Graph-Triggered Rising Bandits

Gianmarco Genalti[1] Marco Mussi[1]
Nicola Gatti[1] Marcello Restelli[1] Matteo Castiglioni[1] Alberto Maria Metelli[1]

Abstract

In this paper, we propose a novel generalization of rested and restless bandits where 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 (degenerate) graphs. Still, the model can represent way more general and interesting scenarios. We first tackle the problem of computing the optimal policy when no specific structure is assumed on the graph, showing that it is NP-hard. Then, we focus on a specific structure, forcing the graph to be composed of a set of fully connected sub-graphs (i.e., cliques), and we prove that the optimal policy can be easily computed in closed form. Subsequently, we move to the learning problem presenting regret minimization algorithms for deterministic and stochastic cases. Our regret bounds highlight the complexity of the learning problem by incorporating instance-dependent terms that encode specific properties of the underlying graph structure. Moreover, we illustrate how the knowledge of the underlying graph is not necessary for achieving the no-regret property.

1. Introduction

In the basic stochastic Multi-Armed Bandit (MAB, Lattimore & Szepesvári, 2020) problem, at each round, the agent is asked to choose an action (a.k.a. arm) among a finite action set observing a reward drawn from an unknown probability distribution. MABs are particularly appealing machine learning models as they provide important theoretical guarantees on the convergence to the optimal solution and the lower and upper bounds to the convergence rate (usually referred to as *regret bounds*). However, the standard MAB model is too naïve for many real-world applications, and understanding which kind of additional problem structure allows recovering good theoretical properties is a central scientific challenge. Examples include *non-stationary* (Gur et al., 2014), *delayed* (Pike-Burke et al., 2018), *linear* (Abbasi-Yadkori et al., 2011), *contextual* (Chu et al., 2011) and *continuous-action spaces* (Kleinberg et al., 2008) bandits.

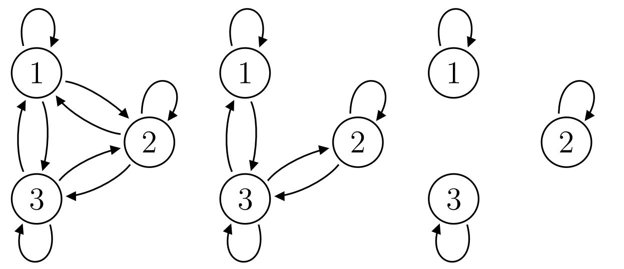
We focus on rather recent MAB structures, called *restless* and *rested* bandits (Tekin & Liu, 2012). In the former, the expected rewards evolve following the time (i.e., as an effect of *nature*); in the latter, the expected reward of an arm evolves as a function of the pulls we perform on that specific arm. In particular, we study a specific shape of the expected reward evolution, namely *rising* (Heidari et al., 2016). In a rising bandit, expected rewards increase according to *monotonic* and *concave* functions. Restless and rested rising bandits can capture many settings of practical interest. Consider, for instance, the scenario in which we have to choose which product to advertise (i.e., our arms), and the reward is the number of sales for such a product. The product we advertise will increase its sales and favor the sales of complementary products. This scenario corresponds to a rested problem in which some elements present a restless behavior.

In this paper, we propose a generalization of restless and rested bandits. We define a novel space of MABs called Graph-Triggered Rising Bandits (GTRBs). A GTRB is represented by a bandit 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 1b shows an example of this scenario. Interestingly, rested and restless bandits are two vertices in the space of GTRB. In particular, restless bandits correspond to the case of a *fully-connected* graph (Figure 1a), while rested ones correspond to the graph with the *self-loops only* (Figure 1c).

Contributions. In this paper, we present Graph-Triggered Rising Bandits (GTRBs), a setting aiming at generalizing

[1] Politecnico di Milano, Milan, Italy.
Correspondence to: G. Genalti <gianmarco.genalti@polimi.it>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).



(a) Restless Setting. (b) This work. (c) Rested Setting.

Figure 1. Examples of 3-armed GTRBs with graph representation.

the rested and restless bandit settings by introducing a graph representing the interaction between the arms. We focus on the case of rising bandits as they represent an interesting case study that will allow us to obtain no-regret algorithms. More in detail, the contributions are as follows.

- In Section 2, we formally present the fundamental notions on the rested and restless bandits and the assumptions related to the rising scenario. Then, we introduce the novel setting of GTRBs and discuss the relevant quantities characterizing an instance, including a representation of the graph based on the adjacency matrix. Finally, we present the learning problem and the performance index to evaluate the algorithms.
- In Section 3, we discuss the notion of optimal policy. We start with a negative result, proving that computing the optimal policy for a *generic graph* is NP-hard (Theorem 1). Then, we characterize the optimal policy for *block-diagonal* adjacency matrices, which can be computed in polynomial time (Theorem 2).
- In Section 4, we discuss the deterministic scenario and we propose two algorithms, the first (`DR-BG-UB`) for block-diagonal matrices and the second (`DR-G-UB`) for general graphs. We analyze their regret guarantees, highlighting the dependence on the graph structure.
- In Section 5, we analyze the seminal algorithm `R-□-UCB` (Metelli et al., 2022), designed for rested and restless stochastic rising bandits that does not require the knowledge of the graph. We analyze its regret guarantees, focusing on the dependence on the characteristics of the underlying graph.

The relevant literature is discussed in Appendix A. The statements' proofs are provided in Appendix B. An extensive numerical validation of the proposed algorithms is provided in Appendix D.

2. Problem Formulation

In this section, we first introduce notions related to Stochastic Rising Bandits (Section 2.1). Then, we present the Graph-Triggered Rising Bandits setting (Section 2.2). Finally, we formalize the learning problem (Section 2.3).

2.1. Notions on Rested and Restless Rising Bandits

Before introducing the setting, we present the fundamental notions concerning stochastic rising rested and restless bandits, and the related assumptions.

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 $k \in \mathbb{N}$, where $[k] := \{1, 2, \dots, k\}$. 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 $N_{i,t} := \sum_{\tau \in [t]} \mathbb{1}\{I_t = i\}$ as the number of pulls of the arm $i \in [k]$ until round t. In this work, we consider two specific types of MAB, namely *restless* and *rested* bandits (Tekin & 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, we consider the case in which the expected reward of a generic arm i evolves according to the number of pulls of such an arm, i.e., $n \leftarrow N_{i,t}$. On the other hand, in the restless case, the expected reward of a generic arm i evolves according to the current time t, i.e., $n \leftarrow t$. In other words, in rested bandits, the reward distribution of an arm evolves only when it is pulled, while in restless bandits, it evolves at each round. 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*[1] for every arm i and $n \in \mathbb{N}$, with their subgaussianity constants uniformly upper bounded by σ^2.

Among the various types of restless and rested bandits available in the literature, we focus on *Stochastic Rising Bandits* (Metelli et al., 2022). They are a specific class of bandits in which the expected reward of each arm evolves in a non-decreasing and concave manner. The following assumption formalizes such a behavior.

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

$$\textit{Non-decreasing:} \quad \gamma_i(n) \geq 0, \tag{1}$$

$$\textit{Concave:} \quad \gamma_i(n-1) \geq \gamma_i(n). \tag{2}$$

The two parts of this assumption allow us to provide theoretical guarantees in both the restless and rested settings.

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

Such guarantees cannot be provided without the concavity assumption (see Theorem 4.2 of Metelli et al., 2022).

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:

$$\Upsilon_{\boldsymbol{\nu}}(M, q) := \sum_{t \in [M-1]} \max_{i \in [k]} \gamma_i(t)^q, \tag{3}$$

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

2.2. Graph-Triggered Rising Bandits

In the Stochastic Rising Bandits, either in the rested or restless fashions, there exists no structure among different actions. In this work, we generalize the rested and restless bandits 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.[2] 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 adjacency matrix $\mathbf{G} \in \{0, 1\}^{k \times k}$. If the matrix contains the value 1 in position (i, j), this implies that a 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:

$$\widetilde{N}_{i,t} = \sum_{\tau \in [t]} \mathbb{1}\{\mathbf{G}_{I_\tau, i} = 1\} = \mathbf{e}_i^\top \mathbf{G}^\top \mathbf{N}_t, \tag{4}$$

where $\mathbf{e}_i$ is a vector belonging to the standard basis of $\mathbb{R}^k$ whose all components are all zero except for the i-th and $\mathbf{N}_t := (N_{1,t}, \ldots, N_{k,t})^\top$ is the vector containing the number of pulls of each arm up to round t. In GTRBs, rewards are sampled from probability distributions whose average rewards vary with the number of triggers, i.e., $n \leftarrow \widetilde{N}_{i,t}$ and, consequently, the expected reward of an arm i evolves as $\mu_i(\widetilde{N}_{i,t})$. Furthermore, we define $t_{i,n} := \sum_{l \in [T]} \mathbb{1}\{N_{i,l} \leq n\}$ as the round in which arm i has been pulled for the n-th time. With $\boldsymbol{t}_{i,t} := (t_{i,n})_{n \leq N_{i,t}}$ we refer to the vector containing all the rounds in which the arm i has been pulled, up to time t. Moreover, we introduce $t^I_{i,n} := \widetilde{N}_{i,t_{i,n}}$, 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. Finally, we introduce the concept of *degree*. In a graph, the degree of a node is the number of edges incident to the node. Formally, given an arm $i \in [k]$, we define:

$$\deg(i) := \mathbf{1}_k^\top \mathbf{G} \mathbf{e}_i.$$

Given the adjacency matrix of a graph $\mathbf{G}$, we define $\bar{k}_1 := |\{i \in [k] : \deg(i) = 1\}|$ as the number of arms having degree of 1. We now observe the relationship between rested and restless bandits and our setting.

