The Raymond and
Beverly Sackler Faculty
of Exact Sciences
Tel Aviv University

Regret Bounds in Rising Rested Multi-Arm Bandit with Linear Drift Models

Thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science
by

Omer Amichay

This work was carried out under the supervision of

Prof. Yishay Mansour

September 2024

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Yishay Mansour. Consistent mentorship, guidance, and support throughout my thesis journey, from the research to the writing of this thesis have been invaluable. In addition i want to thanks my friends and colleagues from room 464 who provided insightful feedback and helpful input throughout the research.

I want to say my heartfelt thanks to my family. Their unwavering belief in me and unconditional support have been the bedrock of my journey. Finally, I am particularly grateful to my wife, Dafna Amichay, who is the wind beneath my wings, here support and encouragement motivated me through this journey.Thank you.

Contents

Chapter 1

Abstract

We consider non-stationary multi-arm bandit (MAB) where the expected reward of each action follows a linear function of the number of times we executed the action. Our main result is a tight regret bound of $\tilde{\Theta}(T^{4/5}K^{3/5})$, by providing both upper and lower bounds. We extend our results to derive instance dependent regret bounds, which depend on the unknown parametrization of the linear drift of the rewards.

Chapter 2

Introduction

The multi-armed bandit (MAB) problem serves as a fundamental framework in decision-making under uncertainty, spanning various domains in machine learning. At the heart of the MAB problem lies the delicate balance between exploration and exploitation, where at each step, the agent decides which arm to pull, seeking to maximize cumulative rewards over time. There are many real world applications for MAB, such as online advertising, recommendations and more. In much of the literature, the rewards of the MAB are assumed to be stochastic and stationary (see, Slivkins (2019); Lattimore and Szepesvári (2020)).

In the non-stationary MAB, the payoff associated with each arm may change over time. An extreme case is an adversarial environment, where the reward sequence is arbitrary. There is a vast literature on this topic, bounding the regret to the best action, in hindsight. The HEDGE Freund and Schapire (1995), in the full information case, and EXP3 Auer et al. (1995), in the bandit case , are two classical algorithms for the adversarial setting.

There are also non-stationary stochastic setting, such as, persistent drift Freund and Mansour (1997), Brownian motion Slivkins and Upfal (2008), uncertainty over timing of rewards distributions changes Garivier and Moulines (2008), bounded reward variation Besbes et al. (2014), and more. A general class of non-stationary stochastic MAB, where the payoffs changes are coordinated or correlate are the rested and restless bandits. In the rested bandit, the payoff of an action depends on the number of time it was executed, and in the restless bandit the reward depends on the time step (see, Gittins et al. (2011)).

The Rested MAB (RMAB) framework captures the phenomenon of monotonically changing efficiency with continued execution of the action. For instance, increased experience in performing a given action may lead to improved effectiveness over time. Under RMAB we have two different settings, depending whether the rewards decrease of increase. The Rotting RMAB problem abstracts rewards which decrease monotonically with the number of pulls (Levine et al. (2017); Seznec et al. (2019)). The Rising RMAB Heidari et al. (2016), captures rewards that increases monotonically.

Rising rested MAB with linear drift is the case that the non-stationary expected rewards of each arm have a linear parameterization. To see the challenges in the model, consider two identical arms with linear increasing rewards. If we play each arm about half the times, we might get a linear regret. In contrast, for a stationary MAB with two identical arms, any policy have zero regret. As a motivation we can consider hyperparameter search over multiple algorithms. As we are modifying parameters of each algorithm, we expect the overall system performance to improve. By selecting between the algorithms we are exploring to find the best overall algorithm. Having a linear rate of increase is a natural modeling assumption, both as a starting step for the theory and a simplistic empirical model.

It is customary to measure the performance of online algorithms using the notion of regret, the performance difference between the online algorithm's cumulative rewards and that of a benchmark. We highlight three common notions of regret's benchmarks, *static*, *dynamic* and *policy* regret. Static regret uses the performance of the best fixed arm as a benchmark. Dynamic regret uses as a benchmark the performance of the best sequence of arms. Policy regret (Arora et al. & 2012) uses the best policy and allows the environment to react to the policy.

2.1 Our results

In this paper, we study the regret of Rising Rested MAB with linear drift, i.e., the expectation of the reward function of each arm is linear with respect to the number of times the arm was played. Our main results:

- We show that the dynamic, static and policy regrets are identical for rising rested MAB with linear drift.
- We design R-ed-EE (Rested Explore Exploit) algorithm, and show a regret bound of $O(T^{\frac{4}{5}}(\Phi K)^{\frac{3}{5}}\ln(\Phi KT)^{\frac{1}{5}})$, where T is the number of time steps, K is the number of actions and Φ bounds the expectation of the arm rewards.
- We design R-ed-AE (Rested Arm Elimination) and HR-re-AE (Halted Rested Arm Elimination) algorithms. Algorithm R-ed-AE gives an instance dependent regret bound. Algorithm HR-re-AE, extends R-ed-AE, and has, in addition to the instance dependent regret, a worse case $O\left(T^{\frac{4}{5}}(\Phi K)^{\frac{3}{5}}\ln\left(\Phi KT^2\right)^{\frac{1}{5}}\right)$ regret.
- We show a lower bound of $\Omega(K^{\frac{3}{5}}T^{\frac{4}{5}})$ for the regret. This, together with R-ed-EE and HR-re-AE shows tight regret of $\tilde{\Theta}(K^{\frac{3}{5}}T^{\frac{4}{5}})$ for the Rising Rested MAB with linear drift problem. (This is in contrast to the classical $\Theta(\sqrt{T})$ regret bounds in stochastic stationary settings).
- We design FIR-ed-EE (Full Information rested Explore Exploit) algorithm, showing a worse case bound of $O(T^{\frac{4}{5}}\Phi^{\frac{3}{5}}\ln(\Phi KT)^{\frac{1}{5}})$ for the full information

rising rested MAB with linear drift algorithm problem. Using the proof from Chapter 6 we show a tight bound of $\tilde{\Theta}(T^{\frac{4}{5}})$.

- We design two algorithm DR-ed-LD (Deterministic Rested Linear Drift), R-ed-LD(Rested Linear Drift), algorithm DR-ed-LD return the optimal policy for the deterministic rested MAB with linear drift problem. Algorithm R-ed-LD solve the non-deterministic rested MAB with linear drift problem using algorithm DR-ed-LD, obtaining a worse case $O\left(T^{\frac{4}{5}}(\Phi K)^{\frac{3}{5}} \ln\left(\Phi K T^2\right)^{\frac{1}{5}}\right)$ regret. Concluding tight regret of $\tilde{\Theta}(T^{\frac{4}{5}}K^{\frac{3}{5}})$.
- We design algorithm R-es-BEE (Restless Block Explore Exploit) for the restless rising MAB with linear drift. Showing a worst case regret of $\tilde{O}(\min\left\{T^{\frac{2}{3}}K^{\frac{1}{3}}, K\sqrt{T}, T\right\})$.

2.2 Related Works

Stochastic stationary MAB There is a vast literature on this topic (see, e.g., Slivkins (2019)).

Non-Stationary MAB the Non-Stationary Markovian MAB was studied by Gittins (1974), who showed that an index policy characterizes the optimal policy for discounted return. Following Gittins (1974), numerous works have explored the domain of rested MAB (e.g.Whittle (1981); Bertsimas and Niño-Mora (2000); Nino-Mora (2001)).

Whittle (1988) introduced the Restless MAB, which also has an optimal index policy. There are works that assume a certain structure to the change in rewards for Restless MAB. Freund and Mansour (1997) consider persistent distributions drift which is a linear drift. Slivkins and Upfal (2008) studies a Brownian motion of the rewards. Besbes et al. (2014) have a regret bounded as a function of the total variation of the rewards. Jia et al. (2023) consider smooth non-stationary bandits.

Tekin and Liu (2012) introduced rested MAB, where the arm's reward distribution depends on the number of times it was executed. A special case of rested MAB is *rotting MAB*, where the expected reward of an arm decreases with the number of pulls of the arm (Levine et al. (2017); Seznec et al. (2019)). Seznec et al. (2020) consider both rested and restless rotting MAB.

Rising rested MAB Most related to our problem is the rising rested MAB problem, where the expected rewards of the arms are monotonically non-decreasing in the number of times the decision maker played the arm. The non-stochastic version of this problem addressed by Heidari et al. (2016); Li et al. (2020) for the static regret. The stochastic case assumes that the payoffs follows a known parametric form. The Best Arm Identification framework of this setting was done in Cella et al. (2021); Mussi et al. (2023). It should be noted that both papers measure the performance of their best arm identification

algorithms as a function of the optimal arm and the arm the algorithm chose. In contrast, we measure the dynamic regret with respect to the entire run of the online algorithm when compared with the optimal sequence of arm selection.

Metelli et al. (2022) considered the stochastic rising MAB both for restless and rested, where the payoffs are increasing and concave. For the rested instance the dynamic regret is equivalent to policy regret. They designed UCB based algorithms with instance dependent expected regret. However their algorithm for rested MAB suffers, in some cases, a linear regret. Specifically, for our lower bound instance their algorithm has a linear regret.

Chapter 3

Problem Setting and Preliminaries

In this section we formalize our model of *Rising Rested MAB with Linear Drift.*

Non-stationary rested K-MAB An instance of a *non-stationary K-MAB* is a tuple (K, T, D), where $\mathcal{K} = \{1, 2, \ldots, K\}$ is the set of K arms, T is the time horizon, and for each arm i and $n \leq T$ we have a reward distribution $D_i(n)$ which is the stochastic reward of arm i when it is performed for the n-th time, and its expectation is $\mu_i(n)$. We assume that the distributions $D_i(n)$ are 1-sub-Gaussian, i.e., $\mathbb{E}_{R\sim D_i(n)}[e^{\lambda(R-\mu_i(n))}] \leq e^{\frac{\lambda^2}{2}}$, for every $\lambda \in \mathbb{R}$. Let $\Phi \geq \max_{i\in\mathcal{K}}(\mu_i(T))$ be an upper bound on the maximum expected reward of any arm. (Recall that the expected rewards are non-decreasing.)

At each time t the learner selects an arm $I_t = i$ and observes a reward $R_t \sim D_i(N_i(t))$, where $N_i(t) = \sum_{\tau=1}^{t} \mathbb{1}(I_\tau = i)$ is the number of times arm $i \in \mathcal{K}$ was pulled up to round t.

Rested MAB with linear drift A Non-stationary rested K-MAB has linear drift, if for each arm $i \in \mathcal{K}$ there are constants L_i and b_i such that $\mu_i(n) = L_i n + b_i$ for every $n \in \{1, 2, \ldots, T\}$. An instance is called *Rising Rested MAB with linear drift* if $L_i \geq 0$ for every $i \in \mathcal{K}$.

Policies and regret A policy π selects for each time t a distribution over the arms, i.e., $\pi(t) \in \Delta(\mathcal{K}) = \{q \in [0,1]^K |\ \sum_i q_i = 1\}$. The **static** regret of policy π is,

$$\Re_{stat} = static - Regret(\pi) = \max_{i\in\mathcal{K}} \sum_{t=1}^{T} \mu_i(t) - \mathbb{E}[\sum_{t=1}^{T} \sum_{i\in\mathcal{K}} \mu_i(N_i(t))\mathbb{I}[\pi(t) = i]],$$

where $\mathbb{I}[\cdot]$ is the indicator function. The **dynamic** regret of policy π is,

$$\Re_{dyn} = Dynamic{-}Regret(\pi) = \mathbb{E}[\sum_{t=1}^{T} \mu_{\pi^*(t)}(N_i(t))] - \mathbb{E}[\sum_{t=1}^{T} \sum_{i \in \mathcal{K}} \mu_i(N_i(t)) \mathbb{I}[\pi(t) = i]],$$

where $\pi^* \in \text{argmax}_\pi \, \mathbb{E}[\sum_{t=1}^{T} \sum_{i \in \mathcal{K}} \mu_i(N_i(t)) \mathbb{I}[\pi(t) = i]]$ is the optimal policy.

Remark 1. *In our setting the rewards are history dependent, so our dynamic regret is in fact a policy regret.*

Characterization of the optimal policy Theorem 44, in the supplementary material, shows that the optimal policy for Rising Rested MAB with Linear Drift selects a single arm and always plays it. Namely, for $i^* = \text{argmax}_{i \in \mathcal{K}} \sum_{t=1}^{T} \mu_i(t)$, Theorem 44 shows that the optimal policy always plays arm i^*. Therefore, in our setting the dynamic regret is equal to the static regret.

Corollary 2. *For Rising Rested MAB with Linear Drift the dynamic regret is equal to the static regret. Namely, the optimal policy plays always arm i^*.*

Notations Let $[k, m] = \{k, \ldots, m\}$ and $[m] = \{1, \ldots, m\}$. Assume we pulled arm $i \in \mathcal{K}$ for $m \in [T]$ times and observed rewards $r_i(1), \ldots, r_i(m)$.

- The estimated reward, given a window of $[n', n' + M] \subseteq [1, m]$, where M is an even integer, is $\hat{\mu}_i^M(n' + \frac{M}{2}) = (1/M) \sum_{n=n'}^{n'+M} r_i(n)$. Note that $\mu_i(n' + \frac{M}{2}) = \mathbb{E}[\hat{\mu}_i^M(n' + \frac{M}{2})]$.
  (Note that for $\hat{\mu}_i^M(\ell)$ we implicitly have $n' = \ell - M/2$.)
- The empirical slope is $\widehat{L}_i^{2M} = (1/M) \left(\hat{\mu}_i^M \left(\frac{3M}{2} \right) - \hat{\mu}_i^M \left(\frac{M}{2} \right) \right)$. Note that $L_i = \mathbb{E}[\widehat{L}_i^{2M}]$.
- Our estimate for the future rewards $\mu_i(n)$, given $2M$ samples for arm i, is $\psi_i^{2M}(n) = \left(\hat{\mu}_i^M \left(\frac{3M}{2} \right) + \hat{\mu}_i^M \left(\frac{M}{2} \right) \right) / 2 + (n - M) \widehat{L}_i^{2M}$. (Note that we might have $n \leq 2M$).
- The cumulative expectation of arm $i \in \mathcal{K}$ during $[n_1, n_2]$ is $s_i(n_1, n_2) = \sum_{n=n_1}^{n_2} \mu_i(n)$, and its estimation using $2M$ sample points is $\hat{s}_i^{2M}(n_1, n_2) = \sum_{n=n_1}^{n_2} \psi_i^{2M}(n)$.

We define the following parameters, which will be useful for our concentration bounds: Let the confidence parameters be $\gamma_n^{2M} = \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2M}} + |n - M| \frac{\sqrt{2 \ln\left(\frac{2}{\delta}\right)}}{M^{1.5}}$, and $\Gamma^{2M}(n_1, n_2) = \sum_{n=n_1}^{n_2} \gamma_n^{2M}$.

