\sqrt{2K}\right)^{-1}$. Similarly to the proof for Corollary. 3.1.1, by assumption 3.3, we have that there exists a finite $\tilde{T}_{max}$ for which,

$$\forall \tilde{T} \geq \tilde{T}_{max} : bal\left(\max_{\theta_1 \neq \theta_2 \in \Theta^2}\left\{det^{\star\downarrow}_{\theta_1,\theta_2}\left(\frac{1}{16}\ln^{-1}\left(\tilde{T}\right)\right)\right\}\right) \leq \epsilon\tilde{T}$$

We denote $T = \left(\sqrt{2K}\right)^{-1}\tilde{T}$, and get,

$$\forall T \geq \frac{\tilde{T}_{max}}{\sqrt{2K}} : K \times bal\left(q\left(T\right)\right) \leq T$$

which implies, under true detection, that $\forall T \geq \tilde{T}_{max}/\sqrt{2K}$, CTO$_{SIM}$ algorithm achieves zero regret.
Since by definition we have $\forall \theta \in \Theta : m^*\left(\frac{1}{KT^2};\theta\right) \leq m^*\left(\frac{1}{KT^2}\right)$, and similarly to what was shown in Thm. 3.1 we have $m^*\left(\frac{1}{KT^2}\right) \leq q\left(T\right)$, we infer that there exists (a finite) T that holds the optimization problem characterization as stated above (i.e., $\forall \tilde{T} \geq T$ the optimization problem holds feasibility).

Misdetection and Expectation
So far, we assumed that each 'saturated' (or 'strongly saturated') arm is truely detected. By definition each 'saturated' (or 'strongly saturated') arm probability of misdetection in any time step is upper bounded by $1/KT^2$. Thereby, after all the arms are 'saturated', the probability of a misdetection in each time step is upper bounded by $1/T^2$. The number of time steps where all the arms are 'saturated' (referred to as the 'saturated step') is trivially bounded by T. Hence, the probability that a misdetection occurs after the 'saturated step' is bounded by $1/T$. Meaning that $\forall T \geq T^*_{SIM}$, CTO$_{SIM}$ achieves zero regret w.p of at least $1 - 1/T$.
Next, we note that,

$$\begin{aligned}
T &= \sum_{i=1}^{K} N_i\left(T\right) \\
&\leq \min_{i\in[K]} N_i\left(T\right) + \left(K-1\right)\max_{i\in[K]} N_i\left(T\right) \\
&\leq \min_{i\in[K]} N_i\left(T\right) + \left(K-1\right) \times bal\left(\min_{i\in[K]} N_i\left(T\right)\right) \\
&\leq K \times bal\left(\min_{i\in[K]} N_i\left(T\right)\right)
\end{aligned}$$

Thereby, by assumption 2.2, we infer that $\min_{i\in[K]} N_i\left(T\right) \overset{T\to\infty}{\longrightarrow} \infty$. Thus, by applying expectation over events (true detection or not), and similarly to Corollary 3.1.1, we get,

$$\mathcal{R}\left(T;\pi^{CTO_{SIM}}\right) \leq \max_{\theta\in\Theta^*}\left\{\mu\left(\min_{i\in[K]}\{N_i\left(T\right)\};\theta\right)\right\}$$

which is $\in o\left(1\right)$, and trivially $\leq \mu_{max}\left(1\right)$.

