Bridging Rested and Restless Bandits with Graph-Triggering: Rising and Rotting

Gianmarco Genalti GIANMARCO.GENALTI@POLIMI.IT
Marco Mussi MARCO.MUSSI@POLIMI.IT
Nicola Gatti NICOLA.GATTI@POLIMI.IT
Marcello Restelli MARCELLO.RESTELLI@POLIMI.IT
Matteo Castiglioni MATTEO.CASTIGLIONI@POLIMI.IT
Alberto Maria Metelli ALBERTOMARIA.METELLI@POLIMI.IT
Politecnico di Milano
Piazza Leonardo da Vinci 32, Milan, 20133, Italy

Abstract

Rested and Restless Bandits are two well-known bandit settings that are useful to model real-world sequential decision-making problems in which the expected reward of an arm evolves over time due to the actions we perform or due to the nature. In this work, we propose Graph-Triggered Bandits (GTBs), a unifying framework to generalize and extend rested and restless bandits. In this setting, the evolution of the arms' expected rewards is governed by a graph defined over the arms. An edge connecting a pair of arms (i, j) represents the fact that a pull of arm i triggers the evolution of arm j, and vice versa. Interestingly, rested and restless bandits are both special cases of our model for some suitable (degenerated) graph. As relevant case studies for this setting, we focus on two specific types of monotonic bandits: rising, where the expected reward of an arm grows as the number of triggers increases, and rotting, where the opposite behavior occurs. For these cases, we study the optimal policies. We provide suitable algorithms for all scenarios and discuss their theoretical guarantees, highlighting the complexity of the learning problem concerning instance-dependent terms that encode specific properties of the underlying graph structure.[1]

Keywords: Multi-Armed Bandits, Rising, Rotting, Rested, Restless

1 Introduction

In the basic stochastic Multi-Armed Bandit (MAB, Lattimore and Szepesvári, 2020) problem, at each round, the learner is asked to choose an action (a.k.a. arm) among a finite action set and, then, it observes a reward drawn from an unknown probability distribution. The simplicity of the MAB framework is both a strength and a limitation. On the one hand, the simple nature of the framework allows for the development of elegant and efficient algorithms that can be exactly characterized and studied from an information-theoretic perspective. On the other hand, the basic MAB model assumes a relatively simplistic environment that may not capture the complexities of real-world situations. As a result, traditional MAB approaches might not be sufficient for more intricate decision-making problems where additional factors come into play. To address these limitations, researchers extended the MAB framework by incorporating additional structures and complexities in order to be

1. A conference version of this work (Genalti et al., 2024), studying Rising GTBs only, appeared at the *International Conference on Machine Learning*.

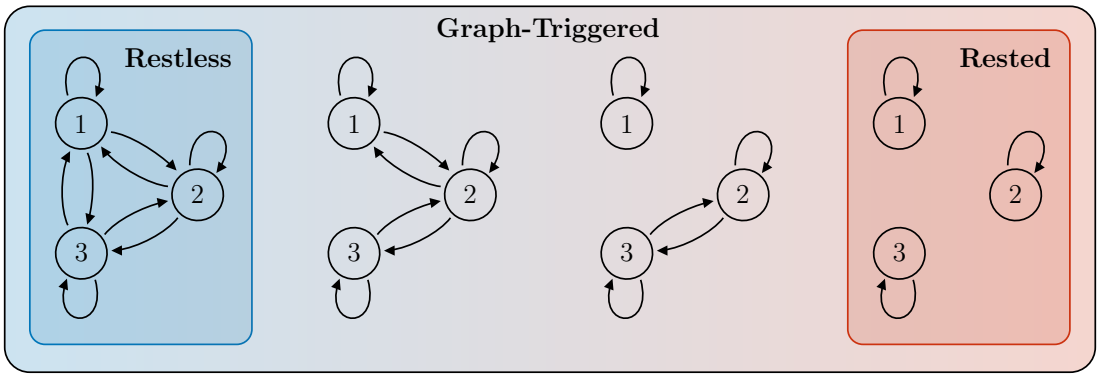


Figure 1: Examples of 3-armed GTBs.

able to handle realistic scenarios. Examples of that are *linear* (Abbasi-Yadkori et al., 2011), *continuous-action spaces* (Kleinberg et al., 2008), and *kernelized* bandits (Chowdhury and Gopalan, 2017), which presents structure over the arms, *non-stationary* bandits (Gur et al., 2014), which allow us to consider evolving environments, *delayed* reward bandits (Pike-Burke et al., 2018), allowing us to consider delayed feedback. Over the different structures available in the literature, we focus on these two specific types of MAB structures, called *restless* and *rested* bandits (Tekin and Liu, 2012). In the former, the expected rewards evolve following the time (i.e., as an effect of *nature*); in the latter, the expected reward of an arm evolves as a function of the pulls we perform on that specific arm.

In this paper, we propose a unified framework to generalize restless and rested bandits. In particular, we define a novel space of MABs called Graph-Triggered Bandits (GTBs). A GTB is represented by a bandit complemented with a *graph* describing the interactions between arms. Specifically, an arm *triggers* the evolution of its own expected reward (as for rested bandit) and the evolution of the "connected" arms. Figure 1 shows an example of this scenario, where the nodes represent arms, and the edges represent interactions. Interestingly, rested and restless bandits are two vertices in the space of GTBs. In particular, restless bandits correspond to the case of a *fully-connected* graph, while rested ones correspond to the graph with the *self-loops only*.

This framework is driven by both *theoretical* and *practical* considerations. Theoretically, it offers a unified approach that generalizes *both rested and restless* bandits. Specifically, our goal is to establish a framework in which the well-known rested and restless bandits emerge as special cases, represented by appropriate (degenerated) graphs (see Figure 1). Practically, restless and rested bandits can model a wide range of real-world situations. For example, consider the scenario where we must choose which product to advertise (represented by our arms), with the reward being the number of sales for that product. On the one hand, with rested bandits, we can handle cases in which the products are all independent. On the other hand, with restless bandits, we can handle scenarios in which all the products interact. However, all the intermediate scenarios, e.g., where advertising a product boosts its sales and also enhances the sales of the subset of complementary products, cannot be handled using restless/rested solutions. Indeed, this scenario is a rested problem with elements exhibiting restless behavior, and our generalization allows us to address such situations.

Contributions. In this paper, we present Graph-Triggered Bandits (GTBs), a setting aiming to generalize and extend rested and restless bandits settings by introducing a graph structure to represent the interaction between the arms. We focus on the cases of rising and rotting bandits, as they represent interesting case studies allowing us to obtain no-regret algorithms. More in detail, the contributions are as follows.

- In Section 2, after having introduced the fundamental notions on the rested and restless bandits, we introduce the novel framework of GTBs and discuss the relevant quantities characterizing an instance, including a representation of the graph based on the connectivity matrix. Then, we present the learning problem and the performance index to evaluate algorithms in this setting.
- In Section 3, we study the *Rising GTBs* scenario. We discuss the optimal policy in this setting, by first providing a negative result, showing that computing the optimal policy is NP-hard for an arbitrary graph (Theorem 1). Then, we characterize the optimal policy for *block-diagonal* connectivity matrices, which can be computed in polynomial time (Theorem 2). Subsequently, we discuss the deterministic scenario, and we propose two algorithms, the first, `DR-BD-UB`, for block-diagonal connectivity matrices and the second, `DR-G-UB`, for general graphs. We analyze their regret guarantees, highlighting the dependence on the graph structure (Theorems 3 and 4). Finally, we analyze the `R-□-UCB` algorithm (Metelli et al., 2022), designed for rested and restless stochastic rising bandits that does not require the knowledge of the graph. We characterize its regret guarantees, focusing on the dependence on the characteristics of the underlying graph (Theorems 6 and 7).
- In Section 4, we study the *Rotting GTBs* scenario. As for the Rising GTBs case, we prove that computing the optimal policy is NP-hard for arbitrary graphs (Theorem 8). Then, we characterize the optimal policy for *block-diagonal* connectivity matrices, which admits a convenient closed-form solution (Theorem 9). Then, we focus on the special case of block-diagonal connectivity matrices, and we study how the `RAW-UCB` algorithm (Seznec et al., 2020) obtains strong regret guarantees with no knowledge of the graph (Theorem 10). Finally, we present a non-learnability result for *all* the Rotting GTBs problem under general matrices (Theorem 11).

The relevant literature is discussed in Section A. The proofs of all the statements are provided in Appendices B and C for the Rising and Rotting GTBs, respectively.[2]

2 Graph-Triggered Bandits

In this section, we present the framework of Graph-Triggered Bandits (GTBs). We start in Section 2.1 by introducing the basic background notions on stochastic rested and restless bandits. Then, in Section 2.2, we formalize the GTBs setting. Finally, in Section 2.3, we formalize the learning problem for the GTBs setting.

2. For Rising GTBs, we report a short version of the proofs. The extended version is provided in (Genalti et al., 2024).

2.1 Notions on Rested and Restless Bandits

Let $T \in \mathbb{N}$ be the learning horizon. We define an instance $\boldsymbol{\nu} = (\nu_i)_{i \in [k]}$ of a k-armed bandit as a vector of probability distributions with support defined over $\mathbb{R}$, where $k \in \mathbb{N}$.[3] The agent interacts with the environment as follows. At every round $t \in [T]$, the agent is asked to select an action I_t among the k available ones and it observes a reward $X_{I_t,t} \sim \nu_{I_t}$. We define $N_{i,t} \coloneqq \sum_{\tau \in [t]} \mathbb{1}\{I_\tau = i\}$ as the number of pulls of the arm $i \in [k]$ until round t. We consider two specific types of MAB, namely *restless* and *rested* bandits (Tekin and Liu, 2012). In both cases, to each arm $i \in [k]$ corresponds a sequence of probability distributions $\boldsymbol{\nu} = (\nu_{i,n})_{i \in [k],\, n \in [T]}$, where the expected reward $\mu_i(n) = \mathbb{E}_{X \sim \nu_{i,n}}[X]$ evolves following an history-dependent quantity $n \in \mathbb{N}$. In the rested scenario, the expected reward of a generic arm i evolves according to the number of pulls of such an arm, i.e., $n \leftarrow N_{i,t}$. Conversely, in the restless case, the expected reward of a generic arm i evolves according to the current time t, i.e., $n \leftarrow t$. This means that, in rested bandits, the reward distribution of an arm evolves only when it is pulled, while in restless bandits, it evolves at each round, no matter the action performed. As customary in this field, we consider expected rewards $\mu_i(n)$ bounded in $[0, 1]$, for every $i \in [k]$ and $n \in [T]$. Finally, we assume distributions to be *subgaussian*[4] for every arm i and $n \in \mathbb{N}$, with their subgaussianity constants upper bounded by σ^2.

2.2 Setting

In rested and restless bandits there exists no structure among different arms. We now present a generalization of rested and restless bandits obtained by adding a structure allowing arms to interact. We consider arms as connected through an undirected graph, that can be either *known* or *unknown* to the agent.[5] If we pull an arm $i \in [k]$, we get its reward, and we *trigger* an evolution of the expected reward of the arm i and of all the arms connected to i. We do not get nor observe rewards from the connected arms (i.e., bandit feedback). Such a graph can be represented by a symmetric Connectivity Matrix (CM) $\mathbf{G} \in \{0, 1\}^{k \times k}$. If the matrix contains the value 1 in row i and column j, this implies that the pull of arm i determines the evolution of the expected reward of arm j. If the matrix contains a 0 in position (i, j), this implies that a pull of arm i does not cause an evolution of the expected reward of arm j. The pull of an arm i always implies the evolution of its own expected reward, formally $\mathbf{G}_{i,i} = 1,\ \forall i \in [k]$. For every round $t \in [T]$ and arm $i \in [k]$, we define the number $\widetilde{N}_{i,t}$ of *triggers* that it has undergone as follows:

$$\widetilde{N}_{i,t} = \sum_{\tau \in [t]} \mathbb{1}\{\mathbf{G}_{I_\tau, i} = 1\} = \mathbf{e}_i^\top \mathbf{G}^\top \mathbf{N}_t, \tag{1}$$

where $\mathbf{e}_i$ is a vector belonging to the canonical basis of $\mathbb{R}^k$ whose all components are all zero except for the i-th and $\mathbf{N}_t \coloneqq (N_{1,t}, \dots, N_{k,t})^\top$ is the vector containing the number of pulls of each arm up to round t. In GTBs, rewards are sampled from probability distributions whose average rewards vary with the number of triggers, i.e., $n \leftarrow \widetilde{N}_{i,t}$ and,

3. Given $k \in \mathbb{N}$, we define $[k] \coloneqq \{1, 2, \dots, k\}$.

4. A (zero-mean) random variable X is σ^2-subgaussian if it holds $\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{\sigma^2 \lambda^2}{2}\right)$ for every $\lambda \in \mathbb{R}$.

5. All the results we present also hold for directed graphs.

consequently, the expected reward of an arm i evolves as $\mu_i(\widetilde{N}_{i,t})$. Furthermore, we define $t_{i,n} \coloneqq \sum_{l\in[T]} \mathbb{1}\{N_{i,l} \leq n\}$ as the round in which arm i has been pulled for the n-th time. With $\boldsymbol{t}_{i,t} \coloneqq (t_{i,n})_{n\leq N_{i,t}}$ we refer to the vector containing all the rounds in which the arm i has been pulled, up to time t. Moreover, we introduce $t^I_{i,n} \coloneqq \widetilde{N}_{i,t_{i,n}}$, namely the *internal time* of the n-th pull of arm i, which is the number of triggers of arm i at the time of the n-th pull. We also define, given the connectivity matrix of a graph $\mathbf{G}$, the notion of $\bar{k}_1 \coloneqq |\{i \in [k] : \deg(i) = 1\}|$ as the number of arms having degree of 1, where $\deg(i) \coloneqq \mathbf{1}_k^\top \mathbf{G}\mathbf{e}_i$ is the degree of a node, i.e., the number of edges incident to the node. We now observe the relationship between rested and restless bandits and our setting.

Remark 1 (Inclusion of Rested and Restless bandits in GTBs) *GTBs include both* rested *and* restless *bandits (Tekin and Liu, 2012). These two settings can be recovered by considering* $\mathbf{G} = \mathbf{I}_k$ *and* $\mathbf{G} = \mathbf{1}_{k\times k}$ *for rested and restless settings, respectively.*[6] *Indeed, a* restless *bandit can be seen as a particular instance of GTB where all arms are triggered at each round, making them change every round independently from which action has been chosen (*$\widetilde{N}_{i,t} = t$*, for every* $i \in [k]$*). Instead, in a* rested *bandit an arm changes its expected reward only when is directly chosen (*$\widetilde{N}_{i,t} = N_{i,t}$*, for every* $i \in [k]$*).*[7]

Block-Diagonal Connectivity Matrix. We now discuss a particular case of GTBs that is interesting from both the practical and analytical point of view. Until now, we considered $\mathbf{G} \in \{0,1\}^{k\times k}$ to be a general binary symmetric matrix. However, we now focus on the specific case in which $\mathbf{G}$ is a *block-diagonal* connectivity matrix, i.e., a matrix in which the main-diagonal blocks are square matrices of all ones, and all off-diagonal blocks are zero matrices. Formally, let $\mathbb{B}_{\widetilde{k}} \subset \{0,1\}^{k\times k}$ be the set of block-diagonal connectivity matrices with exactly $\widetilde{k} \in [k]$ distinct blocks of 1s. We call the *GTBs with block-diagonal connectivity matrix* the set of instances where it holds that $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$, for some $\widetilde{k} \leq k$. We identify with $\mathcal{C}_{\mathbf{G}} = \{C_{m,\mathbf{G}}\}_{m\in[\widetilde{k}]}$ the partition of $[k]$ corresponding to the diagonal blocks of $\mathbf{G}$. In graph theory, a block-diagonal connectivity matrix $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$ corresponds to a cluster graph, i.e., a graph formed from the disjoint union of complete graphs or *cliques* (Shamir et al., 2004). We call $\mathcal{C}_{\mathbf{G}}$ the set of cliques and we indicate with $\widetilde{N}_{C_m,t} \coloneqq \sum_{i\in C_m} N_{i,t}$ the number of times an arm belonging to clique $C_m \in \mathcal{C}_{\mathbf{G}}$ has been pulled, namely the number of triggers of the clique C_m.

2.3 Learning Problem

We define $\mathcal{H}_t = \{(I_l, X_{I_l,l})\}_{l\in[t]}$ as the *history of interactions* at a given round $t \in [T]$. We define a policy $\pi(t)$ as a function $\pi(t) : \mathcal{H}_{t-1} \mapsto I_t$ returning the next action given the history up to that round. For a given instance $\boldsymbol{\nu}$ of a GTB, the performance of a policy π is measured by the means of *expected cumulative reward* throughout T rounds, formally:

$$J_{\boldsymbol{\nu},\mathbf{G},T}(\pi) \coloneqq \mathbb{E}\left[\sum_{t\in[T]} \mu_{I_t}(\widetilde{N}_{I_t,t})\right],$$

6. We denote $\mathbf{I}_k$ the identity matrix of dimension k and $\mathbf{1}_{k\times k}$ the square matrix of dimension k whose entries are all equal to 1.
7. This can be easily seen by looking at Equation (1) considering $\mathbf{G} = \mathbf{I}_k$ and observing that the vector $\mathbf{e}_i$ selects the i-th element of vector $\mathbf{N}_t$.

where the expectation is taken over the randomness of both the environment and the policy/algorithm. A policy is *optimal* for instance $\boldsymbol{\nu}$, a connectivity matrix $\mathbf{G}$, and time horizon T if it maximizes the expected cumulative reward, formally:

$$\pi^*_{\boldsymbol{\nu},\mathbf{G},T} \in \arg\max_{\pi} J_{\boldsymbol{\nu},\mathbf{G},T}(\pi).$$

We denote by $J^*_{\boldsymbol{\nu},\mathbf{G},T} = J_{\boldsymbol{\nu},\mathbf{G},T}(\pi^*_{\boldsymbol{\nu},\mathbf{G},T})$ the expected cumulative reward attained by the optimal policy. We can now define the *expected policy regret* as:

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\pi) = J^*_{\boldsymbol{\nu},\mathbf{G},T} - J_{\boldsymbol{\nu},\mathbf{G},T}(\pi).$$

Therefore, our learning problem is to find a policy π minimizing the expected policy regret $R_{\boldsymbol{\nu},\mathbf{G},T}(\pi)$. Since the optimal policy depends simultaneously on $\boldsymbol{\nu}$, $\mathbf{G}$, and T, from now on, we consider an instance of the GTB problem the triple $(\boldsymbol{\nu}, \mathbf{G}, T)$, instead of the reward distributions $\boldsymbol{\nu}$ only.

Remark 2 (On the Chosen Notion of Regret) *In GTBs, we consider a notion of* policy *regret (Dekel et al., 2012). Indeed, in this setting, diverging from the optimal sequence of actions influences not only instantaneous regret but also leads to a sub-optimal history, implying future regret even when returning to an optimal policy from there on. This notion of regret, which shares similarities with the one of reinforcement learning, is more challenging to optimize.*

3 Rising Graph-Triggered Bandits

Among the various types of restless and rested bandits available in the literature, in this section, we focus on *Rising Bandits* (Heidari et al., 2016; Metelli et al., 2022). We first introduce the assumption of the rising setting and some useful quantities. Then, we discuss the optimality in this setting (Section 3.1). Subsequently, we discuss the regret minimization problem for both the deterministic (Section 3.2) and stochastic (Section 3.3) scenarios.