Chapter 4

Algorithm for Rising Rested MAB with Linear Drift

In this section we derive an algorithm for the Rising Rested MAB with Linear Drift problem and show that the regret is $\tilde{\Theta}\left(K^{3/5}T^{4/5}\right)$. This regret bound is tight, as we show in the lower bound of Chapter 6.

The basic idea behind our algorithm is simple. We use an explore-exploit methodology. We have an exploration period, in which we explore each arm for $2M$ times. Given the observed rewards, we estimate the model parameters for each arm, namely the slope and the intercept point for the linear function of the arm. The surprising outcome is that the dynamic regret bound we get is tight!

In more details. We are sampling $2M$ times each arm $i \in \mathcal{K}$. In order to recover the line for the rewards of arm i we estimate two points on this line (and they would define the parameters of the line). The two points are the rewards after $M/2$ and $3M/2$ uses of arm i. Note that the expected reward for $M/2$ is the average of the expected rewards in $[1, M]$, and the expected reward for $3M/2$ is the average of the expected in $[M+1, 2M]$. We denote our estimates as $\hat{\mu}_i^M(M/2)$ and $\hat{\mu}_i^M(3M/2)$. Given those two estimates, we estimate the slope $\widehat{L}_i^{2M}$ by their difference divided by M. Using $\hat{\mu}_i^M(M/2)$, $\hat{\mu}_i^M(3M/2)$ and $\widehat{L}_i^{2M}$ we estimate $\hat{s}_i^{2M}(2M+1, T-2KM)$ the future cumulative rewards of arm i for the remaining time, i.e., next $(T-2KM)$ time steps. Given our estimates $\hat{s}_i^{2M}(2M+1, T-2KM)$ for each arm i, we choose to play the arm with the highest estimated value of $\hat{s}_i^{2M}(2M+1, T-2KM)$. Essentially we are exploiting of the best arm, given our estimates. (The algorithm appears in Algorithm R-ed-EE.)

Overview of regret analysis We define a good event G, which bounds the deviations between our estimation and the true values. This holds for the slopes, and the expected rewards.

Definition 3. *Let G be the good event that after sampling each arm $2M$ times, for any arm $i \in \mathcal{K}$:*

Algorithm 1 R-ed-EE - Rested Explore Exploit

Input: K, T, M
for $i \in \mathcal{K}$ **do**
  Sample arm i for $2M$ times, and observe $r_i(1), \ldots, r_i(2M)$
  $\hat{\mu}_i^M\left(\frac{M}{2}\right) \leftarrow \frac{1}{M}\sum_{n=1}^{M} r_i(n)$
  $\hat{\mu}_i^M\left(\frac{3M}{2}\right) \leftarrow \frac{1}{M}\sum_{n=M+1}^{2M} r_i(n)$
  $\widehat{L}_i^{2M} \leftarrow \frac{\hat{\mu}_i^M\left(\frac{3M}{2}\right) - \hat{\mu}_i^M\left(\frac{M}{2}\right)}{M}$ ▷ Estimate of the slope
  $\hat{s}_i^{2M}(2M+1, T-2KM) \leftarrow (T-2KM-2M)\left[\frac{\hat{\mu}_i^M\left(\frac{3M}{2}\right) + \hat{\mu}_i^M\left(\frac{M}{2}\right)}{2} + \frac{T-2KM+1}{2}\widehat{L}_i^{2M}\right]$
  ▷ computes the sum $\sum_{t=2M+1}^{T-2KM} \psi_i^{2M}(n)$
end for
$\hat{i}^* \leftarrow \arg\max_{i\in\mathcal{K}} \hat{s}_i^{2M}(2M+1, T-2KM)$ ▷ The arm with the best estimate future reward
for $t \in [2KM+1, T]$ **do**
  Play arm $\hat{i}^*$
end for

$\left|\hat{\mu}_i^M\left(\frac{M}{2}\right) - \mu_i\left(\frac{M}{2}\right)\right| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2M}}$ *and* $\left|\hat{\mu}_i^M\left(\frac{3M}{2}\right) - \mu_i\left(\frac{3M}{2}\right)\right| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2M}}$.

Next, we show that the good event G holds with high probability.

Lemma 4. *The probability of the good even G is at lest $1-2\delta K$*

Proof. From Lemma 35, for each arm $i \in \mathcal{K}$, using the $2M$ samples of arm i, we have that with probability of at least $1-2\delta$:

$$\left|\hat{\mu}_i^M\left(\frac{M}{2}\right) - \mu_i\left(\frac{M}{2}\right)\right| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2M}},$$

$$\left|\hat{\mu}_i^M\left(\frac{3M}{2}\right) - \mu_i\left(\frac{3M}{2}\right)|\right| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2M}},$$

$$\left|\widehat{L}_i^{2M} - L_i|\right| \leq \frac{\sqrt{2\ln\left(\frac{2}{\delta}\right)}}{m^{1.5}}.$$

Using union bound over all K arms we get that G holds with probability of at least $1-2\delta K$. □

We then bound the regret using three different terms: (1) The exploration term, which we bound by $2MK$. (2) The exploitation term, which we bound using the confidence bounds of the good event. (3) The low probability event that G does not hold.

Theorem 5. *For $M = \frac{T^{\frac{4}{5}}\ln(4\Phi KT)^{\frac{1}{5}}}{(\Phi K)^{\frac{2}{5}}}$, Algorithm R-ed-EE guarantees regret $O(T^{\frac{4}{5}}(\Phi K)^{\frac{3}{5}}\ln(\Phi KT)^{\frac{1}{5}})$ for the Rising Rested MAB with Linear Drift.*

Proof. The regret can be partition to three parts. The first part is during times $[1, 2KM]$ when we explore each arm $2M$ times. The second part is during times $[2KM + 1, T]$, assuming the good event G holds. The third part is during times $[2KM + 1, T]$, when the good event G does not hold.

For the first part, i.e., the regret over the first $2M$ samples of each arm, we bound it by $2KM\Phi$.

For the second part, the regret during exploitation stage, assuming the good event G holds. We can upper bound the probability of the good event by one, i.e., $P(G) \leq 1$. Under the good event G, by Lemma 37 and Lemma 38, we have that for each arm $i \in \mathcal{K}$, we bound the estimation error by,

$$|\hat{s}_i^{2M}(2M+1, T-2KM) - s_i(2M+1, T-2KM)| \leq \Gamma^{2M}(2M+1, T-2KM) \leq T^2\sqrt{\frac{1}{2M^3}\ln\left(\frac{2}{\delta}\right)}.$$

Note that the regret of the second part is independent of Φ. This is since our confidence intervals do not depend on the magnitude rewards, i.e., Φ.

The third part bounds the regret when the event G does not hold, i.e., $\bar{G}$ occurs. By Lemma 4 we have $P(\bar{G}) \leq 2K\delta$, so the regret of this part is bounded by $P(\bar{G})T\Phi \leq 2KT\Phi\delta$.

To summarize, we have $\Re \leq 2MK\Phi + \frac{T^2\sqrt{\ln\left(\frac{2}{\delta}\right)}}{\sqrt{2}M^{1.5}} + 2KT\Phi\delta$. Setting $\delta = \frac{1}{2TK\Phi}$ and $M = \frac{T^{\frac{4}{5}}\ln(4KT\Phi)^{\frac{1}{5}}}{(\Phi K)^{\frac{2}{5}}}$, we get that $\Re \leq O(T^{\frac{4}{5}}(\Phi K)^{\frac{3}{5}}\ln(\Phi KT)^{\frac{1}{5}})$. □

Chapter 5

Instance Dependent Rising Rested MAB with Linear Drift

In this section we derive two instance dependent algorithms for Rising Rested MAB with Linear Drift. The first algorithm gives an instance dependent bound, but in some cases of the parameters, this regret might be linear in T. For this reason we have a second algorithm, for which we shows that the regret is almost tight (up to a logarithmic factor) in the worse case.

Our first algorithm, essentially, runs an arm elimination methodology, where it eliminate arms who are, with high probability, sub-optimal arms. Unlike the R-ed-EE algorithm, which samples all arms equally, the arm elimination methodology adjusts the sampling of each arm based on its performance, thereby preventing the over-sampling of sub-optimal arms.

In more detail. The first algorithm R-ed-AE is implementing an arm elimination approach for instance dependent problems. Initially, it starts with the set of all K arms, denoted by $\mathcal{K}'$. In each round the algorithm samples each arm in $i \in \mathcal{K}'$ four times. The algorithm maintains a counter N_i keeping track over the number of times arm i was sampled. (Note that all non-eliminated arms $i \in \mathcal{K}'$ have approximately the same counter N_i.) Using the observations of arm i, the algorithm estimates the line parameters of arm i. The algorithm does so by estimating two points on the line. The two points are the rewards after $N_i/4$ and after $3N_i/4$ uses of arm i. We denote the estimations as $\hat{\mu}_i^{N_i/2}(N_i/4)$ and $\hat{\mu}_i^{N_i/2}(3N_i/4)$, respectively. Using both estimations we estimate the slope of arm i, i.e., L_i, denoted by $L_i^{N_i}$. Using the estimations $\hat{\mu}_i^{N_i/2}(N_i/4)$, $\hat{\mu}_i^{N_i/2}(3N_i/4)$ and $L_i^{N_i}$ we estimate $\hat{s}_i^{N_i}(1,T)$, the expected reward if we played only arm i for the entire T time steps. At the end of each round for every pair of arms $i,j \in \mathcal{K}'$ we evaluate the difference between the estimations $\hat{s}_i^{N_i}(1,T) - \hat{s}_j^{N_i}(1,T)$. If the difference is greater than $2\Gamma^{N_i}(1,T)$, then, arm j is eliminated from $\mathcal{K}'$, where $2\Gamma^{N_i}(1,T)$ is our confidence interval. We are guaranteed that, with high probability, any eliminated arm is indeed sub-optimal.

Algorithm 2 R-ed-AE- Rested Arm Elimination

Input: K, T, δ
Set $N \leftarrow [0]^K$ and $\mathcal{K}' \leftarrow [K]$
For each $i \in \mathcal{K}'$ let B_i an empty array,
$\tau \leftarrow 0$
while $\tau \leq T$ **do**
  for $j \in \mathcal{K}'$ **do**
    for $l \in [4]$ **do**
      sample arm j and set $B_j(N_j) \leftarrow r_j(N_j)$
      $N_j \leftarrow N_j + 1$
    end for
    $\hat{\mu}_j^{\frac{N_j}{2}}(\frac{N_j}{4}) \leftarrow \frac{2}{N_j}\sum_{n=1}^{\frac{N_j}{2}} B_j(t)$
    $\hat{\mu}_j^{\frac{N_j}{2}}(\frac{3N_j}{4}) \leftarrow \frac{2}{N_j}\sum_{\frac{N_j}{2}+1}^{N_j} B_j(t)$
    Set $\widehat{L}_j^{N_j} \leftarrow 2\left(\hat{\mu}_i^{\frac{N_j}{2}}(\frac{3N_j}{4}) - \hat{\mu}_i^{\frac{N_j}{2}}(\frac{N_j}{4})\right)/N_j$
    Set $\hat{s}_j^{N_j}(1,T) \leftarrow T\left[\left(\hat{\mu}_i^{\frac{N_j}{2}}\left(\frac{3N_j}{4}\right) + \hat{\mu}_i^{\frac{N_j}{2}}\left(\frac{N_j}{4}\right)\right)/2 + \left(\frac{T+1}{2} - M\right)\widehat{L}_i^{N_j}\right]$
      ▷ computes the sum $\sum_{t=1}^{T}\psi_i^{N_j}(n)$
  end for
  Set $\tau \leftarrow \tau + 4|\mathcal{K}'|$
  for $i, j \in \mathcal{K}'$ **do**
    if $\hat{s}_i^{N_i}(1,T) - \hat{s}_j^{N_i}(1,T) > 2\Gamma^{N_i}(1,T)$ **then**
      $\mathcal{K}' \leftarrow \mathcal{K}' \setminus \{j\}$
    end if
  end for
end while
return $\mathcal{K}'$

Overview of regret analysis We first define the relevant parameters of the instance dependent analysis. Recall that each arm i, has expected reward define by $L_i m + b_i = m(L_i + b_i/m)$, where m is the number of times we played arm i. We now would like to define the distinguishability between different arms. We view b_i/m as a normalized version of the intercept, and later use $b_i/(T+1)$ where we take $m = T + 1$.

Definition 6. *for any two arm $i, j \in \mathcal{K}$:*
(1) $\Delta_{i,j} = b_i - b_j$ the difference of the intercept points between the arms i, j, and $\widetilde{\Delta}_{i,j} = \Delta_{i,j}/(T+1)$.
(2) $L_{i,j} = L_i - L_j$ the difference of the slopes of the two arms.

For the analysis we first define a good event G, which will hold with high probability. Intuitively, the good event states that all of our estimates are accurate. Formally, the definition of the good event G is as follows.

Definition 7. *Let G be the good event that for all $i \in \mathcal{K}$ and $m \in [T]$:* $\left|\hat{\mu}_i^{\frac{m}{2}}(\frac{m}{4}) - \mu_i(\frac{m}{4})\right| \leq \sqrt{\frac{\ln(\frac{2}{\delta})}{m}}$, $\left|\hat{\mu}_i^{\frac{m}{2}}(\frac{3m}{4}) - \mu_i(\frac{3m}{4})\right| \leq \sqrt{\frac{\ln(\frac{2}{\delta})}{m}}$ *and* $\left|\widehat{L}_i^m - L_i\right| \leq \frac{2\sqrt{\ln(\frac{2}{\delta})}}{m^{1.5}}$.

The following claims that the good event holds with high probability.

Lemma 8. *The probability of the good even G is at lest $1 - 2TK\delta$.*

Proof. From Lemma 35 we get that for any arm $i \in \mathcal{K}$ and $m \in [T]$ with probability of at least $1 - 2\delta$: $\left|\hat{\mu}_i^{\frac{m}{2}}(\frac{m}{4}) - \mu_i(\frac{m}{4})\right| \leq \sqrt{\frac{\ln(\frac{2}{\delta})}{m}}$, $\left|\hat{\mu}_i^{\frac{m}{2}}(\frac{3m}{4}) - \mu_i(\frac{3m}{4})\right| \leq \sqrt{\frac{\ln(\frac{2}{\delta})}{m}}$ and $\left|\widehat{L}_i^m - L_i\right| \leq \frac{2\sqrt{\ln(\frac{2}{\delta})}}{m^{1.5}}$. using union bound over T and K we have that event G is occur with probability of at least $1 - 2TK\delta$. □

From now on we will assume that the good event G holds.

We start by showing that, under the good event G, the best arm i^* is never eliminated. (Note that we fixed the best arm, initially.)