Remark 1 (Inclusion of Rested and Restless bandits in GTRBs)**.** *The GTRB setting includes both* rested *and* restless *bandits (Tekin & 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.*[3] *Indeed, a* restless *bandit can be seen as a particular instance of GTRB 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 $\widetilde{N}_{i,t} = N_{i,t}$.*[4]

Block-Diagonal Adjacency Matrix. We now discuss a particular case of GTRB 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 generic binary symmetric matrix. However, we now focus on the specific case in which $\mathbf{G}$ is a *block-diagonal* 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 matrices with exactly $\widetilde{k} \in [k]$ distinct blocks of 1s. We call the *GTRB with block connectivity* the set of instances where Assumption 1 holds and the adjacency matrix $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$ for some $\widetilde{k} \leq k$. Moreover, 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 particular, when $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$, the graph associated to the adjacency matrix is only composed by completely connected components, or *cliques*. Thus, $\mathcal{C}_{\mathbf{G}}$ represents the set of cliques. Moreover, we indicate with $\widetilde{N}_{C_m,t} := \sum_{i \in C_m} N_{i,t}$ 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] All the results we present also hold for directed graphs.

[3] 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.

[4] This can be easily seen by looking at Equation (4) considering $\mathbf{G} = \mathbf{I}_k$ and observing that the vector $\mathbf{e}_i$ selects the i-th element of vector $\mathbf{N}_t$.

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 GTRB, the performance of a policy π is measured by the means of *expected cumulative reward* throughout T rounds, formally:

$$J_{\boldsymbol{\nu},\mathbf{G},T}(\pi) \coloneqq \mathbb{E}\left[\sum_{t\in[T]} \mu_{I_t}(\widetilde{N}_{I_t,t})\right], \tag{5}$$

where the expectation is taken over the randomness of both the environment and the policy/algorithm. A policy is *optimal* for instance $\boldsymbol{\nu}$, an adjacently 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:

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\pi) = J^*_{\boldsymbol{\nu},\mathbf{G},T} - J_{\boldsymbol{\nu},\mathbf{G},T}(\pi). \tag{6}$$

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 GTRB 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* GTRBs*, we consider a notion of* policy *regret (Dekel et al., 2012). 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. Optimality in GTRB

In this section, we discuss the notion of *optimality*, in our learning problem. We first characterize the complexity of finding the optimal policy followed by the clairvoyant.

Theorem 1 (Complexity of finding the Optimal Policy in GTRBs)**.** *Computing the optimal policy in* GTRBs *with arbitrary 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 "represent" vertexes in a clique. Theorem 1 implies that the class of problems of GTRBs 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.

3.1. Optimality in GTRB with Block Diagonal Connectivity Matrix

We now show how, for this special case of GTRBs with block-diagonal connectivity matrices, the optimal policy can be efficiently computed in a closed-form fashion.

Theorem 2 (Optimal Policy in Rising GTRB with Block Diagonal Connectivity)**.** *For any instance* $(\boldsymbol{\nu}, \mathbf{G}, T)$ *s.t.* $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$*, under Assumption 1, the optimal policy* $\pi^*_{\boldsymbol{\nu},\mathbf{G},T} \in \arg\max_{\pi} J_{\boldsymbol{\nu},\mathbf{G},T}(\pi)$ *is given by:*

$$\pi^*_{\boldsymbol{\nu},\mathbf{G},T}(t) \in \arg\max_{j\in C^*_{\boldsymbol{\nu},\mathbf{G},T}} \mu_j(t), \qquad \forall t \in [T], \tag{7}$$

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 of 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). From now on, for the sake of simplicity in the notation, we will omit explicit references to T.

4. Regret Minimization in Deterministic Settings

In this section, we propose a novel algorithm to solve the *deterministic* GTRB, i.e., all instances of GTRB where $\sigma = 0$. 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

Algorithm 1: DR-BG-UB.

```
  Input : Adjacency matrix $\mathbf{G}$
1 Initialize $N_{i,0} \leftarrow 0, \ \forall i \in [k]$
2 for $t \in [T]$ do
3     Compute $\bar{\mu}_i(t)$ as in Equation (8)
4     Select $I_t \in \arg\max_{i \in [k]} \bar{\mu}_i(t)$
5     Play $I_t$ and observe $\mu_{I_t}(\widetilde{N}_{I_t,t})$
6     $N_{I_t,t} \leftarrow N_{I_t,t-1} + 1$
7     $N_{i,t} \leftarrow N_{i,t-1}, \ \ \forall i \in [k]$
8 end
```

current time t by estimating the first derivative using the last two observations:

$$\begin{aligned}\bar{\mu}_i(t) &:= \mu(t^I_{i,N_{i,t-1}}) + \\ &+ (t - t^I_{i,N_{i,t-1}}) \frac{\mu(t^I_{i,N_{i,t-1}}) - \mu(t^I_{i,N_{i,t-1}-1})}{t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-1}}. \qquad (8)\end{aligned}$$

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

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

As we will see in Section 4.1, this decomposition assumes a particular meaning in instances where $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$, where $(\mathtt{A})$ 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 $(\mathtt{B})$ 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.

4.1. Algorithm for Deterministic GTRBs with Block-Diagonal Matrices

In this part, we introduce DR-BG-UB, an optimistic anytime regret minimization algorithm for the deterministic GTRB setting with block-diagonal matrices, whose pseudocode is provided in Algorithm 1. The algorithm takes as input the matrix $\mathbf{G}$ and employs the estimator presented in Equation (8). 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 3) the $\bar{\mu}_i(t)$ for every $i \in [k]$ as in Equation (8) and plays greedy according to it (line 5).[5]

Regret Analysis. We recall that block-diagonal matrices represent a special case of graph structure for GTRB where the optimal policy can be characterized in a closed form (Theorem 2). The following result provides the regret bound of DR-BG-UB highlighting the impact of the graph topology.

Theorem 3 (Regret Upper Bound for DR-BG-UB in Block-Diagonal Matrices in Deterministic Settings)**.** *Let* $(\boldsymbol{\nu}, \mathbf{G})$ *be an instance of the* GTRB *problem, where* $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$ *and* $\sigma = 0$. *Then, Algorithm 1 suffers a regret bounded by:*

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

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 $(\mathtt{A})$ 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 $(\mathtt{B})$ 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. More details can be found in Remark 5 (Appendix B).

In Theorem 3, 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

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

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.

4.2. Algorithm for Deterministic GTRBs for General Matrices

After having studied the scenario of block-diagonal 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*.

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

$$\mathbf{G}_{i,j} - \mathbf{G}^L_{i,j} \geq 0, \quad \forall i,j \in [k]. \tag{9}$$

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

$$\bar{\mathbf{G}}^L \in \underset{\mathbf{G}^L \quad \text{satisfying Eq. (9)}}{\arg\min} |\mathcal{C}_{\mathbf{G}^L}| \,.$$

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 whose definition recalls the one of Equation (8):

$$\begin{aligned} \bar{\mu}^L_i(t) &:= \mu(t^{I,L}_{i,N_{i,t-1}}) + \\ &+ (t - t^{I,L}_{i,N_{i,t-1}}) \frac{\mu(t^{I,L}_{i,N_{i,t-1}}) - \mu(t^{I,L}_{i,N_{i,t-1}-1})}{t^{I,L}_{i,N_{i,t-1}} - t^{I,L}_{i,N_{i,t-1}-1}}, \end{aligned} \tag{10}$$

where $t^{I,L}_{i,l} := \mathbf{e}_i^\top (\bar{\mathbf{G}}^L)^\top \mathbf{N}_{t_{i,l}}$ 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 an arbitrary $\mathbf{G}$. We introduce 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 (10). In other words, DR-G-UB pretends to be interacting with a bandit with a graph defined by $\bar{\mathbf{G}}^L$.