E. D-CTO$_{UCB}$

E.1. Proof of Thm. 4.1

Decomposing the regret
We note that we can upper bound the regret by,

$$\begin{aligned}\mathcal{R}(T) &= \sum_{i=1}^{K}\sum_{j=1}^{\mathbb{E}[N_i^*(T)]}\mu_i^S(j;\theta_i^*) - \sum_{i=1}^{K}\sum_{j=1}^{\mathbb{E}[N_i^\pi(T)]}\mu_i^S(j;\theta_i^*)\\
&\le \underbrace{\sum_{i\neq a^*}\sum_{j=1}^{\mu^{\star\downarrow}(\Delta_i;\theta_i^*)}\mu^S(j;\theta_i^*)}_{\triangleq\tilde{C}(\Theta^*,\{\mu_i^c\})} + \sum_{j=1}^{T}\mu^S(j;\theta_{a^*}^*) - \sum_{i=1}^{K}\sum_{j=1}^{\mathbb{E}[N_i^\pi(T)]}\mu_i^S(j;\theta_i^*)\\
&= \tilde{C}(\Theta^*,\{\mu_i^c\}) + \sum_{\mathbb{E}\left[N_{a^*}^\pi(T)\right]+1}^{T}\mu^S(j;\theta_{a^*}^*) - \sum_{i\neq a^*}\sum_{j=1}^{\mathbb{E}[N_i^\pi(T)]}\mu_i^S(j;\theta_i^*)\\
&\le \tilde{C}(\Theta^*,\{\mu_i^c\}) + \sum_{\mathbb{E}\left[N_{a^*}^\pi(T)\right]+1}^{T}(\mu_{a^*}^c + \mu(1;\theta_{a^*}^*)) - \sum_{i\neq a^*}\sum_{j=1}^{\mathbb{E}[N_i^\pi(T)]}\mu_i^c\\
&\le \tilde{C}(\Theta^*,\{\mu_i^c\}) + \sum_{i\neq a^*}\mathbb{E}\left[N_i^\pi(T)\right]\times(\Delta_i + \mu(1;\theta_{a^*}^*))\end{aligned}$$

where $\mathbb{E}\left[N_i^*(T)\right]$ is the expected number of pulls of arm i at time T induced by the optimal policy (which maximizes Eq. (1)), and $\mathbb{E}N_i^\pi(T)$ is the expected number of pulls induced by policy π (specifically, we now consider the policy induced by D-CTO$_{UCB}$). The first inequality holds by keeping in mind that π^{max} (which pulls according to argmax rule) is optimal, thus any arm $i\neq a^*$ would not be pulled after yielding expected reward not bigger than $\mu_{a^*}^c$, according to the behavior of $\mu(\cdot;\cdot)$ by assumption 2.2.

Detecting the models
This part is very similar to its counterpart in the CTO$_{DB}$ proof. We start by showing that $m_{\text{diff}}^*(\delta/K)$ is finite. Define,

$$\theta_i^{'}(\tilde{m}) \triangleq \operatorname*{argmin}_{\theta\neq\theta_i^*}\left\{\left|\mathcal{D}(\mu(\cdot;\theta_i^*),1,\tilde{m}) - \mathcal{D}(\mu(\cdot;\theta),1,\tilde{m})\right|\right\}$$

Thus, we have, when we sample only from arm i, and for an even $\tilde{m}$

$$\begin{aligned}P\left(\hat{\theta}_i(\tilde{m})\neq\theta_i^*\right) &= P\left(\exists\theta\neq\theta_i^* : |Z(i,\tilde{m};\theta)| \le |Z(i,\tilde{m};\theta_i^*)|\right)\\
&\le P\left(\left|\left(\sum_{j=1}^{\frac{\tilde{m}}{2}}r_j^i - \sum_{j=\frac{\tilde{m}}{2}+1}^{\tilde{m}}r_j^i\right) - \mathcal{D}(\mu(\cdot;\theta_i^*),1,\tilde{m})\right| > \right.\\
&\qquad\left.\frac{1}{2}\left|\mathcal{D}(\mu(\cdot;\theta_i^*),1,\tilde{m}) - \mathcal{D}\left(\mu\left(\cdot;\theta_i^{'}(\tilde{m})\right),1,\tilde{m}\right)\right|\right)\\
&\le 2\exp\left\{-\frac{1}{8\times Ddet_{\theta_i^*,\theta_i^{'}(\tilde{m})}(\tilde{m})}\right\}\end{aligned}$$

where the first inequality holds by inclusion of events, and the second inequality holds by Eq. (14), the definition of $Ddet_{\theta_i^*,\theta_i^{'}}$, and noting that for an even $\tilde{m}$ we have,

$$\mathbb{E}\left[\sum_{j=1}^{\frac{\tilde{m}}{2}}r_j^i - \sum_{j=\frac{\tilde{m}}{2}+1}^{\tilde{m}}r_j^i\right] = \mathcal{D}(\mu(\cdot;\theta_i^*),1,\tilde{m})$$