Lemma 9. *Under G we never eliminate the optimal arm i^*.*

Proof. Assuming event G hold and $i^* \notin \mathcal{K}'$. We conclude that after N_{i^*} samples of arm i^* the algorithm R-ed-AE eliminated arm i^* using some arm $j \in \mathcal{K}'$, where $N_{i^*} = N_j$. Thus the elimination term was satisfy, namely,

$$\hat{s}_j^{N_{i^*}}(1,T) - \hat{s}_{i^*}^{N_{i^*}}(1,T) > 2\Gamma^{N_{i^*}}(1,T).$$

In addition, from event G we have that for any arm $i \in \mathcal{K}'$ and $t \in [T]$,

$$|\mu_i(t) - \psi_i^{N_{i^*}}(t)| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{N_{i^*}}} + \left|n - \frac{M}{2}\right| \frac{2\sqrt{\ln\left(\frac{2}{\delta}\right)}}{N_{i^*}} = \gamma_t^{N_{i^*}}$$

Therefore,

$$|s_i(1,T) - \hat{s}_i^{N_{i^*}}(1,T)| \leq \Gamma^{N_{i^*}}(1,T).$$

From both inequalities we have,

$$\hat{s}_j^{N_{i^*}}(1,T) - \hat{s}_{i^*}^{N_{i^*}}(1,T) > s_{i^*}(1,T) - \hat{s}_{i^*}^{N_{i^*}}(1,T) + \hat{s}_j^{N_{i^*}}(1,T) - s_j(1,T)$$

Hence,

$$s_j(1,T) > s_{i^*}(1,T).$$

In contradiction to the optimality of arm i^*. □

Next we bound the number of times we can play any sub-optimal arm. The sample size of the sub-optimal arm j depend only on the parameters $\widetilde{\Delta}_{i^*,j}$ and $L_{i^*,j}$.

Lemma 10. *Algorithm R-ed-AE samples a sub-optimal arm j at most* $\left[\frac{16\sqrt{\ln\left(\frac{2}{\delta}\right)}}{2\widetilde{\Delta}_{i^*,j}+L_{i^*,j}}\right]^{\frac{2}{3}}$ *times.*

Proof. Consider the elimination of arm j due to arm i. This implies that so far we have played each of i and j for N_j times each. Then, using that event G holds and using Lemma 37 for N_j, we have the following for the rewards of each of i and j. When we eliminate arm j (by i) we have

$$\hat{s}_i^{N_j}(1,T) - \hat{s}_j^{N_j}(1,T) = \sum_{n=1}^{T} \hat{\mu}_i(t,N_j) - \hat{\mu}_j(n,N_j) \geq \sum_{n=1}^{T} 2\gamma_n^{N_j} = 2\Gamma^{N_j}. \quad (5.1)$$

Due to the definition of $\gamma_n^{N_j}$, we have from Lemma 12 regrading the true rewards,

$$\sum_{n=1}^{T} \mu_i(n) - \mu_j(n) \geq \sum_{n=1}^{T} \hat{\mu}_i(n,N_j) - \hat{\mu}_j(n,N_j) + \sum_{n=1}^{T} 2\gamma_n^{N_j} \geq \sum_{n=1}^{T} 4\gamma_n^{N_j} = 4\Gamma^{N_j}(1,T). \quad (5.2)$$

Since, by definition of i^*, we have that $\sum_{t=1}^{T} \mu_{i^*}(t) \geq \sum_{t=1}^{T} \mu_i(t)$, and $N_j = M/2$, from Lemma 38

$$\Gamma^{N_j}(1,T) \leq T^2 \frac{\sqrt{\ln\left(\frac{2}{\delta}\right)}}{\sqrt{2}M^{1.5}} \leq 2T(T+1)\frac{\sqrt{\ln\left(\frac{2}{\delta}\right)}}{N_j^{1.5}}.$$

Consider,

$$N_j \geq \left[\frac{16(T+1)\sqrt{\ln\left(\frac{2}{\delta}\right)}}{2\Delta_{i^*,j} + (T+1)L_{i^*,j}}\right]^{\frac{2}{3}}.$$

Then,

$$\begin{aligned}
&\Rightarrow N_j^{1.5} \geq \frac{8(T+1)\sqrt{\ln\left(\frac{2}{\delta}\right)}}{\Delta_{i^*,j} + \frac{T+1}{2}L_{i^*,j}} \\
&\Rightarrow \Delta_{i^*,j} + \frac{T+1}{2}L_{i^*,j} \geq 8(T+1)\frac{\sqrt{\ln\left(\frac{2}{\delta}\right)}}{N_j^{1.5}} \\
&\Rightarrow T(\Delta_{i^*,j} + \frac{T+1}{2}L_{i^*,j}) \geq 8T(T+1)\frac{\sqrt{\ln\left(\frac{2}{\delta}\right)}}{N_j^{1.5}} \\
&\sum_{t=1}^{T} \mu_{i^*}(t) - \mu_j(t) \geq 8T(T+1)\frac{\sqrt{\ln\left(\frac{2}{\delta}\right)}}{N_j^{1.5}}.
\end{aligned}$$

This implies that after such N_j arm i^* would have eliminated arm j. If arm i^* eliminated arm j, this can only happen earlier.

Rewriting, the expression for N_j,

$$\left[\frac{16(T+1)\sqrt{\ln\left(\frac{2}{\delta}\right)}}{2\Delta_{i^*,j}+(T+1)L_{i^*,j}}\right]^{\frac{2}{3}} \leq \left[\frac{16\sqrt{\ln\left(\frac{2}{\delta}\right)}}{2\widetilde{\Delta}_{i^*,j}+L_{i^*,j}}\right]^{\frac{2}{3}}.$$

Then, for values of $N_j \geq \left[\frac{16\sqrt{\ln\left(\frac{2}{\delta}\right)}}{2\widetilde{\Delta}_{i^*,j}+L_{i^*,j}}\right]^{\frac{2}{3}}$ we are guaranteed that arm j was already eliminated.

Therefore, the condition for elimination is hold for $\left[\frac{16\sqrt{\ln\left(\frac{2}{\delta}\right)}}{2\widetilde{\Delta}_{i^*,j}+L_{i^*,j}}\right]^{\frac{2}{3}}$. □

Given our bound on the sample size of any sub-optimal arm, we can now bound the overall instance dependent regret. (Recall that the expected rewards are increasing, so we can use T to bound the rewards.)

Theorem 11. *Algorithm R-ed-AE with $\delta = \frac{1}{2\Phi KT^2}$ guarantees regret*

$$\Re \leq \sum_{j\in\mathcal{K}\setminus\{i^*\}}\left[\frac{16\sqrt{\ln\left(4\Phi KT^2\right)}}{2\widetilde{\Delta}_{i^*,j}+L_{i^*,j}}\right]^{\frac{2}{3}}\Phi+1.$$

Proof. Under the good event G, for every time we sample a sub-optimal arm $i \in \mathcal{K}$, we increase the regret by at most Φ. The complement event $\bar{G}$ occurs with probability of at most $\frac{1}{T\Phi}$ and the regret in such case is bounded with $T\Phi$. We combine the two and get

$$\Re \leq \sum_{j\in\mathcal{K}\setminus\{i^*\}}\left[\frac{16\sqrt{\ln\left(4\Phi KT^2\right)}}{2\widetilde{\Delta}_{i^*,j}+L_{i^*,j}}\right]^{\frac{2}{3}}\Phi+2\delta\Phi KT^2 \leq \sum_{j\in\mathcal{K}\setminus\{i^*\}}\left[\frac{16\sqrt{\ln\left(4\Phi KT^2\right)}}{2\widetilde{\Delta}_{i^*,j}+L_{i^*,j}}\right]^{\frac{2}{3}}\Phi+1.$$

□

5.1 Arm Elimination with early stopping

In the case of Rising Rested MAB, there is an issue of playing a sub-optimal arm, even if it is very near to the optimal arm. For example, if we have two optimal arms, it is not the case that we can mix them, and get an optimal reward. This is due to the fact that we have increasing rewards, and in order to achieve the increase we need to concentrate on playing only one of them. So, unlike the case of regular MAB, where playing any mixture of the two optimal arms is optimal, for Rising Rested MAB we have only two optimal sequences, each consisting of playing the same optimal arm always.

When we consider our instance dependent algorithm, R-ed-AE, we will have an issue if we are left at the end with two near optimal arms. More precisely, we

would like to terminate the arm elimination part of the algorithm early, and to force selecting one of the near optimal arms that remain in the set $\mathcal{K}'$.

In addressing such cases, we devised the algorithm HR-ed-AE. The main idea behind the algorithm is using algorithm R-ed-AE as a black box with time horizon of M. After algorithm R-ed-AE terminates we receive $\mathcal{K}'$ a set of arms which are potentially near-optimal arms. From the set $\mathcal{K}'$ algorithm HR-ed-AE picks an arbitrary arm and plays in the remaining time steps.

Algorithm 3 HR-ed-AE- Halted Rested Arm Elimination

Input: K, T, M,δ
Set $\mathcal{K}' \leftarrow$ R-ed-AE(K, KM, δ)
Select arbitrary $a \in \mathcal{K}'$
for $t \in [KM+1, T]$ **do**
    play a
end for

The next theorem shows the improvement of HR-ed-AE algorithm. Namely, it has an additional upper bound on the regret,

Lemma 12. *Fix $x_i, x_j, \hat{x}_i, \hat{x}_j \geq 0$, if $|\hat{x}_i - x_i| \leq \gamma$, $|\hat{x}_j - x_j| \leq \gamma$ and $\hat{x}_i - \hat{x}_j \leq 2\gamma$, then, $x_i - x_j \leq 4\gamma$.*

Theorem 13. *Algorithm HR-ed-AE with $\delta = \frac{1}{2\Phi K T^2}$ and $M = \frac{T^{4/5}\ln\left(\frac{2}{\delta}\right)^{1/5}}{(\Phi K)^{2/5}}$ guarantees regret*

$$\Re \leq O\left(\min\left\{\sum_{j\in\mathcal{K}\setminus\{i^*\}}\left[\frac{16\sqrt{\ln\left(\frac{2}{\delta}\right)}}{2\widetilde{\Delta}_{i^*,j}+L_{i^*,j}}\right]^{\frac{2}{3}}\Phi+1\ ,\ T^{\frac{4}{5}}(\Phi K)^{\frac{3}{5}}\ln\left(\Phi K T^2\right)^{\frac{1}{5}}\right\}\right).$$

Proof. Recall the definition of the good event G as in Definition 7. By Lemma 8, event G holds with probability of at least $1 - 2TK\delta$. From now on we will assume G holds.

We split the analysis to two cases, according to the arms eliminated during R-ed-AE. If at the end of R-ed-AE there is only a single arm remaining in $\mathcal{K}'$, then, under the event G, it is the optimal arm. In this case we will not have any additional regret after R-ed-AE. In this case the regret in the first part of the minimum expression.

From now on assume that two or more arms are in $\mathcal{K}'$, at the end of R-ed-AE. In this case we can bound the regret during R-ed-AE by $KM\Phi$. The main issue is to bound the regret of the selected arm $j \in \mathcal{K}'$ at the end of R-ed-AE.

Since arm j has not been eliminated ($j \in \mathcal{K}'$) in the R-ed-AE, and j have been sampled at least M times, then,

$$\sum_{t=1}^{T}\hat{\mu}_{i^*}^{M}(t) - \hat{\mu}_{j}^{M}(t) < 2\Gamma^{M}(1,T).$$

We derive the following lemma to bounds the difference between two values using their estimations.

Given that arms $i^*, j \in \mathcal{K}'$, we can utilize Lemma 12 and Lemma 38 to derive,

$$\sum_{t=1}^{T} \mu_{i^*}(t) - \mu_j(t) \leq 4\Gamma^M(1,T) \leq 2T^2 \frac{\sqrt{\ln\left(\frac{2}{\delta}\right)}}{M^{1.5}}.$$

Therefore, assuming event G holds the regret in the second part is at most $O(T^{\frac{4}{5}} K^{\frac{3}{5}} \sqrt{\ln(\Phi K T^2)})$.

For the total regret, setting $\delta = 1/(2\Phi K T^2)$ and $M = \frac{T^{4/5} \ln(2/\delta)^{1/5}}{(\Phi K)^{2/5}}$, we have that,

$$\Re \leq MK\Phi + \frac{2T^2\sqrt{\ln(2/\delta)}}{M^{1.5}} + 2\delta\Phi K T^2 \leq 3T^{\frac{4}{5}} (\Phi K)^{\frac{3}{5}} \ln\left(4\Phi K T^2\right)^{\frac{1}{5}} + 1.$$

□

Chapter 6

Lower Bound for Rising Rested MAB with Linear Drift

In this chapter we present a lower bound of $\Omega(\min\{T, K^{\frac{3}{5}}T^{\frac{4}{5}}\})$ for Rising Rested MAB with Linear Drift. We concentrate on the case of $K \leq T^{\frac{1}{3}}$, since for larger K the lower bound becomes $\Omega(T)$. Our lower bound holds for $\Phi = 1$.

The lower bound builds on the fact that we are in a rested MAB model. Note that unlike the standard MAB, we can get a regret even if we have two identical arms. If we initially test them for τ times steps each, then, our regret would be the difference between the cumulative rewards in first τ time steps compared to the last τ steps.

Our lower bound construction would set the parameters such that initially we are unable to decide which arm is indeed better, say for the initial τ steps. Then the regret would be the difference between the cumulative rewards in first τ time steps compared to the last τ steps. More specifically, if one arm has mean rewards of t/T, after t times it is used, and the other has $t/T + t/T^{6/5}$, then the KL divergence over the τ first samples is $\tau^3/T^{12/5}$, which implies that in the first $T^{4/5}$ we will not be able to distinguish between the arms. The regret would be $O(\tau)$, since in the last steps the rewards are almost 1 and initially the rewards are near zero.

We will now give the details of the lower bound construction. The core concept of the lower bound is bounding distance between distribution for specific problem instance. We limit our possible profiles to be one of $K + 1$ profiles, where each profile gives a complete characterization of the rewards of each arm. Specifically, we define the set of profiles as $\mathcal{I} = \{\mathcal{I}_i\}_{i=0}^{K}$, comprising a set of $K+1$ profiles. In profile $\mathcal{I}_j$, $j \in \mathcal{K}$, for a given number of samples $t \in [T]$ and arms $i \neq j$, the expected reward of arm i is defined as $\mu_i(t) = \mathcal{N}(t/T - tK^{\frac{3}{5}}/T^{\frac{6}{5}}, 1)$, where $\mathcal{N}(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2. In profile $\mathcal{I}_j$, arm j has expected reward $\mu_j(t) = \mathcal{N}(t/T, 1)$. So, the optimal arm for profile $\mathcal{I}_j$ is arm j. Additionally, there exists a profile $\mathcal{I}_0$ where for every arm $i \in \mathcal{K}$ and $t \in [T]$ the expected reward is $\mu_i(t) = \mathcal{N}(t/T - tK^{\frac{3}{5}}/T^{\frac{6}{5}}, 1)$. We will select

one of the profiles $\mathcal{I}_j$, $j \in \mathcal{K}$, uniformly at random.

The following theorem demonstrates that the Rising Rested MAB with Linear Drift setting has a lower bound of $\Omega(T^{\frac{4}{5}}K^{\frac{3}{5}})$.

Definition 14.