Regret Analysis. We now provide a regret upper bound for DR-G-UB.

Theorem 4 (Regret Upper Bound for DR-G-UB for General Matrices in Deterministic Settings)**.** *Let* $(\boldsymbol{\nu}, \mathbf{G})$ *be an instance of the* GTRB *problem, where* $\mathbf{G} \in \{0,1\}^{k \times k}$ *and* $\sigma = 0$*. Then,* DR-G-UB *suffers a regret bounded as:*

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

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

Algorithm 2: DR-G-UB.

```
Input :Connectivity matrix $\mathbf{G}$
Initialize $N_{i,0} \leftarrow 0, \ \forall i \in [k]$
Compute maximal sub-matrix $\bar{\mathbf{G}}^L$ from $\mathbf{G}$
for $t \in [T]$ do
    Compute $\bar{\mu}^L_i(t)$ as in Equation (10)
    Select $I_t \in \arg\max_{i \in [k]} \bar{\mu}^L_i(t)$
    Play $I_t$ and observe $\mu^L_{I_t}(\widetilde{N}_{I_t,t})$
    $N_{I_t,t} \leftarrow N_{I_t,t-1} + 1$
    $N_{i,t} \leftarrow N_{i,t-1}, \ \forall i \in [k]$
end
```

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 achieved sublinear regret w.r.t. the optimal policy with a polynomial-time algorithm. This has been made possible by 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.*

5. Regret Minimization in Stochastic Setting

In this section, we focus on the stochastic GTRBs scenario. We characterize the performances of R-□-UCB (Metelli et al., 2022) in the GTRBs setting. 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.

As anticipated in Section 1, we aim at obtaining a computationally efficient algorithm enjoying *sub-linear regret* guarantees. Surprisingly, our analysis shows that R-□-UCB not only enjoys sub-linear regret for any 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

Algorithm 3: R-□-UCB.

```
Input : Sub-gaussianity proxy upper bound $\sigma$,
        confidence levels $\{\delta_t\}_{t\in[T]}$, window size
        parameter $\epsilon \in (0, 1/2)$.
Initialize $N_{i,0} \leftarrow 0,\ \forall i \in [k]$
for $t \in [T]$ do
    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 $X_{I_t,t}$
    $N_{I_t,t} \leftarrow N_{I_t,t-1} + 1$
end
```

arm i, propagates its reward function to the current round t by estimating the first derivative over the last $2h$ samples:

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

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 GTRBs (more details are provided in Appendix C).

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 $1 \le h \le \left\lfloor \frac{N_{i,t-1}}{2} \right\rfloor$, let:*

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

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:

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

The algorithm, whose pseudocode is reported in Algorithm 3, takes as input the subgaussianity constants upper bound σ, sliding window size parameter ϵ, and a sequence of 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, Algorithm 3 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).*

5.1. Regret Analysis for Block-Diagonal Matrices

We analyze its performance in the block-diagonal case before bounding the regret of Algorithm 3 for general matrices.

Theorem 6 (Regret Upper Bound for R-□-UCB in Block-Diagonal Matrices)**.** *Let $(\boldsymbol{\nu}, \mathbf{G})$ be an instance of the* GTRB *problem, where $\mathbf{G} \in \mathbf{B}_{\widetilde{k}}$. 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, Algorithm 3 suffers an expected regret bounded as:*

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

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

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 $\widetilde{\mathcal{O}}(T^{2/3})$. 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 contribution is ϵ. Indeed, ϵ regulates such a trade-off between bias and variance, and this effect can be observed in the complete form 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 contribution depending on graph topology is the one coming from bias (terms (B) and (C)). Indeed, the environment's randomness contribution has been decoupled from estimation bias to get a fully 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 $\frac{1}{1+q}$, 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 could reduce the dependence up to $T^{\frac{q}{1+q}}$, matching the deterministic setting bound.

Comparison with Known Results from Literature. Given that rested and restless rising bandits are special instances of GTRBs, 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 would thus assume the following form:

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

The only other existing result for the rested rising bandit 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 presented in Theorem 6 becomes:

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

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 bandit 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.

5.2. Regret Analysis for General Matrices

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

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

$$\mathbf{G}_{i,j} - \mathbf{G}^U_{i,j} \leq 0, \quad \forall i,j \in [k]. \tag{11}$$

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

$$\bar{\mathbf{G}}^U \in \mathop{\arg\max}_{\mathbf{G}^U \text{ satisfying Eq. (11)}} |\mathcal{C}_{\mathbf{G}^U}| \,.$$

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 case and general matrices.

Theorem 7 (Regret Upper Bound for R-□-UCB in General Matrices)**.** *Let* $(\boldsymbol{\nu}, \mathbf{G})$ *be an instance of the* GTRB *problem, 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, Algorithm 3 suffers an expected regret bounded as:*

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

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

We remark that the result has been obtained by bounding $\widetilde{N}^U_{C^U_m,T} \leq T$ 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, as discussed for the optimal policy (Theorem 1). 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 matrix 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 plays the same policy, regardless of the graph. 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, we point out to 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 graphical 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 scenarios (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 graphical 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 graphical 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 analysis of our algorithms.

6. Discussion and Conclusions

In this paper, we proposed the *graph-triggered rising bandits* (GTRBs), 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 on the stochastic rising bandits, a peculiar type of bandits where the expected rewards are non-decreasing and concave w.r.t. the number of triggers. As a first result, we showed that, in this setting, computing the optimal policy without additional assumptions on the graph structure is NP-Hard. Then, we characterized the optimal policy for the special class of block-diagonal adjacency matrices, and we showed that the optimal policy can be computed in closed form. This special family of instances allowed us to express a strong inter-dependence between the graph topology and the regret bounds and it plays a crucial role also in bounding the regret for the general case. In particular, we presented and studied the performance of two new algorithms, `DR-BG-UB` and `DR-G-UB`, to handle the deterministic scenarios in which the adjacency matrix is block-diagonal and for the general case, respectively. Finally, we studied `R-□-UCB` (Metelli et al., 2022), an algorithm from the SRB literature, and we showed that it can provide regret guarantees also in this more general setting. This work aspires to be a first step in the study of graph-triggered bandits. We started from a special set of instances, namely rising bandits, with the goal of extending this framework to other classes of bandits, e.g., rotting bandits.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgments

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).

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2312–2320, 2011.

Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. Online learning with feedback graphs: Beyond bandits. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, pp. 23–35. PMLR, 2015.

Bubeck, S., Stoltz, G., Szepesvári, C., and Munos, R. Online optimization in x-armed bandits. *Advances in Neural Information Processing Systems (NeurIPS)*, 21, 2008.

Chu, W., Li, L., Reyzin, L., and Schapire, R. E. Contextual bandits with linear payoff functions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15 of *JMLR Proceedings*, pp. 208–214. JMLR, 2011.

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

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

Garivier, A. and Moulines, E. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory (ALT)*, pp. 174–188. Springer, 2011.

Gur, Y., Zeevi, A., and Besbes, O. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 199–207, 2014.

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

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

Jhunjhunwala, P. R., Moharir, S., Manjunath, D., and Gopalan, A. On a class of restless multi-armed bandits with deterministic policies. In *International Conference on Signal Processing and Communications (SPCOM)*, pp. 487–491. IEEE, 2018.

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

Kleinberg, R., Slivkins, A., and Upfal, E. Multi-armed bandits in metric spaces. In *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, pp. 681–690. ACM, 2008.

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

Levine, N., Crammer, K., and Mannor, S. Rotting bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3074–3083, 2017.

Metelli, A. M., Trovo, F., Pirola, M., and Restelli, M. Stochastic rising bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 15421–15457. PMLR, 2022.

Mussi, M., Montenegro, A., Trovò, F., Restelli, M., and Metelli, A. M. Best arm identification for stochastic rising bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2024.

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

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

Seznec, J., Locatelli, A., Carpentier, A., Lazaric, A., and Valko, M. 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*, pp. 2564–2572. PMLR, 2019.

Seznec, J., Ménard, P., Lazaric, A., and Valko, M. 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*, pp. 3784–3794. PMLR, 2020.

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

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

A. Related Works

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

Graphical Relationships in Bandits. The graph-triggered restless bandit 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 & 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 graphical 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 GTRB the feedback is purely bandit.

Rested and Restless Bandits. Restless and rested bandits are a well-established research field. Starting from the seminal paper of Whittle (1988) on restless bandits, several approaches have been proposed over the years to deal with non-stationary bandits (Tekin & Liu, 2012; Raj & 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. In particular, rotting bandits are a family of restless bandits where the reward is assumed to decrease (contrary to rising bandits). 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.

