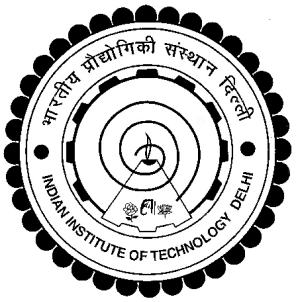


Dept of Electrical Engineering
Indian Institute of Technology Delhi

Course Code : ELD457
Academic Year : 2022 - 2023, Semester - II

Linear Rotting Bandits

Nirjhar Das **2019EE30585**
Het Patel **2019EE10484**
Rohan Sharma **2019EE30121**

Advisor: Prof. Arpan Chattopadhyay

Abstract: In this work, we explore the domain of linear bandits with rotting rewards. Multi-Armed Bandit (MAB) is a well-studied and useful framework sequential decision making under uncertainty. A more general class of bandits is the Linear Bandits, where the rewards are linear in the feature vector encoding each arm. Although linear bandits can capture contextual information, the real world is hardly stationary. Thus, we consider the class of linear bandits whose rewards are non-increasing over time. This problem can be potentially applied in modelling user boredom in recommender systems and in medical drug selection wherein the effectiveness of a drug decreases due to repeated use. Although rotting rewards have been considered in the MAB setting, they have not been studied in the linear case. To the best of our knowledge, this is the first work to consider linear bandits with rotting rewards. We provide a new algorithm that exploits the problem structure and performs better than the SOTA algorithm for non-stationary linear bandits and rotting MAB algorithm. We also attempt to provide a regret analysis for the proposed algorithm and discuss the regret lower bound.

1 Introduction

Multi-Armed Bandits are a class of mathematical models designed to address exploration-exploitation trade-offs involved in sequential learning tasks. These models consider the constraints involved in settings where only limited feedback is provided by the environment. The property of adaptability to changes in the value of actions over time has been one of the most sought-after features in bandit research since its early stages. Adaptability to changes in the "value" of actions over time has been one of the most sought-after features in bandit research. This is because the assumption of stationarity is generally significantly limiting for most realistic scenarios.

However, consideration of non-stationarity is generally very difficult from the perspective of designing and analyzing the performance of strategies for optimal performance. Therefore, several attempts range from considering worst-case scenarios (adversarial bandit setting [1]) to considering certain symmetries in the structure of non-stationary environments [8, 3].

Non-stationary environments are studied broadly in two categories - rested bandits, where the change of rewards results from the agent's actions [7], and restless bandits, where the change occurs independently over time [8]. Furthermore, it is generally assumed that the change in rewards is a non-increasing function of time or the agent's actions depending on the setting considered.

These symmetries offer simplifications that help model several realistic scenarios. For instance, consider the case of online recommendations where items are presented sequentially to a user. On the one hand, this problem can be modeled as a linear MAB problem [7] where the features of each arm can be used to calculate the user' preferences. On the other hand, it may be modeled as an adversarial bandit problem where the worst-case scenario is considered [1]. However, a user's preferences will unlikely remain stable over time, and algorithms for adversarial bandits are too conservative to accurately model the problem's randomness. It is a reasonable assumption that the effectiveness of all advertisements would deteriorate over time because of boredom. This can be modeled as a linear rotting bandit problem in the restless scenario.

In this work, we consider the linear rotting MAB (as in the case of online advertisement), a stochastic linear bandit model where the available actions correspond to arbitrary context vectors whose associated rewards follow a non-stationary linear regression model - such that the expected reward of an arm decreases over time. Building upon previous work in related settings, we shall present an algorithm that exploits the symmetries associated with the problem.

2 Related Works

There is a considerable body of research dedicated to the study of online learning in changing environments. In this section, we shall limit our discussion to stochastic linear bandit models and their analysis in simple non-stationary environments. One of the first works to tackle the problem of non-stationarity in the context of stochastic linear MABs was [2]. Defining d as the problem dimension, B_T as the variation budget, and T as the total time horizon, the authors established a a minimax lower bound of $\Omega(d^{2/3}B_T^{1/3}T^{2/3})$. They propose the SW-UCB algorithm based on the sliding window least square, achieving an optimal dynamic regret of $\tilde{O}(d^{2/3}(B_T+1)^{1/3}T^{2/3})$. [4] proposed to replace the sliding window least square with the weighted least square in the d-LinUCB algorithm and also proved that this approach attains the same dynamic regret. Taking this one step further, [9] showed that instead of weighted penalty or sliding window approaches requiring memory buffers, simple restart strategies on UCB-type algorithms are sufficient to attain the same dynamic regret.

In the context of rotting bandits, [3] was the first paper to discuss the formulation of the rotting bandits model by drawing motivation from real-world scenarios. They also propose a simple algorithm accompanied

by theoretical guarantees. One of the most popular approaches for rotting bandit algorithms is the Filtering On Expanding Window Average (FEWA) algorithm proposed in [5]. The algorithm constructs moving averages of increased windows in order to identify arms that are less likely to return low (rotten) rewards when pulled once more. The algorithm is also shown to achieve an instance independent regret bound of $\mathcal{O}(\sqrt{KT})$ and an instance dependent regret bound of $\mathcal{O}(log(KT))$.