Rising bandits are a specific class of MABs in which the expected reward of each arm evolves in a non-decreasing and concave manner. The following assumption formalizes such behavior.

Assumption 1 (Non-decreasing and Concave Payoffs) *Let $\boldsymbol{\nu}$ be an instance of a rising bandit, then, defining $\gamma_i(n) \coloneqq \mu_i(n+1) - \mu_i(n)$ for every $i \in [k]$ and $n \in [T]$, it holds:*

$$\begin{aligned} &\textit{Non-decreasing:} && \gamma_i(n) \geq 0, \\ &\textit{Concave:} && \gamma_i(n-1) \geq \gamma_i(n). \end{aligned}$$

The two parts of this assumption allow us to provide theoretical guarantees in both the restless and rested settings. Such guarantees cannot be provided without the concavity assumption (see Theorem 4.2 of Metelli et al., 2022). We call *Rising GTBs*, the instances of GTBs in which the expected rewards fulfill Assumption 1.

Instance Characterization. Assumption 1 ensures sufficient structure on the problem to allow for algorithms with provably strong theoretical guarantees. In this scenario, given an instance $\boldsymbol{\nu}$, we define the *total increment* as:

$$\Upsilon_{\boldsymbol{\nu}}(M, q) \coloneqq \sum_{t\in[M-1]} \max_{i\in[k]} \gamma_i(t)^q,$$

where $M \in \mathbb{N}$ and $q \in [0, 1]$. This quantity figures in the (instance-dependent) analysis of algorithms and characterizes the difficulty of learning in instance $\boldsymbol{\nu}$.

3.1 Optimality in Rising GTBs

In this part, we discuss the notion of *optimality* for our learning problem. We first characterize the complexity of finding the optimal policy followed by the clairvoyant when both the expected values and the matrix $\mathbf{G}$ are *known*.

Theorem 1 (Complexity of finding the Optimal Policy in Rising GTBs) *Computing the optimal policy in Rising GTBs with general matrices* $\mathbf{G}$ *is NP-Hard.*

This theorem follows from a reduction to the NP-Hard problem of determining if a large clique in a given graph exists (Karp, 1972). Intuitively, given a graph (V, E), we build an instance in which the cumulative reward is maximum only if the learner plays a sequence of arms that are associated with vertexes in a clique. Theorem 1 implies that the class of problems of Rising GTBs is computationally harder than all restless bandits and rested rising bandits, for which the optimal policy can be computed in polynomial time (Heidari et al., 2016). Moreover, the optimal policy does not admit a simple closed-form representation. Thus, in general, the optimal policy cannot be reduced to a greedy one or to a fixed-arm policy. The result highlights how this definition of optimal policy is closer to the one of MDPs rather than the one of standard bandit settings.

We now show how, for the special case of Rising GTBs with block-diagonal connectivity matrices, the optimal policy can be efficiently computed and admits a closed-form solution.

Theorem 2 (Optimal Policy in Rising GTBs with Block-Diagonal CM) *For any instance* $(\boldsymbol{\nu}, \mathbf{G}, T)$ *of Rising GTBs with* $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$*, the optimal policy* $\pi^*_{\boldsymbol{\nu},\mathbf{G},T} \in \arg\max_\pi J_{\boldsymbol{\nu},\mathbf{G},T}(\pi)$ *is given by:*

$$\pi^*_{\boldsymbol{\nu},\mathbf{G},T}(t) \in \arg\max_{j\in C^*_{\boldsymbol{\nu},\mathbf{G},T}} \mu_j(t), \qquad \forall t \in [T],$$

where $C^*_{\boldsymbol{\nu},\mathbf{G},T}$ *is the "best" cumulative reward clique:*

$$C^*_{\boldsymbol{\nu},\mathbf{G},T} \in \arg\max_{C\in\mathcal{C}_{\mathbf{G}}} \sum_{t\in[T]} \max_{j\in C} \mu_j(t).$$

This result characterizes the optimal policy when the graph linking the actions is only composed only by cliques. In particular, the clairvoyant would play a greedy policy but always inside the same predefined subset of arms composing a clique. Naturally, the chosen clique would be the one having the maximum cumulative reward at the end of the trial. We point out how this policy "combines" the optimal policies from both rising rested bandits (corresponding to always playing the arm with the highest *cumulative* reward), and the optimal policy from rising restless bandits (the *greedy* policy, corresponding to always playing the arm with the highest *instantaneous* reward).

Algorithm 1: DR-BD-UB.

Input : Connectivity matrix $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$

1 **for** $t \in [T]$ **do**
2 Compute $\bar{\mu}_i(t)$ as in Equation (2) , $\forall i \in [k]$
3 Select $I_t \in \arg\max_{i \in [k]} \bar{\mu}_i(t)$
4 Play I_t and observe $\mu_{I_t}(\widetilde{N}_{I_t,t})$
5 **end**

3.2 Deterministic Rising GTBs

In this part, we propose two novel algorithms to learn in *deterministic* Rising GTBs, i.e., all instances of Rising GTBs where $\sigma = 0$. More in detail, in Section 3.2.1, we discuss the block-diagonal CM case, while in Section 3.2.2, we discuss the general scenario. The deterministic scenario allows for a better understanding of the complex structure of this setting since it *ignores* the statistical learning problem.

We start by introducing a novel biased estimator which, for every arm $i \in [k]$, propagates its reward function to the current time t by estimating the first derivative using the last two observations:

$$\bar{\mu}_i(t) \coloneqq \mu(t^I_{i,N_{i,t-1}}) + (t - t^I_{i,N_{i,t-1}}) \frac{\mu(t^I_{i,N_{i,t-1}}) - \mu(t^I_{i,N_{i,t-1}-1})}{t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-1}}. \tag{2}$$

This estimator relies on the concept of *internal time*. Internal times are particularly useful since they can separate the bias in two components:

$$t - t^I_{i,N_{i,t-1}} = \underbrace{(t - t^I_{i,N_{i,t}})}_{\text{(A)}} + \underbrace{(t^I_{i,N_{i,t}} - t^I_{i,N_{i,t-1}})}_{\text{(B)}}.$$

As we will see in Section 3.2.1, this decomposition assumes a particular meaning in instances where $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$, where (A) represents the rested component of the bias, since $t^I_{i,N_{i,t}} = \widetilde{N}_{C_m, t_{i,N_{i,t}}}$ making it equivalent to the bias of a rested bandit where cliques are the arms; and (B) represents the restless component of the bias, since from arm i perspective $t^I_{i,N_{i,t}} = \widetilde{N}_{i,t}$ can be interpreted as the current time inside the clique.

3.2.1 Algorithm for Deterministic Rising GTBs with Block-Diagonal CMs

We now introduce `Deterministic Rising Block-Diagonal Upper Bound` (`DR-BD-UB`), an optimistic anytime regret minimization algorithm for deterministic Rising GTBs with block-diagonal connectivity matrix, whose pseudocode is provided in Algorithm 1. The algorithm takes as input the connectivity matrix $\mathbf{G}$ and employs the estimator presented in Equation (2). Then, after having initialized the counters of the number of pulls, it starts the interaction with the environment. At each round $t \in [T]$, it estimates (line 2) the $\bar{\mu}_i(t)$ for every $i \in [k]$ as in Equation (2) and plays greedy according to it (line 4).[8]

The following result provides the regret bound of `DR-BD-UB`, highlighting the impact of the graph topology.

8. At the beginning, the algorithm is required to play every arm 2 times in a round-robin fashion in order to be able to compute $\bar{\mu}_i(t)$.

Theorem 3 (DR-BD-UB Regret in Det. Rising GTBs with Block-Diagonal CMs) *Let $(\boldsymbol{\nu}, \mathbf{G}, T)$ be an instance of Rising GTB, where $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$ and $\sigma = 0$. Then,* DR-BD-UB *suffers a regret bounded by:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{DR-BD-UB}) \leq \widetilde{\mathcal{O}}\left(\inf_{q\in[0,1]} \left\{ \underbrace{T^q \sum_{C_m\in\mathcal{C}} |C_m| \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C_m,T}}{|C_m|} \right\rceil, q \right)}_{\text{(A) Rested Bias Contribution}} + \right.\right.$$
$$\left.\left. + \underbrace{\sum_{C_m\in\mathcal{C}} |C_m| \widetilde{N}_{C_m,T}^{\frac{q}{1+q}} \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C_m,T}}{|C_m|} \right\rceil, q \right)^{\frac{1}{1+q}}}_{\text{(B) Restless Bias Contribution}} \right\} \right).$$

In this theorem, we report the result as a function of the number of triggers $\widetilde{N}_{C_m,T}$ of the cliques in order to better discuss the properties of the graph. However, this dependence can be removed by simply observing $\widetilde{N}_{C_m,T} \leq T$. This choice allows us to have an interesting discussion on the nature of this result w.r.t. the graph structure. First of all, we observe that we can separate two contributions to the regret: one coming from the rested behavior (part (A) of the bound) determined by the need for identifying the best clique, and the other from the restless behavior needed for identifying the best arm inside the clique (part (B) of the bound). If we compare this result to the bounds in Theorems 4.4 and 5.2 of (Metelli et al., 2022), we can notice how the shapes of the two contributions correspond. We also remark that, in the two corner cases, i.e., rested and restless bandits, the regret bound is actually smaller and corresponds exactly to the bounds presented in (Metelli et al., 2022), even though this is not immediately visible in Theorem 3 because of a mathematical artifact of the proof.[9] In the bound, the graph topology emerges by means of cliques' sizes, that act as multiplicative constants. The major consequence is that having fewer cliques leads, in general, to a better bound. As intuition suggests, the rested scenario can lead to a worst-case bound in the first component (which is, by the way, the one having the greater order in T), and this can be seen by a simple application of Jensen's Inequality, and by noticing that $\Upsilon_{\boldsymbol{\nu}}$ is a concave function:

$$\sum_{C_m\in\mathcal{C}} |C_m| \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C_m,T}}{|C_m|} \right\rceil, q \right) \leq k \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{T}{k} \right\rceil, q \right).$$

We remark that in the two corner cases, one of the two contributions vanishes, even though it cannot be directly seen in Theorem 3. However, since the restless regret has a better order than the rested one, graphs with fewer cliques may lead, in general, to better bounds. Unfortunately, to precisely quantify this property, one would need to know the exact shape of $\Upsilon_{\boldsymbol{\nu}}$ and to solve a difficult optimization problem.

3.2.2 Algorithm for Deterministic Rising GTBs with General Matrices

After having studied the scenario of block-diagonal connectivity matrices, we now consider the case in which $\mathbf{G}$ can be arbitrary. Before introducing the algorithm, we need to define the concept of *block sub-matrix*.

9. More details can be found in Remark 5 (Appendix B).

Algorithm 2: `DR-G-UB`.

Input : Connectivity matrix $\mathbf{G}$
Compute maximal sub-matrix $\bar{\mathbf{G}}^L$ from $\mathbf{G}$
for $t \in [T]$ **do**
  Compute $\bar{\mu}_i^L(t)$ as in Equation (4), $\forall i \in [k]$
  Select $I_t \in \arg\max_{i\in[k]} \bar{\mu}_i^L(t)$
  Play I_t and observe $\mu_{I_t}(\widetilde{N}_{I_t,t})$
end

Definition 1 (Block Sub-matrix) *Let $\mathbf{G} \in \{0,1\}^{k\times k}$ be a general matrix, a block-diagonal matrix $\mathbf{G}^L \in \mathbb{B}_{\widetilde{k}}$ is a* sub-matrix *of $\mathbf{G}$ if it satisfies:*

$$\mathbf{G}_{i,j} - \mathbf{G}^L_{i,j} \geq 0, \quad \forall i,j \in [k]. \tag{3}$$

Moreover, we say that $\bar{\mathbf{G}}^L \in \mathbb{B}_{\widetilde{k}}$ is maximal *if it also satisfies:*

$$\bar{\mathbf{G}}^L \in \underset{\mathbf{G}^L \text{ satisfying Eq. (3)}}{\arg\min} \left|\mathcal{C}_{\mathbf{G}^L}\right|.$$

Informally, $\mathbf{G}^L \in \mathbb{B}_{\widetilde{k}}$ is a sub-matrix of $\mathbf{G}$ if its graph can be obtained by only removing 1s from $\mathbf{G}$. Finally, a maximal sub-matrix has the least number of cliques. Note that such a maximal sub-matrix is, in general, not unique.

For this algorithm, we need to introduce a novel estimator, based on sub-matrices, whose definition recalls the one of Equation (2):

$$\bar{\mu}_i^L(t) \coloneqq \mu(t^{I,L}_{i,N_{i,t-1}}) + (t - t^{I,L}_{i,N_{i,t-1}}) \frac{\mu(t^{I,L}_{i,N_{i,t-1}}) - \mu(t^{I,L}_{i,N_{i,t-1}-1})}{t^{I,L}_{i,N_{i,t-1}} - t^{I,L}_{i,N_{i,t-1}-1}}, \tag{4}$$

where $t^{I,L}_{i,l} \coloneqq \mathbf{e}_i^\top (\bar{\mathbf{G}}^L)^\top \mathbf{N}_{t_{i,l}}$ is the internal time w.r.t. a maximal sub-matrix $\bar{\mathbf{G}}^L$ of the actual matrix $\mathbf{G}$. Given this new estimator, we can generalize Algorithm 1 to attain comparable performance even for a general connectivity matrix $\mathbf{G}$. We introduce a generalization of `DR-BD-UB` called `Deterministic Rising General Upper Bound` (`DR-G-UB`), whose pseudocode is provided in Algorithm 2. The algorithm takes as input a generic matrix $\mathbf{G}$ and computes $\bar{\mathbf{G}}^L$. Then, the algorithm interacts with the environment as before and uses the estimator defined in Equation (4). In other words, `DR-G-UB` pretends to be interacting with a bandit with a graph defined by $\bar{\mathbf{G}}^L$. The following result characterizes the performances of `DR-G-UB`.

Theorem 4 (`DR-G-UB` Regret in Det. Rising GTBs with General Matrices) *Let $(\boldsymbol{\nu}, \mathbf{G}, T)$ be an instance of Rising GTB, where $\mathbf{G} \in \{0,1\}^{k\times k}$ and $\sigma = 0$. Then,* `DR-G-UB` *suffers a regret bounded by:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{DR-G-UB})$$
$$\leq \widetilde{\mathcal{O}}\left(\min_{q\in[0,1]} \left\{ T^q \sum_{C^L_m \in \mathcal{C}_{\bar{\mathbf{G}}^L}} |C^L_m| \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C^L_m,T}}{|C^L_m|} \right\rceil, q \right) + \sum_{C^L_m \in \mathcal{C}_{\bar{\mathbf{G}}^L}} |C^L_m| \widetilde{N}^{\frac{q}{1+q}}_{C^L_m,T} \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C^L_m,T}}{|C^L_m|} \right\rceil, q \right)^{\frac{1}{1+q}} \right\} \right),$$

where $\bar{\mathbf{G}}^L \in \mathbb{B}_{\widetilde{k}}$ is a maximal sub-matrix of $\mathbf{G}$.

This result provides a formal justification to the intuition that the performance of Algorithm 2 can be bounded with the upper bound attained in a less favorable scenario, i.e., a block-diagonal instance that is "closer" to the worst-case instance of a rested bandit. The regret bound of `DR-G-UB` can be found by applying Theorem 3 using the matrix $\bar{\mathbf{G}}^L$.

Remark 3 (Computational Complexity) *Note that, even if the optimal policy in this setting for a general* $\mathbf{G}$ *is NP-hard to be retrieved, with* `DR-G-UB`*, we achieve sublinear regret w.r.t. the optimal policy with a polynomial-time algorithm. This is possible thanks to the ability of* `DR-G-UB` *to identify a convenient matrix* $\bar{\mathbf{G}}^L$ *that is subsequently adopted as a proxy of the real environment in order to play in a computationally efficient manner.*

3.3 Stochastic Rising GTBs

In this part, we focus on the *stochastic* Rising GTBs scenario. We characterize the performances of `R-□-UCB` (Metelli et al., 2022), designed for rising rested and restless bandits, in the Rising GTBs setting for both the block-diagonal CMs (Section 3.3.1) and the general case (Section 3.3.2). We show that such an algorithm achieves good performances for a general $\mathbf{G}$. In particular, we develop a new proof strategy for the regret upper bound that makes graph-dependent terms explicit. We aim at obtaining a computationally efficient algorithm enjoying *sublinear regret* guarantees. Surprisingly, our analysis shows that `R-□-UCB` not only enjoys sublinear regret for any connectivity matrix $\mathbf{G}$, but also that the graph-dependent quantities actually interpolate the regret between the two corner cases. Moreover, we show that there is no need to solve any additional NP-Hard problem before or during the algorithm's executions, letting `R-□-UCB` keep it affordable computational costs, as in the two corner settings. Furthermore, in this case, the algorithm is completely *unaware* of the graph structure.

The algorithm employs a biased estimator which, for every arm i, propagates its reward function to the current round t by estimating the first derivative over the last $2h$ samples:

$$\widehat{\mu}_i^h(t) \coloneqq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(X_{i,t_{i,l}} + (t-l) \frac{X_{i,t_{i,l}} - X_{i,t_{i,l-h}}}{h} \right), \tag{5}$$

where $h \in \mathbb{N}$ is the window size. We report the estimator's concentration rate, which is a function of the window size h. The proof of this result originally appeared in (Metelli et al., 2022). However, it can be extended to Rising GTBs (more details are provided in Appendix B.1).

Lemma 5 (Concentration of Estimator, adapted from Metelli et al. 2022) *For every arm* $i \in [k]$*, every round* $t \in [T]$*, and window width* $1 \leq h \leq \left\lfloor \frac{N_{i,t-1}}{2} \right\rfloor$*, let:*

$$\beta_i^h(t, \delta) \coloneqq \sigma(t - N_{i,t-1} + h - 1) \sqrt{\frac{10 \log \frac{1}{\delta}}{h^3}}.$$

Then, if the window size depends on the number of pulls only $h_{i,t} = h(N_{i,t-1})$ *and if* $\delta_t = t^{-\alpha}$ *for some* $\alpha > 2$*, it holds for every round* $t \in [T]$ *that:*

$$\mathbb{P}\left(\left| \widehat{\mu}_i^{h_{i,t}}(t) - \widetilde{\mu}_i^{h_{i,t}}(t) \right| > \beta_i^{h_{i,t}}(t, \delta_t) \right) \leq 2t^{1-\alpha}.$$

Algorithm 3: R-□-UCB.

Input : Subgaussianity proxy σ, confidence levels $\{\delta_t\}_{t\in[T]}$, window size $\epsilon \in (0, 1/2)$.

for $t \in [T]$ **do**
  Compute $\widehat{\mu}_i^{h_{i,t}}(t)$ as in Equation (5), $\forall i \in [k]$
  Select $I_t \in \arg\max_{i\in[k]} \widehat{\mu}_i^{h_{i,t}}(t) + \beta_i^{h_{i,t}}(t, \delta_t)$
  Play I_t and observe $X_{I_t,t}$
end

Algorithm. The algorithm, whose pseudocode is reported in Algorithm 3, takes as input the subgaussianity proxy σ, the sliding window size parameter ϵ, and a sequence of properly selected confidence levels δ_t, where $t \in [T]$. R-□-UCB relies on the previously defined biased estimator and uses its confidence interval to make decisions in an optimistic manner. R-□-UCB does not require the time horizon T as an input, making it an anytime algorithm. Moreover, the algorithm exploits the sliding window mechanism to deal with the environment's uncertainty while controlling the confidence degree by means of $\{\delta_t\}_{t\in[T]}$. In particular, the window size employed by the algorithm is proportional to parameter $\epsilon \in (0, 1/2)$, in the form of $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$. As we show below, ϵ controls the bias-variance trade-off, where low values for ϵ result in less bias but higher variance, and vice versa.