B. Omitted Proofs

Theorem 1 (Complexity of finding the Optimal Policy in GTRBs)**.** *Computing the optimal policy in* GTRBs *with arbitrary 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 $(v, v') \in E$. Finally, for each arm $a_{v,t} \in \mathcal{A}$, the reward is deterministic and evolves as $\widetilde{\mu}_{a_{v,t}}(n) = \min\{1 + \eta t, \frac{n+1}{t}(1 + \eta t)\}$, where $\eta = T^{-2}$. We call $\widetilde{\boldsymbol{\nu}}$ the set of this functions. It is easy to see that the GTRB 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 $\widetilde{N}_{a_{v_t,t},t} = t - 1$ for every $t \in [T]$. Hence, the reward of the policy $\widetilde{\pi}$ at time t is

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

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 show that for each $t', t \in [T]$ it holds that

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

Indeed, for each $t' < t$, it holds

$$\min\{1+\eta t', \frac{t}{t'}(1+\eta t')\} \leq 1+\eta t' < 1+\eta t.$$

On there other hand, for each $t' > t$ it holds

$$\min\{1+\eta t', \frac{t}{t'}(1+\eta t')\} \leq \frac{t}{t'}(1+\eta t') \leq \frac{t}{t+1}(1+\frac{1}{T}) \leq 1,$$

where in the second inequality we use $\eta = T^{-2}$. Putting together, the previous inequalities imply that for each $t' \neq t$ we have

$$\min\{1+\eta t', \frac{t}{t'}(1+\eta t')\} < 1+\eta t. \tag{12}$$

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

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

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

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

by Equation (12), and ii) $\widetilde{N}_{a_{v,t'},t} = t-1$, 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 $\{v_t\}_{t=1}$ is a clique of size T. □

Theorem 2 (Optimal Policy in Rising GTRB with Block Diagonal Connectivity)**.** *For any instance $(\boldsymbol{\nu}, \mathbf{G}, T)$ s.t. $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$, under Assumption 1, the optimal policy $\pi^*_{\boldsymbol{\nu},\mathbf{G},T} \in \arg\max_\pi J_{\boldsymbol{\nu},\mathbf{G},T}(\pi)$ is given by:*

$$\pi^*_{\boldsymbol{\nu},\mathbf{G},T}(t) \in \arg\max_{j\in C^*_{\boldsymbol{\nu},\mathbf{G},T}} \mu_j(t), \qquad \forall t \in [T], \tag{7}$$

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. Since arms in different cliques are not connected, this corresponds to a rested bandit problem, and we use Proposition 1 from 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 8 (`DR-BG-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:

$$\begin{aligned}
\mu_i(\widetilde{N}_{i,t}) &= \mu_i(t^I_{i,N_{i,t-1}}) + \sum_{j=t^I_{i,N_{i,t-1}}}^{\widetilde{N}_{i,t}-1} \gamma_i(j) \\
&\leq \mu_i(t^I_{i,N_{i,t-1}}) + (\widetilde{N}_{i,t} - t^I_{i,N_{i,t-1}}) \gamma_i(t^I_{i,N_{i,t-1}-1}) \qquad (13) \\
&\leq \mu_i(t^I_{i,N_{i,t-1}}) + (t - t^I_{i,N_{i,t-1}}) \gamma_i(t^I_{i,N_{i,t-1}-1}) \qquad (14)
\end{aligned}$$

where line (13) follows from Assumption 1, and line (14) is obtained from observing that $\widetilde{N}_{i,t} \leq t$. Concerning the bias, when $N_{i,t-1} \geq 2$, we have:

$$\begin{aligned}
\bar{\mu}_i(t) - \mu_i(\widetilde{N}_{i,t}) &\leq \mu_i(t^I_{i,N_{i,t-1}}) - \mu_i(\widetilde{N}_{i,t}) + (t - t^I_{i,N_{i,t-1}}) \frac{\mu_i(t^I_{i,N_{i,t-1}}) - \mu_i(t^I_{i,N_{i,t-1}-1})}{t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-1}} \qquad (15) \\
&\leq (t - t^I_{i,N_{i,t-1}}) \frac{\mu_i(t^I_{i,N_{i,t-1}}) - \mu_i(t^I_{i,N_{i,t-1}-1})}{t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-1}} \qquad (16) \\
&\leq (t - t^I_{i,N_{i,t-1}}) \gamma_i(t^I_{i,N_{i,t-1}-1}), \qquad (17)
\end{aligned}$$

where line (16) follows from observing that $\mu_i(t^I_{i,N_{i,t-1}}) \leq \mu_i(\widetilde{N}_{i,t})$, and line (17) 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 (Regret Upper Bound for `DR-BG-UB` in Block-Diagonal Matrices in Deterministic Settings)**.** *Let $(\boldsymbol{\nu}, \mathbf{G})$ be an instance of the* GTRB *problem, where $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$ and $\sigma = 0$. Then, Algorithm 1 suffers a regret bounded by:*

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

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

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

where $i_t^* \in \arg\max_{i \in C^*_{\boldsymbol{\nu},\mathbf{G}}} \mu_i(t)$ for all $t \in [T]$. Then:

$$\begin{aligned}
R^{\texttt{DR-BG-UB}}_{\boldsymbol{\nu},\mathbf{G}} &= \sum_{t=1}^{T} \mu_{i_t^*}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t}) \\
&= \sum_{t=1}^{T} \mu_{i_t^*}(t) \pm \bar{\mu}_{I_t}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t}) \\
&\leq \sum_{t=1}^{T} \min\{1, \bar{\mu}_{I_t}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t})\} && (18) \\
&\leq \sum_{t=1}^{T} \min\{1, (t - t^I_{I_t,N_{I_t,t-1}})\gamma_{I_t}(t^I_{I_t,N_{I_t,t-1}-1})\} && (19) \\
&= \sum_{t=1}^{T} \min\{1, (t \pm t^I_{I_t,N_{I_t,t}} - t^I_{I_t,N_{I_t,t-1}})\gamma_{I_t}(t^I_{I_t,N_{I_t,t-1}-1})\} \\
&\leq \sum_{t=1}^{T} \min\{1, (t - t^I_{I_t,N_{I_t,t}})\gamma_{I_t}(t^I_{I_t,N_{I_t,t-1}-1})\} + \sum_{t=1} \min\{1, (t^I_{I_t,N_{I_t,t}} - t^I_{I_t,N_{I_t,t-1}})\gamma_{I_t}(t^I_{I_t,N_{I_t,t-1}-1})\} && (20) \\
&= 4k + \underbrace{\sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \sum_{i \in C_m} \sum_{j=3}^{N_{j,T}} \min\{1, (t - t^I_{i,j})\gamma(t^I_{i,j-2})\}}_{\text{(a)}} + \underbrace{\sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \sum_{i \in C_m} \sum_{j=3}^{N_{j,T}} \min\{1, (t^I_{i,j} - t^I_{i,j-1})\gamma(t^I_{i,j-2})\}}_{\text{(b)}}, \\
& && (21)
\end{aligned}$$

where lines (18) and (19) follow from Lemma 8, line (20) from the fact that $\min\{1, x+y\} \leq \min\{1, x\} + \min\{1, y\}$ for any $x, y \geq 0$, line (21) from a decomposition over the cliques, the arms in the cliques and the number of pulls of every arm.

Let us bound the two terms separately, let $q \in [0, 1]$:

$$\begin{aligned}
\text{(a)} &\leq \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \sum_{i \in C_m} \sum_{j=3}^{N_{j,T}} \min\{1, T\gamma(t^I_{j,j-2})\} && (22) \\
&\leq \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \sum_{i \in C_m} \sum_{j=3}^{N_{j,T}} T^q \gamma(t^I_{i,j-2})^q && (23) \\
&\leq \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} T^q |C_m| \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C_m,T}}{|C_m|} \right\rceil, q \right), && (24)
\end{aligned}$$

where line (22) follows by bounding $t - t^I_{i,j} \leq T$ for every i and every j, line (23) from the inequality $\min\{1, x\} \leq \min\{1, x\}^q \leq x^q$ for $q \in [0, 1]$, line (24) from Lemma 13.