3 Notations

In linear bandits, at time t, the decision maker is presented with a set of arms $\mathcal{X}_t \subseteq \mathbb{R}^d$, and she has to play an arm $\mathbf{x}_t \in \mathcal{X}_t$. Upon playing the arm, she receives a reward $r_t = \mathbf{x}^\intercal \theta_t + \eta_t$ where $\theta_t \in \mathbb{R}^d$ is an unknown preference vector that varies with time, and η_t is an additive noise. The dynamic regret over time horizon T is defined as

$$R(T) = \left[\sum_{t=1}^{T} \max_{\mathbf{x} \in \mathcal{X}_t} \mathbf{x}^\intercal \theta_t\right] - \sum_{t=1}^{T} \mathbf{x}_t^\intercal \theta_t \tag{1}$$

. Further, let $\mathbf{x}_t^* = \arg\max_{\mathbf{x} \in \mathcal{X}_t} \mathbf{x}^\intercal \theta_t$. The sets $\mathcal{F}_t = \{\mathbf{x}_1, r_1, \dots \mathbf{x}_{t-1}, r_{t-1}, \mathbf{x}_t\}$ form a filtration. Now, we make the following assumptions:

Assumption 1: We assume that η_t is R-sub Gaussian, that is, it satisfies $\mathbb{E}[\eta_t | \mathcal{F}_t] = 0$ and $\mathbb{E}[e^{\lambda \eta_t} | \mathcal{F}_t] \leq \exp(\lambda^2 R^2/2)$.

Assumption 2: Let us denote by $A \in \mathbb{R}^{K \times d}$ the matrix whose rows represent the K arms. Thus, $A = [\mathbf{a}_1, \dots \mathbf{a}_K]^\intercal$ where $a_i \in \mathbb{R}^d$ is the i-th arm. Now, we assume that for all arms $\mathbf{x} \in \mathcal{X}_t$, $\|\mathbf{x}\| \leq L$ and for all time t, $\|\theta_t\| \leq S$, where $L, S > 0$ are constants.

Assumption 3: Further, we assume that the variation on θ_t is restricted by a budget B_T, that is,

$$\sum_{t=1}^{T-1} \|\theta_t - \theta_{t+1}\| \leq B_T \tag{2}$$

These are standard assumptions in non-stationary linear bandits and are necessary to provide regret guarantees.

4 Problem Formulation

We deal with the problem of linear bandits with rotting rewards. Now there can be two cases defined as follows:

1. **Rested Rotting**: The reward decreases only for the arm that is pulled. In other words if arm $\mathbf{a}$ is pulled at time t, then, $\mathbf{a}^\intercal \theta_t \leq \mathbf{a}^\intercal \theta_{t+1}$ and $\mathbf{x}^\intercal \theta_t = \mathbf{x}^\intercal \theta_{t+1}$ $\forall$ $\mathbf{x} \neq \mathbf{a}$.

2. **Restless Rotting**: The reward is non-increasing for all arms irrespective of which arm is pulled. Thus, $\mathbf{x}^\intercal \theta_t \leq \mathbf{x}^\intercal \theta_{t+1}$. For the arm matrix A, this can be written as:

$$A\theta_t \leq A\theta_{t-1}, \ \forall \ t \tag{3}$$

where the inequality is coordinate-wise.

This is a new condition we place on the problem and thus our problem is a special case of non-stationary linear bandits. However, this additional condition can be specifically useful in modelling non-stationary effects which are non-increasing in trend. For example, we can consider the arms to encode movies and the preference vector to model user preference over time. A user watching movies over a long period will generally feel bored to watch more movies. If the reward is in getting the user to watch a recommended movie, then obviously it is decreasing. Moreover, with this additional assumption, we hope to provide a regret upper bound better than that of the non-stationary linear bandit.

In the following section, we present the analysis of the problem along with the main technical results and an algorithm for the problem.

5 Analysis

5.1 Characterization

Lemma 5.1 *It is possible for a problem instance of linear rested rotting bandits with arms such that* $\mathbf{x}_i^\intercal \mathbf{x}_j = 0 \forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$ *(uncorrelated arms) to be equivalent to a problem instance of rested rotting MAB problem.*

Proof: Using the Gram-Schmidt orthonormalization process, we can say that there exists an orthonormal basis $\Psi = (\psi_1, \psi_2 \dots \psi_d)$ such that for $\mathbf{x}_i \in \mathcal{X}$; there exists $z_i \in \mathbb{R}$ and an $i \in [d]$ such that $\mathbf{x}_i = z_i \cdot \psi_i$. Suppose we project θ_t on Ψ. Let this be written as $\theta_t = \sum_{i=1}^{d} \xi_i^t \cdot \psi_i$, $\xi_i \in \mathbb{R}\ \forall\, i \in [d]$. Then,

$$\begin{aligned} \mu_j^t \equiv \langle \theta_t, \mathbf{x}_j \rangle &= \langle \sum_{i=1}^{d} \xi_i^t \cdot \psi_i,\ z_j \cdot \psi_j \rangle \\ &= \sum_{i=1}^{d} \langle \xi_i^t \cdot \psi_i,\ z_j \cdot \psi_j \rangle = \xi_j^t \cdot z_j \end{aligned} \tag{4}$$