Remark 4 (Computational Complexity) *At each round,* R-□-UCB *only needs to update the estimator and the related confidence bounds for every arm, which can be done in a time linear in the number of arms at every step. For an efficient update, we refer the reader to (Mussi et al., 2024, Appendix C).*

3.3.1 Regret for Stochastic Rising GTBs with Block-Diagonal CMs

We now analyze the performances of R-□-UCB in the block-diagonal CMs case.

Theorem 6 (R-□-UCB Regret in Rising GTBs with Block-Diagonal CMs) *Let $(\boldsymbol{\nu}, \mathbf{G}, T)$ be an instance of Rising GTB, where $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$. Let $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$ for $\epsilon \in (0, 1/2)$ and $\delta_t = t^{-\alpha}$ for $\alpha > 2$. Then,* R-□-UCB *suffers an expected regret bounded by:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\text{R-□-UCB}) \leq \widetilde{\mathcal{O}}\left(\min_{q\in[0,1]} \left\{ \underbrace{(\sigma T)^{\frac{2}{3}}}_{\text{(A) Variance Contribution}} + \underbrace{\bar{k}_1 T^q \Upsilon_{\boldsymbol{\nu}}\left(\left\lceil \frac{T}{\bar{k}_1} \right\rceil, q\right)}_{\text{(B) Rested Bias Contribution}} + \underbrace{T^{\frac{2q}{1+q}} \sum_{C_m \in \mathcal{C}_{\mathbf{G}} : |C_m| > 1} |C_m| \Upsilon_{\boldsymbol{\nu}}\left(\left\lceil \frac{T}{|C_m|} \right\rceil, q\right)^{\frac{1}{1+q}}}_{\text{(C) Restless Bias Contribution}} \right\} \right),$$

where $\bar{k}_1$ is the number of cliques in $\mathbf{G}$ containing only one action.

Existence of a Bias-Variance Trade-off. In the regret upper bound, we can observe three distinct contributions. First, (A) represents the variance contribution, which is the regret suffered by the algorithm due to the stochastic nature of the environment. This contribution is due to the estimator's concentration properties and sets a minimum order of regret to $\widetilde{\mathcal{O}}((\sigma T)^{2/3})$. This term is independent of the total increment $\Upsilon_{\boldsymbol{\nu}}$ but, differently from

the others, is the only contribution depending on σ. The contribution due to the estimator's bias is split into two distinct parts. The term (B) represents the rested contribution, which scales with the number of blocks containing only one arm. The term (C), instead, represents the restless contribution that scales with the number and the sizes of cliques. The bias contributions depend explicitly on the shape of average reward functions by total increment $\Upsilon_{\boldsymbol{\nu}}$. The only term common to variance and bias contributions is ϵ. Indeed, ϵ regulates such a trade-off between bias and variance, and this effect can be observed in the complete expression of the regret upper bound in Appendix B. The variance contribution depends linearly on ϵ^{-1}; thus, a smaller window size implies a higher variance in the estimate. On the contrary, the bias tends to increase with ϵ: this is expected since a larger window means including older samples in the estimate.

Dependence on Graph Topology. In the regret upper bound of Theorem 6, the only contributions depending on graph topology are the bias ones (terms (B) and (C)). Indeed, the environment's randomness contribution has been decoupled from the estimation bias to get a tractable stochastic structure. We observe how the different behaviors of arms not connected with the others (size-1 cliques, corresponding to rested arms) and arms belonging to larger cliques. The regret scales as T^q in rested arms, but the dependence on the total increment $\Upsilon_{\boldsymbol{\nu}}$ is linear. Instead, for cliques with size greater than 1, regret scales as $T^{\frac{2q}{1+q}}$, which is greater than in rested contribution, but scales with $\Upsilon_{\boldsymbol{\nu}}$ to the power of $\frac{1}{1+q}$, that is indeed a better dependence. Moreover, each clique contributes differently, based on its size. Overall, the higher the size, the higher the contribution is, since the linear term is dominant w.r.t. the inverse term inside the total increment $\Upsilon_{\boldsymbol{\nu}}$. Another interesting dependence is the one on ϵ^{-1} for the restless contribution, which can be observed in the complete form of the bound in Appendix B. For connected arms, stochasticity and graph topology produce an interaction. Indeed, if one could design an estimator with strong concentration properties for connected arms, this would simplify the analysis of the restless contribution, eliminating the bad dependence on stochasticity. With such an estimator, we conjecture we could reduce the dependence up to $T^{\frac{q}{1+q}}$, matching the deterministic setting bound.[10]

Comparison with Known Results from Literature. Given that rested and restless rising bandits are special instances of Rising GTBs, we now comment on how the presented bound links to existing results when Algorithm 3 is run over one of those instances. We start from the rested scenario, i.e., when $\mathbf{G} = \mathbf{I}_k$. Then, we would have $\bar{k}_1 = k$ and an empty summation in the restless bias contribution. The bound of Theorem 6 would thus assume the following form:

$$R_{\boldsymbol{\nu},\mathbf{I}_k,T}(\texttt{R-}\square\texttt{-UCB}) \leq \widetilde{\mathcal{O}}\left(\min_{q\in[0,1]} \left\{ (\sigma T)^{\frac{2}{3}} + kT^q \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{T}{k} \right\rceil, q \right) \right\} \right).$$

The only other existing result for the rested rising bandits setting is the one of Theorem 4.4 of (Metelli et al., 2022), which is matched up to constants by ours. In the restless scenario, i.e., when $\mathbf{G} = \mathbf{1}_{k\times k}$, we have a unique clique of size k, and $\bar{k}_1 = 0$. Thus, the bound we

10. The lower bounds for rising rested and restless bandits are still an open problem.

presented in Theorem 6 becomes:

$$R_{\boldsymbol{\nu},\mathbf{1}_{k\times k},T}(\texttt{R-}\square\texttt{-UCB}) \leq \widetilde{\mathcal{O}}\left(\min_{q\in[0,1]}\left\{(\sigma T)^{\frac{2}{3}} + kT^{\frac{2q}{1+q}}\Upsilon_{\boldsymbol{\nu}}\left(\left\lceil\frac{T}{k}\right\rceil, q\right)^{\frac{1}{1+q}}\right\}\right).$$

Once again, this result matches (up to constants) the result from Theorem 5.3 of (Metelli et al., 2022), the current state-of-the-art for the restless rising bandits problem. To conclude, we generalize the stochastic rising rested/restless bandit setting, with regret bounds that are tight w.r.t. the known results for the two corner scenarios.

3.3.2 Regret for Stochastic Rising GTBs with General Matrices

We are now ready to generalize the result of Theorem 6 to general matrices in $\mathbf{G} \in \{0,1\}^{k\times k}$. Before that, we have to first introduce the notion of *block super-matrix*.

Definition 2 (Block Super-matrix) *Let $\mathbf{G} \in \{0,1\}^{k\times k}$ be a general matrix, a block-diagonal matrix $\mathbf{G}^U \in \mathbb{B}_{\widetilde{k}}$ is a* super-matrix *of $\mathbf{G}$ if it satisfies:*

$$\mathbf{G}_{i,j} - \mathbf{G}^U_{i,j} \leq 0, \quad \forall i,j \in [k]. \tag{6}$$

Moreover, we say that $\bar{\mathbf{G}}^U \in \mathbb{B}_{\widetilde{k}}$ is minimal *if it also satisfies:*

$$\bar{\mathbf{G}}^U \in \underset{\mathbf{G}^U \text{ satisfying Eq. (6)}}{\arg\max} \left|\mathcal{C}_{\mathbf{G}^U}\right|.$$

This concept of minimal super-matrix plays an analogous role as the maximal sub-matrix in Theorem 4. We now have all the elements to present the upper bound on the regret for the stochastic Rising GTBs case and general matrices.

Theorem 7 (R-□-UCB Regret in Rising GTBs with General Matrices) *Let $(\boldsymbol{\nu}, \mathbf{G}, T)$ be an instance of Rising GTB, where $\mathbf{G} \in \{0,1\}^{k\times k}$. Let $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$ for $\epsilon \in (0, 1/2)$ and $\delta_t = t^{-\alpha}$ for $\alpha > 2$. Then, R-□-UCB suffers an expected regret bounded by:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{R-}\square\texttt{-UCB}) \leq \widetilde{\mathcal{O}}\left(\min_{q\in[0,1]}\left\{(\sigma T)^{\frac{2}{3}} + T^q \bar{k}_1 \Upsilon_\nu\left(\frac{T}{\bar{k}_1}, q\right) + T^{\frac{2q}{1+q}}\sum_{C^U_m} |C^U_m| \Upsilon_\nu\left(\frac{T}{|C^U_m|}, q\right)^{\frac{1}{1+q}}\right\}\right),$$

where $\bar{\mathbf{G}}^U$ is the minimal super-matrix of $\mathbf{G}$.

This result is obtained by bounding $\widetilde{N}_{C^U_m,T} \leq T$ for every $C^U_m \in \mathcal{C}_{\bar{\mathbf{G}}^U}$ to remove any stochastic quantity from the regret, but a more precise bound can be provided by finding the worst-case allocation of the triggers among the cliques (as discussed for the similar result in Theorem 4). However, this would require solving a challenging optimization problem that does not admit any closed-form solution. This result is similar to the one presented in Theorem 4, with the only difference being that the dependence on graph topology is linked to the minimal super-matrix. In principle, the result holds for any super-matrix of $\mathbf{G}$. Still, in the stochastic setting, the upper bound for the rested scenario is better than the one for the restless scenario. Hence, a block-diagonal CM with as many cliques as possible will, in most cases, lead to better bounds.

About the Knowledge of G. In the stochastic scenario, we avoid extracting the super-matrix structure from the graph before executing the algorithm, as it always plays the same policy, regardless of the graph structure. Indeed, Algorithm 3 *does not require the knowledge on the graph*: the algorithm plays as if the true matrix is the identity one (i.e., a rested instance). To justify this behavior in an intuitive way, have to look at Theorems 4.4 and 5.3 of (Metelli et al., 2022): in stochastic scenarios, the *rested* contribution to regret's upper bound has a better dependence on T w.r.t. the restless one. Moreover, our optimistic estimator computed by assuming a less connected graph will always be higher than the one computed from any more densely connected graph. Thus, by playing a purely rested policy, we are always sure to over-estimate the true reward (i.e., optimism holds) and we are guaranteed that the rested contribution to the regret is maximized w.r.t. the restless contribution. The final form of the regret bound is obtained by including the minimal super-matrix as a pessimistic proxy of the effect of connected arms (informally, the minimal super-matrix represents the maximum possible contribution to the regret that is due to the arms connections). We point out that Algorithm 3 does not require the minimal super-matrix as an input, as it is needed only in the analysis. For this reason, one could reformulate the following result by removing the dependence on the minimal super-matrix and including a minimization over the set of all super-matrices. As a side effect, this dramatically reduces the computational burden w.r.t. the deterministic setting at the cost of a slightly higher regret bound.

Comparison with Deterministic Regret Bounds. In deterministic scenario (Theorems 3 and 4), the restless contributions are always of smaller order compared to the rested one, which is the contrary of what we observe in stochastic settings (Theorems 6 and 7). Due to this reason, in Algorithm 2, the regret bound scales with the maximal sub-matrix instead of the minimal super-matrix. In the deterministic setting, the maximal sub-matrix represents the maximum possible contribution to the regret that is due to the *absence* of arms connections. In principle, we could remove the necessity for graph knowledge also in the deterministic setting by simply playing as in a rested scenario (i.e., run Algorithm 1 by setting $\mathbf{G} = \mathbf{I}_k$). This would be sensibly sub-optimal since any graph connection can be used to obtain a strictly better regret bound. This is not the case for the stochastic setting, where over-estimating the number of connections (e.g., by playing as in a restless scenario) may result in a non-optimistic estimator, compromising the theoretical soundness of our algorithms.

4 Rotting Graph-Triggered Bandits

Rotting bandits (Levine et al., 2017) are an important family of evolving rewards bandits where, contrary to what happens in rising bandits, the reward functions are not allowed to grow. In this section, we explore how the graph-triggering mechanism interacts with the non-increasing reward function assumption. We characterize the optimal policies and the challenges in finding them (Section 4.1). Then, we study the regret minimization problem for this setting in the presence of stochastic noise (Section 4.2).[11] Before that, we start

11. For Rotting GTBs, we skip the deterministic case, as all the interesting results we want to show are visible also in the presence of noise.

by stating the main setting assumption and presenting the quantities characterizing this specific kind of bandits.

Assumption 2 (Non-increasing Payoffs) *Let $\boldsymbol{\nu}$ be an instance of a rotting bandit, then, defining $\gamma_i(n) \coloneqq \mu_i(n+1) - \mu_i(n)$ for every $i \in [k]$ and $n \in [T]$, it holds:*

$$\textit{Non-increasing:} \qquad \gamma_i(n) \leq 0.$$

This assumption allows for strong theoretical guarantees in both the restless and rested settings, as it has been shown in the literature (see, e.g., Heidari et al., 2016; Levine et al., 2017; Seznec et al., 2019, 2020). Notably, for rotting bandits, we are not required to have a concavity/convexity assumption.

Instance Characterization. In this scenario, given an instance $\boldsymbol{\nu}$, we define the *total decrement* as:

$$V_{\boldsymbol{\nu}}(M) \coloneqq \sum_{n \in [M-1]} \max_{i \in [k]} \gamma_i(n),$$

where $M \in \mathbb{N}$. Moreover, we define the *maximum per-round variation* as:

$$L \coloneqq \max_{i \in [k]} \max_{n \in [T]} |\gamma_i(n)|,$$

with $\mu_i(-1) \coloneqq \max_{i \in [k]} \mu_i(0)$. These quantities figure in the instance-dependent guarantees of algorithms operating in this setting and characterize the difficulty of learning for the instance $\boldsymbol{\nu}$. In particular, $V_{\boldsymbol{\nu}}(T)$ is required to properly bound the regret in restless rotting bandits (see, e.g., Seznec et al. 2020), while L appears in the minimax regret bound of rested rotting bandits, as shown in the setting's lower bound of $\Omega(kL)$ by Heidari et al. (2016).

4.1 Optimality in Rotting GTBs

Under the standard literature's assumptions, we are now ready to characterize our optimal policies. We first show, as for Rising GTBs, a negative result on the complexity of finding the optimal policy for a clairvoyant who knows all about our Rotting GTBs instance.

Theorem 8 (Complexity of finding the Optimal Policy in Rotting GTBs) *Computing the optimal policy in Rotting GTBs with general matrices $\mathbf{G}$ is NP-Hard.*

The proof of this result follows a similar logic to the one of Theorem 1. Given this result, we now proceed by studying the block-diagonal connectivity matrices scenario, which composes an interesting class of Rotting GTBs. We now characterize the optimal policy in the block-diagonal connectivity scenario and the total cumulative reward that it obtains.

Theorem 9 (Optimal Policy in Rotting GTBs with Block-Diagonal CM) *For any instance $(\boldsymbol{\nu}, \mathbf{G}, T)$ of Rotting GTBs s.t. $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$, the optimal policy $\pi^*_{\boldsymbol{\nu},\mathbf{G},T} \in \arg\max_\pi J_{\boldsymbol{\nu},\mathbf{G},T}(\pi)$ is given by:*

$$\pi^*_{\boldsymbol{\nu},\mathbf{G},T}(t) \in \arg\max_{j \in [k]} \mu_j(\widetilde{N}^*_{j,t}), \qquad \forall t \in [T],$$

Algorithm 4: `RAW-UCB` (Seznec et al., 2020)

Input : Subgaussianity proxy σ, confidence levels $\{\delta_t\}_{t\in[T]}$.

1 **for** $t \in [T]$ **do**
2 | Compute $\widehat{\mu}_i^h(t)$ as in Equation (8), $\quad \forall i \in [k],\ h \in [N_{i,t}]$
3 | Select $I_t \in \arg\max_{i\in[k]} \min_{h\le N_{i,t}} \widehat{\mu}_i^h(t) + c(h, \delta_t)$
4 | Play I_t and observe $X_{I_t,t}$
5 **end**

where $\widetilde{N}^*_{j,t}$ is the number of times arm j has been triggered by the optimal policy up to time t. Moreover, we have:

$$J^*_{\boldsymbol{\nu},\mathbf{G},T} = \sum_{C_m \in \mathcal{C}_\mathbf{G}} \sum_{n=1}^{N^*_{C_m,T}} \max_{i\in C_m} \mu_i(n), \tag{7}$$

where $N^*_{C_m,T}$ is the number of times the optimal policy pulls an action belonging to clique C_m before T, i.e., $N^*_{C_m,T} = \widetilde{N}^*_{i,T}$, for every $i \in C_m$.

Interestingly, the optimal policy is *greedy* in every rotting bandit with block-diagonal CM: this extends known results for rested and restless rotting bandits, where the optimal policy was already be proven to be greedy (Heidari et al., 2016; Levine et al., 2017). Equation (7) provides a closed form of the total reward obtained by the optimal policy, which will come in handy in the next part.

4.2 Stochastic Rotting GTBs

In this part, we discuss the regret minimization problem for the stochastic Rotting GTBs. We first study an algorithm, namely `RAW-UCB` (Seznec et al., 2020), which is able to achieve sublinear regret in the block-diagonal connectivity scenario (Section 4.2.1). Then, we show that, under the literature's standard assumptions, we cannot learn for general matrices (Section 4.2.2).

4.2.1 Algorithm for Stochastic Rotting GTBs with Block-Diagonal CMs

We now show that the `RAW-UCB` algorithm (Seznec et al., 2020), whose pseudocode is provided in Algorithm 4, provides sublinear regret guarantees in the Rotting GTB setting with block-diagonal connectivity matrices. `RAW-UCB` does not require any knowledge on $\mathbf{G}$ and allows for efficient computation.[12]

The behavior of `RAW-UCB` is characterized as follows. At each round $t \in [T]$, the algorithm computes a family of estimators for every action (line 2). In particular, for every action $i \in [k]$ and for every window size $h_{i,t} \in [N^\pi_{i,t-1}]$, it computes:

$$\widehat{\mu}_i^h(t) := \frac{1}{h} \sum_{s=1}^{t-1} \mathbb{1}_{\{I_t = i \ \wedge\ N_{i,s} > N_{i,t-1} - h\}} X_{i,s}. \tag{8}$$

12. More details on the computationally efficient version of `RAW-UCB`, namely `EFF-RAW-UCB`, can be found in (Seznec et al., 2020).

Then, for every action, the chosen window size is the one minimizing the upper confidence bound $\widehat{\mu}_i^h(t) + c(h, \delta_t)$ where $c(h, \delta_t) := \sqrt{2\sigma^2 \log(2\delta_t^{-1})/h}$ (line 3).[13] Proving the algorithm's guarantees in the rested and restless setting requires characterizing how it concentrates, as has been done for the base case in Lemma 2 of (Seznec et al., 2020). We extend this result to the Rotting GTBs setting, devising a concentration bound involving the number of triggers of an action. the result can be found in Lemma 21 (Appendix C). This result will play a key role in the regret analysis of `RAW-UCB` in the Rotting GTBs setting. We are now ready to state the regret upper bound of `RAW-UCB` in the Rotting GTBs for block-diagonal connectivity matrices.