By assumption 4.1, there exists a finite, even, $\tilde{m}$ for which,

$$\max_{\theta_1 \neq \theta_2 \in \Theta^2} \left\{ Ddet_{\theta_1, \theta_2}\left(\tilde{m}\right) \right\} \leq \frac{1}{8} \ln^{-1}\left(\frac{2K}{\delta}\right)$$

Thus, if we plug $\tilde{m}$ back to the above equation we get,

$$2\exp\left\{ -\frac{1}{8 \times Ddet_{\theta_i^*, \theta_i'(\tilde{m})}\left(\tilde{m}\right)} \right\} \leq \frac{\delta}{K}$$

Thus, we have a finite $\tilde{m}$ that satisfies the constraints of Prob. (6) for $p = \delta/K$, and by definition $m^*_{\text{diff}}\left(\delta/K\right) \leq \tilde{m}$. i.e., $m^*_{\text{diff}}\left(\delta/K\right)$ is finite.

Bounding number of pulls

We wish to bound $\mathbb{E}\left[N_i^\pi\left(T\right)\right]$ for all $i \neq a^*$. Remember that in the Detection part, we pull each arm $m^*_{\text{diff}}\left(\delta/K\right)$ times, hence,

$$N_i^\pi\left(T\right) = m^*_{\text{diff}}\left(\delta/K\right) + \sum_{t=K \times m^*_{\text{diff}}(\delta/K)+1}^{T} 1_{\{i(t)=i\}}$$

where $1_{\{\cdot\}}$ is the indicator function. Similarly to the proof of UCB1 ((Auer et al., 2002a)) we have,

$$N_i^\pi\left(T\right) \leq l_i + \sum_{t=1}^{\infty} \sum_{s=m^*_{\text{diff}}(\delta/K)}^{t-1} \sum_{s_i=l_i}^{t-1} 1_{\left\{\hat{\mu}^c_{a^*}(s)+\mu\left(s;\theta^*_{a^*}\right)+c_{t,s} \leq \hat{\mu}^c_i(s_i)+\mu\left(s_i;\theta^*_i\right)+c_{t,s_i}\right\}}$$

where for some $\epsilon_i \in (0, \Delta_i)$, we denote $l_i = \max\left\{ m^*_{\text{diff}}\left(\delta/K\right), \mu^{\star\downarrow}\left(\epsilon_i;\theta^*_i\right), m_{sep}\left(\ln\left(T\right)\right) \right\}$, and we note that we assume that we have detected the true underlying rotting models (holds w.p of at least $1-\delta$ as shown above).

The above indicator function holds when at least one of the following holds,

$$\begin{cases} \hat{\mu}^c_{a^*}\left(s\right) \leq \mu^c_{a^*} - c_{t,s} \\ \hat{\mu}^c_i\left(s_i\right) \geq \mu^c_i + c_{t,s_i} \\ \mu^c_{a^*} + \mu\left(s;\theta^*_{a^*}\right) < \mu^c_i + \mu\left(s_i;\theta^*_i\right) + 2c_{t,s_i} \end{cases}$$

Plugging $c_{t,s}$ and c_{t,s_i}, and using Eq. (14), we have,

$$\begin{cases} P\left(\hat{\mu}^c_{a^*}\left(s\right) \leq \mu^c_{a^*} - c_{t,s}\right) = t^{-4} \\ P\left(\hat{\mu}^c_i\left(s_i\right) \geq \mu^c_i + c_{t,s_i}\right) = t^{-4} \end{cases}$$

And for $s_i \geq l_i$ we have,

$$\begin{aligned} \mu^c_{a^*} + \mu\left(s;\theta^*_{a^*}\right) - \mu^c_i - \mu\left(s_i;\theta^*_i\right) - 2c_{t,s_i} &\geq \mu^c_{a^*} - \mu^c_i - \mu\left(s_i;\theta^*_i\right) - 2c_{t,s_i} \\ &\geq \mu^c_{a^*} - \mu^c_i - \epsilon_i - 2c_{t,s_i} \\ &= \left(\Delta_i - \epsilon_i\right) - 2c_{t,s_i} \\ &\geq 0 \end{aligned}$$

where the first inequality holds by assumption 2.2, the second inequality by $s_i \geq \mu^{\star\downarrow}\left(\epsilon_i;\theta^*_i\right)$, and the third inequality by $s_i \geq m_{sep}\left(\ln\left(T\right)\right)$ and the definition of $m_{sep}\left(\ln\left(T\right)\right)$.