Since the mean reward of arm j depends on ξ_j (j-th coordinate of θ_t in Ψ-space) only, it can be possible to have a problem instance where the decrease in the rewards is independent across arms. Then this problem instance can be mapped to a rested rotting MAB problem where pulling arm i reveals nothing about arm j for $i \neq j$.

Note that for the condition of 5.1 to hold, we $K \leq d$. This is not a particularly interesting case as in real life, the arm set is very large and we want to take advantage of the encoding of the arms in a lower dimensional space. Moreover, [5, 6] solve this problem with optimal regret.

Lemma 5.2 *In a linear rotting bandit setting, if the arm set contains correlated (not uncorrelated) arms then the problem is necessarily restless. This is generally the case*

Proof: We prove by contradiction. Suppose the problem is rested. Now, since arms are not uncorrelated, the arms are linearly dependent, which means $\exists\ (\alpha_1, \dots, \alpha_K)$, with some α_i's non-zero such that, $\sum_{i=1}^{K} \alpha_i \mathbf{a}_i = \mathbf{0}$. Let us pick an arm $\mathbf{a}_m$ such that $\alpha_m \neq 0$. Let $\mathbf{a}_m = \sum_{p=1}^{d} z_p^m \cdot \psi_p$ and let $z_1^m \neq 0$ without loss of generality.

Now, let $\theta_t = \sum_{p=1}^{d} \xi_p^t \cdot \psi_p$. So we have $\langle \theta_t, \mathbf{a}_m \rangle = \sum_{p=1}^{d} \xi_p^t \cdot z_p^m$. Let $\xi_1^{t+1} < \xi_1^t$ without loss of generality. Thus, $\langle \theta_{t+1}, \mathbf{a}_m \rangle < \langle \theta_t, \mathbf{a}_m \rangle$. Now, if we have $z_1^j = 0\ \forall\ j \neq m$, $j \in [K]$, then this change in θ_t gives rise to a rested rotting bandit. However, this condition can not hold since then, we have

$$\alpha_m \cdot \mathbf{a}_m = \sum_{i=1, i \neq m}^{K} -\alpha_i \cdot \mathbf{a}_i \tag{5}$$

Along ψ_1, this equation becomes:

$$\begin{aligned} \alpha_m \cdot z_1^m &= -\sum_{i=1, i\neq m}^{K} \alpha_i \cdot z_1^i = 0 \\ &\implies z_1^m = 0 \because \alpha_m \neq 0 \\ &\implies \text{Contradiction } \because z_1^m \neq 0 \end{aligned} \tag{6}$$

Thus, we see that only restless rotting linear bandits are possible in the case of correlated arms.

Hereafter, we focus on the restless linear rotting bandit case for correlated arms, as we believe that this models the real world problems better. Moreover, this case is less restrictive and thus can be useful in more scenarios.

5.2 An Optimization Perspective

Since we already know that the non-stationary effect is such that the reward means are decreasing, a possible approach can be to use this condition as a constraint to frame an optimization problem whose solution gives the estimates of θ_s for s in some range. Specifically, consider the following optimization problem,

$$\begin{aligned} \min_{\{\hat{\theta}_s : t-w \leq s < t\}} \quad & \frac{1}{2} \sum_{s=t-w}^{t-1} \Big(\langle \mathbf{x}_s, \hat{\theta}_s \rangle - r_s \Big)^2 + \lambda \sum_{s=t-w}^{t-1} \|\hat{\theta}_s\|^2 \\ \text{s.t.} \quad & A\hat{\theta}_s \geq A\hat{\theta}_{s+1} \quad \forall\ s \in \{t-w, \ldots, t-2\} \end{aligned} \tag{7}$$

The $\hat{\theta}_s, t-w \leq s \leq t$ at the optimal point gives a set of estimates which obey the rotting condition as well as tries to minimize the least squared error over a window of size w (a hyperparameter). Although an analytical solution to this problem is intractable, we can find numerical solution, though at a computational cost. The estimate $\hat{\theta}_{t-1}$ can then be used to devise a simple algorithm for picking an arm: $\mathbf{x}_t = \arg\max_{\mathbf{x} \in \mathcal{X}_t} \mathbf{x}^\intercal \hat{\theta}_{t-1}$.