Theorem 10 (`RAW-UCB` Regret in Rotting GTBs with Block-Diagonal CM) *Let $(\boldsymbol{\nu}, \mathbf{G}, T)$ be an instance of the Rotting GTBs, where $\mathbf{G} \in \mathbf{B}_{\widetilde{k}}$. Let $\delta_t = t^{-\alpha}$ for $\alpha \geq 5$. Then, `RAW-UCB` suffers an expected regret bounded as:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{RAW-UCB}) \leq \widetilde{\mathcal{O}}\Bigg(\underbrace{k\left(\sigma\sqrt{\log T} + V_{\boldsymbol{\nu}}(T)\right)}_{\text{(A) Variance Contribution}} + \underbrace{L\sum_{C_m \in \mathcal{C}_{\mathbf{G}}} |C_m|^2 + kL + \sigma \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \left(\sqrt{\frac{|C_m|}{k}T}\right)}_{\text{(B) Rested Contribution}} + \underbrace{(\alpha\sigma)^{\frac{2}{3}} \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \left(V_T^{\pi} \frac{|C_m|}{k} T^2\right)^{\frac{1}{3}}}_{\text{(C) Restless Contribution}} \Bigg).$$

Dependence on Graph Topology. In Theorem 10, we observe the same phenomenon of Theorem 6. Indeed, we have three components: (A) representing the fixed regret contribution that comes from the noise, (B) representing the contribution to the regret coming from the rested nature of the problem (i.e., the sub-optimality accrued by choosing an action in the wrong clique), and (C) representing the contribution coming from the restless nature of the problem, instead (i.e., the sub-optimality accrued by not choosing the best action in a clique). The separation between the latter two components becomes clear in the proof of the result:

$$\begin{aligned}
\mathbb{E}[R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{RAW-UCB})] &= \sum_{t=1}^{T} \mu_{i_t^*}(\widetilde{N}_{i_t^*,t}^*) - \mu_{I_t}(\widetilde{N}_{I_t,t}^{\pi}) \pm \max_{i \in C_{I_t}} \mu_i(\widetilde{N}_{I_t,t}^{\pi}) \\
&= \underbrace{\sum_{t=1}^{T} (\mu_{i_t^*}(\widetilde{N}_{i_t^*,t}^*) - \max_{i \in C_{I_t}} \mu_i(\widetilde{N}_{I_t,t}^{\pi}))}_{\leq \text{(B)} + k\sigma\sqrt{\log T}} + \underbrace{\sum_{t=1}^{T} (\max_{i \in C_{I_t}} \mu_i(\widetilde{N}_{I_t,t}^{\pi}) - \mu_{I_t}(\widetilde{N}_{I_t,t}^{\pi}))}_{\leq \text{(C)} + kV_{\boldsymbol{\nu}}(T)}.
\end{aligned}$$

In rotting bandits, there is a clear hierarchy between the difficulties of statistical learning in the rested and the restless settings. Rested rotting bandits are easier than their restless counterpart, and this is reflected also in our bound: when the number of cliques is higher, and the GTB is closer to a rested instance, the regret bound is lower since the weight of (A) is higher.

13. At the beginning of the execution, we need a round-robin pull of the arms to initialize the estimators.

Comparison with Known Results from Literature. `RAW-UCB` has already been proven to be nearly optimal in both the rested and the restless scenario (Seznec et al., 2020). However, as an artifact of the analysis, we cannot retrieve the exact same bounds by plugging $\mathbf{G} = \mathbf{I}_k$, or $\mathbf{G} = \mathbf{1}_{k\times k}$ in our expression. A similar consideration to the one of Remark 5 can also be done for rotting bandits. We conjecture that `RAW-UCB` is actually nearly-optimal also in the intermediate Rotting GTB instances with block-diagonal connectivity matrices, and this claim is supported by the fact that there is no room for improvement in neither of the two contributions in the corner cases. We also conjecture that the dependence on $L\sum_{C_m\in\mathcal{C}_\mathbf{G}}|C_m|^2$ may be an artifact of the analysis, being the output of a delicate pigeonhole principle argument used to prove Theorem 10 (see Appendix C). We leave the task of finding a graph-dependent lower bound for this setting as a fascinating open problem.

4.2.2 Non-learnability for Stochastic Rotting GTBs with General Matrices

We now move to the scenario of stochastic Rotting GTBs for general connectivity matrices, and we present an impossibility result for this case.

Theorem 11 (Regret Lower Bound for Rotting GTBs with General Matrices) *For every $\mathbf{G}\in\{0,1\}^{k\times k}$ that is not block-diagonal, there exists an instance of Rotting GTB $(\boldsymbol{\nu},\mathbf{G},T)$ s.t., for every policy π, it holds:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\pi)\geq\frac{T}{12}.$$

This negative result poses a several limitation to what can be obtained from Rotting GTBs in the general scenario since, if we do not consider additional assumptions on the reward functions, no algorithm can obtain sublinear regret.

5 Discussion and Conclusions

In this paper, we proposed *graph-triggered bandits* (GTBs), a generalization of the rested and restless bandit settings, where the expected rewards of the different arms evolve by means of a graph. We focused and compared two special families of bandits, namely *rising* and *rotting* bandits, where the expected rewards of an arm evolve in a monotonic fashion with the number of *triggers* the specific arm received. We showed that computing the optimal policy without additional assumptions on the matrix is NP-Hard in both the rising and rotting scenarios. Then, for both these classes, we showed how, instead, for block-diagonal connectivity matrices, we can find the optimal policy in polynomial time and have a convenient closed-form expression. From the algorithmic perspective, we showed how it is possible to achieve sublinear regret for both of these special instances of MABs with block-diagonal connectivity matrices. On the other hand, for general matrices, we have an interesting distinction. Indeed, for Rising GTBs we were able to achieve sublinear regret, while for Rotting GTBs (in the standard scenario, without additional assumption on the behavior of expected rewards, e.g., on second-order derivatives), we proved that we cannot learn. This work aspires to be a first step in the study of GTBs and should be integrated by studying the statistical complexity of learning through lower bounds, and by considering more general models, e.g., smoothly evolving bandits.

Acknowledgments and Disclosure of Funding

Funded by the European Union – Next Generation EU within the project NRPP M4C2, Investment 1.3 DD. 341 - 15 March 2022 – FAIR – Future Artificial Intelligence Research – Spoke 4 - PE00000013 - D53C22002380006, and by the EU Horizon project ELIAS (European Lighthouse of AI for Sustainability, No. 101120237).

Appendix A. Related Works

In this appendix, we discuss the relevant literature for the GTB setting. We divide this appendix into two parts. First, we discuss the relevant works concerning graph structures. Then, we discuss the literature related to restless and rested bandits, with particular attention to rotting and rising bandits.

Graph Relationships in Bandits. The graph-triggered bandits setting has been introduced in this work. Thus, no prior literature is available on this setting. However, we mention similar settings that appeared in the last years. Herlihy and Dickerson (2023) propose the networked restless bandit setting. Despite some similarities with our setting, e.g., the presence of a graph among arms, their action space and learning objectives radically differ from ours, and thus the two settings are not comparable. In (Jhunjhunwala et al., 2018), a restless bandit setting is proposed in which the graph structure is not explicit in the formulation; however, the authors develop a graph representation of the policies in the deterministic scenario. Their algorithm builds and exploits a graph in an online fashion. Once again, we cannot properly compare this setting to ours, despite some sparse similarities. Finally, we mention bandits with graph feedback (Alon et al., 2015). Despite this setting being conceptually different from ours since arms do not interact, we report it here just because it features graph topology-dependent bounds. We remark that in this case, the graph has not to be intended as a structure for arms interactions but rather as a feedback structure for the learner, in GTBs the feedback is purely bandit.

Rested and Restless Bandits. Restless and rested bandits are a well-established research field. Starting from the seminal paper by Whittle (1988) on restless bandits, several approaches have been proposed over the years to deal with non-stationary bandits (Tekin and Liu, 2012; Raj and Kalyani, 2017). Then, specialization of these settings such as *rising* (Metelli et al., 2022; Mussi et al., 2024) and *rotting* (Levine et al., 2017) has been introduced. Over the last years, several works tackled rotting bandits (Levine et al., 2017; Seznec et al., 2019). Remarkably, (Seznec et al., 2020) provide a single algorithm for dealing with both rested and restless rotting bandits but show that in the rotting setting, achieving sub-linear regret is not possible when there are both rested and restless arms in the same instance. We remark that for any two-armed rotting bandit where one arm is rested and the other is restless, we can construct an (asymmetric) matrix $\mathbf{G}$ such that the instance can be mapped to a graph-triggered rotting bandit instance. This highlights a crucial difference between rotting and rising bandits for what concerns graph-triggering. Recently, literature studied a broader class of restless bandits called *smooth* bandits, which generalizes both rotting and rising bandits (Manegueu et al., 2021; Jia et al., 2023).

Appendix B. Proofs on Rising Bandits

In this appendix, we report a short version of the proofs of Rising GTBs. The extended version is provided in (Genalti et al., 2024).

Theorem 1 (Complexity of finding the Optimal Policy in Rising GTBs) *Computing the optimal policy in Rising GTBs with general matrices* $\mathbf{G}$ *is NP-Hard.*

Proof We reduce from a decision problem related to finding cliques in graphs. In particular, given a graph (V, E) and $\widetilde{M} \in \mathbb{N}$, it is NP-Hard to determine if there exists a clique of size $\widetilde{M}$ (Karp, 1972). In the following, we design an instance of our problem such that the reward of the optimal policy is at least $\sum_{t=1}^{T}(1 + \frac{t}{T^2})$ if and only if there exists a clique of size $\widetilde{M} = T$.

Construction. Given a graph (V, E), we build an instance such that the horizon is T. Our set of actions can be constructed by assigning an action to every node and time step couple, i.e., $\mathcal{A} = \{a_{v,t}\}_{v \in V,\ t \in [T]}$. We define the matrix $\widetilde{\mathbf{G}}$ is such that for any $v, v' \in V$ and $t, t' \in [T]$, it holds $G_{a_{v,t},a_{v',t'}} = 1$ if $(v, v') \in E$, and $G_{a_{v,t},a_{v',t'}} = 0$ otherwise. Finally, for each arm $a_{v,t} \in \mathcal{A}$, the reward is deterministic and evolves as $\mu_{a_{v,t}}(n) = \min\{1 + \eta t, \frac{n}{t}(1 + \eta t)\}$, where $\eta = T^{-2}$. We call $\widetilde{\boldsymbol{\nu}}$ the set of these functions. It is easy to see that the GTB instance $(\widetilde{\boldsymbol{\nu}}, \widetilde{\mathbf{G}}, T)$ satisfies Assumption 1.

if. We show that if there exists a clique $C^\star = \{v_1, \dots, v_T\}$ of size T, then there exists a policy with a cumulative reward of at least $\sum_{t=1}^{T}(1 + \eta t)$. Consider the policy $\widetilde{\pi}$ s.t. $\widetilde{\pi}(t) = a_{v_t,t}$. It is easy to see that $\widetilde{N}_{a_{v_t,t},t} = t$ for every $t \in [T]$. Hence, the reward of the policy $\widetilde{\pi}$ at time t is

$$\mu_{a_{v_t,t}}(\widetilde{N}_{a_{v_t,t},t}) = \min\{1 + \eta t, \frac{t}{t}(1 + \eta t)\} = 1 + \eta t.$$

Thus, $J_{\widetilde{\boldsymbol{\mu}},\widetilde{\mathbf{G}},T}(\widetilde{\pi}) = \sum_{t=1}^{T}(1 + \eta t)$ and the claim is proven.

only if. We show that if there is a policy $\widetilde{\pi}$ s.t. $J_{\widetilde{\boldsymbol{\mu}},\widetilde{\mathbf{G}},T}(\widetilde{\pi}) \geq \sum_{t=1}^{T}(1 + \eta t)$, then there exists a clique of size T.

First, we observe that for each $t', t \in [T]$ it holds that

$$\max_{t' \in [T]} \min\{1 + \eta t', \frac{t}{t'}(1 + \eta t')\} = 1 + \eta t. \tag{9}$$

This implies that, at any round t, the best obtainable reward is

$$\begin{aligned}
\max_{t' \in [T]} \max_{v \in V} \max_{l \leq t} \mu_{a_{v,t'}}(l) &= \max_{t' \in [T]} \max_{v \in V} \widetilde{\mu}_{a_{v,t'}}(t) \\
&= \max_{t' \in [T]} \min \left\{ 1 + \eta t', \frac{t}{t'}(1 + \eta t') \right\} \\
&= \min \left\{ 1 + \eta t, \frac{t}{t}(1 + \eta t) \right\} = 1 + \eta t.
\end{aligned}$$

Since by assumption there is a policy with reward at least $\sum_{t=1}^{T}(1 + \eta t)$, then there is a policy such that at each round $t \in [T]$ the reward is exactly $1 + \eta t$.

Consider a round $t \in [T]$. Let $a_{v,t'}$ be the arm played by the policy at this round. It must be the case that: i) $t' = t$, otherwise

$$\mu_{a_{v,t'}}(\widetilde{N}_{a_{v,t'},t}) \leq \mu_{a_{v,t'}}(t) < 1 + \eta t$$

by Equation (9), and ii) $\widetilde{N}_{a_{v,t'},t} = t$, otherwise

$$\mu_{a_{v,t'}}(\widetilde{N}_{a_{v,t'},t}) \leq \frac{t-1}{t}(1+\eta t) < 1 + \eta t.$$

Let $a_{v_t,t}$ be the arm chosen at round t. Then, each arm in $\{a_{v_t,t}\}_{t\in[T]}$, is chosen while having exactly $t-1$ triggers. By the definition of $\widetilde{\mathbf{G}}$ this directly implies that $\{v_t\}_{t=1}^{T}$ is a clique of size T. ■

Theorem 2 (Optimal Policy in Rising GTBs with Block-Diagonal CM) *For any instance $(\boldsymbol{\nu}, \mathbf{G}, T)$ of Rising GTBs with $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$, the optimal policy $\pi^*_{\boldsymbol{\nu},\mathbf{G},T} \in \arg\max_\pi J_{\boldsymbol{\nu},\mathbf{G},T}(\pi)$ is given by:*

$$\pi^*_{\boldsymbol{\nu},\mathbf{G},T}(t) \in \arg\max_{j \in C^*_{\boldsymbol{\nu},\mathbf{G},T}} \mu_j(t), \qquad \forall t \in [T],$$

where $C^*_{\boldsymbol{\nu},\mathbf{G},T}$ is the "best" cumulative reward clique:

$$C^*_{\boldsymbol{\nu},\mathbf{G},T} \in \arg\max_{C \in \mathcal{C}_{\mathbf{G}}} \sum_{t\in[T]} \max_{j\in C} \mu_j(t).$$

Proof For each clique $C_m \in \mathcal{C}_{\mathbf{G}}$, we substitute the reward function of every arm $i \in C_m$ with $\mu^*_i(t) = \max_{i\in C_m} \mu_i(t)$, for every $t \in [T]$. Now, since all arms sharing the same clique have the same reward function, our instance is equivalent to a $\widetilde{k}$-armed bandit problem, where $\widetilde{k}$ is the number of cliques. Since arms in different cliques are not connected, this corresponds to a rested bandit problem, and we use Proposition 1 of (Heidari et al., 2016) to get that the optimal policy would only pull the best action in terms of cumulative reward at the end of the time horizon T. To conclude the proof, we remark that playing greedily inside a clique corresponds exactly to play on the reward function defined above, which dominates the initial problem, and so the maximum cumulative reward is exactly the one attained in the problem with $\widetilde{k}$ arms. ■

Lemma 12 (`DR-BD-UB` Estimator's Instantaneous Bias) *For every arm $i \in [k]$, every round $t > 1$, let us define:*

$$\bar{\mu}_i(t) \coloneqq \mu_i(t^I_{i,N_{i,t-1}}) + (t - t^I_{i,N_{i,t-1}}) \frac{\mu_i(t^I_{i,N_{i,t-1}}) - \mu_i(t^I_{i,N_{i,t-1}-1})}{t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-1}},$$

then, $\bar{\mu}_i(t) \geq \mu_i(t^I_{i,N_{i,t-1}})$ and, if $N_{i,t-1} \geq 2$ it holds that:

$$\bar{\mu}_i(t) - \mu_i(\widetilde{N}_{i,t}) \leq (t - t^I_{i,N_{i,t}-1})\gamma_i(t^I_{i,N_{i,t-1}-1}).$$

Proof Let us start by observing the following equality holding:

$$\mu_i(\widetilde{N}_{i,t}) = \mu_i(t^I_{i,N_{i,t-1}}) + \sum_{j=t^I_{i,N_{i,t-1}}}^{\widetilde{N}_{i,t}-1} \gamma_i(j).$$

We have:

$$\begin{aligned}
\mu_i(\widetilde{N}_{i,t}) &= \mu_i(t^I_{i,N_{i,t-1}}) + \sum_{j=t^I_{i,N_{i,t-1}}}^{\widetilde{N}_{i,t}-1} \gamma_i(j) \\
&\leq \mu_i(t^I_{i,N_{i,t-1}}) + (\widetilde{N}_{i,t} - t^I_{i,N_{i,t-1}})\gamma_i(t^I_{i,N_{i,t-1}-1}) && (10)\\
&\leq \mu_i(t^I_{i,N_{i,t-1}}) + (t - t^I_{i,N_{i,t-1}})\gamma_i(t^I_{i,N_{i,t-1}-1}), && (11)
\end{aligned}$$

where line (10) follows from Assumption 1, and line (11) is obtained from observing that $\widetilde{N}_{i,t} \leq t$. Concerning the bias, when $N_{i,t-1} \geq 2$, we have:

$$\begin{aligned}
\bar{\mu}_i(t) - \mu_i(\widetilde{N}_{i,t}) &\leq \mu_i(t^I_{i,N_{i,t-1}}) - \mu_i(\widetilde{N}_{i,t}) + (t - t^I_{i,N_{i,t-1}})\frac{\mu_i(t^I_{i,N_{i,t-1}}) - \mu_i(t^I_{i,N_{i,t-1}-1})}{t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-1}} && (12)\\
&\leq (t - t^I_{i,N_{i,t-1}})\frac{\mu_i(t^I_{i,N_{i,t-1}}) - \mu_i(t^I_{i,N_{i,t-1}-1})}{t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-1}} && (13)\\
&\leq (t - t^I_{i,N_{i,t-1}})\gamma_i(t^I_{i,N_{i,t-1}-1}), && (14)
\end{aligned}$$

where line (13) follows from observing that $\mu_i(t^I_{i,N_{i,t-1}}) \leq \mu_i(\widetilde{N}_{i,t})$, and line (14) derives from bounding $\frac{\mu_i(t^I_{i,N_{i,t-1}}) - \mu_i(t^I_{i,N_{i,t-1}-1})}{t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-1}} \leq \gamma_i(t^I_{i,N_{i,t-1}-1})$ thanks to Assumption 1. ■