- *Distributions:*

$$q_t \sim \mathcal{N}(\frac{t}{T} - \frac{tK^{\frac{3}{5}}}{T^{\frac{6}{5}}}, 1), \quad Q_m = q_1 \times \ldots \times q_m.$$
$$p_t \sim \mathcal{N}(\frac{t}{T}, 1), \qquad\qquad P_m = p_1 \times \ldots \times p_m.$$

- *We define a set $\mathcal{I}$ of $K+1$ profiles of rested bandits problems with linear drift, where for each $j \in \mathcal{K}$ $\forall t \in T$ the profile $\mathcal{I}_j$ is:*

$$\mathcal{I}_j = \begin{cases} \mu_i(t) = q_t & \text{if } i \neq j \\ \mu_i(t) = p_t & \text{if } i = j \end{cases}.$$

We also define a special profile $\mathcal{I}_0$:

$$\mathcal{I}_0 = \Big\{ \mu_i(t) = q_t \quad .$$

Lemma 15. *For every instance $j \in \mathcal{K}$, for any event $A \subset \Omega^*$, we have $|Z_0^*(A) - Z_j^*(A)| \leq 10\frac{\tau^{1.5}K^{\frac{3}{5}}}{T^{\frac{6}{5}}}$, i.e., $\|Z_0^* - Z_j^*\|_1 \leq 10\frac{\tau^{1.5}K^{\frac{3}{5}}}{T^{\frac{6}{5}}}$.*

Lemma 16. *When using instance $\mathcal{I}_j$ assuming $N_j(T) \leq T - (\frac{K}{4} - 8)\tau$ the regret is lower bounded with $\frac{K^{\frac{3}{5}}T^{\frac{4}{5}}}{40}$. i.e., $\mathbb{E}_j^*[\Re] \geq \frac{K^{\frac{3}{5}}T^{\frac{4}{5}}}{40}$.*

Theorem 17. *Rising Rested MAB with Linear Drift problem is lower bounded with $\Omega(T^{\frac{4}{5}}K^{\frac{3}{5}})$, for $K \geq 36$, and $\Omega(T^{\frac{4}{5}})$ otherwise.*

Proof. For $K \geq 36$ Fix an arm j satisfying the follow property:

$$\mathbb{E}_0[N_j(\frac{1}{4}K\tau)] \leq \tau. \tag{6.1}$$

Since there are at least $\frac{3}{4}K$ such arms at the $\mathcal{I}_0$ instance, this implies that if we select arm j uniformly at random we have $3/4$ probability that it holds, i.e., $\Pr_{j\in\mathcal{K}}[\mathbb{E}[N_j(\frac{1}{4}K\tau)] \leq \tau |\ \mathcal{I}_0] \geq \frac{3}{4}$. If $\mathbb{E}_0[N_j(\frac{1}{4}K\tau)] \leq \tau$, then by Markov inequality we have,

$$\Pr[N_j(\frac{1}{4}K\tau)] \leq 8\tau | \mathcal{I}_0] \geq 7/8.$$

We will now refine our definition of the sample space. For each arm a, define the t-round sample space $\Omega_a^t = [0,1]^t$, where each outcome corresponds to a particular realization of the tuple $(r_a(s) : s \in [t])$. Then, the partial first $\frac{1}{4}K\tau$ sample space we considered before can be expressed as $\Omega = \prod_{a\in\mathcal{K}} \Omega_a^{\frac{1}{4}K\tau}$.

We consider a "reduced" sample space in which arm j played at most 8τ times:

$$\Omega_j^* = \Omega_j^{8\tau} \times \prod_{a \neq j} \Omega_a^{\frac{1}{4}K\tau}.$$

We will be interested in the event $A = \{N_j(\frac{1}{4}K\tau)] \leq 8\tau\}$. Note that the event A is determined by Ω_j^*.

For each profile $\mathcal{I}_j$, we define distribution Z_j^* on Ω_j^* as follows:

$$Z_j^*(A) = \Pr[A|\mathcal{I}_j] \text{ for each } A \subset \Omega_j^*.$$

We want to bound the difference between the distributions $Z_0^*(A)$ and $Z_j^*(A)$. Using Lemma 15, we have that,

$$|Z_0^*(A) - Z_j^*(A)| \leq 10\frac{\tau^{1.5}K^{\frac{3}{5}}}{T^{\frac{6}{5}}}.$$

For $\tau = \frac{T^{\frac{4}{5}}}{12K^{\frac{2}{5}}}$ we obtain that $|Z_0^*(A) - Z_j^*(A)| \leq \frac{1}{4}$. Therefore, for arm j for which property (6.1) holds, we attain

$$Pr[N_j(\frac{1}{4}K\tau) \geq 8\tau|\ \mathcal{I}_j] \leq \frac{1}{4} + Pr[N_j(\frac{1}{4}K\tau) \geq 8\tau|\ \mathcal{I}_0] \leq 3/8.$$

This implies that $Pr[N_j(\frac{1}{4}K\tau) \leq 8\tau|\ \mathcal{I}_j] \geq 5/8$.

Again, assuming that for arm j property (6.1) holds over instance $\mathcal{I}_0$, when playing "full" game over instance $\mathcal{I}_j$, with probability at least $5/8$ arm j will be sample at most $T - (\frac{K}{4} - 8)\tau$ times, i.e., $N_j(T)|\ I_j \leq T - (\frac{K}{4} - 8)\tau$.

From Lemma 16, since with probability at least $5/8$ we have $N_j(T) \leq T - (\frac{K}{4} - 8)\tau$, then we have that, $\mathbb{E}_j[\Re] \geq \frac{5}{8}\frac{K^{\frac{3}{5}}T^{\frac{4}{5}}}{40}$. Therefore, for a uniform random profile $\mathcal{I}_j$ we have an expected of $\mathbb{E}[\Re] \geq \frac{K^{\frac{3}{5}}T^{\frac{4}{5}}}{64}$.

For $K = 2$. Let the two arms be a_q and a_p when $r_p(t) \sim \mathcal{N}(\frac{t}{T}, 1)$ and $r_q(t) \sim \mathcal{N}(\frac{t}{T} - \frac{t}{\sqrt{2}T^{\frac{6}{5}}}, 1)$. Applying Pinsker's inequality and KL-divergent we get that,

$$\begin{aligned} &2\left\|P_\tau - Q_\tau\right\|_1^2 KL(P_\tau||Q_\tau) \leq \sum_{t=1}^{\tau}\left(\frac{t}{\sqrt{2}T^{\frac{6}{5}}}\right)^2 \leq \frac{\tau^3}{2T^{\frac{12}{5}}} \\ &\Rightarrow \left\|P_\tau - Q_\tau\right\|_1 \leq \frac{\tau^{1.5}}{2T^{\frac{6}{5}}}. \end{aligned}$$

Therefore, for $\tau \leq T^{\frac{4}{5}}$ we have that $\left\|P_\tau - Q_\tau\right\|_1 \leq \frac{1}{2}$, meaning that the two distributions are indistinguishable if both sampled less than τ times. Hence, to distinguish between the arms the player have to sample one arm at least $T^{\frac{4}{5}}$ times. The probability of sampling the bad arm $T^{\frac{4}{5}}$ times is $\frac{1}{2}$. Giving a regret

of,

$$\begin{aligned}\Re &\leq 0.5\sum_{t=1}^{T^{\frac{4}{5}}}\frac{T-t+1}{T}-\frac{t}{T}+\frac{t}{\sqrt{2}T^{\frac{6}{5}}}\\ &\geq 0.5T^{\frac{4}{5}}-\frac{T^{\frac{8}{5}}+T^{\frac{4}{5}}}{2}\frac{1}{T}\\ &= 0.5(T^{\frac{4}{5}}-T^{\frac{3}{5}}+T^{-\frac{1}{5}})\\ &\geq 0.25T^{\frac{4}{5}}.\end{aligned}$$

□

6.1 Impossibility result when the horizon T is unknown:

Theorem 18. *If the horizon T is unknown to the learner, there exists a problem instance where the regret is $\Omega(T)$.*

Proof. Consider the follow the following example. A rising rested MAB with linear drift with 2 arms, a_1 and a_2. Where for any time step $t \in [T]$ the true values of the arms are, $\mu_{a_1}(t) = 1$ and $\mu_{a_2}(t) = Lt$, where $L = O(1/T)$.

For $L < 2/T$ the optimal policy is to always play arm 1, and if $L > 2/T$ the optimal policy is to always play arm 2.

given any online algorithm at time $t = 1/L$ we have two events.

1. If the algorithm plays arm 1 less than $1/2L$ times, then we set $T = 1/L$. The the algorithm has at most reward of $0.625/L + 1/4$ while the optimal policy (playing always arm 1) has reward of $1/L$. Hence having a linear regret in T.

2. If the algorithm plays arm 1 more than $1/2L$ times, then we set $T = 3/L$. The algorithm has at most reward of $3.625/L$ while the optimal policy (playing always arm 2) has reward of $4.5/L$, also having a linear regret in T.

Therefore, for any online algorithm if the horizon is unknown regret is $\Omega(T)$. □

6.2 Missing proofs

Lemma 15. *For every instance $j \in \mathcal{K}$, for any event $A \subset \Omega^*$, we have $|Z_0^*(A) - Z_j^*(A)| \leq 10\frac{\tau^{1.5}K^{\frac{3}{5}}}{T^{\frac{6}{5}}}$, i.e., $\|Z_0^* - Z_j^*\|_1 \leq 10\frac{\tau^{1.5}K^{\frac{3}{5}}}{T^{\frac{6}{5}}}$.*

Proof. Applying Pinsker's inequality and KL-divergent we will have for any event $A \subset \Omega^*$:

$$
\begin{aligned}
2(Z_0^*(A) - Z_j^*(A))^2 &\leq KL(Z_0^*||Z_j^*) \\
&= \sum_{a\in\mathcal{K}} \sum_{t=1}^{N_a(\frac{1}{4}K\tau)} KL(Z_0^{a,t}||Z_j^{a,t}) \\
&= \sum_{a\neq j} \sum_{t=1}^{N_a(\frac{1}{4}K\tau)} KL(Z_0^{a,t}||Z_j^{a,t}) + \sum_{t=1}^{N_j(\frac{1}{4}K\tau)} KL(Z_0^{j,t}||Z_j^{j,t}) \\
&\leq \sum_{t=1}^{8\tau} \left(\frac{tK^{\frac{3}{5}}}{T^{\frac{6}{5}}}\right)^2 \leq 180\frac{\tau^3 K^{\frac{6}{5}}}{T^{\frac{12}{5}}}.
\end{aligned}
$$

In the inequality we use the assumption $N_j(\frac{1}{4}K\tau) \leq 8\tau$ and that for any $a \neq j$ $Z_0^{a,t}$ and $Z_j^{a,t}$ are the same. Therefore,

$$
|Z_0^*(A) - Z_j^*(A)| \leq 10\frac{\tau^{1.5}K^{\frac{3}{5}}}{T^{\frac{6}{5}}}.
$$

□

Definition 19.

- *Let the vector $v = \{v_i\}_{i\in\mathcal{K}}$, where $v_i = N_i(T)$ represent the number of times arm i was sampled, note that $\|v\|_1 = T$.*
- *Let the vector v' be the vector v where we zero the entry for arm j.*
- *V a set of vectors v such that the entries $v_j \leq T - (\frac{K}{4} - 8)\tau$. i.e., $V = \{v \in \mathbb{N}^k : \|v\|_1 = T, v_j \leq T - (\frac{K}{4} - 8)\tau\}$*

Lemma 16. *When using instance $\mathcal{I}_j$ assuming $N_j(T) \leq T - (\frac{K}{4} - 8)\tau$ the regret is lower bounded with $\frac{K^{\frac{3}{5}}T^{\frac{4}{5}}}{40}$. i.e., $\mathbb{E}_j^*[\Re] \geq \frac{K^{\frac{3}{5}}T^{\frac{4}{5}}}{40}$.*

Proof. In this proof we will use Definition 19. The regret when using instance $\mathcal{I}_j$ assuming $E[N_j(T)] \leq T - (\frac{K}{4} - 8)\tau$. Formally, we define $\mathbb{E}_j^*[\Re] = \mathbb{E}[\Re|\ \mathcal{I}_j, E[N_j(T)] \leq T - (\frac{K}{4} - 8)\tau]$.

$$
\begin{aligned}
\mathbb{E}_j^*[\Re] &\geq \min_{v\in V} \left\{ \sum_{t=1}^{T} \frac{t}{T} - \left(\sum_{i\in\mathcal{K}\setminus\{j\}} \sum_{t=1}^{v_i} \left[\frac{t}{T} - \frac{tK^{\frac{3}{5}}}{T^{\frac{6}{5}}} \right] + \sum_{t=1}^{v_j} \frac{t}{T} \right) \right\} \\
&= \min_{0\leq\beta\leq T-(\frac{K}{4}-8)\tau} \left\{ \sum_{t=1}^{T} \frac{t}{T} - \left(\sum_{t=1}^{T-\beta} \left[\frac{t}{T} - \frac{tK^{\frac{3}{5}}}{T^{\frac{6}{5}}} \right] + \sum_{t=1}^{\beta} \frac{t}{T} \right) \right\} \\
&= \min_{0\leq\beta\leq T-(\frac{K}{4}-8)\tau} \left\{ \sum_{t=\beta+1}^{T} \frac{t}{T} - \sum_{t=1}^{T-\beta} \frac{t}{T} + \sum_{t=1}^{T-\beta} \frac{tK^{\frac{3}{5}}}{T^{\frac{6}{5}}} \right\} \\
&= \min_{0\leq\beta\leq T-(\frac{K}{4}-8)\tau} \left\{ \beta - \frac{\beta^2}{T} + \frac{K^{\frac{3}{5}}T^{\frac{4}{5}}}{2} + \frac{K^{\frac{3}{5}}\beta^2}{2T^{\frac{6}{5}}} + \frac{K^{\frac{3}{5}}\beta}{T^{\frac{1}{5}}} + \frac{K^{\frac{3}{5}}}{2T^{\frac{1}{5}}} - \frac{K^{\frac{3}{5}}\beta}{2T^{\frac{6}{5}}} \right\}.
\end{aligned}
$$

In the first equality we use that $\sum_{i\neq j}\sum_{t=1}^{v_i}\left[\frac{t}{T}-\frac{tK^{\frac{3}{5}}}{T^{\frac{6}{5}}}\right]\leq\sum_{t=1}^{\|\acute{v}\|_1}\left[\frac{t}{T}-\frac{tK^{\frac{3}{5}}}{T^{\frac{6}{5}}}\right]$.

Given $f(\beta)=\beta-\frac{\beta^2}{T}+\frac{K^{\frac{3}{5}}T^{\frac{4}{5}}}{2}+\frac{K^{\frac{3}{5}}\beta^2}{2T^{\frac{6}{5}}}+\frac{K^{\frac{3}{5}}\beta}{T^{\frac{1}{5}}}+\frac{K^{\frac{3}{5}}}{2T^{\frac{1}{5}}}-\frac{K^{\frac{3}{5}}\beta}{2T^{\frac{6}{5}}}$, lets calculate the first and second moments of f:

$$f'(\beta)=1-\frac{2\beta}{T}+\frac{K^{\frac{3}{5}}\beta}{T^{\frac{6}{5}}}+\frac{K^{\frac{3}{5}}}{T^{\frac{1}{5}}}-\frac{K^{\frac{3}{5}}}{2T^{\frac{6}{5}}},$$

$$f''(\beta)=-\frac{2}{T}+\frac{K^{\frac{3}{5}}}{T^{\frac{6}{5}}}\leq-\frac{2}{T}+\frac{T^{\frac{1}{5}}}{T^{\frac{6}{5}}}=-\frac{1}{T},$$

where we use that $K\ll T^{\frac{1}{3}}$.