5.3 Index Algorithms

Here we discuss some possible index-based algorithms that solve the problem. In algorithm 1 we list a few algorithms we formulated. We base our algorithm on RestartUCB [9] which uses a simple epoch based algorithm that restarts after every fixed number of steps. This algortihm achieves the SOTA performance both in practice and in theory. We modify the algorithm to suit our case. The main idea is to find a stricter bound for the confidence set. By results from [2], we see that the confidence interval for non-stationary linear bandits is defined as follows:

$$\begin{aligned} |\mathbf{x}^T(\hat{\theta}_t - \theta_t)| \leq\ & \beta_t \|\mathbf{x}\|_{V_{t-1}^{-1}} \\ & + L^2 \sqrt{\frac{(t-t_0)d}{\lambda}} \sum_{s=t_0}^{t-1} \|\theta_s - \theta_{s+1}\| \end{aligned} \tag{8}$$

from which we derive the Upper Confidence Bound as

$$\begin{aligned} \mathbf{x}^T\theta_t \leq\ & \mathbf{x}^T\hat{\theta}_t + \beta_t \|\mathbf{x}\|_{V_{t-1}^{-1}} \\ & + L^2 \sqrt{\frac{(t-t_0)d}{\lambda}} \sum_{s=t_0}^{t-1} \|\theta_s - \theta_{s+1}\| \end{aligned} \tag{9}$$

Define $I(\mathbf{x},t) \equiv \mathbf{x}^T\hat{\theta}_t + \beta_t\|\mathbf{x}\|_{V_{t-1}^{-1}}$. Thus, we have the following for time t and $t+1$:

$$\begin{aligned}
\mathbf{x}^T\theta_t &\leq I(\mathbf{x},t) + L^2\sqrt{\frac{(t-t_0)d}{\lambda}}\sum_{s=t_0}^{t-1}\|\theta_s - \theta_{s+1}\| \\
\mathbf{x}^T\theta_{t+1} &\leq I(\mathbf{x},t+1) + L^2\sqrt{\frac{(t-t_0+1)d}{\lambda}}\sum_{s=t_0}^{t}\|\theta_s - \theta_{s+1}\| \\
&= I(\mathbf{x},t+1) + L^2\sqrt{\frac{(t-t_0+1)d}{\lambda}}\sum_{s=t_0}^{t-1}\|\theta_s - \theta_{s+1}\| \\
&\quad + L^2\sqrt{\frac{(t-t_0+1)d}{\lambda}}\|\theta_t - \theta_{t+1}\|
\end{aligned} \tag{10}$$

Now, by rotting condition, we have $\mathbf{x}^T\theta_{t+1} \leq \mathbf{x}^T\theta_t$. Thus, if we have $I(\mathbf{x},t+1) > I(\mathbf{x},t)$, we have that the confidence bound of the t-th time step is smaller than the confidence bound at $t+1$-th time step. As a result, if we use the index $U(\mathbf{x},t) = \min_{s\in[t_0,t]} I(\mathbf{x},s)$, we have a tighter upper bound on the arm $\mathbf{x}$.

Thus, we pick the arm $j_t = \arg\max_j U(\mathbf{a}_j,t)$. This step brings in optimism while definition $U(\cdot,\cdot)$ ensures that the upper bounds on the arms are as tight as possible. This can possibly be beneficial as will be evident from the analysis below. Now, if at time t, we have $U(\mathbf{a}_j,t-1) > I(\mathbf{a}_j,t)$ we need not necessarily update $U(\mathbf{a}_j,t)$ as we see that the confidence bound at time t also accumulates the sum of the θ drift terms $\|\theta_{s-1} - \theta_s\|$. Thus, we use the update rule shown in 1 for policy 2 to update $U(\cdot,\cdot)$.

5.4 Regret Upper Bound

Let $\mathbf{x}_t^*$ be the best arm at time t and $\mathbf{x}_t$ be the arm actually chosen by the algorithm. Let κ_i's be the time instants when the indices get updated within an epoch and $w_i = \kappa_i - \kappa_{i-1}$. We have $\kappa_0 = 1$ and $\kappa_h = H$. Also let $P(t_1,t_2) = \sum_{s=t_1}^{t_2-1}\|\theta_s - \theta_{s+1}\|$. Finally let $\tau_t = \max\{\kappa_i : \kappa_i \leq t\}$. Then, regret at time t is given by

$$(\mathbf{x}_t^* - \mathbf{x}_t)^T\theta_t \tag{11}$$

Now let us list down some inequalities of interest:

$$\begin{aligned}
&\mathbf{x}^T\hat{\theta}_t - \beta_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}} - L^2\sqrt{\frac{td}{\lambda}}P(0,t-1) \leq \mathbf{x}^T\theta_t \leq \mathbf{x}^T\hat{\theta}_t + \beta_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}} + L^2\sqrt{\frac{td}{\lambda}}P(0,t-1) \\
&\mathbf{x}^T\theta_{t_1} \leq \mathbf{x}^T\theta_{t_2}\ \forall t_1 > t_2 \\
&\mathbf{a}_i^T\hat{\theta}_{t_i} + \beta_{\tau_t-1}\|\mathbf{a}_i\|_{V_{\tau_t-1}^{-1}} \leq \mathbf{a}_i^T\hat{\theta}_t + \beta_{t-1}\|\mathbf{a}_i\|_{V_{t-1}^{-1}} + \varepsilon_t \\
&U(\mathbf{x},t) = I(\mathbf{x},\tau_t)\forall\mathbf{x}\in\mathcal{X} \\
&\mathbf{x}_t^T\hat{\theta}_{\tau_t} + \beta_{\tau_t-1}\|\mathbf{x}_t\|_{V_{\tau_t-1}^{-1}} \geq \mathbf{x}_t^{*T}\hat{\theta}_{\tau_t} + \beta_{\tau_t-1}\|\mathbf{x}_t^*\|_{V_{\tau_t-1}^{-1}} \ \because \mathbf{x}_t = \mathbf{x}_{\tau_t} \text{ by algorithm}
\end{aligned} \tag{12}$$