Theorem 3 (`DR-BD-UB` Regret in Det. Rising GTBs with Block-Diagonal CMs) *Let $(\boldsymbol{\nu}, \mathbf{G}, T)$ be an instance of Rising GTB, where $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$ and $\sigma = 0$. Then,* `DR-BD-UB` *suffers a regret bounded by:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{DR-BD-UB}) \leq \widetilde{\mathcal{O}}\Bigg(\inf_{q\in[0,1]} \Bigg\{ \underbrace{T^q \sum_{C_m \in \mathcal{C}} |C_m| \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C_m,T}}{|C_m|} \right\rceil, q \right)}_{\text{(A) Rested Bias Contribution}} + \underbrace{\sum_{C_m \in \mathcal{C}} |C_m| \widetilde{N}_{C_m,T}^{\frac{q}{1+q}} \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C_m,T}}{|C_m|} \right\rceil, q \right)^{\frac{1}{1+q}}}_{\text{(B) Restless Bias Contribution}} \Bigg\} \Bigg).$$

Proof Let $C^*_{\boldsymbol{\nu},\mathbf{G},T} \in \mathcal{C}_{\mathbf{G}}$ be the optimal clique of the instance. We analyze the following expression:

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{DR-BD-UB}) = \sum_{t=1}^{T} \mu_{i^*_t}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t}),$$

where $i_t^* \in \arg\max_{i \in C^*_{\boldsymbol{\nu},\mathbf{G},T}} \mu_i(t)$ for all $t \in [T]$. Then, we can decompose the regret in two meaningful components:

$$
\begin{aligned}
R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{DR-BD-UB}) &= \sum_{t=1}^{T} \mu_{i_t^*}(t) \pm \bar{\mu}_{I_t}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t}) \\
&\leq \sum_{t=1}^{T} \min\{1, \bar{\mu}_{I_t}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t})\} && (15) \\
&\leq \sum_{t=1}^{T} \min\{1, (t - t^I_{I_t,N_{I_t,t-1}})\gamma_{I_t}(t^I_{I_t,N_{I_t,t-1}-1})\} && (16) \\
&= \sum_{t=1}^{T} \min\{1, (t \pm t^I_{I_t,N_{I_t,t}} - t^I_{I_t,N_{I_t,t-1}})\gamma_{I_t}(t^I_{I_t,N_{I_t,t-1}-1})\} \\
&\leq \sum_{t=1}^{T} \min\{1, (t - t^I_{I_t,N_{I_t,t}})\gamma_{I_t}(t^I_{I_t,N_{I_t,t-1}-1})\} + && (17) \\
&\quad + \sum_{t=1} \min\{1, (t^I_{I_t,N_{I_t,t}} - t^I_{I_t,N_{I_t,t-1}})\gamma_{I_t}(t^I_{I_t,N_{I_t,t-1}-1})\} && (18) \\
&= 4k + \underbrace{\sum_{C_m \in \mathcal{C}_\mathbf{G}} \sum_{i \in C_m} \sum_{j=3}^{N_{j,T}} \min\{1, (t - t^I_{i,j})\gamma(t^I_{i,j-2})\}}_{\text{(a)}} + \\
&\quad + \underbrace{\sum_{C_m \in \mathcal{C}_\mathbf{G}} \sum_{i \in C_m} \sum_{j=3}^{N_{j,T}} \min\{1, (t^I_{i,j} - t^I_{i,j-1})\gamma(t^I_{i,j-2})\}}_{\text{(b)}},
\end{aligned}
$$

where lines (15) and (16) follow from Lemma 12, line (18) from the fact that $\min\{1, x+y\} \leq \min\{1, x\} + \min\{1, y\}$ for any $x, y \geq 0$.

These two terms represent the rested and the restless contribution to the regret, and we can bound them using similar techniques as in (Metelli et al., 2022). ■

Remark 5 (Regret Bound in Rested and Restless Rising Bandits) *When we are in a purely rested (resp. restless) scenario, the contribution term associated to the restless (resp. rested) scenario vanish, and we get the same regret orders from (Metelli et al., 2022). In particular, we can avoid splitting the minimum in Equation* (18) *and instead notice that in a rested setting we have* $t - t^I_{I_t,N_{I_t,t-1}} = t - N_{I_t,t-1}$*, and thus we can bound the cumulative regret as we bound the term (a). Instead, in a restless setting we have* $t - t^I_{I_t,N_{I_t,t-1}} = t - t_{I_t,N_{I_t,t-1}}$*, and thus we can bound the cumulative regret as we bound the term (b).*

Theorem 4 (`DR-G-UB` Regret in Det. Rising GTBs with General Matrices) *Let* $(\boldsymbol{\nu}, \mathbf{G}, T)$ *be an instance of Rising GTB, where* $\mathbf{G} \in \{0,1\}^{k \times k}$ *and* $\sigma = 0$*. Then,* `DR-G-UB`

suffers a regret bounded by:

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{DR-G-UB})$$
$$\leq \widetilde{\mathcal{O}}\left(\min_{q\in[0,1]} \left\{ T^q \sum_{C_m^L \in \mathcal{C}_{\bar{\mathbf{G}}^L}} |C_m^L| \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C_m^L,T}}{|C_m^L|} \right\rceil, q \right) + \sum_{C_m^L \in \mathcal{C}_{\bar{\mathbf{G}}^L}} |C_m^L| \widetilde{N}_{C_m^L,T}^{\frac{q}{1+q}} \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C_m^L,T}}{|C_m^L|} \right\rceil, q \right)^{\frac{1}{1+q}} \right\} \right),$$

where $\bar{\mathbf{G}}^L \in \mathbb{B}_{\widetilde{k}}$ *is a maximal sub-matrix of* $\mathbf{G}$.

Proof The theorem can be proved by showing that estimator's bias is always larger when internal times are decreased. For every arm $i \in [k]$ we define:

$$f_i(t;\ x, y) = \mu_i(x) + (t-x)\frac{\mu_i(x) - \mu_i(y)}{x-y}, \tag{19}$$

for every triplet of natural numbers $y \leq x \leq t \leq T$. Note that $\bar{\mu}_i(t) = f_i(t;\ t^I_{i,N_{i,t-1}}, t^I_{i,N_{i,t-1}-1})$, so if we can show that f_i is decreasing in both x and y, we can prove the claim. We start with the second argument: fix t and x, then for any y:

$$\begin{aligned} f_i(t;x,y) - f_i(t;x,y-1) &= (t-x)\left(\frac{\sum_{j=y}^{x-1}\gamma_i(j)}{x-y} - \frac{\sum_{j=y-1}^{x-1}\gamma_i(j)}{x-y+1} \right) \\ &= \frac{\sum_{j=y}^{x-1}\gamma_i(j) - (x-y)\gamma_i(y-1)}{(x-y)(x-y+1)} \leq 0, \end{aligned} \tag{20}$$

where line (20) follows from Assumption 1. With slightly more calculations we show that f_i is also decreasing in the first argument, fix t and y, then for any x:

$$f_i(t;\ x,y) - f_i(t;\ x-1,y) \leq 0. \tag{21}$$

Now we observe that, for every $i \in [k]$ and every $t \in [T]$, we have $t^I_{i,N_{i,t}} \geq t^{I,L}_{i,N_{i,t}}$. This is a consequence of Definition 1, since:

$$t^I_{i,N_{i,t}} - t^{I,L}_{i,N_{i,t}} = \sum_{j=1}^{t} (G_{I_t,i} - \bar{G}^L_{I_t,i}) \geq 0.$$

As a consequence of this, we have:

$$f_i(t;\ t^I_{i,N_{i,t-1}}, t^I_{i,N_{i,t}-1}) \leq f_i(t;\ t^{I,L}_{i,N_{i,t-1}}, t^{I,L}_{i,N_{i,t}-1}), \tag{22}$$

and

$$\mu_i(t^I_{i,N_{i,t}}) \geq \mu_i(t^{I,L}_{i,N_{i,t}}). \tag{23}$$

The proof can be concluded in the same way as for Theorem 3. ■

Lemma 13 (Estimator's Instantaneous Bias) *For every arm $i \in [k]$, every round $t \in [T]$, and window width $1 \leq h \leq \left\lfloor \frac{N_{i,t-1}}{2} \right\rfloor$, let us define:*

$$\widetilde{\mu}_i^h(t) \coloneqq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(t_{i,l}^I) + (t-l) \frac{\mu_i(t_{i,l}^I) - \mu_i(t_{i,l-h}^I)}{h} \right),$$

otherwise if $h = 0$, we set $\widetilde{\mu}_i^h(t) \coloneqq +\infty$. Then, $\widetilde{\mu}_i^h(t) \geq \mu_i(t_{i,N_{i,t-1}})$ and, if $N_{i,t-1} \geq 2$ it holds that:

$$\widetilde{\mu}_i^h(t) - \mu_i(\widetilde{N}_{i,t}) \leq \frac{(2t - 2N_{i,t-1} + h - 1)(t_{i,N_{i,t-1}}^I - t_{i,N_{i,t-1}-2h+1}^I)}{2h} \gamma_i(t_{i,N_{i,t-1}-2h+1}^I).$$

Proof Let us start by observing the following equality holding for every $l \in \{2, \dots, N_{i,t-1}\}$:

$$\mu_i(\widetilde{N}_{i,t}) = \mu_i(t_{i,l}^I) + \sum_{j=t_{i,l}^I}^{\widetilde{N}_{i,t}-1} \gamma_i(j).$$

By averaging over a window of length h, we obtain:

$$\begin{aligned}
\mu_i(\widetilde{N}_{i,t}) &= \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(t_{i,l}^I) + \sum_{j=t_{i,l}^I}^{\widetilde{N}_{i,t}-1} \gamma_i(j) \right) \\
&\leq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(t_{i,l}^I) + (\widetilde{N}_{i,t} - t_{i,l}^I) \gamma_i(t_{i,l}^I - 1) \right) && (24) \\
&\leq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(t_{i,l}^I) + \frac{\widetilde{N}_{i,t} - t_{i,l}^I}{t_{i,l}^I - t_{i,l-h}^I} \sum_{j=t_{i,l-h}^I}^{t_{i,l}^I-1} \gamma_i(j) \right) && (25) \\
&\leq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(t_{i,l}^I) + (t-l) \frac{\mu_i(t_{i,l}^I) - \mu_i(t_{i,l-h}^I)}{h} \right) =: \widetilde{\mu}_i^h(t), && (26)
\end{aligned}$$

where lines (24) and (25) follow from Assumption 1, and line (26) is obtained from observing that $t_{i,l}^I \geq l$, $\widetilde{N}_{i,t} \leq t$ and $t_{i,l}^I - t_{i,l-h}^I \geq h$.

Concerning the bias, when $N_{i,t-1} \geq 2$, we have:

$$\begin{aligned}
\widetilde{\mu}_i^h(t) - \mu_i(\widetilde{N}_{i,t}) &= \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(t_{i,l}^I) + (t-l) \frac{\mu_i(t_{i,l}^I) - \mu_i(t_{i,l-h}^I)}{h} \right) - \mu_i(\widetilde{N}_{i,t}) \\
&\leq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} (t-l) \frac{\mu_i(t_{i,l}^I) - \mu_i(t_{i,l-h}^I)}{h} && (27) \\
&= \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} (t-l) \frac{\mu_i(t_{i,l}^I) - \mu_i(t_{i,l-h}^I)}{t_{i,l}^I - t_{i,l-h}^I} \frac{t_{i,l}^I - t_{i,l-h}^I}{h}
\end{aligned}$$

$$\leq \frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}(t-l)\gamma_i(t^I_{i,l-h})\frac{t^I_{i,l}-t^I_{i,l-h}}{h} \tag{28}$$

$$\leq \frac{t^I_{i,N_{i,t-1}}-t^I_{i,N_{i,t-1}-2h+1}}{h^2}\gamma_i(t^I_{i,N_{i,t-1}-2h+1})\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}(t-l) \tag{29}$$

$$= \frac{(2t-2N_{i,t-1}+h-1)(t^I_{i,N_{i,t-1}}-t^I_{i,N_{i,t-1}-2h+1})}{2h}\gamma_i(t^I_{i,N_{i,t-1}-2h+1}), \tag{30}$$

where line (27) follows from observing that $\mu_i(t^I_{i,l}) \leq \mu_i(\widetilde{N}_{i,t})$, line (28) derives from Assumption 1 and bounding $\frac{\mu_i(t^I_{i,l})-\mu_i(t^I_{i,l-h})}{t^I_{i,l}-t^I_{i,l-h}} \leq \gamma_i(t^I_{i,l-h})$, line (29) is obtained by bounding $t^I_{i,l}-t^I_{i,l-h} \leq t^I_{i,N_{i,t-1}}-t^I_{i,N_{i,t-1}-2h+1}$ and $\gamma_i(t^I_{i,l-h}) \leq \gamma_i(t^I_{i,N_{i,t-1}-2h+1})$, and line (30) follows from computing the summation.

■

Lemma 14 (Bound on Estimator's Cumulative Bias for Block-Diagonal CMs) *Let $(I_t)_{t=1}$ be a sequence of actions. For every action $i \in [k]$, every round $t \in [T]$, let window width $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$. Let $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$ be a block diagonal matrix, then for every $q \in [0,1]$, we have:*

$$\sum_{t=1}^{T}\min\left\{1, \widetilde{\mu}_{I_t}^{h_{I_t,t}}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t})\right\} \leq$$

$$\leq 2k + \bar{k}_1 T^q \left\lceil \frac{1}{1-2\epsilon} \right\rceil \Upsilon_{\boldsymbol{\nu}}\left(\left\lceil (1-2\epsilon)\frac{T}{\bar{k}_1}\right\rceil, q\right) +$$

$$+ T^{\frac{2q}{1+q}}(1+\log(\epsilon T))^{\frac{q}{1+q}}\left\lceil \frac{1}{\epsilon}\right\rceil\left\lceil \frac{1}{1-2\epsilon}\right\rceil \sum_{C_m \in \mathcal{C}_{\mathbf{G}}:|C_m|>1}|C_m|\Upsilon_{\boldsymbol{\nu}}\left(\left\lceil (1-2\epsilon)\frac{T}{|C_m|}\right\rceil, q\right)^{\frac{1}{1+q}},$$

where $\mathcal{C}$ is the set of blocks of matrix $\mathbf{G}$, and $\bar{k}_1 \leq k$ is the number of blocks of size 1.

Proof The statement can be proven by decomposing over the cliques and then over the arms, splitting cliques with only one arm from the others:

$$\sum_{t=1}^{T}\min\left\{1, \widetilde{\mu}_{I_t}^{h_{I_t,t}}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t})\right\} \leq 2k + \underbrace{\sum_{\substack{C_m \in \mathcal{C}_{\mathbf{G}}:|C_m|=1 \\ C_m=\{i\}}}\sum_{j=3}^{N_{i,T}}\min\left\{1, \widetilde{\mu}_i^{h_{i,t_{i,j}}}(t_{i,j}) - \mu_i(j)\right\}}_{\text{(a)}} +$$

$$+ \underbrace{\sum_{C_m \in \mathcal{C}_{\mathbf{G}}:|C_m|>1}\sum_{i \in C_m}\sum_{j=3}^{N_{i,T}}\min\left\{1, \widetilde{\mu}_i^{h_{i,t_{i,j}}}(t_{i,j}) - \mu_i(t^I_{i,j})\right\}}_{\text{(b)}}.$$

The two terms can be bound in a similar way as in (Metelli et al., 2022), as the rested and the restless component, respectively. ■

Theorem 6 (R-□-UCB Regret in Rising GTBs with Block-Diagonal CMs) *Let $(\boldsymbol{\nu}, \mathbf{G}, T)$ be an instance of Rising GTB, where $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$. Let $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$ for $\epsilon \in (0, 1/2)$ and $\delta_t = t^{-\alpha}$ for $\alpha > 2$. Then, R-□-UCB suffers an expected regret bounded by:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{R-□-UCB}) \leq \widetilde{\mathcal{O}}\Bigg(\min_{q \in [0,1]} \Bigg\{ \underbrace{(\sigma T)^{\frac{2}{3}}}_{\text{(A) Variance Contribution}} + \underbrace{\bar{k}_1 T^q \Upsilon_{\boldsymbol{\nu}}\left(\left\lceil \frac{T}{\bar{k}_1} \right\rceil, q \right)}_{\text{(B) Rested Bias Contribution}} + \underbrace{T^{\frac{2q}{1+q}} \sum_{C_m \in \mathcal{C}_{\mathbf{G}} : |C_m| > 1} |C_m| \Upsilon_{\boldsymbol{\nu}}\left(\left\lceil \frac{T}{|C_m|} \right\rceil, q \right)^{\frac{1}{1+q}}}_{\text{(C) Restless Bias Contribution}} \Bigg\} \Bigg),$$

where $\bar{k}_1$ is the number of cliques in $\mathbf{G}$ containing only one action.

Proof Let us define the good events $\mathcal{E}_t = \bigcap_{i \in [k]} \mathcal{E}_{i,t}$ that correspond to the event in which all confidence intervals hold:

$$\mathcal{E}_{i,t} := \left\{ \left| \widehat{\mu}_i^{h_{i,t}}(t) - \widetilde{\mu}_i^{h_{i,t}}(t) \right| \leq \beta_i^{h_{i,t}}(t) \right\} \qquad \forall i \in [T],\, i \in [k].$$

We have to analyze the following expression:

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{DR-BD-UB}) = \mathbb{E}\left[\sum_{t=1}^{T} \mu_{i_t^*}(t) - \mu_{I_t}(t) \right],$$

where $i_t^* \in \arg\max_{i \in C^*_{\boldsymbol{\nu},\mathbf{G},T}} \mu_i(t)$ for all $t = 1$. We decompose according to the good events $\mathcal{E}_t$:

$$\begin{aligned} R_{\boldsymbol{\nu},\mathbf{G},T}(\pi^{\texttt{DR-BD-UB}}) &= \sum_{t=1}^{T} \mathbb{E}\left[\left(\mu_{i_t^*}(t) - \mu_{I_t}(t) \right) \mathbb{1}\{\mathcal{E}_t\} \right] + \sum_{t=1}^{T} \mathbb{E}\left[\left(\mu_{i_t^*}(t) - \mu_{I_t}(t) \right) \mathbb{1}\{\neg\mathcal{E}_t\} \right] \\ &\leq \sum_{t=1}^{T} \mathbb{E}\left[\left(\mu_{i_t^*}(t) - \mu_{I_t}(t) \right) \mathbb{1}\{\mathcal{E}_t\} \right] + \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\{\neg\mathcal{E}_t\} \right], \end{aligned}$$

where we exploited $\mu_{i_t^*}(t) - \mu_{I_t}(t) \leq 1$ in the inequality. The second summation can be bounded using standard arguments, recalling that $\alpha > 2$:

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\{\neg\mathcal{E}_t\} \right] &\leq 1 + \sum_{i \in [k]} \sum_{t=2}^{T} \mathbb{P}\left(\neg\mathcal{E}_{i,t} \right) \\ &\leq 1 + \frac{2k}{\alpha - 2}. \end{aligned}$$

where the first inequality is obtained with $\mathbb{P}(\neg\mathcal{E}_1) \leq 1$ and a union bound over $[k]$. Recalling $\mathbb{P}(\neg\mathcal{E}_{i,t})$ was bounded in Lemma 5, we bound the summation with the integral and obtain the second inequality.