We got that the second derivative of f is always negative. Therefore, f is concave, and the β value that minimize the expiration is at one of the edges of the domain $[0,T-\frac{K\tau}{4}+\tau]$, so it is sufficient to compare the value in $\beta=0$ and $\beta=T-(\frac{K}{4}-8)\tau$.

for $\beta=0$ we have

$$\mathbb{E}_j^*[\Re]=\frac{K^{\frac{3}{5}}T^{\frac{4}{5}}}{2}+\frac{K^{\frac{3}{5}}}{2T^{\frac{1}{5}}}.$$

for $\beta=T-\frac{K\tau}{4}+8\tau$,

$$\begin{aligned}
\mathbb{E}_j^*[\Re]&=\frac{(\frac{K\tau}{4}-8\tau)(2T-\frac{K\tau}{4}+8\tau+1)}{2T}-\frac{(\frac{K\tau}{4}-8\tau)(\frac{K\tau}{4}-8\tau+1)}{2T}\\
&\quad+\frac{K^{\frac{3}{5}}(\frac{K\tau}{4}-8\tau)(\frac{K\tau}{4}-8\tau+1)}{2T^{\frac{6}{5}}}\\
&=\frac{(\frac{K\tau}{4}-8\tau)(T-\frac{K\tau}{4}+8\tau)}{T}+\frac{K^{\frac{3}{5}}\overbrace{(\frac{K\tau}{4}-\tau)^2}^{\geq 0}}{2T^{\frac{6}{5}}}+\frac{K^{\frac{8}{5}}\tau}{8T^{\frac{6}{5}}}-\frac{4K^{\frac{3}{5}}\tau}{T^{\frac{6}{5}}}\\
&\geq\frac{K\tau}{4}-\frac{K^2\tau^2}{16T}+\frac{4K\tau^2}{T}-8\tau-\frac{64\tau^2}{T}+\frac{K^{\frac{8}{5}}\tau}{8T^{\frac{6}{5}}}-\frac{4K^{\frac{3}{5}}\tau}{T^{\frac{6}{5}}}\\
&=\frac{K^{\frac{3}{5}}T^{\frac{4}{5}}}{36}-\frac{K^{\frac{6}{5}}T^{\frac{3}{5}}}{2304}+\frac{K^{\frac{1}{5}}T^{\frac{3}{5}}}{36}-\frac{3T^{\frac{4}{5}}}{4K^{\frac{2}{5}}}-\frac{2T^{\frac{3}{5}}}{5K^{\frac{4}{5}}}+\frac{K^{\frac{6}{5}}}{96T^{\frac{2}{5}}}-\frac{4K^{\frac{1}{5}}}{12T^{\frac{2}{5}}}\\
&\geq\frac{K^{\frac{3}{5}}T^{\frac{4}{5}}}{40},
\end{aligned}$$

where in the last inequality we use that $K\leq T^{\frac{1}{3}}$.

Finally we get:

$$\mathbb{E}_j^*[\Re]\geq\frac{K^{\frac{3}{5}}T^{\frac{4}{5}}}{40}.$$

□

Chapter 7

Full information rising restless MAB with linear drift

In this chapter we address the full information setting, where in each time step we observe for each arm $i \in \mathcal{K}$ the reward $r_i(N_i(t))$. We use similar technics to those in Chapters 4, therefore we assume that Lemma 37 and Lemma 38 hold as well for algorithms FIR-ed-EE. Showing both upper and lower bounds we show a tight regret bound of $\tilde{\Theta}(T^{\frac{4}{5}})$.

Algorithm FIR-ed-EE solves the full information restless MAB with linear drift problem. As algorithm R-ed-EE , algorithm FIR-ed-EE use the explore-exploit methodology, and bound the expected regret with $O(T^{\frac{4}{5}}\Phi^{\frac{3}{5}}\ln\left(\frac{2}{\delta}\right)^{\frac{1}{5}})$, reducing factor of $K^{3/5}$ from the bandit feedback problem.

The main idea of FIR-ed-EE is explore for $2MK$ times and for the rest of the time exploit. For the exploration faze FIR-ed-EE use algorithm FE twice and observe the reward of each arm for $2M$ time. In the exploitation faze FIR-ed-EE estimate two point on the line of the reward, using those estimation it calculate the slope and the cumulative future rewards of the arms and choose to play the arm with the maximal cumulative future rewards.

In more details, Algorithm FE have as input (K, a, b) where a is the first time step to sample and b is the last. FE then sample each arm in a round robin manner, observe the reward of each arm and then sum a weighted average of the rewards of each arm. Then FE return an estimation of the expected reward in time $(b+a)/2$ for each arm in $\mathcal{K}$. Thereby algorithm FIR-ed-EE have estimation for both points in time steps $M/2$ and $3M/2$. From now on the algorithm act the same as algorithm R-ed-EE, calculating estimation of the slope and playing the arm with the maximal future cumulative reward.

Definition 20. *Let G be the good event that after using FE 2 times, for any arm $i \in \mathcal{K}$: $\left|\hat{\mu}_i^M\left(\frac{M}{2}\right) - \mu_i\left(\frac{M}{2}\right)\right| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2K(M-1)}}$ and $\left|\hat{\mu}_i^M\left(\frac{3M}{2}\right) - \mu_i\left(\frac{3M}{2}\right)\right| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2K(M-1)}}$.*

Algorithm 4 FIR-ed-EE - Full Information Rested Explore Exploit

Input: K, T, M
Set $\hat{\mu}^{K(M-1)}\left(\frac{M}{2}\right) = FE(K, 1, M)$
Set $\hat{\mu}^{K(M-1)}\left(\frac{3M}{2}\right) = FE(K, M+1, 2M)$
for $i \in \mathcal{K}$ **do**
    $\widehat{L}_i^{2K(M-1)} \leftarrow \frac{\hat{\mu}_i^{K(M-1)}\left(\frac{3M}{2}\right) - \hat{\mu}_i^{K(M-1)}\left(\frac{M}{2}\right)}{M}$
    $\hat{s}_i^{2K(M-1)}(2M+1, T-2KM+2M) \leftarrow \sum_{t=\frac{M}{2}+1}^{T-2KM+\frac{M}{2}} \hat{\mu}_i^{K(M-1)}\left(\frac{3M}{2}\right) + t\hat{L}_i^{2M}$
end for
$\hat{i}^* \leftarrow \arg\max_{i \in \mathcal{K}} \hat{s}_i^{2K(M-1)}(2M+1, T-2KM+2M)$
for $t \in [2KM+1, T]$ **do**
    Play arm $\hat{i}^*$
end for

Algorithm 5 FE - Full Explore

Input: K, a, b
Define $\hat{\mu}$ Array of zeros
for $n \in [a, b]$ **do**
    for arm $i \in \mathcal{K}$ **do**
        Sample arm i and observe $\{r_1^j(n+1), \ldots, r_{i-1}^j(n+1), r_i^j(n), \ldots r_K^j(n)\}$
        for $j \in \mathcal{K}$ **do**
            if $n = a$ and $j \geq i$ **then** Set $w_j = \frac{1}{j}$
            else if $n = b$ and $j < i$ **then** Set $w_j = 0$
            else: Set $w_j = \frac{1}{K}$
            end if
        end for
        $\hat{\mu} = \hat{\mu} + [w_1 r_1^j(n+1), \ldots, w_{i-1} r_{i-1}^j(n+1), w_i r_i^j(n), \ldots, w_K r_K^j(n)]$
    end for
end for
Return $\frac{1}{b-a+1}\hat{\mu}$

Lemma 21. *The probability of the good even G is at lest $1-2\delta K$*

Proof. Lets observe the way Algorithm FE handle each arm $i \in \mathcal{K}$. In each round $t \in [a,b]$, the algorithm observes the reward in time step t for i times and observes the reward at time step $t+1$ for $k-i$ times. Therefor FE observe the reward at time step a for i times, K times for time steps $t \in [a+1,b]$ and $K-i$ for time steps $b+1$. Then algorithm FE sum the rewards with weights for each time step and arm. In this way the expectation of the cumulative weighted reward for each time step is the expected reward in that time step. the only exception is FE doesn't sum the rewards for time steps $b+1$.

Have that,

$$\hat{\mu}_i = \frac{1}{b-a+1}\left(\sum_{j=1}^{i}\frac{1}{i}r_i^j(a) + \sum_{n=1}^{b}\frac{1}{K}\sum_{j=1}^{K}r_i^j(n)\right).$$

Hence,

$$\mathbb{E}[\hat{\mu}_i] = \frac{1}{b-a+1}\left(\mathbb{E}[\sum_{j=1}^{i}\frac{1}{i}r_i^j(a)] + \sum_{n=a+1}^{b}\mathbb{E}[\frac{1}{K}\sum_{j=1}^{K}r_i^j(n)]\right) = \mu_i(\frac{b+a}{2}).$$

Now we have that each arm was sampled at least $2K(M-1)$ times, using Lemma 33 we have that with probability of at least $1-2\delta$:

$$\left|\hat{\mu}_i^M\left(\frac{M}{2}\right) - \mu_i\left(\frac{M}{2}\right)\right| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2K(M-1)}},$$

$$\left|\hat{\mu}_i^M\left(\frac{3M}{2}\right) - \mu_i\left(\frac{3M}{2}\right)|\right| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2K(M-1)}}$$

Using union bound over all K arms we get that G holds with probability of at least $1-2\delta K$. □

Theorem 22. *for $\delta = \frac{1}{2TK\Phi}$ and $M = \frac{T^{\frac{4}{5}}\ln\left(\frac{2}{\delta}\right)^{\frac{1}{5}}}{\Phi^{\frac{2}{5}}K}$ Algorithm FIR-ed-EE guarantees regret $O(T^{\frac{4}{5}}\Phi^{\frac{3}{5}}\ln\left(\frac{2}{\delta}\right)^{\frac{1}{5}})$.*

Proof. Similar to the proof of Theorem 5 we can partition to three parts. The first part is the exploration faze in time $[1, 2KM]$ which in the worst case the regret is $2KM\Phi$.

The second part is in the time $[2KM+1, T]$ when the good event G holds and by Lemma 37 and Lemma 38, we have that for each arm $i \in \mathcal{K}$, we bound

the estimation error by,

$$\begin{aligned}|\hat{s}_i^{2K(M-1)}(2M+1, T-2KM+2M) - s_i(2M+1, T-2KM+2M)| &\leq \Gamma^{2K(M-1)}(1,T) \\ &\leq T^2\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2(K(M-1))^3}} \\ &\leq T^2\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{(MK)^3}}.\end{aligned}$$

The third part bounds the regret when the event G does not hold, i.e., $\bar{G}$ occurs. By Lemma 4 we have $P(\bar{G}) \leq 2K\delta$, so the regret of this part is bounded by $P(\bar{G})T\Phi \leq 2KT\Phi\delta$.

And we have that $\Re_{FIR-ed-EE} \leq 2M\Phi + T^2\sqrt{\frac{1}{(MK)^3}\ln\left(\frac{2}{\delta}\right)} + 2KT\Phi\delta$, and for $\delta = \frac{1}{2TK\Phi}$ and $M = \frac{T^{\frac{4}{5}}\ln\left(\frac{2}{\delta}\right)^{\frac{1}{5}}}{\Phi^{\frac{2}{5}}K}$ the regret is $O(T^{\frac{4}{5}}\Phi^{\frac{3}{5}}\ln(\Phi KT)^{\frac{1}{5}})$. □

Theorem 23. *The regret for full information rising rested MAB with linear drift is lower bounded with* $\Omega(T^{\frac{4}{5}})$

Proof. The proof of Theorem 17 for $K = 2$ holds for the full information case up to a constant of 2. Therefore the expected regret bounded with $\Omega(T^{\frac{4}{5}})$. □

Chapter 8

Rested MAB with Linear Drift

In the follow chapter we derive two algorithms for the rested MAB with linear drift problem and show a tight regret bound of $\tilde{\Theta}\left(K^{3/5}T^{4/5}\right)$.

Remark In this chapter the algorithm, theorems and proofs are similar to those in Chapters 4 and 6 except for minor changes, due to that we will assume that Lemma 4, Lemma 37 and Lemma 38 hold as well for algorithms DR-ed-LD and R-ed-LD.

The following are a few definitions which are related to the fact that we have both rising and rotting bandits.

Definition 24. *Given arm set $\mathcal{K}$ and horizon T.*

- *$\mathcal{K}_n$ and $\mathcal{K}_p$ are subsets of the set $\mathcal{K}$, where $\mathcal{K}_n = \{i \mid i \in \mathcal{K},\ L_i < 0\}$ is the subset of arms with negative slope in $\mathcal{K}$, and $\mathcal{K}_p = \{i \mid i \in \mathcal{K},\ L_i \geq 0\}$ is the subset of arms with non-negative slope in $\mathcal{K}$.*
- *$V_K^T = \left\{v | v \in \{\mathbb{N} \cup \{0\}\}^K, \|v\|_1 = T\right\}$ as the set of non-negative integer vectors of size K and norm 1 of T.*
- *$Q_{\mathcal{K}}^T \in \{\mathbb{R} \times \mathcal{K}\}^T$ is the optimal action selection. $Q_{\mathcal{K}}^T$ record tuples of expected value for arms $i \in \mathcal{K}$ in time steps $t \in [T]$ and the associated arm, i.e., $\{(\mu_i(t), i)\}$. $Q_{\mathcal{K}}^T$ sort the tuples using the first value $\mu_i(t)$ where the larger value come first, and keep only the first T items.*
- *$\Phi \geq \max_{i \in \mathcal{K}}(\max(\mu_i(T), \mu_i(1)))$ be an upper bound on the maximum expected reward of any arm.*

Notations for rising and rotting bandits.

- We denote a rising (res., rotting) rested MAB with linear drift game with arm set $\mathcal{K}$ and horizon h as $Ris(\mathcal{K}, h)$ (res. $Rot(\mathcal{K}, h)$).

- From Lemma 43, for any vector $v \in V_K^T$ we refer π_v as it respective policy. Policy π_v is a deterministic policy and for any arm $i \in \mathcal{K}$ policy π_v samples arm i for $v(i)$ times.

  In-addition for any deterministic policy π we denote $v_\pi \in V_K^T$ as it respective vector if π is the respective policy of v_π.

- For array A, $A[a,b]$ denotes the sub-array of A that includes entries $A[i]$ for $i \in [a,b]$.