where $\alpha = L^2\sqrt{d/\lambda}$. Let $L(\mathbf{x},t) = \langle \mathbf{x}, \hat{\theta}_{\tau_t}\rangle - \beta_{\tau_t - 1}\|\mathbf{x}\|_{V_{\tau_t-1}^{-1}}\}$ and $I'(\mathbf{x},t) \equiv \langle \mathbf{x}, \hat{\theta}_t\rangle - \beta_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}}$. Let us further define the following event:

$$\begin{aligned} C_t^1 &= \{\forall \mathbf{x} \in \mathcal{X} : I'(\mathbf{x},t) - \varepsilon_t \le L(\mathbf{x}, t-1)\} \\ C_t^2 &= \{\forall \mathbf{x} \in \mathcal{X} : I(\mathbf{x},t) + \varepsilon_t \ge U(\mathbf{x}, t-1)\} \\ C_t^3 &= \{I(\mathbf{x}_{\tau_t}^*, t) \le I(\mathbf{x}_t, t)\} \\ C_t &= C_t^1 \cap C_t^2 \cap C_t^3 \end{aligned} \tag{13}$$

This set can be thought of as a *Good Set*.

The instantaneous regret is given by

$$\begin{aligned} r_t &= (\mathbf{x}_t^* - \mathbf{x}_t)^T \theta_t \\ &\le I(\mathbf{x}_{\tau_t}^*, t) - I'(\mathbf{x}_t, t) + 2\sqrt{\tau_t} P(0, \tau_t) + 2\varepsilon_t \quad [\because I(\mathbf{x}_{\tau_t}^*, t) \le I(\mathbf{x}_t, t)] \\ &\le 2\left(\beta_{t-1}\|\mathbf{x}_t\|_{V_{t-1}^{-1}} + \alpha\sqrt{\tau_t} P(0,\tau_t) + \varepsilon_t\right) \\ &\le 2\left(\beta_{t-1}\|\mathbf{x}_t\|_{V_{t-1}^{-1}} + \alpha\sqrt{\tau_t} P(0,H-1) + \varepsilon_t\right) \end{aligned} \tag{14}$$

The epoch regret of the j-th epoch is thus given by $R(\mathcal{E}_j) = \sum_{t=t_0}^{t_0+H-1} r_t$. We thus have,

$$\begin{aligned} R(\mathcal{E}_j) &\le \sum_{t=t_0}^{t_0+H-1} 2\left(\beta_{t-1}\|\mathbf{x}_t\|_{V_{t-1}^{-1}} + \alpha\sqrt{\tau_t} P(0,H-1) + \varepsilon_t\right) \\ &\le 2\alpha P(\mathcal{E}) \sum_{t=t_0}^{t_0+H-1} \sqrt{\tau_t} + 2\sum_{t=t_0}^{t_0+H-1} (\beta_{t-1}\|\mathbf{x}_t\|_{V_{t-1}^{-1}} + \varepsilon_t) \\ &\le \beta_H \sqrt{2dH\log(1 + HL^2/\lambda d)} + 2\int_0^H \varepsilon_t dt + 2\alpha P(\mathcal{E}) \sum_{t=t_0}^{t_0+H-1} \sqrt{\tau_t} \end{aligned} \tag{15}$$

Taking $\varepsilon_t = c/\sqrt{t}$ where t is the time index within the epoch, we have,

$$R(\mathcal{E}_j) \le \beta_H \sqrt{2dH\log(1 + HL^2/\lambda d)} + c\sqrt{H} + 2\alpha P(\mathcal{E}_j) \sum_{t=t_0}^{t_0+H-1} \sqrt{\tau_t} \tag{16}$$

The total regret is now:

$$\begin{aligned} R_T &= \sum_{i=1}^{\lceil T/H \rceil} R(\mathcal{E}_i) \\ &\le \beta_H \frac{T}{\sqrt{H}} \sqrt{2d\log(1 + HL^2/\lambda d)} + \frac{cT}{\sqrt{H}} + \sum_{\text{all epochs } \mathcal{E}} 2\alpha P(\mathcal{E}) \sum_{t=t_0}^{t_0+H-1} \sqrt{\tau_t} \\ &= \beta_H \frac{T}{\sqrt{H}} \sqrt{2d\log(1 + HL^2/\lambda d)} + \frac{cT}{\sqrt{H}} + \sum_{\text{all epochs } \mathcal{E}} 2\alpha P(\mathcal{E}) \sum_{i=1}^{h} w_i \sqrt{\kappa_{i-1}} \end{aligned} \tag{17}$$

The main challenge is in analyzing the random variables τ_t. At worst, we can use $\tau_t \leq H$ and get back the regret of non-stationary linear bandits.