The rest of the analysis can be conducted under the good event $\mathcal{E}_t$, recalling that $B_i(t) \equiv \widehat{\mu}_i^{h_{i,t}}(t) + \beta_i^{h_{i,t}}(t)$. Let $t \in [T]$, and we exploit the optimism, i.e., $B_{i_t^*}(t) \leq B_{I_t}(t)$:

$$\begin{aligned}\mu_{i^*}(t) - \mu_{I_t}(t) + B_{I_t}(t) - B_{I_t}(t) &\leq \min\left\{1, \underbrace{\mu_{i_t^*}(t) - B_{i_t^*}(t)}_{\leq 0} + B_{I_t}(t) - \mu_{I_t}(t)\right\} \\ &\leq \min\left\{1, B_{I_t}(t) - \mu_{I_t}(t)\right\}.\end{aligned}$$

Now, we work on the term inside the minimum:

$$\begin{aligned}B_{I_t}(t) - \mu_{I_t}(t) &= \widehat{\mu}_{I_t}^{h_{I_t,t}}(t) + \beta_{I_t}^{h_{I_t,t}}(t) - \mu_{I_t}(t) &(31)\\ &\leq \underbrace{\widetilde{\mu}_{I_t}^{h_{I_t,t}}(t) - \mu_{I_t}(t)}_{\text{(a)}} + \underbrace{2\beta_{I_t}^{h_{I_t,t}}(t)}_{\text{(b)}}, &(32)\end{aligned}$$

where line (31) follows from the definition of $B_i(t)$ and line (32) from the good event $\mathcal{E}_t$. We make use of Lemma 14 and Lemma 18 to bound the summations over t of (a) and (b), respectively.

Putting all together, we obtain:

$$\begin{aligned}&R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{R-}\square\texttt{-UCB}) \\ &\leq 1 + \frac{2k}{\alpha - 2} + 5k + \frac{k}{\epsilon} + \frac{3k}{\epsilon}(2\sigma T)^{\frac{2}{3}}\left(10\alpha \log T\right)^{\frac{1}{3}} + \\ &\quad + T^{\frac{2q}{1+q}}(1 + \log(\epsilon T))^{\frac{q}{1+q}} \left\lceil \frac{1}{\epsilon} \right\rceil \left\lceil \frac{1}{1-2\epsilon} \right\rceil k \Upsilon_{\boldsymbol{\mu}}\left(\left\lceil (1-2\epsilon)\frac{T}{k} \right\rceil, q\right)^{\frac{1}{1+q}} + \\ &\quad + 2k + \bar{k}_1 T^q \left\lceil \frac{1}{1-2\epsilon} \right\rceil \Upsilon_{\boldsymbol{\nu}}\left(\left\lceil (1-2\epsilon)\frac{T}{\bar{k}_1} \right\rceil, q\right) + \\ &\quad + T^{\frac{2q}{1+q}}(1 + \log(\epsilon T))^{\frac{q}{1+q}} \left\lceil \frac{1}{\epsilon} \right\rceil \left\lceil \frac{1}{1-2\epsilon} \right\rceil \sum_{C_m \in \mathcal{C}_{\mathbf{G}} : |C_m| > 1} |C_m| \Upsilon_{\boldsymbol{\nu}}\left(\left\lceil (1-2\epsilon)\frac{T}{|C_m|} \right\rceil, q\right)^{\frac{1}{1+q}}.\end{aligned}$$

■

Lemma 15 (Bound on Estimator's Cumulative Bias for General Matrices) *Let $\{I_t\}_{t=1}$ be a sequence of actions. For every action $i \in [k]$, every round $t \in [T]$, let window width $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$. Let $\mathbf{G} \in \{0,1\}^{k \times k}$, then for every $q \in [0,1]$, we have*

$$\begin{aligned}&\sum_{t=1}^{T} \min\left\{1, \widetilde{\mu}_{I_t}^{h_{I_t},t}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t})\right\} \leq \\ &\qquad \leq 2k + \bar{k}_1 T^q \left\lceil \frac{1}{1-2\epsilon} \right\rceil \Upsilon_{\boldsymbol{\nu}}\left(\left\lceil (1-2\epsilon)\frac{T}{\bar{k}_1} \right\rceil, q\right) + \\ &\qquad\quad + T^{\frac{2q}{1+q}}(1 + \log(\epsilon T))^{\frac{q}{1+q}} \left\lceil \frac{1}{\epsilon} \right\rceil \left\lceil \frac{1}{1-2\epsilon} \right\rceil (k - \bar{k}_1) \Upsilon_{\boldsymbol{\nu}}\left(\left\lceil (1-2\epsilon)\frac{T}{k - \bar{k}_1} \right\rceil, q\right)^{\frac{1}{1+q}},\end{aligned} \tag{33}$$

where $\bar{k}_1 \leq k$ *is the number of arms having degree of* 1*, i.e.,* $\bar{k}_1 \coloneqq |\{i \in [k] : deg(i) = 1\}|$.

Proof The proof follows similar steps as Lemma 14. We decided to split arms based on their degree; in particular, we bound separately the bias due to arms having a degree of 1 (i.e., they are only triggered by themselves).

$$\sum_{t=1}^{T} \min\left\{1, \widetilde{\mu}_{I_t}^{h_{I_t,t}}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t})\right\}$$
$$\leq 2k + \underbrace{\sum_{i\in[k]} \sum_{j=3}^{N_{i,T}} \min\left\{1, \widetilde{\mu}_i^{h_{i,t_{i,j}}}(t_{i,j}) - \mu_i(j)\right\}}_{\deg^-(i)=1\text{ (a)}} + \underbrace{\sum_{i\in[k]} \sum_{j=3}^{N_{i,T}} \min\left\{1, \widetilde{\mu}_i^{h_{i,t_{i,j}}}(t_{i,j}) - \mu_i(t_{i,j}^I)\right\}}_{\deg^-(i)>1\text{ (b)}}.$$

As a consequence of Definition 1, we observe that:

$$t^I_{i,N_{i,t}} - t^{I,U}_{i,N_{i,t}} = \sum_{j=1}^{t} (G_{I_t,i} - \bar{G}^U_{I_t,i}) \leq 0.$$

As a consequence of this, we have that, for every $i \in [k]$ and for every $t \in [T]$:

$$\widetilde{N}_{i,t} \leq \widetilde{N}^U_{i,t}, \tag{34}$$

where $\widetilde{N}^U_{i,t} \coloneqq \mathbf{e}_i^\top (\bar{\mathbf{G}}^U)^\top \mathbf{N}_t$. Then, following similar steps as in (Metelli et al., 2022), we can bound the two components separately and make the dependency on the upper block-diagonal matrix explicit. ■

Theorem 7 (R-□-UCB Regret in Rising GTBs with General Matrices) *Let* $(\boldsymbol{\nu}, \mathbf{G}, T)$ *be an instance of Rising GTB, where* $\mathbf{G} \in \{0,1\}^{k\times k}$*. Let* $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$ *for* $\epsilon \in (0, 1/2)$ *and* $\delta_t = t^{-\alpha}$ *for* $\alpha > 2$*. Then,* R-□-UCB *suffers an expected regret bounded by:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{R-□-UCB}) \leq \widetilde{\mathcal{O}}\left(\min_{q\in[0,1]}\left\{(\sigma T)^{\frac{2}{3}} + T^q \bar{k}_1 \Upsilon_\nu\left(\frac{T}{\bar{k}_1}, q\right) + T^{\frac{2q}{1+q}} \sum_{C_m^U} |C_m^U| \Upsilon_\nu\left(\frac{T}{|C_m^U|}, q\right)^{\frac{1}{1+q}}\right\}\right),$$

where $\bar{\mathbf{G}}^U$ *is the minimal super-matrix of* $\mathbf{G}$*.*

Proof The proof follows similar steps of the proof of Theorem 6, but uses Lemma 15 (instead of Lemma 14) to bound cumulative estimator's bias.

As in Theorem 6, we decompose the regret in two components and instead make use of Lemma 15 and Lemma 18 to bound the summations over t of the two components, respectively. Putting all together, we obtain:

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{R-□-UCB}) \leq 1 + \frac{2k}{\alpha - 2} + 5k + \frac{k}{\epsilon} + \frac{3k}{\epsilon}(2\sigma T)^{\frac{2}{3}} (10\alpha \log T)^{\frac{1}{3}} +$$

$$+ 2k + \bar{k}_1 T^q \left\lceil \frac{1}{1-2\epsilon} \right\rceil \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil (1-2\epsilon)\frac{T}{\bar{k}_1} \right\rceil, q \right) +$$
$$+ T^{\frac{2q}{1+q}} (1+\log(\epsilon T))^{\frac{q}{1+q}} \left\lceil \frac{1}{\epsilon} \right\rceil \left\lceil \frac{1}{1-2\epsilon} \right\rceil \cdot \sum_{\substack{C_m^U \in \mathcal{C}_{\bar{\mathbf{G}}^U} \\ |C_m^U|>1}} |C_m| \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil (1-2\epsilon)\frac{T}{|C_m|} \right\rceil, q \right)^{\frac{1}{1+q}}.$$

■

B.1 Technical Lemmas

Lemma 16 (Lemma C.1 of Metelli et al. 2022) *Let $M \geq 3$, and let $f : \mathbb{N} \to \mathbb{R}$, and $\beta \in (0,1)$. Then it holds that:*

$$\sum_{j=3}^{M} f(\lfloor \beta j \rfloor) \leq \left\lceil \frac{1}{\beta} \right\rceil \sum_{l=\lfloor 3\beta \rfloor}^{\lfloor \beta M \rfloor} f(l).$$

Lemma 17 (Lemma C.2 of Metelli et al. 2022) *Under Assumption 1, it holds that:*

$$\max_{\substack{(N_{i,T})_{i\in[k]} \\ N_{i,T}\geq 0, \sum_{i\in[k]} N_{i,T}=T}} \sum_{i\in[k]} \sum_{l=1}^{N_{i,T}-1} \gamma_i(l)^q \leq k \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{T}{k} \right\rceil, q \right).$$

Lemma 5 (Concentration of Estimator, adapted from Metelli et al. 2022) *For every arm $i \in [k]$, every round $t \in [T]$, and window width $1 \leq h \leq \left\lfloor \frac{N_{i,t-1}}{2} \right\rfloor$, let:*

$$\beta_i^h(t,\delta) \coloneqq \sigma(t - N_{i,t-1} + h - 1)\sqrt{\frac{10 \log \frac{1}{\delta}}{h^3}}.$$

Then, if the window size depends on the number of pulls only $h_{i,t} = h(N_{i,t-1})$ and if $\delta_t = t^{-\alpha}$ for some $\alpha > 2$, it holds for every round $t \in [T]$ that:

$$\mathbb{P}\left(\left| \widehat{\mu}_i^{h_{i,t}}(t) - \widetilde{\mu}_i^{h_{i,t}}(t) \right| > \beta_i^{h_{i,t}}(t, \delta_t) \right) \leq 2t^{1-\alpha}.$$

Proof Using a Doob's *optional skipping* argument (Doob, 1953; Bubeck et al., 2008), and noting that, at round t, $t_{i,l}^I$ is a stopping time for every arm $i \in [k]$ and pull number $l \in \{1, \dots, N_{i,t-1}\}$ w.r.t. the filtration $\mathcal{F}_{\tau-1} = \sigma(I_1, X_1, \dots, I_{\tau-1}, X_{\tau-1}, I_\tau)$, we can proceed to prove this lemma as in (Metelli et al., 2022) also for GTB. ■

Lemma 18 (Bound on Estimator's Variance, Theorem 4.4 of Metelli et al. 2022) *Let $(I_t)_{t\in[T]}$ be a sequence of actions such that:*

$$\left| \widehat{\mu}_{I_t}^{h_{I_t,t}}(t) - \widetilde{\mu}_{I_t}^{h_{I_t,t}}(t) \right| \leq \beta_{I_t}^{h_{I_t,t}}(t, t^{-\alpha}), \ \forall t \in [T], \tag{35}$$

where $\alpha > 2$. For every action $i \in [k]$, every round $t \in [T]$, let window width $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$, then, we have:

$$\sum_{t=1}^{T} \min\left\{1, 2\beta_{I_t}^{h_{I_t,t}}(t, t^{-\alpha})\right\} \leq k\left(3 + \frac{1}{\epsilon}\right) + \frac{3k}{\epsilon}(2\sigma T)^{\frac{2}{3}}(10\alpha \log T)^{\frac{1}{3}}. \tag{36}$$

Appendix C. Proofs on Rotting Bandits

Theorem 8 (Complexity of finding the Optimal Policy in Rotting GTBs) *Computing the optimal policy in Rotting GTBs with general matrices $\mathbf{G}$ is NP-Hard.*

Proof We reduce from a decision problem related to finding independent sets in graphs. In particular, given a graph (V, E) and $\widetilde{M} \in \mathbb{N}$, it is NP-Hard to determine if there exists an independent set of size $\widetilde{M}$ (Karp, 1972). In the following, we design an instance of our problem such that the reward of the optimal policy is at least T if and only if there exists an independent set of size $\widetilde{M} = T$.

Construction. Given a graph (V, E), we build an instance such that the horizon is T. Our set of actions can be constructed by assigning an action to every node, i.e., $\mathcal{A} = \{a_v\}_{v \in V}$. We define the matrix $\widetilde{\mathbf{G}}$ is such that for any $v, v' \in V$, it holds $G_{a_v, a_{v'}} = 1$ if $(v, v') \in E$, and $G_{a_v, a_{v'}} = 0$ otherwise. Finally, for each arm $a_v \in \mathcal{A}$, the reward is deterministic and evolves as $\mu_{a_v,t}(n) = \max\{2 - n, 0\}$. We call $\widetilde{\boldsymbol{\nu}}$ the set of these functions. It is easy to see that the GTB instance $(\widetilde{\boldsymbol{\nu}}, \widetilde{\mathbf{G}}, T)$ satisfies Assumption 2.

if. We show that if there exists an independent set $I^\star = \{v_1, \dots, v_T\}$ of size T, then there exists a policy with a cumulative reward of at least T. Consider the policy $\widetilde{\pi}$ s.t. $\widetilde{\pi}(t) = a_{v_t}$. It is easy to see that $\widetilde{N}_{a_{v_t},t} = 1$ for every $t \in [T]$. Hence, the reward of the policy $\widetilde{\pi}$ at time t is

$$\mu_{a_{v_t}}(\widetilde{N}_{a_{v_t},t}) = 1.$$

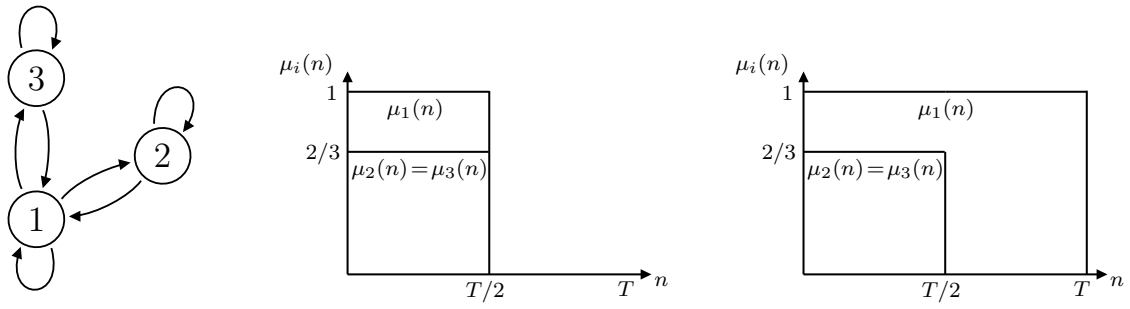
Thus, $J_{\widetilde{\boldsymbol{\mu}}, \widetilde{\mathbf{G}}, T}(\widetilde{\pi}) = T$ and the claim is proven.

only if. We show that if there is a policy $\widetilde{\pi}$ s.t. $J_{\widetilde{\boldsymbol{\mu}}, \widetilde{\mathbf{G}}, T}(\widetilde{\pi}) \geq T$, then there exists an independent set of size T. First, we observe that at any round t the best obtainable reward is 1. Since, by assumption, there is a policy with a reward of at least T; then there is a policy such that at each round $t \in [T]$, the reward is exactly 1.

Let a_{v_t} be the arm played by the policy at round $t \in [T]$. Then, consider a round $t \in [T]$. Since the reward of the arm a_{v_t} must be 1, it must be the case that $\mu_{a_{v_t}}(\widetilde{N}_{a_{v_t},t}) = 1$ and $\widetilde{N}_{a_{v_t},t} = 1$. By the definition of $\widetilde{\mathbf{G}}$ this directly implies that $\{v_t\}$ is not connected to any $v_{t'}$, $t' < t$, and that $v_{t'} \neq v_t$ for any $t' < t$. Hence, $\{v_t\}_{t \in [T]}$ is an independent set of size T, proving the claim. ■

Lemma 19 *Let $\mathcal{G}_n = (V_n, E_n)$ be a graph. Then, either the graph possesses block-diagonal connectivity and the nodes can be partitioned in M disjoint clique, i.e., $V = \bigcup_{m=1}^{M} C_m$, or there exist three nodes v_1, v_2 and v_3 s.t. e_{v_1,v_2}, $e_{v_2,v_3} \in E$ and $e_{v_1,v_3} \notin E$.*

Proof We proceed by induction. The statement holds for $n \leq 3$. We assume that $\mathcal{G}_n$ is an arbitrary graph satisfying the statement. Then we add one node v_{n+1} and obtain $\mathcal{G}_{n+1} = (V_{n+1}, E_{n+1})$.



(a) Graph considered for both $\boldsymbol{\nu}$ and $\boldsymbol{\nu}'$.

(b) Expected values $\mu_i(\cdot)$ for $\boldsymbol{\nu}$.

(c) Expected values $\mu_i(\cdot)$ for $\boldsymbol{\nu}'$.

Figure 2: Instances used in the proof of Theorem 11.

If $\mathcal{G}_n$ is block-diagonal connected. We list all the possible scenarios:

- If $e_{v_{n+1},i} \notin E_{n+1}$ for every $i \in V_n$ then the node v_{n+1} is a single-element clique and the new graph is block-diagonal.
- If $e_{v_{n+1},i} \in E_{n+1}$ for every $i \in C_m$, and $e_{v_{n+1},j} \notin E_{n+1}$ for every $j \in V_n \setminus C_m$, then the node v_{n+1} is added to the clique C_m and the new graph is block-diagonal.
- If $e_{v_{n+1},i} \in E_{n+1}$ for some $i \in C_m$ and $e_{v_{n+1},j} \notin E_{n+1}$ for some $j \in C_m$, then $e_{v_{n+1},i}$, $e_{i,j} \in E$ and $e_{v_{n+1},j} \notin E$.
- If $e_{v_{n+1},i} \in E_{n+1}$ for some $i \in C_m$ and $e_{v_{n+1},j} \in E_{n+1}$ for some $j \in C_{m'}$, then $e_{v_{n+1},i}$, $e_{v_{n+1},j} \in E$ and $e_{i,j} \notin E$.

If there exists three nodes v_1, v_2 and v_3 s.t. e_{v_1,v_2}, $e_{v_2,v_3} \in E$ and $e_{v_1,v_3} \notin E$. There is no way to connect v_1 and v_3 by adding a node, thus the statement still holds for $\mathcal{G}_{n+1}$. ■

Theorem 11 (Regret Lower Bound for Rotting GTBs with General Matrices) *For every $\mathbf{G} \in \{0,1\}^{k\times k}$ that is not block-diagonal, there exists an instance of Rotting GTB $(\boldsymbol{\nu}, \mathbf{G}, T)$ s.t., for every policy π, it holds:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\pi) \geq \frac{T}{12}.$$

Proof Consider the deterministic rotting scenario, i.e., where $\sigma = 0$. Consider two instances $\boldsymbol{\nu}$ and $\boldsymbol{\nu}'$ of 3-armed rotting bandit with graph structure as depicted in Figure 2. The graph which represents the connection of the arms is represented in Figure 2a. The expected rewards at the different number of triggers n is depicted in Figure 2b for instance $\boldsymbol{\nu}$ and in Figure 2c for instance $\boldsymbol{\nu}'$. For both instances, arms 2 and 3 present an expected reward equal to $2/3$ for the first $T/2$ triggers, and then the expected reward becomes 0. On the other hand, the two instances differ in the behavior of the expected reward of arm 1. Indeed, such reward is 1 until we trigger the arm $T/2$ times for instance $\boldsymbol{\nu}$ and for all the T triggers for instance $\boldsymbol{\nu}'$.