- We denote the n'th largest reward value in a game with arm set $\mathcal{K}$ and horizon h with $Select(\{\mu_i(t)\}_{i\in\mathcal{K},t\in[h]}, n)$, and the action of that arm with $ArgSelect(\{\mu_i(t)\}_{i\in\mathcal{K},t\in[h]}, n)$.

The first algorithm is DR-ed-LD, a deterministic algorithm which receives for each arms in $\mathcal{K}$ its slope and the true reward at time step 1.

The main idea behind DR-ed-LD is separating the set of arms $\mathcal{K}$ to two subsets, $\mathcal{K}_n$ the subset of the descending arms and $\mathcal{K}_p$ the subset of the rising arms. Then, for each $h \in [T] \cup \{0\}$, the algorithm finds the optimal policy and its reward for the games $Rot(\mathcal{K}_n, h)$ and $Ris(\mathcal{K}_p, T-h)$. Next, the algorithm finds $h \in [T]\cup\{0\}$ such that, the optimal cumulative reward of games $Rot(\mathcal{K}_n, h)$ and $Ris(\mathcal{K}_p, T-h)$ is the highest. Finally, the algorithm return the concatenation of the respective vectors of the optimal policies for games $Rot(\mathcal{K}_n, h)$ and $Ris(\mathcal{K}_p, T-h)$.

The following theorem states the correctness of our algorithm.

Theorem 25. *Algorithm DR-ed-LD returns the vector representation of optimal policy for the deterministic rewards rested MAB with linear drift game with arm set $\mathcal{K}$ and horizon T.*

Proof. Let π_v be a deterministic optimal policy for the problem and $v \in V_K^T$ as it vector representation. (There is a deterministic optimal policy because we can represented the problem using a MDP.)

For contradiction, assume that $[v_n,\ v_p]$ in the algorithm is a vector representation of a sub-optimal policy. Then,

$$\begin{aligned}\sum_{i\in\mathcal{K}}\sum_{t=1}^{v(i)}\mu_i(t) &= \sum_{i\in\mathcal{K}_n}\sum_{t=1}^{v(i)}\mu_i(t) + \sum_{i\in\mathcal{K}_p}\sum_{t=1}^{v(i)}\mu_i(t)\\ &> \sum_{i\in\mathcal{K}_n}\sum_{t=1}^{v_n(i)}\mu_i(t) + \sum_{i\in\mathcal{K}_p}\sum_{t=1}^{v_p(i)}\mu_i(t)\\ &= \max_{h\in[T]\cup\{0\}}[\sum_{t=1}^{h} Select(\{\mu_i(n)\}_{i\in\mathcal{K}_n,n\in[h]}, t) + \max_{a\in\mathcal{K}_p}\left\{\sum_{t=1}^{T-h}\mu_a(t)\right\}].\end{aligned}$$

The second equality follows from the algorithm. From lines $17-19$ we have that $\sum_{i\in\mathcal{K}_n}\sum_{t=1}^{v_n(i)}\mu_i(t)+\sum_{i\in\mathcal{K}_p}\sum_{t=1}^{v_p(i)}\mu_i(t) = \max_{h\in[0,T]}\{g_v[h] + f_v(T-h)\}$. The way the algorithm choose $g_v[h]$ is using a realization of the *select* function over arm

Algorithm 6 DR-ed-LD: Deterministic Rested Linear Drift

Input: $\mathcal{K}$, T, L, $\mu(1)$
Define A, g_k, g_v arrays
$g_v[0] = 0$; $f_v[0] = 0$
Set $\mathcal{K}_p = \{i \in \mathcal{K} |\ L_i \geq 0\}$ ▷ The set of rising arms.
Set $\mathcal{K}_n = \mathcal{K} \setminus \mathcal{K}_p$ ▷ The set of rotting arms.
Set $f_k = [\text{argmax}_{i \in \mathcal{K}_p} \left\{ \sum_{n=1}^{t} \mu_i(1) + (n-1)L_i \right\}]_{t=0}^{T}$ ▷ $f_k(t)$ Hold the optimal arm for the game $Ris(K_p, t)$.
Set $f_v = [\sum_{n=1}^{t} \mu_{f_k(t)}(t)]_{t=1}^{T}$ ▷ $f_v(t)$ hold the optimal cumulative reward value for the game $Ris(K_p, t)$
for $i \in [\mathcal{K}_n]$ **do**
    $A[i] = \mu_i(1)$
end for
for $t \in [T]$ **do**
    $v = \max_i A[i]$ ▷ The t'th largest reward value from the arms in $\mathcal{K}_n$.
    $g_k[t] = \arg\max_i A[i]$ ▷ The arm with the t'th largest reward value from the arms in $\mathcal{K}_n$.
    $A[g_k[t]] = v - L_{g_k[t]}$ ▷ update the value of arm $g_k[t]$ to the value of the next time step.
    $g_v[t] = g_v[t-1] + v$ ▷ Hold the optimal cumulative reward value for the game $Rot(K_n, t)$.
end for
$h = \underset{x \in [0,T]}{\text{argmax}} \{g_v[x] + f_v(T-x)\}$ ▷ Find h which maximize the cumulative reward of $Ris(K_p, T-h)$ and $Rot(K_n, h)$.
Set v_n when $\forall i \in \mathcal{K}_n$: $v_n(i) = \sum_{t=1}^{h} \mathbb{1}\{g_k[x] = i\}$ ▷ The vector representation for the optimal policy for the game $Rot(K_n, h)$.
Set $v_p = e_{f_k[T-h]}(T-h)$ ▷ The vector representation for the optimal policy for the game $Ris(K_p, T-h)$.
return $[v_n,\ v_p]$. ▷ Concatenation of the to vectors.

set $\mathcal{K}_n$ and horizon h. In the algorithm $f_v(T-h)$ is the maximal cumulative reward for playing one arm from $i \in \mathcal{K}_p$ for horizon $T-h$.

If so, for $h = \sum_{i \in \mathcal{K}_n} v(i)$ at least one of the follow inequalities hold.

$$\sum_{t=1}^{h} \sum_{i \in \mathcal{K}_n} \mu_i(N_i(t)) \mathbb{1}\{\pi_v(t) = i\} = \sum_{i \in \mathcal{K}_n} \sum_{t=1}^{v(i)} \mu_i(t) > \sum_{t=1}^{h} Select(\{\mu_i(n)\}_{i \in \mathcal{K}_n, n \in [h]}, t), \tag{8.1}$$

$$\sum_{i \in \mathcal{K}_p} \sum_{t=1}^{v(i)} \mu_i(t) > \max_{a \in \mathcal{K}_p} \left\{ \sum_{t=1}^{T-h} \mu_a(t) \right\}. \tag{8.2}$$

If inequality 8.1 hold, then for some $l \in [h]$,

$$\sum_{t=1}^{l-1} \sum_{i \in \mathcal{K}} \mu_i(N_i(t)) \mathbb{1}\{\pi_v(t) = i\} = \sum_{t=1}^{l-1} Select(\{\mu_i(n)\}_{i \in \mathcal{K}_n, n \in [h]}, t),$$

and

$$\sum_{i \in \mathcal{K}} \mu_i(N_i(l)) \mathbb{1}\{\pi_v(l) = i\} > Select(\{\mu_i(n)\}_{i \in \mathcal{K}_n, n \in [h]}, l),$$

in contradiction to that $Select(\{\mu_i(n)\}_{i \in \mathcal{K}_n, n \in [h]}, l)$ is the largest l'th reward value.

From inequality 8.2 and Corollary 2 we know that for a rising rested MAB with linear drift the optimal policy is playing only one arm. Therefore $\max_{a \in \mathcal{K}_p} \left\{ \sum_{t=1}^{T-h} \mu_a(t) \right\}$ is the maximal value for playing arms K_n for $T-h$ step, thereby inequality 8.2 do not hold.

Namely, policy $\pi_{[v_n, v_p]}$ is a vector representation of optimal policy. □

The second algorithm is R-ed-LD, an algorithm for the non-deterministic rested MAB with linear drift problem. Same As algorithm R-ed-EE, algorithm R-ed-LD use an explore-exploit methodology as well. The exploration part in both algorithms stay the same so we will not elaborate on that part.

In the exploitation part, we have for each arm $i \in \mathcal{K}$ the estimations of the expectation of the rewards in time step $M/2$ and $3M/2$, denoted with $\hat{\mu}_i^M(M/2)$ and $\hat{\mu}_i^M(3M/2)$ respectively. In addition we have the estimation of the slope $\hat{L}_i^{2M}$. Using those estimations algorithm R-ed-LD use algorithm DR-ed-LD with parameters, $\mathcal{K}$ as the set of arms, $T - 2KM$ as the horizon, $\hat{L}^{2M}$ as the array of the slops and $\hat{\mu}^M(\frac{3M}{2}) + \frac{M}{2}\widehat{L}^{2M}$ as the array of the reward values in time step $2M+2$. Algorithm DR-ed-LD returns a vector representation for a policy of game with horizon $T - 2kM$. R-ed-LD plays this policy for the rest of the time.

The good event G is define as before.

Definition 26. *Event G as in Definition 3.*

Algorithm 7 R-ed-LD: Rising Linear Drift

Input: K,T,M
Set $\hat{L}$, $\hat{\mu}^M(\frac{M}{2})$,$\hat{\mu}^M(\frac{3M}{2})$ Array
for $i \in \mathcal{K}$ **do**
    Sample arm i for $2M$ times, and observe $r_i(1), \ldots, r_i(2M)$
    $\hat{\mu}_i^M\left(\frac{M}{2}\right) \leftarrow \frac{1}{M}\sum_{n=1}^{M} r_i(n)$
    $\hat{\mu}_i^M\left(\frac{3M}{2}\right) \leftarrow \frac{1}{M}\sum_{n=1}^{M} r_i(n)$
    $\widehat{L}_i^{2M} \leftarrow \frac{\hat{\mu}_i^M\left(\frac{3M}{2}\right)-\hat{\mu}_i^M\left(\frac{M}{2}\right)}{M}$
end for
$v = DR-ed-LD(K, T-2KM, \widehat{L}, \hat{\mu}^M(\frac{3M}{2}) + \frac{M}{2}\widehat{L}^{2M})$
for $i \in \mathcal{K}$ **do**
    play arm i $v(i) times$
end for

Theorem 27. *For* $M = \frac{T^{\frac{4}{5}}\ln(4\Phi KT)^{\frac{1}{5}}}{(\Phi K)^{\frac{2}{5}}}$*, Algorithm R-ed-LD guarantees regret of* $O(T^{\frac{4}{5}}(\Phi K)^{\frac{3}{5}}\ln(\Phi KT)^{\frac{1}{5}})$ *for Rested MAB with Linear Drift.*

Proof. The regret can be partition to three parts. The first part is during times $[1, 2KM]$ when we explore each arm $2M$ times. The second part is during times $[2KM+1, T]$, assuming the good event G holds. The third part is during times $[2KM+1, T]$, when the good event G does not hold.

For the first part, i.e., the regret over the first $2M$ samples of each arm, we bound the regret by $2KM\Phi$.

For the second part, the regret during exploitation stage, assuming the good event G holds. Under good event G, by Lemma 37 and Lemma 39, we have that for any vector $v \in \{\mathbb{N}\cup\{0\}\}^K$ with $\|v\|_1 \leq T-2KM$, we bound the estimation error by,

$$\begin{aligned}\left|\sum_{i\in\mathcal{K}}\sum_{n=1}^{v(i)}\hat{\mu}_i(t) - \sum_{i\in\mathcal{K}}\sum_{n=1}^{v(i)}\mu_i(t)\right| &\leq \sum_{i\in\mathcal{K}}\Gamma^{2M}(2M+1, 2M+v(i)) \\ &\leq \Gamma^{2M}(2M+1, T-2KM) \\ &\leq T^2\sqrt{\frac{1}{2M^3}\ln\left(\frac{2}{\delta}\right)}.\end{aligned}$$

Note that the regret of the second part is independent of Φ. This is since our confidence intervals do not depend on the magnitude rewards, i.e., Φ.

The third part, we bound the regret when the event G does not hold, i.e., $\bar{G}$ occurs. By Lemma 4 we have $P(\bar{G}) \leq 2K\delta$, so the regret of this part is bounded by $P(\bar{G})T\Phi \leq 2KT\Phi\delta$.

To summarize, we have $\Re \leq 2MK\Phi + \frac{T^2\sqrt{\ln\left(\frac{2}{\delta}\right)}}{\sqrt{2}M^{1.5}} + 2KT\Phi\delta$. Setting $\delta = \frac{1}{2TK\Phi}$ and $M = \frac{T^{\frac{4}{5}}\ln(4KT\Phi)^{\frac{1}{5}}}{(\Phi K)^{\frac{2}{5}}}$, we get that $\Re \leq O(T^{\frac{4}{5}}(\Phi K)^{\frac{3}{5}}\ln(\Phi KT)^{\frac{1}{5}})$. □

Theorem 28. *Rested K-MAB with linear drift is lower bounded with* $\Omega(T^{\frac{4}{5}}K^{\frac{3}{5}})$.

Proof. We know that Rested Rising K-MAB with linear drift is a sub problem of Rested K-MAB with linear drift. Thereby from Theorem 17 we have that Rested K-MAB with linear drift is lower bounded with $\Omega(T^{\frac{4}{5}}K^{\frac{3}{5}})$. □

Chapter 9

Restless rising MAB with linear drift

In this chapter we tackle the restless rising MAB with linear drift problem.

Remark 29. *The restless rising MAB with linear drift and restless rotting MAB with linear drift are identical problems. Considering the rotting framework with costs is equal to the rising setting with rewards, and vice versa.*

Auer et al. (2002) introduced algorithm $Exp3.S$, which bound the dynamic regret under adversarial setting with expected regret of $O(\sqrt{SKT\ln(KT)})$, where K is the number of arms, T is the horizon and S is the number of times the optimal arm has been changed in the game regarded to the expectancy of the arms. (*i.e.*, $S = \sum_{t=1}^{T-1} \mathbb{1}[\text{argmax}_{i\in\mathcal{K}} \{\mu_i(t)\} \neq \text{argmax}_{i\in\mathcal{K}} \{\mu_i(t+1)\}]$).

Noticing that restless rising MAB with linear drift problem is a sub-problem of the adversarial setting. Due to the linearity of the arms we know that the number of times the optimal arm can change is bounded with $K-1$. Therefore $Exp3.S$ bound the restless rising MAB with linear drift problem with $O(K\sqrt{T\ln(KT)})$.

However when $K > T^{1/4}$ we can obtain better regret bound which we will achieve using algorithm R-es-BEE. The main idea of algorithm R-es-BEE is dividing the game to $\log(T/2)$ growing blocks, in each block the algorithm use an explore-exploit methodology in which explore each arm M times. Given the observations the algorithm estimate the slope of the linear function of the arm and the expectation reward of a point in the block on the arm.