Then, let $y \in [0, \frac{1}{2}]$, and $q \coloneqq \frac{y}{1-y}$:

$$\begin{aligned}
\text{(b)} &\leq \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \sum_{i \in C_m} \sum_{j=3}^{N_{i,T}} (t^I_{i,j} - t^I_{i,j-1})^y \gamma(t^I_{i,j-2})^y && (25) \\
&\leq \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \sum_{i \in C_m} \left(\sum_{j=3}^{N_{i,T}} (t^I_{i,j} - t^I_{i,j-1}) \right)^y \left(\sum_{j=3}^{N_{i,T}} \gamma(t^I_{i,j-2})^{\frac{y}{1-y}} \right)^{1-y} && (26) \\
&\leq \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \sum_{i \in C_m} \widetilde{N}^y_{C_m,T} \left(\sum_{j=3}^{N_{i,T}} \gamma(j-2)^{\frac{y}{1-y}} \right)^{1-y} && (27)
\end{aligned}$$

$$\leq \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \widetilde{N}^{y}_{C_m,T} |C_m|^{y} \left(\sum_{i \in C_m} \sum_{j=3}^{N_{i,T}} \gamma(j-2)^{\frac{y}{1-y}} \right)^{1-y} \tag{28}$$

$$\leq \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \widetilde{N}^{y}_{C_m,T} |C_m| \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C_m,T}}{|C_m|} \right\rceil, \frac{y}{1-y} \right)^{1-y}, \tag{29}$$

where line (25) follows from the inequality $\min\{1, x\} \leq \min\{1, x\}^q \leq x^q$ for $q \in [0, 1]$, line (26) is obtained from Hölder's inequality with exponents $\frac{1}{y} \geq 1$ and $\frac{1}{1-y} \geq 1$ respectively, line (27) follows from bounding $\sum_{j=3}^{N_{i,T}} (t^I_{i,j} - t^I_{i,j-1}) \leq \widetilde{N}_{C_m,T}$ and by Assumption 1, line (28) is obtained by an application of Jensen's Inequality, and line (29) follows from Lemma 13.

By setting $q = \frac{y}{1-y} \in [0, 1]$, and putting all together:

$$R^{\texttt{DR-BG-UB}}_{\boldsymbol{\nu},\mathbf{G}} \leq 4k + \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} T^q |C_m| \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C_m,T}}{|C_m|} \right\rceil, q \right) + \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \widetilde{N}^{\frac{q}{1+q}}_{C_m,T} |C_m| \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C_m,T}}{|C_m|} \right\rceil, q \right)^{\frac{1}{1+q}}.$$

□

Remark 5 (Regret Bound in Rested and Restless Rising Bandits)**.** *When we are in a purely rested/restless scenario, the contribution term associated to the restless/rested scenario vanish, and we get the same regret orders from (Metelli et al., 2022). In particular, we can avoid to split the minimum in (20) 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 (Regret Upper Bound for `DR-G-UB` for General Matrices in Deterministic Settings)**.** *Let $(\boldsymbol{\nu}, \mathbf{G})$ be an instance of the* GTRB *problem, where $\mathbf{G} \in \{0,1\}^{k \times k}$ and $\sigma = 0$. Then,* `DR-G-UB` *suffers a regret bounded as:*

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

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:

$$f_i(t;\ x, y) = \mu_i(x) + (t - x) \frac{\mu_i(x) - \mu_i(y)}{x - y}, \tag{30}$$

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 previous claim. We start with the second argument: fix t and x, then for any y:

$$\begin{aligned} f_i(t; x, y) - f_i(t; x, y-1) &= (t - x) \left(\frac{\sum_{j=y}^{x-1} \gamma_i(j)}{x - y} - \frac{\sum_{j=y-1}^{x-1} \gamma_i(j)}{x - y + 1} \right) \\ &= \frac{\sum_{j=y}^{x-1} \gamma_i(j) - (x - y)\gamma_i(y - 1)}{(x - y)(x - y + 1)} \leq 0, \end{aligned} \tag{31}$$

where line (31) 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:

$$f_i(t;\ x, y) - f_i(t;\ x-1, y) = \tag{32}$$

$$= \gamma_i(x-1) + (t-x)\frac{\sum_{j=y}^{x-1}\gamma_i(j)}{x-y} - (t-x+1)\frac{\sum_{j=y}^{x-2}\gamma_i(j)}{x-1-y}$$

$$= \gamma_i(x-1) + \frac{(t-x)(x-1-y)\sum_{j=y}^{x-1}\gamma_i(j) - (x-y)(t-x+1)\sum_{j=y}^{x-2}}{(x-y)(x-1-y)}$$

$$= \gamma_i(x-1) + \frac{\sum_{j=y}^{x-2}\gamma_i(j)[(t-x)(x-1-y) - (x-y)(t-x+1)] + (t-x)(x-1-y)\gamma_i(x-1)}{(x-y)(x-1-y)} \tag{33}$$

$$= \gamma_i(x-1)\left(1 + \frac{t-x}{x-y}\right) + \gamma_i(x-2)\frac{x-y-1}{x-y-1}\frac{(t-x)(x-1-y) - (x-y)(t-x+1)}{x-y}$$

$$= \gamma_i(x-1)\frac{t-y}{x-y} - \gamma_i(x-2)\frac{t-y}{x-y}$$

$$\leq \frac{t-y}{x-y}(\gamma_i(x-1) - \gamma_i(x-2)) \leq 0, \tag{34}$$

where line (33) follows from observing that $(t-x)(x-1-y)-(x-y)(t-x+1) \leq 0$, line (34) follows from Assumption 1. We proved that the estimator is decreasing with internal times.

Now we observe that, for every i and every t, we have $t^I_{i,N_{i,t}} \geq t^{I,L}_{i,N_{i,t}}$. This is a trivial 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

$$f_i(t;\ t^I_{i,N_{i,t-1}}, t^I_{i,N_{i,t}-1}) \leq f_i(t;\ t^{I,L}_{i,N_{i,t-1}}, t^{I,L}_{i,N_{i,t}-1}), \tag{35}$$

and

$$\mu_i(t^I_{i,N_{i,t}}) \geq \mu_i(t^{I,L}_{i,N_{i,t}}). \tag{36}$$

Finally, given the optimal policy as a sequence $(i^*_t)_{t\in[T]}$, we bound the regret:

$$\begin{aligned}
R_{\boldsymbol{\nu},\mathbf{G}}(\texttt{DR-G-UB}) &= \sum_{t=1}^{T}\mu_{i^*_t}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t}) \\
&= \sum_{t=1}^{T}\mu_{i^*_t}(t) - \mu_{I_t}(t^I_{I_t,N_{I_t,t}}) \\
&= \sum_{t=1}^{T}\mu_{i^*_t}(t) \pm \bar{\mu}_{I_t}(t) - \mu_{I_t}(t^I_{I_t,N_{I_t,t}}) \\
&\leq \sum_{t=1}^{T}\bar{\mu}_{I_t}(t) - \mu_{I_t}(t^I_{I_t,N_{I_t,t}}) \\
&\leq \sum_{t=1}^{T}\bar{\mu}^L_{I_t}(t) - \mu_{I_t}(t^{I,L}_{I_t,N_{I_t,t}}),
\end{aligned} \tag{37}$$

where line (37) follows from (35) and (36). The proof can be concluded the same way as in Theorem 3. □

Lemma 9 (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:*

$$\widetilde{\mu}^h_i(t) := \frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}\left(\mu_i(t^I_{i,l}) + (t-l)\frac{\mu_i(t^I_{i,l}) - \mu_i(t^I_{i,l-h})}{h}\right),$$

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:

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

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

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

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

$$\begin{aligned}
\mu_i(\widetilde{N}_{i,t}) &= \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(t^I_{i,l}) + \sum_{j=t^I_{i,l}}^{\widetilde{N}_{i,t}-1} \gamma_i(j) \right) \\
&\leq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \Big(\mu_i(t^I_{i,l}) + (\widetilde{N}_{i,t} - t^I_{i,l}) \gamma_i(t^I_{i,l} - 1) \Big) && (38) \\
&\leq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(t^I_{i,l}) + \frac{\widetilde{N}_{i,t} - t^I_{i,l}}{t^I_{i,l} - t^I_{i,l-h}} \sum_{j=t^I_{i,l-h}}^{t^I_{i,l}-1} \gamma_i(j) \right) && (39) \\
&\leq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(t^I_{i,l}) + (t - l) \frac{\mu_i(t^I_{i,l}) - \mu_i(t^I_{i,l-h})}{h} \right) \eqqcolon \widetilde{\mu}_i^h(t), && (40)
\end{aligned}$$

where lines (38) and (39) follow from Assumption 1, and line (40) is obtained from observing that $t^I_{i,l} \geq l$, $\widetilde{N}_{i,t} \leq t$ and $t^I_{i,l} - t^I_{i,l-h} \geq h$.