We recall that the clairvoyant is aware of both the graph $\mathbf{G}$ and the expected values $\mu_i(n)$, for every $i \in [k]$ and $n \in [T]$. We can easily compute the total reward for the best policy possible π^* for instance $\boldsymbol{\nu}$:

$$J_{\boldsymbol{\nu},\mathbf{G},T}(\pi^*) = \frac{2}{3}T, \tag{37}$$

which corresponds to pull arms 2 and 3 only (both for $T/2$ times), and for instance $\boldsymbol{\nu}'$:

$$J_{\boldsymbol{\nu}',\mathbf{G},T}(\pi^*) = T, \tag{38}$$

which corresponds to pull always pull arm 1 for T times. We highlight that optimal policy π^* is different for the two instances.

We now need to introduce some additional notations that will be used in the proof. We call $\mathbb{E}_{\boldsymbol{\nu}}\left[N_i^R(n)\right]$ the expected number of pulls for arm i generating reward (i.e., for which the expected reward is different from 0) up to time n for instance $\boldsymbol{\nu}$. We now start by observing that, up to the round $T/2$, the two instances are exactly the same, so every policy π will have the same behavior in expectation. Given that, we observe that for both the instances we have the same reward, equal to:

$$\begin{aligned} J_{\boldsymbol{\nu},\mathbf{G},T/2}(\pi) = J_{\boldsymbol{\nu}',\mathbf{G},T/2}(\pi) &= \mathbb{E}_{\boldsymbol{\nu}}\left[N_1^R\left(\frac{T}{2}\right)\right] + \frac{2}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] + \frac{2}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right] \\ &= \frac{T}{2} - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right], \end{aligned}$$

where the last equality follows from $\mathbb{E}_{\boldsymbol{\nu}}\left[N_1^R\left(\frac{T}{2}\right)\right] + \mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] + \mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right] = T/2$. This result is valid for both $\boldsymbol{\nu}$ and $\boldsymbol{\nu}'$, as the policy will behave in the same way, and so $\mathbb{E}_{\boldsymbol{\nu}}\left[N_i^R\left(\frac{T}{2}\right)\right] = \mathbb{E}_{\boldsymbol{\nu}'}\left[N_i^R\left(\frac{T}{2}\right)\right]$, for every $i \in [3]$, as the policy on the two instances $\boldsymbol{\nu}$ and $\boldsymbol{\nu}'$ are not distinguishable the first $T/2$ rounds.

We now have to understand what will happen from $T/2$ to T in the best case possible.

Instance $\boldsymbol{\nu}$ We can easily see how for arm 1 we have terminated the pulls which generate reward, so we have to pull arms 2 and 3. We can now compute the remaining triggers generating reward for arm 2 in the second half of the rounds:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R(T)\right] - \mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] &\leq \underbrace{\frac{T}{2}}_{\substack{\text{Triggers} \\ \text{initially} \\ \text{available}}} \underbrace{- \mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right]}_{\substack{\text{Already} \\ \text{used}}} \\ &\quad \underbrace{- \left(\frac{T}{2} - \mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] - \mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right]\right)}_{\substack{\text{Triggers used} \\ \text{from arm 1}}} \\ &\leq \mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right] \end{aligned}$$

We can do the same reasoning for arm 3 and, for symmetry, we get:

$$\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R(T)\right] - \mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right] \leq \mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right]$$

We now consider a policy using all these triggers, and we compute the expected cumulative reward:

$$\begin{aligned}
J_{\boldsymbol{\nu},\mathbf{G},T}(\pi) &\leq J_{\boldsymbol{\nu},\mathbf{G},T/2}(\pi) + \frac{2}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] + \frac{2}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right] \\
&\leq \frac{T}{2} - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right] + \frac{2}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] + \frac{2}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right] \\
&\leq \frac{T}{2} + \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] + \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right].
\end{aligned}$$

Instance $\boldsymbol{\nu}'$ Instead, for instance $\boldsymbol{\nu}'$, we can easily see that the best choice from $T/2$ to T is to always pull arm 1 for all the $T/2$ rounds, receiving a reward of 1 each time. Given that, we have:

$$\begin{aligned}
J_{\boldsymbol{\nu}',\mathbf{G},T}(\pi) &\leq J_{\boldsymbol{\nu}',\mathbf{G},T/2}(\pi) + \frac{T}{2} \\
&= T - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right].
\end{aligned}$$

Regret Moving to the regret, we have for instance $\boldsymbol{\nu}$:

$$\begin{aligned}
R_{\boldsymbol{\nu},\mathbf{G},T}(\pi) &= J_{\boldsymbol{\nu},\mathbf{G},T}(\pi^*) - J_{\boldsymbol{\nu},\mathbf{G},T}(\pi) \\
&\geq \frac{2}{3}T - \frac{T}{2} - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right] \\
&\geq \frac{T}{6} - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right],
\end{aligned}$$

while for instance $\boldsymbol{\nu}'$:

$$\begin{aligned}
R_{\boldsymbol{\nu}',\mathbf{G},T}(\pi) &= J_{\boldsymbol{\nu}',\mathbf{G},T}(\pi^*) - J_{\boldsymbol{\nu}',\mathbf{G},T}(\pi) \\
&\geq \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] + \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right].
\end{aligned}$$

We can now compute a lower bound on the regret:

$$\begin{aligned}
R_T(\mathfrak{A}) &= \max\left\{R_{\boldsymbol{\nu},\mathbf{G},T}(\pi), R_{\boldsymbol{\nu}',\mathbf{G},T}(\pi)\right\} \\
&\geq \frac{1}{2}\left(R_{\boldsymbol{\nu},\mathbf{G},T}(\pi) + R_{\boldsymbol{\nu}',\mathbf{G},T}(\pi)\right) \\
&= \frac{1}{2}\left(\frac{T}{6} - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right] + \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] + \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right]\right) \\
&= \frac{T}{12}.
\end{aligned}$$

This proof holds for the specific graph structure we discussed here. However, by joining this result with the one of Lemma 19, we can generalize this result for every non-block-diagonal connectivity matrix. ■

Theorem 9 (Optimal Policy in Rotting GTBs with Block-Diagonal CM) *For any instance $(\boldsymbol{\nu}, \mathbf{G}, T)$ of Rotting GTBs s.t. $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$, the optimal policy $\pi^*_{\boldsymbol{\nu},\mathbf{G},T} \in \arg\max_\pi J_{\boldsymbol{\nu},\mathbf{G},T}(\pi)$ is given by:*

$$\pi^*_{\boldsymbol{\nu},\mathbf{G},T}(t) \in \arg\max_{j\in[k]} \mu_j(\widetilde{N}^*_{j,t}), \qquad \forall t \in [T],$$

where $\widetilde{N}^*_{j,t}$ is the number of times arm j has been triggered by the optimal policy up to time t. Moreover, we have:

$$J^*_{\boldsymbol{\nu},\mathbf{G},T} = \sum_{C_m \in \mathcal{C}_\mathbf{G}} \sum_{n=1}^{N^*_{C_m,T}} \max_{i\in C_m} \mu_i(n), \tag{7}$$

where $N^*_{C_m,T}$ is the number of times the optimal policy pulls an action belonging to clique C_m before T, i.e., $N^*_{C_m,T} = \widetilde{N}^*_{i,T}$, for every $i \in C_m$.

Proof For every Rotting GTB instance, we create an alternative instance which is better, in terms of total cumulative reward, than the original instance. Then we show that playing greedy in the original instance yields the same cumulative reward of the optimal policy from the alternative instance.

For each clique $C_m \in \mathcal{C}_\mathbf{G}$, we substitute the reward function of every arm $i \in C_m$ with $\mu^*_i(n) = \max_{i\in C_m} \mu_i(n)$ for every $n \in [T]$. This way, whenever an action is chosen it is guaranteed to always yield the same reward as any other possible action inside the same clique. We create an alternative instance $(\widetilde{\boldsymbol{\nu}}, \widetilde{\mathbf{G}}, T)$ by collapsing all the actions inside the same clique into a single meta-action, resulting in a $\widetilde{k}$-armed rested rotting bandit problem, where the set of actions corresponds the set of cliques of the original instance. We use Proposition 2 of (Heidari et al., 2016) to get that the optimal policy in the alternative instance is to play, at every round, the action with the highest instantaneous reward. Such policy achieves a total reward, in the alternative instance, of:

$$J^*_{\widetilde{\boldsymbol{\nu}},\widetilde{\mathbf{G}},T} = \sum_{C_m \in \mathcal{C}_\mathbf{G}} \sum_{n=1}^{N^*_{C_m,T}} \max_{i\in C_m} \mu_i(n),$$

We now show that playing the greedy policy in the original instance yields an equal total cumulative reward. Playing greedily in the original instance we get:

$$\begin{aligned}
J^*_{\boldsymbol{\nu},\mathbf{G},T} &= \sum_{t=1}^{T} \max_{i\in[k]} \mu_i(\widetilde{N}^*_{i,t}) \\
&= \sum_{C_m \mathcal{C}_\mathbf{G}} \sum_{t=1}^{T} \mathbb{1}_{\{I^*_t \in C_m\}} \max_{i\in[k]} \mu_i(\widetilde{N}^*_{i,t}) \\
&= \sum_{C_m \mathcal{C}_\mathbf{G}} \sum_{t=1}^{T} \mathbb{1}_{\{I^*_t \in C_m\}} \max_{i\in C_m} \mu_i(\widetilde{N}^*_{i,t}) \\
&= \sum_{C_m \in \mathcal{C}_\mathbf{G}} \sum_{n=1}^{N^*_{C_m,T}} \max_{i\in C_m} \mu_i(n)
\end{aligned}$$

$$= J^*_{\widetilde{\boldsymbol{\nu}},\widetilde{\mathbf{G}},T} \geq J_{\boldsymbol{\nu},\mathbf{G},T}(\pi) \quad \forall \pi.$$

The performance of the optimal policy in the alternative instance is matched by the greedy policy played in the original instance. The proof is concluded by observing that the optimal total cumulative reward of the alternative instance cannot be lower than the total reward of any policy π in the original instance, since the alternative instance has pointwise higher reward functions for every action. ■

C.1 Upper Bounding the Regret of RAW-UCB

We start by defining the expectation version of the estimator defined in Equation (8) as $\bar{\mu}_i^h(t) \coloneqq \frac{1}{h}\sum_{s=1}^{t-1} \mathbb{1}_{\{I_t=i \,\wedge\, N_{i,s}>N_{i,t-1}-h\}}\mu_i(\widetilde{N}_{i,s})$. Before moving on, we recall the following result, which also introduces the notion of *good event* ξ_t^α.

Proposition 20 (Bound on the Probability of Bad Event, Seznec et al. 2020) *Let $\delta_t = 2t^{-\alpha}$, and*

$$\xi_t^\alpha \coloneqq \left\{\forall i \in [k],\ \forall n \leq t-1,\ \forall h \leq n,\ |\widehat{\mu}_i^h(t) - \bar{\mu}_i^h(t)| \leq c(h,\delta_t)\right\},$$

for $c(h,\delta_t) \coloneqq \sqrt{2\sigma^2 \log(2\delta_t^{-1})/h}$. Then

$$\mathbb{P}\left(\bar{\xi_t^\alpha}\right) \leq Kt^{2-\alpha}. \tag{39}$$

Lemma 21 (Overestimation under the Good Event) *Under ξ_t^α, if action I_t is selected by Algorithm 4, for every $h \in [N_{i,t-1}]$ we have:*

$$\bar{\mu}_{I_t}^h \geq \max_{i\in[k]} \mu_i(\widetilde{N}_{i,t-1}^\pi) - 2c(h,\delta_t), \tag{40}$$

where $\widetilde{N}_{i,t-1}^\pi$ is the number of triggers of action i provoked by playing with Algorithm 4 up until time t.

Proof This proof is adapted from the one of Lemma 1 of (Seznec et al., 2020). Let $h_{i,t}^{\min} \in \arg\min_{h\leq N_{i,t-1}} \widehat{\mu}_i^h(t) + c(h,\delta_t)$.

Let $i_t^\pi \in \arg\max_{i\in[k]} \mu_i(\widetilde{N}_{i,t-1}^\pi)$ be the best available action at time t. From the rotting assumption, we know that:

$$\max_{i\in[k]} \mu_i(\widetilde{N}_{i,t-1}^\pi) = \mu_{i_t^\pi}(\widetilde{N}_{i,t-1}^\pi) \leq \bar{\mu}_{i_t^\pi}^1(t) \leq \ldots \leq \bar{\mu}_{i_t^\pi}^{h_{i_t^\pi,t}^{\min}}(t).$$

Under ξ_t^α, we have:

$$\bar{\mu}_{i_t^\pi}^{h_{i_t^\pi,t}^{\min}}(t) \leq \widehat{\mu}_{I_t}^{h_{I_t,t}^{\min}}(t) + c(h_{I_t,t}^{\min},\delta_t).$$

We now use the definition of $h_{I_t,t}^{\min}$:

$$\widehat{\mu}_{I_t}^{h_{I_t,t}^{\min}}(t) + c(h_{I_t,t}^{\min},\delta_t) \leq \widehat{\mu}_{I_t}^h(t) + c(h,\delta_t).$$

Again, we use ξ_t^α:

$$\widehat{\mu}_{I_t}^h(t) + c(h, \delta_t) \leq \bar{\mu}_{I_t}^h(t) + 2c(h, \delta_t).$$

Putting all together, we obtain the statement. ■

Theorem 10 (RAW-UCB Regret in Rotting GTBs with Block-Diagonal CM) *Let $(\boldsymbol{\nu}, \mathbf{G}, T)$ be an instance of the Rotting GTBs, where $\mathbf{G} \in \mathbf{B}_{\widetilde{k}}$. Let $\delta_t = t^{-\alpha}$ for $\alpha \geq 5$. Then,* RAW-UCB *suffers an expected regret bounded as:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{RAW-UCB}) \leq \widetilde{\mathcal{O}}\Bigg(\underbrace{k\left(\sigma\sqrt{\log T} + V_{\boldsymbol{\nu}}(T)\right)}_{\text{(A) Variance Contribution}} + \underbrace{L\sum_{C_m \in \mathcal{C}_\mathbf{G}} |C_m|^2 + kL + \sigma \sum_{C_m \in \mathcal{C}_\mathbf{G}} \left(\sqrt{\frac{|C_m|}{k}T}\right)}_{\text{(B) Rested Contribution}} + \underbrace{(\alpha\sigma)^{\frac{2}{3}} \sum_{C_m \in \mathcal{C}_\mathbf{G}} \left(V_T^\pi \frac{|C_m|}{k} T^2\right)^{\frac{1}{3}}}_{\text{(C) Restless Contribution}} \Bigg).$$

Proof Let us proceed to decompose the regret:

$$\begin{aligned}
R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{RAW-UCB}) &= \sum_{t=1}^T \left(\mu_{i_t^*}(\widetilde{N}_{i_t^*,t}^*) - \mu_{I_t}(\widetilde{N}_{I_t,t}^\pi)\right) \\
&= \sum_{t=1}^T \left(\mu_{i_t^*}(\widetilde{N}_{i_t^*,t}^*) - \mu_{I_t}(\widetilde{N}_{I_t,t}^\pi) \pm \max_{i \in C_{I_t}} \mu_i(\widetilde{N}_{I_t,t}^\pi)\right) \\
&= \underbrace{\sum_{t=1}^T (\mu_{i_t^*}(\widetilde{N}_{i_t^*,t}^*) - \max_{i \in C_{I_t}} \mu_i(\widetilde{N}_{I_t,t}^\pi))}_{\text{(b)}} + \underbrace{\sum_{t=1}^T (\max_{i \in C_{I_t}} \mu_i(\widetilde{N}_{I_t,t}^\pi) - \mu_{I_t}(\widetilde{N}_{I_t,t}^\pi))}_{\text{(c)}}
\end{aligned}$$

Before bounding the two terms, we observe the following:

$$\sum_{t=1}^T \max_{i \in C_{I_t}} \mu_i(\widetilde{N}_{I_t,t}^\pi) = \sum_{C_m \in \mathcal{C}_\mathbf{G}} \sum_{n=1}^{N_{C_m,T}^\pi} \max_{i \in C} \mu_i(n). \tag{41}$$

Equation (41) is a consequence of Equation (7) (Theorem 9), when applied to the restless bandit problems obtained by each clique when considered alone. We have for (c):

$$\begin{aligned}
\text{(c)} &= \sum_{t=1}^T \max_{i \in C_{I_t}} \left(\mu_i(\widetilde{N}_{I_t,t}^\pi) - \mu_{I_t}(\widetilde{N}_{I_t,t}^\pi)\right) \\
&\overset{\text{Eq. (41)}}{=} \sum_{C_m \in \mathcal{C}_\mathbf{G}} \sum_{n=1}^{N_{C_m,T}^\pi} \left(\max_{i \in C} \mu_i(n) - \mu_{I_{t_{C,n}}}(n)\right)
\end{aligned}$$

$$\overset{(\lozenge)}{\leq} 6kV_{\boldsymbol{\nu}}(T) + 4(8\alpha\sigma)^{\frac{2}{3}} \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \left(V_{\boldsymbol{\nu}}(T)|C_m|(N^{\pi}_{C_m,T})^2 \log T\right)^{\frac{1}{3}} + \\ + 2(2\sqrt{2}\alpha\sigma)^{\frac{1}{3}} \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \left(V_{\boldsymbol{\nu}}(T)^2|C_m|^2 N^{\pi}_{C_m,T}\sqrt{\log T}\right)^{\frac{1}{3}}$$

The last inequality is obtained by observing that, fixing the number of times a pull is selected, we have a nested restless bandit problem having as the time horizon the number of times the clique is pulled $N^{\pi}_{C_m,T}$. `RAW-UCB` plays greedily in each clique independently. Thus, when an action belonging to clique C is selected, it is the same action that an instance of `RAW-UCB` would have played in a restless rotting bandit composed only by the actions belonging to C.[14] In the step marked with $(\lozenge)$, this equivalence allows us to bound (c) with the summation of regret bounds of the algorithm for smaller restless bandits defined for the cliques, by using Theorem 1 from (Seznec et al., 2020) and bounding $\log N^{\pi}_{C_m,T} \leq \log T$ and $V_{\boldsymbol{\nu}}(N^{\pi}_{C_m,T}) \leq V_{\boldsymbol{\nu}}(T)$ for every $C_m \in \mathcal{C}_{\mathbf{G}}$. The last term is dominated by the other two in every quantity, and is thus omitted in the final bound.