Let us try to analyze the value of $w_i = \kappa_i - \kappa_{i-1}$.

$$\begin{aligned} \mathbb{P}[w_i \geq n | \kappa_{i-1}] &= \mathbb{P}\left[\bigcap_{s=\kappa_{i-1}}^{\kappa_{i-1}+n-1} C_s\right] \\ \implies \mathbb{P}[w_i < n | \kappa_{i-1}] &= \mathbb{P}\left[\bigcup_{s=\kappa_{i-1}}^{\kappa_{i-1}+n-1} \overline{C}_s^1 \cup \overline{C}_s^2 \cup \overline{C}_s^3\right] \\ &\leq \sum_{s=\kappa_{i-1}}^{\kappa_{i-1}+n-1} \sum_{j=1}^{3} \mathbb{P}[\overline{C}_s^j] \end{aligned} \tag{18}$$

Thus, for $\kappa_{i-1} < t \leq \kappa_i$, the necessary equations are:

$$\begin{aligned} V_t = V_{\kappa_{i-1}} + \sum_{s=\kappa_{i-1}}^{t} \mathbf{x}_{\kappa_i} \mathbf{x}_{\kappa_i}^T, \qquad S_t = S_{\kappa_{i-1}} + \sum_{s=\kappa_{i-1}}^{t} r_s \mathbf{x}_{\kappa_i}, \qquad r_s = \langle \mathbf{x}_{\kappa_i}, \theta_s \rangle + \eta_s \\ \hat{\theta}_t = V_{t-1}^{-1} S_{t-1}, \qquad \hat{\theta}_{\kappa_{i-1}} = V_{\kappa_{i-1}-1}^{-1} S_{\kappa_{i-1}-1} \end{aligned} \tag{19}$$

By using some matrix identities, assumin that C_{t-1} holds, we get the following recursion at time t with τ_{t-1} written as τ,

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \frac{V_{t-2}^{-1} \mathbf{x}_\tau \mathbf{x}_\tau^T (\theta_{t-1} - \hat{\theta}_{t-1})}{1 + \|\mathbf{x}_\tau\|_{V_{t-2}^{-1}}^2} + \left(1 - \|\mathbf{x}_\tau\|_{V_{t-2}^{-1}}^2\right) \eta_{t-1} V_{t-2}^{-1} \mathbf{x}_\tau \tag{20}$$

Further, since C_{t-1} holds, we have that $\forall \mathbf{y} \in \mathcal{X}$, we have,

$$\begin{aligned} \mathbf{y}^T \hat{\theta}_{t-1} + \beta_{t-2} \|\mathbf{y}\|_{V_{t-2}^{-1}} + \varepsilon_{t-2} \geq U(\mathbf{y}, \tau) \\ \mathbf{y}^T \hat{\theta}_{t-1} - \beta_{t-2} \|\mathbf{y}\|_{V_{t-2}^{-1}} - \varepsilon_{t-2} \leq L(\mathbf{y}, \tau) \end{aligned} \tag{21}$$

Let us denote by $z(\mathbf{y}, t-1) = \beta_{t-2} \|\mathbf{y}\|_{V_{t-2}^{-1}} + \varepsilon_{t-2}$. Now, multiplying by $\mathbf{y}^T$ in eq.(20), we get,

$$\mathbf{y}^T \hat{\theta}_t = \mathbf{y}^T \hat{\theta}_{t-1} + \frac{\mathbf{y}^T V_{t-2}^{-1} \mathbf{x}_\tau \mathbf{x}_\tau^T (\theta_{t-1} - \hat{\theta}_{t-1})}{1 + \|\mathbf{x}_\tau\|_{V_{t-2}^{-1}}^2} + \left(1 - \|\mathbf{x}_\tau\|_{V_{t-2}^{-1}}^2\right) \eta_{t-1} \mathbf{y}^T V_{t-2}^{-1} \mathbf{x}_\tau \tag{22}$$

Let us further use the notation $\omega(\mathbf{y}, t-1) = \frac{\mathbf{y}^T V_{t-2}^{-1} \mathbf{x}_\tau \mathbf{x}_\tau^T (\theta_{t-1} - \hat{\theta}_{t-1})}{1 + \|\mathbf{x}_\tau\|_{V_{t-2}^{-1}}^2}$ and $\delta(\mathbf{y}, t-1) = (1 - \|\mathbf{x}_\tau\|_{V_{t-2}^{-1}}^2) \mathbf{y}^T V_{t-2}^{-1} \mathbf{x}_\tau$.