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

$$\begin{aligned}
\widetilde{\mu}_i^h(t) - \mu_i(\widetilde{N}_{i,t}) &= \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(t^I_{i,l}) + (t - l) \frac{\mu_i(t^I_{i,l}) - \mu_i(t^I_{i,l-h})}{h} \right) - \mu_i(\widetilde{N}_{i,t}) \\
&\leq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} (t - l) \frac{\mu_i(t^I_{i,l}) - \mu_i(t^I_{i,l-h})}{h} && (41) \\
&= \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} (t - l) \frac{\mu_i(t^I_{i,l}) - \mu_i(t^I_{i,l-h})}{t^I_{i,l} - t^I_{i,l-h}} \frac{t^I_{i,l} - t^I_{i,l-h}}{h} \\
&\leq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} (t - l) \gamma_i(t^I_{i,l-h}) \frac{t^I_{i,l} - t^I_{i,l-h}}{h} && (42) \\
&\leq \frac{t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-2h+1}}{h^2} \gamma_i(t^I_{i,N_{i,t-1}-2h+1}) \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} (t - l) && (43) \\
&= \frac{(2t - 2N_{i,t-1} + h - 1)(t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-2h+1})}{2h} \gamma_i(t^I_{i,N_{i,t-1}-2h+1}), && (44)
\end{aligned}$$

where line (41) follows from observing that $\mu_i(t^I_{i,l}) \leq \mu_i(\widetilde{N}_{i,t})$, line (42) 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 (43) 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 (44) follows from computing the summation.

□

Lemma 10 (Bound on Estimator's Cumulative Bias for Block Diagonal 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 $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$ be a block diagonal matrix, then for every $q \in [0, 1]$, we have*

$$\begin{aligned} &\sum_{t=1}^{T} \min \left\{ 1, \widetilde{\mu}_{I_t}^{h_{I_t,t}}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t}) \right\} \leq \\ &\qquad \leq 2k + \bar{k}_1 T^q \left\lceil \frac{1}{1-2\epsilon} \right\rceil \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil (1-2\epsilon) \frac{T}{\bar{k}_1} \right\rceil, q \right) + \\ &\qquad + T^{\frac{2q}{1+q}} (1 + \log(\epsilon T))^{\frac{q}{1+q}} \left\lceil \frac{1}{\epsilon} \right\rceil \left\lceil \frac{1}{1-2\epsilon} \right\rceil \sum_{C_m \in \mathcal{C}_{\mathbf{G}} : |C_m| > 1} |C_m| \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil (1-2\epsilon) \frac{T}{|C_m|} \right\rceil, q \right)^{\frac{1}{1+q}}, \end{aligned} \tag{45}$$

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. We proceed decomposing over the cliques and then over the arms, splitting cliques with only one arm from the others:

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

We start from bounding the first term:

$$\text{(a)} \leq \sum_{\substack{C_m \in \mathcal{C}_{\mathbf{G}} : |C_m| = 1 \\ C_m = \{i\}}} \sum_{j=3}^{N_{i,T}} \min \left\{ 1, \frac{(2t_{i,j} - 2(j-1) + h_{i,t_{i,j}} - 1) 2h_{i,t}}{2h_{i,t}} \gamma_i(t^I_{i,(j-1)-2h_{i,t_{i,j}}+1}) \right\} \tag{46}$$

$$\leq \sum_{\substack{C_m \in \mathcal{C}_{\mathbf{G}} : |C_m| = 1 \\ C_m = \{i\}}} \sum_{j=3}^{N_{i,T}} \min \left\{ 1, T \gamma_i(t^I_{i,j-2\lfloor \epsilon(j-1) \rfloor}) \right\} \tag{47}$$

$$\leq \sum_{\substack{C_m \in \mathcal{C}_{\mathbf{G}} : |C_m| = 1 \\ C_m = \{i\}}} \sum_{j=3}^{N_{i,T}} \min \left\{ 1, T \gamma_i(\lfloor (1-2\epsilon) j \rfloor) \right\} \tag{48}$$

$$\leq T^q \sum_{\substack{C_m \in \mathcal{C}_{\mathbf{G}} : |C_m| = 1 \\ C_m = \{i\}}} \sum_{j=3}^{N_{i,T}} \gamma_i(\lfloor (1-2\epsilon) j \rfloor)^q \tag{49}$$

$$\leq T^q \left\lceil \frac{1}{1-2\epsilon} \right\rceil \sum_{\substack{C_m \in \mathcal{C}_{\mathbf{G}} : |C_m| = 1 \\ C_m = \{i\}}} \sum_{j=3\lfloor 3(1-2\epsilon) \rfloor}^{\lfloor (1-2\epsilon) N_{i,T} \rfloor} \gamma_i(j)^q \tag{50}$$

$$\leq \bar{k}_1 T^q \left\lceil \frac{1}{1-2\epsilon} \right\rceil \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil (1-2\epsilon) \frac{T}{\bar{k}_1} \right\rceil, q \right), \tag{51}$$

where line (46) follows from Lemma (9) and the fact that, for cliques with a single arm, internal times are equivalent to the number of pulls (*i.e.*, $t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-2h+1} = 2h$), line (47) follows from Assumption 1, by $h_{i,t_{i,j}} = \lfloor \epsilon N_{i,t_{i,j}-1} \rfloor$ and by bounding $2t_{i,j} - 2(j-1) + h_{i,t_{i,j}} - 1 \leq T$, line (48) by Assumption 1, line (49) from the inequality $\min\{1, x\} \leq \min\{1, x\}^q \leq x^q$ for $q \in [0, 1]$, line (50) from Lemma 12, and line (51) from Lemma 13.

We now proceed on bounding the second term:

$$\text{(b)} \leq \sum_{C_m \in \mathcal{C}_{\mathbf{G}}:|C_m|>1} \sum_{i \in C_m} \sum_{j=3}^{N_{i,T}} \min\left\{1, \frac{(2t_{i,j} - 2(j-1) + h_{i,t_{i,j}} - 1)(t^I_{i,j-1} - t^I_{i,j-2h_{i,t}+1})}{2h_{i,t}} \gamma_i(t^I_{i,(j-1)-2h_{i,t_{i,j}}+1})\right\} \tag{52}$$

$$\leq \sum_{C_m \in \mathcal{C}_{\mathbf{G}}:|C_m|>1} \sum_{i \in C_m} \sum_{j=3}^{N_{i,T}} \min\left\{1, \frac{T\widetilde{N}_{C_m,T}}{\lfloor \epsilon(j-1) \rfloor} \gamma_i(t^I_{i,j-2\lfloor \epsilon(j-1) \rfloor})\right\} \tag{53}$$

$$\leq \sum_{C_m \in \mathcal{C}_{\mathbf{G}}:|C_m|>1} \sum_{i \in C_m} \sum_{j=3}^{N_{i,T}} \min\left\{1, \frac{T\widetilde{N}_{C_m,T}}{\lfloor \epsilon(j-1) \rfloor} \gamma_i(\lfloor (1-2\epsilon)j \rfloor)\right\} \tag{54}$$

$$\leq T^z \sum_{C_m \in \mathcal{C}_{\mathbf{G}}:|C_m|>1} \sum_{i \in C_m} \widetilde{N}^z_{C_m,T} \sum_{j=3}^{N_{i,T}} \left(\frac{\gamma_i(\lfloor (1-2\epsilon)j \rfloor)}{\lfloor \epsilon(j-1) \rfloor}\right)^z \tag{55}$$

$$\leq T^z \sum_{C_m \in \mathcal{C}_{\mathbf{G}}:|C_m|>1} \sum_{i \in C_m} \widetilde{N}^z_{C_m,T} \left(\sum_{j=3}^{N_{i,T}} \frac{1}{\lfloor \epsilon(j-1) \rfloor}\right)^z \left(\sum_{j=3}^{N_{i,T}} \gamma_i(\lfloor (1-2\epsilon)j \rfloor)^{\frac{z}{1-z}}\right)^{1-z} \tag{56}$$

$$\leq T^z \left\lceil \frac{1}{\epsilon} \right\rceil \left\lceil \frac{1}{1-2\epsilon} \right\rceil \sum_{C_m \in \mathcal{C}_{\mathbf{G}}:|C_m|>1} \sum_{i \in C_m} \widetilde{N}^z_{C_m,T} \left(\sum_{j=\lfloor 2\epsilon \rfloor}^{\lfloor \epsilon(N_{i,T}-1) \rfloor} \frac{1}{j}\right)^z \left(\sum_{j=\lfloor 3(1-2\epsilon) \rfloor}^{\lfloor (1-2\epsilon)N_{i,T} \rfloor} \gamma_i(j)^{\frac{z}{1-z}}\right)^{1-z} \tag{57}$$

$$\leq T^z (1+\log(\epsilon T))^z \left\lceil \frac{1}{\epsilon} \right\rceil \left\lceil \frac{1}{1-2\epsilon} \right\rceil \sum_{C_m \in \mathcal{C}_{\mathbf{G}}:|C_m|>1} \sum_{i \in C_m} \widetilde{N}^z_{C_m,T} \left(\sum_{j=\lfloor 3(1-2\epsilon) \rfloor}^{\lfloor (1-2\epsilon)N_{i,T} \rfloor} \gamma_i(j)^{\frac{z}{1-z}}\right)^{1-z} \tag{58}$$