In more details the sizes of the blocks will denoted with $B = \left\{2^{\log(MK)}, \ldots, 2^{\log(\frac{T}{2})}\right\}$, in each block $b \in [|B|]$ the algorithm sample each arm $i \in \mathcal{K}$ for M time steps with interval of size K. The first block is entirely dedicated to exploration, for any other block $b \in [2, |B|]$ for each arm $i \in \mathcal{K}$ the algorithm use the observations from the current exploration faze to estimate the expected value of the arm at time $l_i^b = B_b + \frac{MK+K+2i+2}{2}$. We denote those values with $\hat{\mu}_i^M(l_i^b)$. In addition the algorithm use the estimations $\hat{\mu}_i^M(l_i^b)$ and $\hat{\mu}_i^M(l_i^1)$ to estimate the slope, which we denote with $\hat{L}_i^{2M}(b)$. For the remaining time steps $t \in [B_b + 2MK + 1, B_{b+1}]$

in the block the algorithm estimate the value of each arm in time step t and play the arm with the highest value.

Algorithm 8 R-es-BEE - Restless Blocks Explore Exploit

Input: K, T, M
Set : $B = [2^{\log(MK)}, 2^{\log(MK)+1}, .., 2^{\log(\frac{T}{2})}]$
Set: $B_0 = 0$
for $b \in [0, |B|]$ **do**
    Set: $\hat{\mu}_i^M(b) = 0 : \ \forall\, i \in \mathcal{K}$
    Set: $l_i^b = B_b + \frac{MK+K+2i+2}{2} : \ \forall\, i \in \mathcal{K}$
    Set: $time = 0$
    for $t \in [0, M-1]$ **do**
        for $i \in \mathcal{K}$ **do**
            sample arm i, $\hat{\mu}_i^M(l_i^b) \leftarrow \hat{\mu}_i^M(l_i^b) + \frac{r_i(B_b+tK+(i-1))}{M}$
            $time = time + 1$
        end for
    end for
    Set: $\hat{L}_i^{2M}(b) = \frac{\hat{\mu}_i^M(l_i^b) - \hat{\mu}_i^M(l_0)}{B_b} \ \forall\, i \in \mathcal{K}$
    for $t \in [time + 1, time + B_b - MK]$ **do**
        Set $\hat{i}^*(t) = \underset{i \in \mathcal{K}}{\operatorname{argmax}} \{\hat{\mu}_i^M(l_i^b) + (t - l_i^b)\hat{L}_i^{2M}(b)(t - l_i^b)\}$
        sample arm $\hat{i}^*(t)$
        $time = time + 1$
    end for
end for

Definition 30. *Let G be the good event, which for every block $b \in [0, |B|]$ for any arm $i \in \mathcal{K}$,*

$$|\mu_i(l_i^b) - \hat{\mu}_i^M(l_i^b)| \leq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2M}}$$

Lemma 31. *The good event G hold with probably of at least $1 - K\log(T)\delta$*

Proof. From the Lemma 33 we have that for arm $i \in \mathcal{K}$ and $b \in [0, |B|]$ with probability of at least $1 - \delta$,

$$|\mu_i(l_i^b) - \hat{\mu}_i^M(l_i^b)| \leq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2M}}$$

Using union bound over K and $|B| + 1 = \log(T)$ we get that G holds with probability of at least $1 - K\log(T)\delta$. □

Theorem 32. *Algorithm R-es-BEE with parameter $M = \frac{T^{\frac{2}{3}}\left(2\ln\left(\frac{1}{\delta}\right)\right)^{\frac{1}{3}}}{(K\log(T)\Phi)^{\frac{2}{3}}}$ guarantees regret of $O((K\Phi)^{\frac{1}{3}}T^{\frac{2}{3}}\ln(2K\log(T)T\Phi)^{\frac{1}{3}})$ for restless rising MAB with Linear Drift problem.*

Proof. Define $i^*(t)$ as the arm with the highest expected reward for each time step $t \in [T]$.

We partition the regret to three parts. The first part is the exploration phase in each block, which in the worst case the regret is $KM\log(T)\Phi$.

The second part is at the exploitation phase in the blocks when the good event G holds. We bound the regret for each block $b \in [|B|]$.

For any arm $i \in \mathcal{K}$ we can bound the estimation error of $\hat{L}_i^{2M}(b)$ by,

$$|L_i - \hat{L}_i^{2M}(b)| = |\frac{\mu_i(l_i^b) - \mu_i(l_0)}{B_b} - \frac{\hat{\mu}_i^M(l_i^b) - \hat{\mu}_i^M(l_0)}{B_b}| \leq \sqrt{\frac{2\log\left(\frac{2}{\delta}\right)}{MB_b^2}}$$

.

Thereby we can bound the estimations error of the algorithm for any time step $t \in [B_b + 2KM + 1, B_{b+1}]$,

$$\begin{aligned}
&|\mu_i(t) - \hat{\mu}_i^M(l_i^b) + (t - l_i^b)\hat{L}_i^{2M}(b)(t - l_i^b)| \\
&\leq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2M}} + \left(t - B_b - \frac{MK + K + 2i + 2}{2}\right)\sqrt{\frac{2\log\left(\frac{2}{\delta}\right)}{MB_b^2}} \\
&\leq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2M}} + \left(B_{b+1} - B_b - \frac{MK + K + 2i + 2}{2}\right)\sqrt{\frac{2\log\left(\frac{2}{\delta}\right)}{MB_b^2}} \\
&\leq 3\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{M}}.
\end{aligned}$$

Using Lemma 12 we get that $\mu_{i^*}(t) - \mu_{\hat{i}^*}(t) \leq 6\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{M}}$, hence,

$$\sum_{b\in[|B|]}\sum_{t=B_b+MK+1}^{B_{b+1}} \mu_{i^*}(t) - \mu_{\hat{i}^*}(t) \leq \sum_{b\in[|B|]}\sum_{t=B_b+MK+1}^{B_{b+1}} 6\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{M}} \leq 6T\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{M}}.$$

Therefore, in this part the algorithm suffer regret of at most $6T\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{M}}$.

The third part bounds the regret when the event G does not hold, i.e., $\bar{G}$ occurs. By Lemma 31 we have $P(\bar{G}) \leq 2K\log(T)\delta$, so the regret of this part is bounded by $P(\bar{G})T\Phi \leq 2KT\log(T)\Phi\delta$.

All together we get that $\Re \leq MK\log(T)\Phi + 6T\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{M}} + 2KT\log(T)\Phi\delta$. For $M = \frac{T^{\frac{2}{3}}\left(\ln\left(\frac{1}{\delta}\right)\right)^{\frac{1}{3}}}{(K\log(T)\Phi)^{\frac{2}{3}}}$ and $\delta = \frac{1}{2KT\log(T)\Phi}$ we get a regret bound of $O((K\Phi)^{\frac{1}{3}}T^{\frac{2}{3}}\ln\left(KT\log(T)\Phi\right)^{\frac{1}{3}})$

□

Open question We leave this chapter with an open question of finding a tight for the rising restless MAB with linear drift problem. Sow the current lower bound is $\sqrt{KT}$, *i.e.*,the lower bound for the classic MAB.

Chapter 10

Discussion

In this work we addressed the rising rested MAB with linear drift problem, i.e., the expectation of the reward function of each arm is linear non-decreasing with respect to the number of time the arm was played. We bounded the dynamic regret of the problem with both upper and lower bounds, deriving a tight dynamic regret of $\tilde{\Theta}(T^{4/5}K^{3/5})$. In addition we provided a tight bound for the full information feedback setting, which reduce the $K^{3/5}$ factor to a $\tilde{\Theta}(T^{4/5})$ regret bound. Furthermore, we showed an instance dependent regret bound algorithm, which can achieve better regret in some parameters, and in the worst case its regret is still near optimal.

Regarding the rotting rested MAB with linear drift. The work of Seznec et al. (2019) considered the rotting MAB problem, and showed dynamic regret upper bound of $\tilde{O}(\sqrt{KT})$. which solve the rotting rested MAB with linear drift problem with a tight regret bound as well.

We extended our techniques to handle a mixture of the rested rotting and rising bandits which have tight dynamic regret of $\tilde{\Theta}(T^{4/5}K^{3/5})$ as well.

As for the restless setting we show a $\tilde{O}(\min\{K\sqrt{T}, K^{1/3}T^{2/3}, T\})$ upper bound in face of the known $\sqrt{KT}$ lower bound, leaving an open problem.

It is evident from our results, that in terms of regret, the rising rested MAB with linear drift problem has a higher regret than the rotting rested MAB with linear drift problem or even the stationary stochastic MAB. Here is an informal intuition why this happens. In the rising rested MAB with linear drift problem, when we fix an optimal arm, the higher reward values of the optimal arm are obtained when the number of samples are close to T. This implies that each time the learner does not play the optimal arm it might incurs a significant regret. For this reason we bound the regret by the number of times we pull the non-optimal arms, and cannot multiply it by a sub-optimality gap, as done in stationary stochastic MAB. This is the major source to the significantly higher regret in our setting, which is inherent, as our lower bound shows.

Bibliography

R. Arora, O. Dekel, and A. Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. *arXiv preprint arXiv:1206.6400*, 2012.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*, pages 322–331. IEEE, 1995.

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

D. Bertsimas and J. Niño-Mora. Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operations Research*, 48(1):80–90, 2000.

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

L. Cella, M. Pontil, and C. Gentile. Best model identification: A rested bandit formulation. In *International Conference on Machine Learning*, pages 1362–1372. PMLR, 2021.

Y. Freund and Y. Mansour. Learning under persistent drift. In *Computational Learning Theory: Third European Conference, EuroCOLT'97 Jerusalem, Israel, March 17–19, 1997 Proceedings 3*, pages 109–118. Springer, 1997.

Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

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

J. Gittins. A dynamic allocation index for the sequential design of experiments. *Progress in statistics*, pages 241–266, 1974.

J. Gittins, K. Glazebrook, and R. Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

H. Heidari, M. J. Kearns, and A. Roth. Tight policy regret bounds for improving and decaying bandits. In *IJCAI*, pages 1562–1570, 2016.

S. Jia, Q. Xie, N. Kallus, and P. I. Frazier. Smooth non-stationary bandits. In *International Conference on Machine Learning*, pages 14930–14944. PMLR, 2023.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

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

Y. Li, J. Jiang, J. Gao, Y. Shao, C. Zhang, and B. Cui. Efficient automatic cash via rising bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

A. M. Metelli, F. Trovo, M. Pirola, and M. Restelli. Stochastic rising bandits. In *International Conference on Machine Learning*, pages 15421–15457. PMLR, 2022.

M. Mussi, A. Montenegro, F. Trovó, M. Restelli, and A. M. Metelli. Best arm identification for stochastic rising bandits. *arXiv preprint arXiv:2302.07510*, 2023.

J. Nino-Mora. Restless bandits, partial conservation laws and indexability. *Advances in Applied Probability*, 33(1):76–98, 2001.

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

J. Seznec, P. Menard, A. Lazaric, and M. Valko. A single algorithm for both restless and rested rotting bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3784–3794. PMLR, 2020.

A. Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

A. Slivkins and E. Upfal. Adapting to a changing environment: the brownian restless bandits. In *COLT*, pages 343–354, 2008.

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

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

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

Appendix A

Concentration Lemmas

Lemma 33. *Hoeffding's inequality: Let $X_1, X_2, \ldots, X_m$ independent random variables such for every $i \in [m]$ $X_i \in [0,1]$ and $\mu_i = \mathbb{E}[X_i]$, for $\delta \geq 0$*

$$Pr\left[\left|\frac{1}{m}\sum_{i\in[m]}(X_i - \mu_i)\right| \geq \delta\right] \leq 2\exp\left(-2m\delta^2\right).$$

Lemma 34. *In Rising Rested MAB with Linear Drift for any arm $i \in \mathcal{K}$, $m \in \mathbb{N}$ even number, $n' \in \mathbb{N}$ and set of samples $r_i = \{r_i(n'), \ldots, r_i(n'+m)\}$ of arm i, with probability of at least $1-\delta$,*

$$\left|\hat{\mu}_i^m(n' + \frac{m}{2}) - \mu_i(n' + \frac{m}{2})\right| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2m}}.$$

Proof. From the definition we have that $\mu_i(n' + \frac{m}{2}) = \mathbb{E}\left[\hat{\mu}_i^m(n' + \frac{m}{2})\right] = \mathbb{E}\left[\sum_{t=n'}^{n'+m} r_i(n)\right]$, applying Lemma 33 give as,

$$\left|\hat{\mu}_i^m(n' + \frac{m}{2}) - \mu_i(n' + \frac{m}{2})\right| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2m}}.$$

□

Lemma 35. *In Rising Rested MAB with Linear Drift for any arm $i \in \mathcal{K}$, $M \in \mathbb{N}$ even number and set of sample $r_i = \{r_i(1), \ldots, r_i(2M)\}$ of arm i, with probability of at least $1-2\delta$ we have,*

$$\left|\hat{\mu}_i^M\left(\frac{3M}{2}\right) - \mu_i\left(\frac{3M}{2}\right)\right| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2M}},$$

$$\left|\hat{\mu}_i^M\left(\frac{M}{2}\right) - \mu_i\left(\frac{M}{2}\right)\right| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2M}},$$

$$\left|\widehat{L}_i^{2M} - L_i\right| \leq \frac{\sqrt{2\ln\left(\frac{2}{\delta}\right)}}{M^{1.5}}.$$

Proof. Using Lemma 34 we have that with probability of $1-\delta$ each $\left|\hat{\mu}_i^M\left(\frac{3M}{2}\right) - \mu_i\left(\frac{3M}{2}\right)\right| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2M}}$ and $|\hat{\mu}_i^M\left(\frac{M}{2}\right) - \mu_i\left(\frac{M}{2}\right)| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2M}}$, using union bound we get that with probability of at least $1-2\delta$ both inequality are holds, and,

$$|\widehat{L}_i^{2M} - L_i| = \frac{|[\hat{\mu}_i^M\left(\frac{3M}{2}\right) - \hat{\mu}_i^M\left(\frac{M}{2}\right)] - [\mu_i\left(\frac{3M}{2}\right) - \mu_i\left(\frac{M}{2}\right)]|}{M}] \leq \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{M}}/M.$$

□

Appendix B

General Lemmas and claims

Lemma 36. *In Rising Rested MAB with Linear Drift, if* $\left|\hat{\mu}_i^M\left(\frac{3M}{2}\right)-\mu_i\left(\frac{3M}{2}\right)\right| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2M}}$, $\left|\hat{\mu}_i^M\left(\frac{M}{2}\right)-\mu_i\left(\frac{M}{2}\right)\right| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2M}}$ *and* $\left|\widehat{L}_i^{2M}-L_i\right| \leq \frac{\sqrt{2\ln\left(\frac{2}{\delta}\right)}}{M^{1.5}}$ *holds for some arm* $i \in \mathcal{K}$*, then for any* $n \in [T]$*,*

$$\left|\psi_i^{2M}(n)-\mu_i(n)\right| \leq \gamma_n^{2M}.$$

Proof. We have that:

$$\mu_i(n)=\frac{\mu_i\left(\frac{3M}{2}\right)+\mu_i\left(\frac{M}{2}\right)}{2}+|n-M|\,L_i.$$