Thus, combining the above observations, we get,

$$\begin{aligned} \mathbb{E}\left[N_i^\pi\left(T\right)\right] &\leq l_i + \sum_{t=1}^{\infty} \sum_{s=m^*_{\text{diff}}(\delta/K)}^{t-1} \sum_{s_i=l_i}^{t-1} \left(P\left(\hat{\mu}^c_{a^*} \leq \mu^c_{a^*} - c_{t,s}\right) + P\left(\hat{\mu}^c_i \geq \mu^c_i + c_{t,s_i}\right)\right) \\ &\leq l_i + \frac{\pi^2}{3} \end{aligned}$$

Denoting $C\left(\Theta^*, \{\mu^c_i\}\right) = \tilde{C}\left(\Theta^*, \{\mu^c_i\}\right) + \sum_{i \neq a^*} \frac{\pi^2}{3}\left(\Delta_i + \mu\left(1;\theta^*_{a^*}\right)\right)$, and plugging back into the upper bound on the regret, we achieve the stated result.

F. Example 2.1

The example that we next show how the different assumptions hold for, is the case where $\Theta = \{\theta_1, \theta_2, ..\}$ is a discrete (possibly infinite) set, and $\forall \theta \in \Theta : 0.01 \leq \theta \leq 0.49$. The rewards are sub-Gaussians random variables, with expected rewards given by $\mu(n; \theta_i^*)$, where $\theta_i^* \in \Theta$, and constant variances σ^2.

F.1. Assumption 3.1

The assumption is given by,

$$\max_{\theta_1 \neq \theta_2 \in \Theta^2} \left\{ det^{\star\downarrow}_{\theta_1,\theta_2}(\epsilon) \right\} \leq M(\epsilon) < \infty, \quad \forall \epsilon > 0$$

Without a loss of generality, assume $\theta_2 > \theta_1$. We have for large enough n,

$$\begin{aligned}
det_{\theta_1,\theta_2}(n) &= \frac{n\sigma^2}{\left(\sum_{j=1}^{n} j^{-\theta_1} - \sum_{j=1}^{n} j^{-\theta_2}\right)^2} \\
&\leq \frac{n\sigma^2}{\left(c_1 n^{1-\theta_1} - c_1 - c_2 n^{1-\theta_2}\right)^2} \\
&= \frac{n\sigma^2}{c_1^2 n^{2-2\theta_1} + c_2^2 n^{2-2\theta_2} - 2c_1c_2 n^{2-\theta_1-\theta_2} - 2c_1^2 n^{1-\theta_1} + 2c_1c_2 n^{1-\theta_2} + c_1^2} \\
&\leq \frac{n\sigma^2}{\tilde{c} n^{2-2\theta_1}} \\
&= \frac{\bar{c}}{n^{1-2\theta_1}}
\end{aligned}$$

where $\{c_1, c_2, \tilde{c}, \bar{c}\}$ are constants (independent of n). The first inequality holds by bounding the sums by integrals and keeping in mind that $\theta_2 > \theta_1$ combined with $0.01 \leq \theta \leq 0.49$. The second inequality holds from large enough n (leading exponent, depends only on $\{\theta_1, \theta_2\}$, but finite). Next, we have,

$$\frac{\bar{c}}{n^{1-2\theta_1}} < \epsilon \Longrightarrow n > \left(\frac{\bar{c}}{\epsilon}\right)^{\frac{1}{1-2\theta_1}} > \left(\frac{\bar{c}}{\epsilon}\right)^{50}$$

We have that $\forall \epsilon > 0$, the RHS is finite, thereby, by definition, $\max_{\theta_1 \neq \theta_2 \in \Theta^2} \left\{ det^{\star\downarrow}_{\theta_1,\theta_2}(\epsilon) \right\}$ is finite, i.e., the assumption holds.