$$\leq T^z (1+\log(\epsilon T))^z \left\lceil \frac{1}{\epsilon} \right\rceil \left\lceil \frac{1}{1-2\epsilon} \right\rceil \sum_{C_m \in \mathcal{C}_{\mathbf{G}}:|C_m|>1} \widetilde{N}^z_{C_m,T} |C_m|^z \left(\sum_{i \in C_m} \sum_{j=\lfloor 3(1-2\epsilon) \rfloor}^{\lfloor (1-2\epsilon)N_{i,T} \rfloor} \gamma_i(j)^{\frac{z}{1-z}}\right)^{1-z} \tag{59}$$

$$\leq T^z (1+\log(\epsilon T))^z \left\lceil \frac{1}{\epsilon} \right\rceil \left\lceil \frac{1}{1-2\epsilon} \right\rceil \sum_{C_m \in \mathcal{C}_{\mathbf{G}}:|C_m|>1} \widetilde{N}^z_{C_m,T} |C_m| \Upsilon_{\boldsymbol{\nu}} \left((1-2\epsilon) \left\lfloor \frac{\widetilde{N}_{C_m,T}}{|C_m|} \right\rfloor, \frac{z}{1-z}\right)^{1-z} \tag{60}$$

$$\leq T^{2z} (1+\log(\epsilon T))^z \left\lceil \frac{1}{\epsilon} \right\rceil \left\lceil \frac{1}{1-2\epsilon} \right\rceil \sum_{C_m \in \mathcal{C}_{\mathbf{G}}:|C_m|>1} |C_m| \Upsilon_{\boldsymbol{\nu}} \left((1-2\epsilon) \left\lfloor \frac{T}{|C_m|} \right\rfloor, \frac{z}{1-z}\right)^{1-z}, \tag{61}$$

where line (52) follows from the bias bound of Lemma 9, line (53) is obtained from bounding $(2t_{i,j} - 2(j-1) + h_{i,t_{i,j}} - 1)(t^I_{i,j-1} - t^I_{i,j-2h_{i,t}+1}) \leq 2T\widetilde{N}_{i,T}$ and using the definition of $h_{i,t}$, line (54) derives from observing that $\gamma_i(t_{i,j}) \leq \gamma_i(j)$ for Assumption 1, line (55) from the inequality $\min\{1, x\} \leq \min\{1, x\}^z \leq x^z$ for $z \in [0, 1/2]$, line (56) is obtained from Hölder's inequality with exponents $\frac{1}{z} \geq 1$ and $\frac{1}{1-z} \geq 1$ respectively, line (57) is an application of Lemma 12 to independently to both inner summations, line (58) derives from bounding the harmonic sum, i.e., $\sum_{\lfloor 2\epsilon \rfloor}^{\lfloor \epsilon(N_{i,T}-1) \rfloor} \frac{1}{j} \leq 1 + \log(\epsilon(N_{i,T}-1)) \leq 1 + \log(\epsilon T)$, line (59) follows from Jensen's inequality, line (60) is obtained from Lemma 13, line (61) by bounding $\widetilde{N}_{i,T} \leq T$. By recalling $q = \frac{z}{1-z} \in [0, 1]$, we obtain:

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

□

Theorem 6 (Regret Upper Bound for `R-□-UCB` in Block-Diagonal Matrices)**.** *Let* $(\boldsymbol{\nu}, \mathbf{G})$ *be an instance of the* GTRB *problem, where* $\mathbf{G} \in \mathbf{B}_{\widetilde{k}}$*. 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, Algorithm 3 suffers an expected regret bounded as:*

$$R_{\boldsymbol{\nu},\mathbf{G}}(\texttt{R-□-UCB})$$

$$\leq \widetilde{\mathcal{O}}\Bigg(\min_{q\in[0,1]} \Bigg\{ \underbrace{(\sigma T)^{\frac{2}{3}}}_{\text{(A) Variance Contribution}} + \underbrace{\bar{k}_1 T^q \Upsilon_{\boldsymbol{\nu}}\left(\left\lceil \frac{T}{\bar{k}_1}\right\rceil, q\right)}_{\text{(B) Rested Bias Contribution}} +$$

$$\underbrace{+ T^{\frac{2q}{1+q}} \sum_{C_m\in\mathcal{C}_{\mathbf{G}}:|C_m|>1} |C_m| \Upsilon_{\boldsymbol{\nu}}\left(\left\lceil\frac{T}{|C_m|}\right\rceil, q\right)^{\frac{1}{1+q}} \Bigg\}}_{\text{(C) Restless Bias Contribution}}\Bigg),$$

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:

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

We have to analyze the following expression:

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

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 $\mathcal{E}_t$:

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

where we exploited $\mu_{i_t^*}(t) - \mu_{I_t}(t) \leq 1$ in the inequality. Now, we bound the second summation, recalling that $\alpha > 2$:

$$\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\{\neg\mathcal{E}_t\}\right] = \sum_{t=1}^{T} \mathbb{P}\left(\neg\mathcal{E}_t\right) \leq 1 + \sum_{t=2}^{T} \mathbb{P}\left(\neg \bigcap_{i\in[k]} \mathcal{E}_{i,t} \right) = 1 + \sum_{t=2}^{T} \mathbb{P}\left(\bigcup_{i\in[k]} \neg\mathcal{E}_{i,t} \right) \leq 1 + \sum_{i\in[k]} \sum_{t=2}^{T} \mathbb{P}\left(\neg\mathcal{E}_{i,t}\right),$$

where the first inequality is obtained with $\mathbb{P}(\neg\mathcal{E}_1) \leq 1$ and the second with 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 to get:

$$\sum_{i\in[k]} \sum_{t=2}^{T} \mathbb{P}\left(\neg\mathcal{E}_{i,t}\right) \leq \sum_{i\in[k]} \sum_{t=2}^{T} 2t^{1-\alpha} \leq 2k \int_{x=1}^{+\infty} x^{1-\alpha} dx = \frac{2k}{\alpha - 2}.$$

From now on, we will proceed the analysis 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:

$$B_{I_t}(t) - \mu_{I_t}(t) = \widehat{\mu}_{I_t}^{h_{I_t,t}}(t) + \beta_{I_t}^{h_{I_t,t}}(t) - \mu_{I_t}(t) \tag{62}$$

$$\leq \underbrace{\widetilde{\mu}_{I_t}^{h_{I_t,t}}(t) - \mu_{I_t}(t)}_{\text{(a)}} + \underbrace{2\beta_{I_t}^{h_{I_t,t}}(t)}_{\text{(b)}}, \tag{63}$$

where line (62) follows from the definition of $B_i(t)$ and line (63) from the good event $\mathcal{E}_t$. We make use of Lemma 10 and Lemma 14 to bound the summations over t of (a) and (b), respectively.

Putting all together, we obtain:

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

□

Lemma 11 (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*

$$\begin{aligned} &\sum_{t=1}^{T}\min\left\{1, \widetilde{\mu}_{I_t}^{h_{I_t,t}}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t})\right\} \leq \\ &\qquad \leq 2k + \bar{k}_1 T^q\left\lceil\frac{1}{1-2\epsilon}\right\rceil\Upsilon_{\boldsymbol{\nu}}\left(\left\lceil(1-2\epsilon)\frac{T}{\bar{k}_1}\right\rceil, q\right) + \\ &\qquad + T^{\frac{2q}{1+q}}(1+\log(\epsilon T))^{\frac{q}{1+q}}\left\lceil\frac{1}{\epsilon}\right\rceil\left\lceil\frac{1}{1-2\epsilon}\right\rceil(k-\bar{k}_1)\Upsilon_{\boldsymbol{\nu}}\left(\left\lceil(1-2\epsilon)\frac{T}{k-\bar{k}_1}\right\rceil, q\right)^{\frac{1}{1+q}}, \end{aligned} \tag{64}$$

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 10. We decide to split arms based on their degree, in particular we bound separately the bias due to arms having degree of 1 (*i.e.*, they only are triggered by themselves).