Thus, eq.(22)now becomes $\mathbf{y}^T \hat{\theta}_t = \mathbf{y}^T \hat{\theta}_{t-1} + \omega(\mathbf{y}, t-1) + \delta(\mathbf{y}, t-1) \eta_{t-1}$. Thus, we have by using inequalities in eq.(21),

$$U(\mathbf{y}, \tau) - z(\mathbf{y}, t-1) \leq \mathbf{y}^T \hat{\theta}_t - \omega(\mathbf{y}, t-1) - \delta(\mathbf{y}, t-1) \eta_{t-1} \leq L(\mathbf{y}, \tau) + z(\mathbf{y}, t-1) \tag{23}$$

From this, we get that,

$$\begin{aligned} U(\mathbf{y}, \tau) - z(\mathbf{y}, t-1) + z(\mathbf{y}, t) \leq I(\mathbf{y}, t) + \varepsilon_t - \omega(\mathbf{y}, t-1) - \delta(\mathbf{y}, t-1) \eta_{t-1} \\ L(\mathbf{y}, \tau) + z(\mathbf{y}, t-1) - z(\mathbf{y}, t) \geq I'(\mathbf{y}, t) - \varepsilon_t - \omega(\mathbf{y}, t-1) - \delta(\mathbf{y}, t-1) \eta_{t-1} \end{aligned} \tag{24}$$

Algorithm 1 Strict UCB

Require: Epoch size H, Arm matrix $A \in \mathbb{R}^{K\times d}$, Regularizer λ, Tolerance ε, Update Frequency h

Set Epoch counter $j = 1$
for $j = 1, 2, \ldots, \lceil T/H \rceil$ **do**
  $\tau = (j-1)H$
  $V_\tau = \lambda I_d$, $S_\tau = \mathbf{0}_d$
  Initialize $U(i,\tau) = L(i,\tau) = \beta_\tau \|\mathbf{a}_i\|_{V_\tau^{-1}}$ $\forall\ i \in [K]$
  for $t = \tau+1, \tau+2, \ldots \tau+H-1$ **do**
    Calculate $\hat{\theta}_t = V_{t-1}^{-1} S_{t-1}$
    Calculate β_{t-1} with $t_0 = \tau$
    Define $I(i,t) = \langle \mathbf{a}_i, \hat{\theta}_t \rangle + \beta_{t-1} \|\mathbf{a}_i\|_{V_{t-1}^{-1}}$ $\forall\ i \in [K]$
    Update with either policy 1, 2 or 3 (to be decided later which one to use):
    Policy 1: $\forall\ i \in [K],\ U(i,t) = \begin{cases} U(i,t-1), & \text{if } U(i,t-1) < I(i,t) + \varepsilon \\ I(i,t) & \text{otherwise} \end{cases}$
    Policy 2: $\big(U(i,t), L(i,t)\big) = \begin{cases} \big(I(i,t), I'(i,t)\big), & \text{if } C_t \text{ doesn't hold} \\ \big(U(i,t-1), L(i,t-1)\big) & \text{otherwise} \end{cases}$
    Play arm i_t where $i_t = \arg\max_{i\in[K]} U(i,t)$
    Receive reward r_t
    Update for Policy 1 and 2: $V_t = V_{t-1} + \mathbf{a}_{i_t}\mathbf{a}_{i_t}^T$, $S_t = S_{t-1} + r_t \mathbf{a}_{i_t}$
  end for
  Update $j = j + 1$
end for

Our goal is to obtain $\mathbb{P}[I(\mathbf{y},t) + \varepsilon_t > U(\mathbf{y},\tau) | C_{t-1}]$ and likewise for the lower tail. From eq.(24), we have,

$$\begin{aligned} I(\mathbf{y},t) + \varepsilon_t &\geq U(\mathbf{y},\tau) - z(\mathbf{y},t-1) + z(\mathbf{y},t) + \omega(\mathbf{y},t-1) + \delta(\mathbf{y},t-1)\eta_{t-1} \\ I'(\mathbf{y},t) - \varepsilon_t &\leq L(\mathbf{y},\tau) + z(\mathbf{y},t-1) - z(\mathbf{y},t) + \omega(\mathbf{y},t-1) + \delta(\mathbf{y},t-1)\eta_{t-1} \end{aligned} \tag{25}$$