F.2. Assumptions 3.2 and 3.3

The two assumptions are very similar, where Assumption 3.3 is stricter, hence, we will show that it holds for that one. The assumption is given by,

$$bal\left(\max_{\theta_1 \neq \theta_2 \in \Theta^2} \left\{ det^{\star\downarrow}_{\theta_1,\theta_2}\left(\frac{1}{16}\ln^{-1}(\zeta)\right) \right\}\right) \in o(\zeta)$$

Without a loss of generality assume $\theta_2 > \theta_1$. Following the same arguments as in F.1, we have that for large enough n,

$$det_{\theta_1,\theta_2}(n) \leq \frac{\bar{c}}{n^{1-2\theta_1}}$$

where $\bar{c}$ is a constant (independent of n). Next, we have,

$$\frac{\bar{c}}{n^{1-2\theta_1}} < \frac{1}{16}\ln^{-1}(\zeta) \Longrightarrow n > \left(16\bar{c}\ln(\zeta)\right)^{\frac{1}{1-2\theta_1}} > \left(16\bar{c}\ln(\zeta)\right)^{50}$$

Meaning that ζ large enough,

$$\max_{\theta_1 \neq \theta_2 \in \Theta^2} \left\{ det^{\star\downarrow}_{\theta_1,\theta_2}\left(\frac{1}{16}\ln^{-1}(\zeta)\right) \right\} < \left(16\bar{c}\ln(\zeta)\right)^{50}$$

Next, we have,

$$\alpha^{-0.1} \leq x^{-0.49} \Longrightarrow \alpha \geq x^{4.9}$$

Hence, $bal\left(x\right) = x^{4.9}$. Since $bal\left(\cdot\right)$ is monotonically increasing, we have that for ζ large enough,

$$bal\left(\max_{\theta_1 \neq \theta_2 \in \Theta^2}\left\{det^{\star\downarrow}_{\theta_1,\theta_2}\left(\frac{1}{16}\ln^{-1}\left(\zeta\right)\right)\right\}\right) < \hat{c}\ln^{245}\left(\zeta\right)$$

where $\hat{c}$ is a constant (independent of ζ). Finally, we note that,

$$\lim_{\zeta\to\infty}\frac{\ln^{245}\left(\zeta\right)}{\zeta} = 0$$

Thus we infer that the assumption holds.

F.3. Assumption 4.1

The assumption is given by,

$$\max_{\theta_1 \neq \theta_2 \in \Theta^2}\left\{Ddet^{\star\downarrow}_{\theta_1,\theta_2}\left(\epsilon\right)\right\} \leq B\left(\epsilon\right) < \infty, \quad \forall\epsilon > 0$$

Without a loss of generality, assume $\theta_2 > \theta_1$. We have for large enough n,

$$\begin{aligned}
Ddet_{\theta_1,\theta_2}\left(n\right) &= \frac{n\sigma^2}{\left(\left(\sum_{j=1}^{\left\lfloor\frac{n}{2}\right\rfloor} n^{-\theta_1} - \sum_{j=\left\lfloor\frac{n}{2}\right\rfloor+1}^{n} j^{-\theta_1}\right) - \left(\sum_{j=1}^{\left\lfloor\frac{n}{2}\right\rfloor} n^{-\theta_2} - \sum_{j=\left\lfloor\frac{n}{2}\right\rfloor+1}^{n} j^{-\theta_2}\right)\right)^2} \\
&\leq \frac{n\sigma^2}{\left(c_1\left(-1+2\left\lfloor\frac{n}{2}\right\rfloor^{1-\theta_1} - n^{1-\theta_1}\right) - c_2\left(2\left(\left\lfloor\frac{n}{2}\right\rfloor+1\right)^{1-\theta_2} - n^{1-\theta_2}\right)\right)^2} \\
&\leq \frac{n\sigma^2}{\tilde{c}n^{2-2\theta_1}} \\
&= \frac{\tilde{c}}{n^{1-2\theta_1}}
\end{aligned}$$

where $\{c_1, c_2, \tilde{c}\}$ are constants (independent of n). The inequalities hold by the same arguments as in F.1. Again, following the same logic as the end of F.1, we have that the assumption holds.