Together with the definition of $\psi_i^{2M}(n)$ we get that,

$$\begin{aligned}\left|\psi_i^{2M}(n)-\mu_i(n)\right| &= \left|\frac{\hat{\mu}_i^M(\frac{3M}{2})+\hat{\mu}_i^M(\frac{M}{2})}{2}+|n-M|\,\widehat{L}_i^{2M}-\frac{\mu_i(\frac{3M}{2})+\mu_i(\frac{M}{2})}{2}-|n-M|\,L_i\right| \\ &\leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2M}}+|n-M|\,\frac{\sqrt{2\ln\left(\frac{2}{\delta}\right)}}{M^{1.5}}=\gamma_n^{2M}.\end{aligned}$$

□

Lemma 37. *In Rising Rested MAB with Linear Drift, if* $\left|\hat{\mu}_i^M\left(\frac{3M}{2}\right)-\mu_i\left(\frac{3M}{2}\right)\right| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2M}}$, $\left|\hat{\mu}_i^M\left(\frac{M}{2}\right)-\mu_i\left(\frac{M}{2}\right)\right| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2M}}$ *and* $\left|\widehat{L}_i^{2M}-L_i\right| \leq \frac{\sqrt{2\ln\left(\frac{2}{\delta}\right)}}{M^{1.5}}$ *holds for some arm* $i \in \mathcal{K}$*, then*

$$\left|\hat{s}_i^{2M}(n_1,n_2)-s_i(n_1,n_2)\right| \leq \Gamma^{2M}(n_1,n_2).$$

Proof. Using Lemma 36 we have that $\left|\psi_i^{2M}(n)-\mu_i(n)\right| \leq \gamma_n^{2M}$ and we get that,

$$\left|\hat{s}_i^{2M}(n_1,n_2)-s_i(n_1,n_2)\right| \leq \sum_{n=n_1}^{n_2}\left|\psi_i^{2M}(n)-\mu_i(n)\right| \leq \sum_{n=n_1}^{n_2}\gamma_n^{2M}=\Gamma^{2M}(n_1,n_2).$$

□

Lemma 38. *For any $n_1, n_2 \in [T]$,*

$$\Gamma^{2M}(n_1, n_2) \leq \Gamma^{2M}(1, T) \leq T^2 \frac{\sqrt{\ln\left(\frac{2}{\delta}\right)}}{\sqrt{2}M^{1.5}}.$$

Proof.

$$\begin{aligned}
\Gamma^{2M}(n_1, n_2) \leq \Gamma^{2M}(1, T) &= \sum_{n=1}^{T} \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2M}} + |n - M| \frac{\sqrt{2\ln\left(\frac{2}{\delta}\right)}}{M^{1.5}} \\
&= 2M \left[\sqrt{\frac{\ln(\frac{2}{\delta})}{2M}} + \frac{M+1}{2} \frac{\sqrt{2\ln(\frac{2}{\delta})}}{M^{1.5}} \right] + (T - 2M) \left[\sqrt{\frac{\ln(\frac{2}{\delta})}{2M}} + \left(\frac{T+1}{2} \right) \frac{\sqrt{2\ln(\frac{2}{\delta})}}{M^{1.5}} \right] \\
&= (2M^2 + M) \frac{\sqrt{2\ln\left(\frac{2}{\delta}\right)}}{M^{1.5}} + (T - 2M) \left(\frac{T + M + 1}{2} \right) \frac{\sqrt{2\ln\left(\frac{2}{\delta}\right)}}{M^{1.5}} \\
&\leq T^2 \frac{\sqrt{\ln\left(\frac{2}{\delta}\right)}}{\sqrt{2}M^{1.5}}.
\end{aligned}$$

□

Lemma 39. *In Rested MAB with Linear Drift, if for some arm $i \in \mathcal{K}$, $h \in [T]$ and vector $v \in \mathbb{N} \cup \{0\}$ such $\|v\|_1 = h$, $\left|\hat{\mu}_i^M\left(\frac{3M}{2}\right) - \mu_i\left(\frac{3M}{2}\right)\right| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2M}}$, $|\hat{\mu}_i^M\left(\frac{M}{2}\right) - \mu_i\left(\frac{M}{2}\right)| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2M}}$ and $\left|\widehat{L}_i^{2M} - L_i\right| \leq \frac{\sqrt{2\ln\left(\frac{2}{\delta}\right)}}{M^{1.5}}$ holds , then*

$$\left| \sum_{i \in \mathcal{K}} \sum_{n=1}^{v(i)} \hat{\mu}_i(t) - \sum_{i \in \mathcal{K}} \sum_{n=1}^{v(i)} \mu_i(t) \right| \leq \Gamma^{2M}(1, h).$$

Proof. Using Lemma 36 we have that $\left|\psi_i^{2M}(n) - \mu_i(n)\right| \leq \gamma_n^{2M}$ and we get that,

$$\begin{aligned}
\left| \sum_{i \in \mathcal{K}} \sum_{n=1}^{v(i)} \hat{\mu}_i(t) - \sum_{i \in \mathcal{K}} \sum_{n=1}^{v(i)} \mu_i(t) \right| &\leq \sum_{i \in \mathcal{K}} \sum_{n=1}^{v(i)} \left|\psi_i^{2M}(n) - \mu_i(n)\right| \\
&\leq \sum_{n=1}^{h} \left|\psi_i^{2M}(n) - \mu_i(n)\right| \\
&\leq \sum_{n=}^{h} \gamma_n^{2M} = \Gamma^{2M}(1, h).
\end{aligned}$$

□

Claim 40.

$$\hat{s}_i^{2M}(n_1, n_2) = (n_2 - n_1 + 1) \left[\left(\hat{\mu}_i^M \left(\frac{3M}{2} \right) + \hat{\mu}_i^M \left(\frac{M}{2} \right) \right) / 2 + \left(\frac{n_2 + n_1}{2} - M \right) \widehat{L}_i^{2M} \right].$$

Proof.

$$\begin{aligned}\hat{s}_i^{2M}(n_1,n_2) &= \sum_{t=n_1}^{n_2} \psi_i^{2M}(n) \\ &= \sum_{t=n_1}^{n_2} \left(\hat{\mu}_i^M\left(\frac{3M}{2}\right) + \hat{\mu}_i^M\left(\frac{M}{2}\right)\right)/2 + (n-M)\widehat{L}_i^{2M} \\ &= (n_2-n_1+1)\left[\left(\hat{\mu}_i^M\left(\frac{3M}{2}\right) + \hat{\mu}_i^M\left(\frac{M}{2}\right)\right)/2 + \left(\frac{n_2+n_1}{2} - M\right)\widehat{L}_i^{2M}\right].\end{aligned}$$

□

Lemma 12. *Fix $x_i, x_j, \hat{x}_i, \hat{x}_j \geq 0$, if $|\hat{x}_i - x_i| \leq \gamma$, $|\hat{x}_j - x_j| \leq \gamma$ and $\hat{x}_i - \hat{x}_j \leq 2\gamma$, then, $x_i - x_j \leq 4\gamma$.*

Proof. $x_i - x_j \leq \hat{x}_i - \hat{x}_j + 2\gamma \leq 4\gamma$. □

Appendix C

Dynamic vs static regret

Definition 41.

1. *Let V be a set of vectors, where $V = \left\{v \in \{\mathbb{N} \cup \{0\}\}^K | \ \|v\|_1 = T\right\}$.*
2. *For $i \in \mathcal{K}$, let $e_i \in \mathbb{R}^K$ be the unit vector with 1 at the i'th entry and 0 at the rest.*

Lemma 42. *For any two linear monotonic increasing functions $f(x) = ax+b$ and $g(x)+cx+d$ and any $\alpha, \beta \in \mathbb{N} \cup \{0\}$, assuming w.l.o.g. $\sum_{x=1}^{\alpha+\beta} f(x) \geq \sum_{x=1}^{\alpha+\beta} g(x)$, then the follow holds,*

$$\sum_{x=1}^{\alpha+\beta} f(x) \geq \sum_{x=1}^{\alpha} f(x) + \sum_{x=1}^{\beta} g(x).$$

Proof. Note that,

$$\sum_{x=1}^{\alpha+\beta} f(x) \geq \sum_{x=1}^{\alpha} f(x) + \sum_{x=1}^{\beta} g(x) \qquad \Longleftrightarrow \qquad \sum_{x=\alpha+1}^{\alpha+\beta} f(x) \geq \sum_{x=1}^{\beta} g(x). \tag{C.1}$$

1. If for any $x \in [\alpha + \beta]$ we have $f(x) \geq g(x)$, we get that,

$$\sum_{x=\alpha+1}^{\alpha+\beta} f(x) \geq \sum_{x=1}^{\beta} f(x) \geq \sum_{x=1}^{\beta} g(x),$$

   and we are done.

2. Else, f and g have intersection point $z = \frac{d-b}{a-c}$. In the case that for any $x \in [\lfloor z \rfloor]$ we have $f(x) \geq g(x)$ and for any $x \in [\lceil z \rceil, \alpha + \beta]$ we have $f(x) \leq g(x)$.

2.1. If $\beta \leq z$,

$$\sum_{x=\alpha+1}^{\alpha+\beta} f(x) \geq \sum_{x=1}^{\beta} f(x) \geq \sum_{x=1}^{\beta} g(x),$$

and we are done.

2.2. If $\beta \geq z$,

$$\sum_{x=\beta+1}^{\alpha+\beta} g(x) \geq \sum_{x=\beta+1}^{\alpha+\beta} f(x) \geq \sum_{x=1}^{\alpha} f(x)$$

since $f(x)$ is non-decreasing. Hence,

$$\sum_{x=1}^{\alpha+\beta} f(x) \geq \sum_{x=1}^{\alpha+\beta} g(x) = \sum_{x=\beta+1}^{\alpha+\beta} g(x) + \sum_{x=1}^{\beta} g(x) \geq \sum_{x=1}^{\alpha} f(x) + \sum_{x=1}^{\beta} g(x),$$

and we are done.

2.3. Otherwise for $x \in [\lfloor z \rfloor]$ we have $f(x) \leq g(x)$ and for $x \in [\lceil z \rceil, \alpha+\beta]$ we have $f(x) \geq g(x)$.

2.3.1. If $\alpha \leq z$,

$$\sum_{x=\beta+1}^{\alpha+\beta} g(x) \geq \sum_{x=1}^{\alpha} g(x) \geq \sum_{x=1}^{\alpha} f(x)$$
$$\Rightarrow \sum_{x=1}^{\alpha+\beta} f(x) \geq \sum_{x=1}^{\alpha+\beta} g(x) = \sum_{x=\beta+1}^{\alpha+\beta} g(x) + \sum_{x=1}^{\beta} g(x) \geq \sum_{x=1}^{\alpha} f(x) + \sum_{x=1}^{\beta} g(x).$$

2.3.2. If $\alpha \geq z$,

$$\sum_{x=\alpha+1}^{\alpha+\beta} f(x) \geq \sum_{x=\alpha+1}^{\alpha+\beta} g(x) \geq \sum_{x=1}^{\beta} g(x).$$

□

Lemma 43. *In the Rested MAB problem with arm set $\mathcal{K}$ and horizon T, a trajectory of a policy can be represented as vector $v \in \{\mathbb{N} \cup \{0\}\}^K$ with $\|v\|_1 = T$.*

Proof. The return of any trajectory dependent only on the number of time the policy sampled each arm. Fix a trajectory $a = \{a_1, a_2, \ldots, a_T\}$, lets define the vector $v \in \mathbb{R}^K$ as follow: for any $i \in \mathcal{K}$, the i'th entry $v_i = \sum_{t=1}^{T} \mathbb{1}(a_t = i) = N_i(T)$, and we get that

$$\sum_{t=1}^{T} \mu_t(N_t(a_t)) = \sum_{i \in \mathcal{K}} \sum_{t=1}^{v_i} \mu_i(t).$$

□

Note: thanks to Lemma 43 From now on when we talk about a vector in relation to the policy, we will mean the vector representing the policy run and vice versa.

Theorem 44. *The optimal policy for Rising Rested MAB with Linear Drift is playing always arm i^*, i.e., for any $v \in V$:*

$$\sum_{i\in\mathcal{K}}\sum_{t=1}^{d_i}\mu_i(t) \leq \max_{i\in\mathcal{K}}\sum_{t=1}^{T}\mu_i(t).$$

Proof. define $i^* = \underset{i\in\mathcal{K}}{\operatorname{argmax}}\sum_{t=1}^{T}\mu_i(t)$ as the best arm.

For any vector $v \in V \setminus \{e_i \cdot T\}_{i\in\mathcal{K}}$, from Lemma 42 we know that there exist $j, l \in \mathcal{K}$ when $l \neq j$ and $v_j, v_l \geq 1$ such:

$$\begin{aligned}
&\sum_{n=1}^{v_j+v_l}\mu_l(n) \geq \sum_{n=1}^{v_j}\mu_j(n) + \sum_{n=1}^{v_l}\mu_l(n)\\
\Rightarrow &\sum_{i\in\mathcal{K}}\sum_{n=1}^{v_i}\mu_i(n) \leq \sum_{i\in\mathcal{K}\setminus\{j\}}\sum_{n=1}^{v_i}\mu_i(n) + \sum_{n=v_l+1}^{v_l+v_j}\mu_l(n).
\end{aligned}$$

Hence, for any vector $v \in V \setminus \{e_i \cdot T\}_{i\in\mathcal{K}}$, if we apply Lemma 42 at most K times we will get that there exists $j \in \mathcal{K}$ s.t.:

$$\sum_{i\in\mathcal{K}}\sum_{n=1}^{v_i}\mu_i(n) \leq \sum_{n=1}^{T}\mu_j(n).$$

Since $i^* = \underset{i\in\mathcal{K}}{\operatorname{argmax}}\sum_{n=1}^{T}\mu_i(n)$ we get that,

$$\sum_{n=1}^{T}\mu_{i^*}(n) \geq \max_{v\in V}\left\{\sum_{i\in\mathcal{K}}\sum_{n=1}^{v_i}\mu_i(n)\right\}.$$

□

הפקולטה למדעים
מדויקים ע"ש ריימונד
ובברלי סאקלר
אוניברסיטת תל אביב

חסמי חרטה במודל בנדיט בו רווחי הזרועות עולים באופן לינארי לאחר בחירתן

חיבור זה הוגש כחלק מהדרישות לקבלת התואר
"מוסמך אוניברסיטה" באוניברסיטת תל אביב
על ידי

עומר עמיחי

העבודה הוכנה בהדרכתו של

פרופ׳ ישי מנצור

ספטמבר 2024

תקציר:

אנחנו בוחנים את מודל בנדיט רב זרועות לא נייח, כאשר תוחלות הרווחים של הפעולות נעים לפי פונקצית קו ישר של מספר הפעמים שדגמנו את אותה פעולה. התוצאה המרכזית שלנו היא חסם חרטה הדוק של $\widetilde{\Theta}\left(T^{4/5}K^{3/5}\right)$, בעזרת הוכחתם של חסם עליון וחסם תחתון. אנחנו מרחיבים את התוצאות שלנו לחסמי חרטה לפי התלות במופעים אשר תלויים בפרמטרים של הפעולות הליניאריות, כאשר הם לא ידועים.