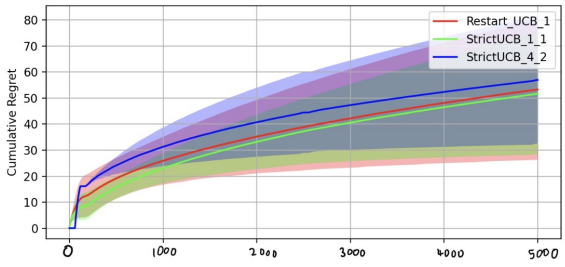
5.4.1 Recursion on $\hat{\theta}_t$

$$\begin{aligned} \hat{\theta}_t &= \hat{\theta}_{t-1} + \frac{V_{t-2}^{-1}\mathbf{x}_\tau \mathbf{x}_\tau^T(\theta_{t-1} - \hat{\theta}_{t-1})}{1 + \|\mathbf{x}_\tau\|^2_{V_{t-2}^{-1}}} + \left(1 - \|\mathbf{x}_\tau\|^2_{V_{t-2}^{-1}}\right)\eta_{t-1} V_{t-2}^{-1}\mathbf{x}_\tau \\ &= \hat{\theta}_\tau + \sum_{s=\tau}^{t-1} \frac{V_{s-1}^{-1}\mathbf{x}_\tau \mathbf{x}_\tau^T(\theta_s - \hat{\theta}_s)}{1 + \|\mathbf{x}_\tau\|^2_{V_{s-1}^{-1}}} + \sum_{s=\tau}^{t-1}\left(1 - \|\mathbf{x}_\tau\|^2_{V_{s-1}^{-1}}\right)\eta_s V_{s-1}^{-1}\mathbf{x}_\tau \end{aligned} \tag{26}$$

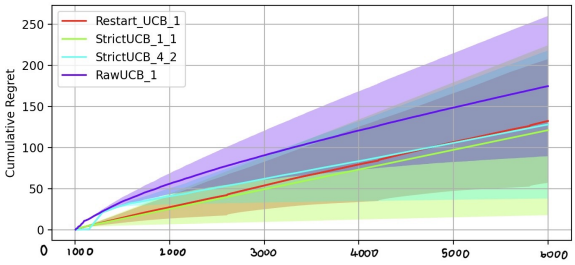
5.5 Regret Lower Bound

We modify the result in [6] to get the folllowing loose lower bound :
For any strategy π, there exists a rotting variation budget bandit scenario with path length $P_T \geq \sigma\frac{1}{8T}$ such that

$$\mathrm{E}[R_T(\pi)] = \frac{\sigma^{2/3}}{16\sqrt{2}} d^{1/6} P_T^{1/3} T^{2/3} = O(d^{1/6} P_T^{1/3} T^{2/3})$$



(a) Simulation for Gaussian noise for 5000 time steps. Our algorithm is shown in green and blue. StrictUCB_1_1 cooresponds to the Policy 1 and StrictUCB_4_2 corresponds to Policy 2. We compare against benchmark RestartUCB.



(b) Simulation for Gaussian noise for 5000 time steps. Our algorithm is shown in green and blue. StrictUCB_1_1 cooresponds to the Policy 1 and StrictUCB_4_2 corresponds to Policy 2. We compare against benchmark RestartUCB and RAW-UCB.

In comparison, the non-stationary lower bound[2] is

$$\mathrm{E}[R_T(\pi)] = O(d^{2/3} P_T^{1/3} T^{2/3})$$

Thus, the rotting bandit scenario has the same order of lower bound as non-stationary bandits. Thus, we can not improve on the non-stationary algorithms in terms of the order of the performance, but we can improve the constants observed in the order of the algorithms such as RestartUCB etc.

6 Simulation

In the simulations below we compare the algorithms designed for non-stationary linear and non-linear bandits. In the simulations we used parameters

From the simulations we see that RawUCB performs significantly worse. This is to be expected since it does not consider the linear structure of the problem. Further, we see that **our algorithm performs better** than all algorithms, even though only slightly.

7 Conclusion

In conclusion, this research paper addressed the problem of linear bandits with rotting rewards, considering the case where the rewards decrease over time.This work is the first to consider linear bandits with rotting rewards, filling a gap in the existing literature.
The study introduces multiple new algorithms that exploits the problem structure and outperforms existing state-of-the-art algorithms for non-stationary linear bandits and rotting multi-armed bandit problems. An partial regret analysis is provided for the proposed algorithms. Further, we prove the order of the lower bound for our scenario is the same as that of non-stationary bandits. The simulations conducted demonstrated the superior performance of the proposed algorithm compared to other algorithms designed for non-stationary bandit problems. Overall, this research contributes to the understanding and improvement of sequential decision-making under uncertainty in the context of linear bandits with non-decreasing rewards over time.

8 Acknowledgement

We would like to express our sincere gratitude to Prof. Arpan Chattopadhyay for his invaluable support and guidance throughout the research process. His expertise and insights have significantly contributed to the development and completion of this project.

References

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

[2] Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1079–1087. PMLR, 2019.

[3] Nir Levine, Koby Crammer, and Shie Mannor. Rotting bandits. *Advances in neural information processing systems*, 30, 2017.

[4] Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems*, 32, 2019.

[5] Julien Seznec, Andrea Locatelli, Alexandra Carpentier, Alessandro Lazaric, and Michal Valko. Rotting bandits are no harder than stochastic ones. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, 16–18 Apr 2019.

[6] Julien Seznec, Pierre Menard, Alessandro Lazaric, and Michal Valko. A single algorithm for both restless and rested rotting bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3784–3794. PMLR, 2020.

[7] Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

[8] Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298, 1988.

[9] Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 746–755. PMLR, 2020.