We now focus on (b):

$$\begin{aligned}
\text{(b)} &= \sum_{t=1}^{T} (\mu_{i_t^*}(\widetilde{N}^*_{i_t^*,t}) - \max_{i \in C_{I_t}} \mu_i(\widetilde{N}^{\pi}_{I_t,t})) \\
&\overset{(40)}{=} \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \sum_{n=1}^{N^*_{C_m,T}} \max_{i \in C} \mu_i(n) - \sum_{t=1}^{T} \max_{i \in C_{I_t}} \mu_i(\widetilde{N}^{\pi}_{I_t,t}) \\
&\overset{(41)}{=} \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \sum_{n=1}^{N^*_{C_m,T}} \max_{i \in C} \mu_i(n) - \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \sum_{n=1}^{N^{\pi}_{C_m,T}} \max_{i \in C} \mu_i(n)
\end{aligned}$$

The term (b) only depends on the difference between the allocation of pulls among the cliques between the optimal policy and the algorithm's policy. Thus, it makes sense to split the cliques into two sets, namely OP and UP: the first will contain the OverPulled cliques, the second the UnderPulled cliques, which are cliques pulled by `RAW-UCB` more than the optimal policy and the cliques pulled less, respectively.

$$\begin{aligned}
&\sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \sum_{n=1}^{N^*_{C_m,T}} \max_{i \in C} \mu_i(n) - \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \sum_{n=1}^{N^{\pi}_{C_m,T}} \max_{i \in C} \mu_i(n) \\
&\qquad = \sum_{C_m \in \mathrm{UP}} \sum_{n=N^{\pi}_{C_m,T}+1}^{N^*_{C_m,T}} \max_{i \in C} \mu_i(n) - \sum_{C_m \in \mathrm{OP}} \sum_{n=N^*_{C_m,T}+1}^{N^{\pi}_{C_m,T}} \max_{i \in C} \mu_i(n)
\end{aligned}$$

We now introduce the auxiliary quantity $\mu^+_T(\pi) \coloneqq \max_{i \in [k]} \mu_i(\widetilde{N}^{\pi}_{i,T})$. We also observe that the two terms in the RHS have the same number of addends, since the number of overpulls

14. The UCB is different since the clique-specific instance of `RAW-UCB` would have used *internal times* instead of the external time t, however, the order is preserved and the decision is the same.

must be equal to the number of underpulls. Finally, we define $h_{C,T}$ as the number of overpulls of clique C.

$$\begin{aligned}
&\sum_{C_m\in\mathrm{UP}}\sum_{n=N^{\pi}_{C_m,T}}^{N^{*}_{C_m,T}-1}\max_{i\in C}\mu_i(n)-\sum_{C_m\in\mathrm{OP}}\sum_{n=N^{*}_{C_m,T}}^{N^{\pi}_{C_m,T}-1}\max_{i\in C}\mu_i(n)\\
&\qquad\le\sum_{C_m\in\mathrm{UP}}\sum_{n=N^{\pi}_{C_m,T}}^{N^{*}_{C_m,T}-1}\mu^{+}_T(\pi)-\sum_{C_m\in\mathrm{OP}}\sum_{n=N^{*}_{C_m,T}}^{N^{\pi}_{C_m,T}-1}\max_{i\in C}\mu_i(n)\\
&\qquad=\sum_{C_m\in\mathrm{OP}}\sum_{n=N^{*}_{C_m,T}}^{N^{\pi}_{C_m,T}-1}(\mu^{+}_T(\pi)-\max_{i\in C}\mu_i(n))\\
&\qquad=\sum_{C_m\in\mathrm{OP}}\sum_{h=0}^{h_{C,T}-1}(\mu^{+}_T(\pi)-\max_{i\in C}\mu_i(N^{*}_{C_m,T}+h)).
\end{aligned}$$

We can now decompose the last summation by the means of events $\{\xi^{\alpha}_t\}_t$:

$$\begin{aligned}
(\mathrm{b}_\xi)&\le\sum_{C_m\in\mathrm{OP}}\sum_{h=0}^{h_{C,T}-1}\mathbb{1}\{\xi^{\alpha}_{t^{\pi}_{C,N^{*}_{C_m,T}+h}}\}(\mu^{+}_T(\pi)-\max_{i\in C}\mu_i(N^{*}_{C_m,T}+h))\\
&\le\sum_{C_m\in\mathrm{OP}^{\xi}}\sum_{h=0}^{h^{\xi}_{C,T}}(\mu^{+}_T(\pi)-\max_{i\in C}\mu_i(N^{*}_{C_m,T}+h)),
\end{aligned}$$

where $h^{\xi}_{C,T}\coloneqq\max\{h\le h_{C,T}:\xi_{t^{\pi}_{C,N^{*}_{C_m,T}+h}}\}$ is the largest number of overpulls a clique undergoes before time $t^{\pi}_{C,N^{*}_{C_m,T}+h}\le T$ under the events ξ^{α}_t, and $\mathrm{OP}^{\xi}\coloneqq\{C_m\in\mathrm{OP}:h^{\xi}_{C,T}\ge 1\}$. We call, for short, $\widetilde{t}_{C,h}$ the time at which clique C is overpulled for the h-th time i.e., $t^{\pi}_{C,N^{*}_{C_m,T}+h}$, and observe that

$$\begin{aligned}
&\sum_{h=0}^{h^{\xi}_{C,T}}\max_{i\in C}\mu_i(N^{*}_{C_m,T}+h)\\
&\qquad=\sum_{h=0}^{h^{\xi}_{C,T}}\mathbb{1}\{h\ne h_{C,t_{j,N^{\pi}_{j,T}}}\forall j\in[k]\}\max_{i\in C}\mu_i(N^{*}_{C_m,T}+h)+\sum_{j\in C}\max_{i\in C}\mu_i(N^{*}_{C_m,T}+h_{C,t_{j,N^{\pi}_{j,T}}})\\
&\qquad=\sum_{i\in C}\sum_{h=0}^{h^{\xi}_{i,T}-1}\mu_i(N^{\pi}_{C,\widetilde{t}_{C,h}})+\sum_{j\in C}\max_{i\in C}\mu_i(\widetilde{N}^{\pi}_{j,t_{j,N^{\pi}_{j,T}}})\\
&\qquad\overset{(8)}{=}\sum_{i\in C}(h^{\xi}_{i,T}-1)\bar{\mu}_i^{h^{\xi}_{i,T}-1}(\widetilde{t}_{C,h^{\xi}_{i,T}})+\sum_{j\in C}\max_{i\in C}\mu_i(\widetilde{N}^{\pi}_{j,t_{j,N^{\pi}_{j,T}}})\\
&\qquad\overset{(40)}{\ge}\sum_{i\in C}(h^{\xi}_{i,T}-1)\left(\max_{i\in[k]}\mu_i(\widetilde{N}^{\pi}_{i,T})-2c(h^{\xi}_{i,T}-1,\delta_{\widetilde{t}_{C,h^{\xi}_{i,T}}})\right)+\sum_{j\in C}\max_{i\in C}\mu_i(\widetilde{N}^{\pi}_{j,t_{j,N^{\pi}_{j,T}}})
\end{aligned}$$

$$\geq (h^{\xi}_{C,T} - |C_m|) \max_{i\in[k]} \mu_i(\widetilde{N}^{\pi}_{i,T}) - 2\sum_{i\in C}(h^{\xi}_{i,T}-1)c(h^{\xi}_{i,T}-1,\delta_T) + \sum_{j\in C}\max_{i\in C}\mu_i(\widetilde{N}^{\pi}_{j,t_{j,N^{\pi}_{j,T}}})$$
$$= (h^{\xi}_{C,T} - |C_m|)\mu^{+}_{T}(\pi) - 2\sum_{i\in C}(h^{\xi}_{i,T}-1)c(h^{\xi}_{i,T}-1,\delta_T) + \sum_{j\in C}\max_{i\in C}\mu_i(\widetilde{N}^{\pi}_{j,t_{j,N^{\pi}_{j,T}}}).$$

Plugging this observation into the previous, we get:

$$\begin{aligned}
(\mathrm{b}_\xi) &\leq \sum_{C_m\in \mathrm{OP}^{\xi}} \left(|C_m|\mu^{+}_{T}(\pi) - \sum_{j\in C}\max_{i\in C}\mu_i(\widetilde{N}^{\pi}_{j,t_{j,N^{\pi}_{j,T}}}) + 2\sum_{i\in C}(h^{\xi}_{i,T}-1)c(h^{\xi}_{i,T}-1,\delta_T)\right)\\
&= \sum_{C_m\in \mathrm{OP}^{\xi}} \left(\sum_{i\in C}(\mu^{+}_{T}(\pi) - \max_{j\in C}\mu_j(\widetilde{N}^{\pi}_{i,t_{i,N^{\pi}_{i,T}}})) + 2\sum_{i\in C}(h^{\xi}_{i,T}-1)c(h^{\xi}_{i,T}-1,\delta_T)\right)\\
&\overset{(\star)}{\leq} 2k\sigma\sqrt{\log T} + L\sum_{C_m\in\mathcal{C}_{\mathbf{G}}}|C_m|^2 + 2\sum_{C_m\in \mathrm{OP}^{\xi}}\left(\sum_{i\in C}(h^{\xi}_{i,T}-1)c(h^{\xi}_{i,T}-1,\delta_T)\right)\\
&\leq 2k\sigma\sqrt{\log T} + L\sum_{C_m\in\mathcal{C}_{\mathbf{G}}}|C_m|^2 + 2\sum_{C_m\in \mathrm{OP}^{\xi}}\left(\sigma\sum_{i\in C}\sqrt{(h^{\xi}_{i,T}-1)\log T}\right)\\
&\leq 2k\sigma\sqrt{\log T} + L\sum_{C_m\in\mathcal{C}_{\mathbf{G}}}|C_m|^2 + 2\sum_{C_m\in \mathrm{OP}}\left(\sigma\sqrt{\log T}\sum_{i\in C}\sqrt{(h_{i,T}-1)}\right)\\
&\overset{(J)}{\leq} 2k\sigma\sqrt{\log T} + L\sum_{C_m\in\mathcal{C}_{\mathbf{G}}}|C_m|^2 + 2\sum_{C_m\in\mathcal{C}_{\mathbf{G}}}\left(\sigma\sqrt{|C_m|N^{\pi}_{C_m,T}\log T}\right).
\end{aligned}$$

The step marked with $(\star)$ is justified by the following considerations. Let $i\in C$, we shorten the notation for the time at which clique C is triggered for the $(\widetilde{N}^{\pi}_{i,t_{i,N^{\pi}_{i,T}}} - m)$-th time as $t_{i,-m} \coloneqq t_{C,\widetilde{N}^{\pi}_{i,t_{i,N^{\pi}_{i,T}}}-m}$. In other words, after this time the clique C is only chosen m times before the action $i\in C$ is pulled for the last time. Consider the $|C_m|$ times the clique C is chosen before pulling i for the last time: then, due to the pigeonhole principle, at least one action belonging to the clique should appear at least two times before the last pull. Without loss of generality, we assume that only one action appears exactly two times, and call the first appearance time $t_{i,-m}$ and the second $t_{i,-m'}$ (note that $m'\leq m\leq |C_m|$). We now observe that:

$$\begin{aligned}
\sum_{i\in C}\left(\mu^{+}_{T}(\pi) - \max_{j\in C}\mu_j(\widetilde{N}^{\pi}_{i,t_{i,N^{\pi}_{i,T}}})\right) &= \sum_{i\in C}\left(\mu^{+}_{T}(\pi) - \max_{j\in C}\left\{\mu_j(\widetilde{N}^{\pi}_{i,t_{i,N^{\pi}_{i,T}}}) \pm \mu_j(\widetilde{N}^{\pi}_{i,t_{i,N^{\pi}_{i,T}}} - m)\right\}\right)\\
&\leq \sum_{i\in C}\left(\mu^{+}_{T}(\pi) - \max_{j\in C}\mu_j(\widetilde{N}^{\pi}_{i,t_{i,N^{\pi}_{i,T}}} - m) + mL\right)\\
&\leq \sum_{i\in C}\left(\mu^{+}_{T}(\pi) - \max_{j\in C}\mu_j(\widetilde{N}^{\pi}_{i,t_{i,N^{\pi}_{i,T}}} - m)\right) + |C_m|^2 L,
\end{aligned}$$

We can now prove the step $(\star)$ by bounding

$$\sum_{i\in C}\max_{j\in C}\mu_j(\widetilde{N}^{\pi}_{i,t_{i,N^{\pi}_{i,T}}} - m) \geq \sum_{i\in C}\mu_{I_{t_{i,-m}}}(\widetilde{N}^{\pi}_{i,t_{i,N^{\pi}_{i,T}}} - m)$$

$$
\begin{aligned}
&= \sum_{i \in C} \bar{\mu}^1_{I_{t_{i,-m}}}(t_{i,-m'}) \\
&\overset{(40)}{\geq} \sum_{i \in C} \left(\max_{j \in [k]} \mu_j(\widetilde{N}^\pi_{j,t_{i,-m'}}) - 2c(1, \delta_{t_{i,-m'}}) \right) \\
&\overset{(\dagger)}{\geq} \sum_{i \in C} \left(\mu^+_T(\pi) - 2c(1, \delta_T) \right),
\end{aligned}
$$

where (†) is a consequence of $\widetilde{N}^\pi_{j,t_{i,-m'}} \leq \widetilde{N}^\pi_{j,T}$ for every $j \in [k]$.

Finally, in the step marked with (J) we used Jensen inequality to find the worst allocation of overpull among the actions in the same clique, which is the uniform one, i.e., $h_{i,T} \leq N^\pi_{C_m,T}/|C_m|$.

To conclude the proof, we need to find out what happens under $\bar{\xi}_t{}^\alpha$:

$$
\begin{aligned}
(\mathrm{b}_{\bar{\xi}}) &\leq \sum_{C_m \in \mathrm{OP}} \sum_{h=0}^{h_{C,T}-1} \mathbb{1}\{\bar{\xi}^\alpha_{t^\pi_{C,N^*_{C_m,T}+h}}\} (\mu^+_T(\pi) - \max_{i \in C} \mu_i(N^*_{C_m,T} + h)) \\
&= \sum_{C_m \in \mathrm{OP}} \sum_{h=0}^{h_{C,T}-1} \sum_{t=1}^{T} \mathbb{1}\{\bar{\xi}^\alpha_{t^\pi_{C,N^*_{C_m,T}+h}} \wedge t^\pi_{C,N^*_{C_m,T}+h} = t\} (\mu^+_T(\pi) - \max_{i \in C} \mu_i(N^*_{C_m,T} + h)) \\
&\overset{(*)}{\leq} \sum_{t=1}^{T} \mathbb{1}\{\bar{\xi}^\alpha_{t^\pi_{C,N^*_{C_m,T}+h}}\} Lt \left(\sum_{C_m \in \mathrm{OP}} \sum_{h=0}^{h_{C,T}-1} \mathbb{1}\{t^\pi_{C,N^*_{C_m,T}+h} = t\} \right) \\
&\leq \sum_{t=1}^{T} \mathbb{1}\{\bar{\xi}^\alpha_{t^\pi_{C,N^*_{C_m,T}+h}}\} Lt.
\end{aligned}
$$

In the step marked with $(\star)$, we use the fact that at time t overpulling a clique can yield at most Lt regret. The last step is a consequence that for each round t we can have at most 1 overpull. We conclude the bound by using Proposition 20:

$$
\mathbb{E}[(\mathrm{b}_{\bar{\xi}})] \leq \sum_{t=1}^{T} \mathbb{P}\left(\bar{\xi}^\alpha_t\right) Lt \overset{(39)}{\leq} \sum_{t=1}^{T} kLt^{3-\alpha} \overset{(\alpha \geq 5)}{\leq} 2kL.
$$

We then observe that, given an arbitrary concave function g, we have

$$
\sum_{C_m \in \mathcal{C}_\mathbf{G}} g(|C_m| N^\pi_{C_m,T}) \leq \sum_{C_m \in \mathcal{C}_\mathbf{G}} g\left(\frac{|C_m|}{k} T\right).
$$

This can be applied to component (b) with $g(\cdot) = \sqrt{\cdot}$ and to component (c) with $g(\cdot) = (\cdot)^{\frac{2}{3}}$. The statement of the theorem can be obtained by summing up all the components, i.e.,

$$
\mathbb{E}[R_T(\pi)] \leq \mathbb{E}\left[(\mathrm{b}_\xi) + (\mathrm{b}_{\bar{\xi}}) + (\mathrm{c})\right].
$$

■

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2312–2320, 2011.

Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, pages 23–35. PMLR, 2015.

Sébastien Bubeck, Gilles Stoltz, Csaba Szepesvári, and Rémi Munos. Online optimization in x-armed bandits. *Advances in Neural Information Processing Systems (NeurIPS)*, 21, 2008.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 844–853. PMLR, 2017.

Ofer Dekel, Ambuj Tewari, and Raman Arora. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the International Conference on Machine Learning (ICML)*. Omnipress, 2012.

J.L. Doob. *Stochastic Processes*. Probability and Statistics Series. Wiley, 1953.

Gianmarco Genalti, Marco Mussi, Nicola Gatti, Marcello Restelli, Matteo Castiglioni, and Alberto Maria Metelli. Graph-triggered rising bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 15351–15380. PMLR, 2024.

Yonatan Gur, Assaf Zeevi, and Omar Besbes. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 199–207, 2014.

Hoda Heidari, Michael J Kearns, and Aaron Roth. Tight policy regret bounds for improving and decaying bandits. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1562–1570, 2016.

Christine Herlihy and John P. Dickerson. Networked restless bandits with positive externalities. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 11997–12004. AAAI Press, 2023.

Prakirt Raj Jhunjhunwala, Sharayu Moharir, D Manjunath, and Aditya Gopalan. On a class of restless multi-armed bandits with deterministic policies. In *International Conference on Signal Processing and Communications (SPCOM)*, pages 487–491. IEEE, 2018.

Su Jia, Qian Xie, Nathan Kallus, and Peter I Frazier. Smooth non-stationary bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 14930–14944. PMLR, 2023.

Richard M Karp. Reducibility among combinatorial problems. *Complexity of Computer Computations*, pages 85–103, 1972.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, pages 681–690. ACM, 2008.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Nir Levine, Koby Crammer, and Shie Mannor. Rotting bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3074–3083, 2017.

Anne Gael Manegueu, Alexandra Carpentier, and Yi Yu. Generalized non-stationary bandits. *arXiv preprint arXiv:2102.00725*, 2021.

Alberto Maria Metelli, Francesco Trovo, Matteo Pirola, and Marcello Restelli. Stochastic rising bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 15421–15457. PMLR, 2022.

Marco Mussi, Alessandro Montenegro, Francesco Trovò, Marcello Restelli, and Alberto Maria Metelli. Best arm identification for stochastic rising bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2024.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvári, and Steffen Grünewälder. Bandits with delayed, aggregated anonymous feedback. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 4102–4110. PMLR, 2018.

Vishnu Raj and Sheetal Kalyani. Taming non-stationary bandits: A bayesian approach. *arXiv preprint arXiv:1707.09727*, 2017.

Julien Seznec, Andrea Locatelli, Alexandra Carpentier, Alessandro Lazaric, and Michal Valko. Rotting bandits are no harder than stochastic ones. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, 2019.

Julien Seznec, Pierre Ménard, Alessandro Lazaric, and Michal Valko. A single algorithm for both restless and rested rotting bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pages 3784–3794. PMLR, 2020.

Ron Shamir, Roded Sharan, and Dekel Tsur. Cluster graph modification problems. *Discrete Applied Mathematics*, 144(1-2):173–182, 2004.

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

Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25(A):287–298, 1988.