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

We start from bounding the first term:

$$\text{(a)} \leq \sum_{\substack{i\in[k]\\ \deg^-(i)=1}}\sum_{j=3}^{N_{i,T}}\min\left\{1, \frac{(2t_{i,j} - 2(j-1) + h_{i,t_{i,j}} - 1)2h_{i,t}}{2h_{i,t}}\gamma_i(t_{i,(j-1)-2h_{i,t_{i,j}}+1})\right\} \tag{65}$$

$$\leq \sum_{\substack{i\in[k]\\ \deg^-(i)=1}}\sum_{j=3}^{N_{i,T}}\min\left\{1, T\gamma_i(t^I_{i,j-2\lfloor\epsilon(j-1)\rfloor})\right\} \tag{66}$$

$$\leq \sum_{\substack{i\in[k]\\ \deg^-(i)=1}} \sum_{j=3}^{N_{i,T}} \min\left\{1, T\gamma_i(\lfloor(1-2\epsilon)j\rfloor)\right\} \tag{67}$$

$$\leq T^q \sum_{\substack{i\in[k]\\ \deg^-(i)=1}} \sum_{j=3}^{N_{i,T}} \gamma_i(\lfloor(1-2\epsilon)j\rfloor)^q \tag{68}$$

$$\leq T^q \left\lceil \frac{1}{1-2\epsilon} \right\rceil \sum_{\substack{i\in[k]\\ \deg^-(i)=1}} \sum_{j=3\lfloor 3(1-2\epsilon)\rfloor}^{\lfloor(1-2\epsilon)N_{i,T}\rfloor} \gamma_i(j)^q \tag{69}$$

$$\leq T^q \left\lceil \frac{1}{1-2\epsilon} \right\rceil \sum_{\substack{i\in[k]\\ \deg^-(i)=1}} \sum_{j=3\lfloor 3(1-2\epsilon)\rfloor}^{\lfloor(1-2\epsilon)N_{i,T}\rfloor} \gamma_i(j)^q \tag{70}$$

$$\leq \bar{k}_1 T^q \left\lceil \frac{1}{1-2\epsilon} \right\rceil \Upsilon_{\boldsymbol{\nu}}\left(\left\lceil (1-2\epsilon)\frac{T}{\bar{k}_1} \right\rceil, q \right), \tag{71}$$

where line (65) follows from Lemma (9) and the fact that, for cliques with a single arm, internal times are equivalent to the number of pulls (*i.e.*, $t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-2h+1} = 2h$), line (66) follows from Assumption 1, by $h_{i,t_{i,j}} = \lfloor \epsilon N_{i,t_{i,j}-1} \rfloor$ and by bounding $2t_{i,j} - 2(j-1) + h_{i,t_{i,j}} - 1 \leq T$, line (68) by Assumption 1, line (69) from the inequality $\min\{1,x\} \leq \min\{1,x\}^q \leq x^q$ for $q \in [0,1]$, line (70) from Lemma 12, and line (71) from Lemma 13.

As a trivial 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 and for every t:

$$\widetilde{N}_{i,t} \leq \widetilde{N}^U_{i,t}, \tag{72}$$

where $\widetilde{N}^U_{i,t} \coloneqq \mathbf{e}_i^\top (\bar{\mathbf{G}}^U)^\top \mathbf{N}_t$.

We now proceed on bounding the second term:

$$\text{(b)} \leq \sum_{\substack{i\in[k]\\ \deg^-(i)>1}} \sum_{j=3}^{N_{i,T}} \min\left\{1, \frac{(2t_{i,j} - 2(j-1) + h_{i,t_{i,j}} - 1)(t^I_{i,j-1} - t^I_{i,j-2h_{i,t}+1})}{2h_{i,t}} \gamma_i(t^I_{i,(j-1)-2h_{i,t_{i,j}}+1}) \right\} \tag{73}$$

$$\leq \sum_{\substack{i\in[k]\\ \deg^-(i)>1}} \sum_{j=3}^{N_{i,T}} \min\left\{1, \frac{T\widetilde{N}_{i,T}}{\lfloor \epsilon(j-1) \rfloor} \gamma_i(t^I_{i,j-2\lfloor\epsilon(j-1)\rfloor}) \right\} \tag{74}$$

$$\leq \sum_{\substack{i\in[k]\\ \deg^-(i)>1}} \sum_{j=3}^{N_{i,T}} \min\left\{1, \frac{T\widetilde{N}_{i,T}}{\lfloor \epsilon(j-1) \rfloor} \gamma_i(\lfloor(1-2\epsilon)j\rfloor) \right\} \tag{75}$$

$$\leq \sum_{\substack{C^U_m \in \mathcal{C}_{\bar{\mathbf{G}}^U}\\ |C^U_m|>1}} \sum_{i\in C^U_m} \sum_{j=3}^{N_{i,T}} \min\left\{1, \frac{T\widetilde{N}^U_{C_m,T}}{\lfloor \epsilon(j-1) \rfloor} \gamma_i(\lfloor(1-2\epsilon)j\rfloor) \right\} \tag{76}$$

$$\leq T^z \sum_{\substack{C^U_m \in \mathcal{C}_{\bar{\mathbf{G}}^U}\\ |C^U_m|>1}} \sum_{i\in C^U_m} (\widetilde{N}^U_{C_m,T})^z \sum_{j=3}^{N_{i,T}} \left(\frac{\gamma_i(\lfloor(1-2\epsilon)j\rfloor)}{\lfloor \epsilon(j-1) \rfloor} \right)^z \tag{77}$$

$$\leq T^z \sum_{\substack{C_m^U \in \mathcal{C}_{\bar{\mathbf{G}}^U}\\ |C_m^U|>1}} \sum_{i \in C_m^U} (\widetilde{N}^U_{C_m,T})^z \left(\sum_{j=3}^{N_{i,T}} \frac{1}{\lfloor \epsilon(j-1) \rfloor} \right)^z \left(\sum_{j=3}^{N_{i,T}} \gamma_i(\lfloor (1-2\epsilon)j \rfloor)^{\frac{z}{1-z}} \right)^{1-z} \tag{78}$$

$$\leq T^z \left\lceil \frac{1}{\epsilon} \right\rceil \left\lceil \frac{1}{1-2\epsilon} \right\rceil \sum_{\substack{C_m^U \in \mathcal{C}_{\bar{\mathbf{G}}^U}\\ |C_m^U|>1}} \sum_{i \in C_m^U} (\widetilde{N}^U_{C_m,T})^z \left(\sum_{j=\lfloor 2\epsilon \rfloor}^{\lfloor \epsilon(N_{i,T}-1) \rfloor} \frac{1}{j} \right)^z \left(\sum_{j=\lfloor 3(1-2\epsilon) \rfloor}^{\lfloor (1-2\epsilon)N_{i,T} \rfloor} \gamma_i(j)^{\frac{z}{1-z}} \right)^{1-z} \tag{79}$$

$$\leq T^z (1+\log(\epsilon T))^z \left\lceil \frac{1}{\epsilon} \right\rceil \left\lceil \frac{1}{1-2\epsilon} \right\rceil \sum_{\substack{C_m^U \in \mathcal{C}_{\bar{\mathbf{G}}^U}\\ |C_m^U|>1}} \sum_{i \in C_m^U} (\widetilde{N}^U_{C_m,T})^z \left(\sum_{j=\lfloor 3(1-2\epsilon) \rfloor}^{\lfloor (1-2\epsilon)N_{i,T} \rfloor} \gamma_i(j)^{\frac{z}{1-z}} \right)^{1-z} \tag{80}$$

$$\leq T^z (1+\log(\epsilon T))^z \left\lceil \frac{1}{\epsilon} \right\rceil \left\lceil \frac{1}{1-2\epsilon} \right\rceil \sum_{\substack{C_m^U \in \mathcal{C}_{\bar{\mathbf{G}}^U}\\ |C_m^U|>1}} |C_m^U|^z \left(\sum_{i \in C_m^U} \sum_{j=\lfloor 3(1-2\epsilon) \rfloor}^{\lfloor (1-2\epsilon)N_{i,T} \rfloor} \gamma_i(j)^{\frac{z}{1-z}} \right)^{1-z} \tag{81}$$

$$\leq T^{2z} (1+\log(\epsilon T))^z \left\lceil \frac{1}{\epsilon} \right\rceil \left\lceil \frac{1}{1-2\epsilon} \right\rceil \sum_{\substack{C_m^U \in \mathcal{C}_{\bar{\mathbf{G}}^U}\\ |C_m^U|>1}} |C_m^U| \Upsilon_{\boldsymbol{\nu}} \left((1-2\epsilon) \left\lfloor \frac{\widetilde{N}^U_{C_m,T}}{|C_m^U|} \right\rfloor, \frac{z}{1-z} \right)^{1-z}, \tag{82}$$

where line (73) follows from the bias bound of Lemma 9, line (74) is obtained from bounding $(2t_{i,j} - 2(j-1) + h_{i,t_{i,j}} - 1)(t^I_{i,j-1} - t^I_{i,j-2h_{i,t}+1}) \leq 2T\widetilde{N}_{i,T}$ and using the definition of $h_{i,t}$, line (75) derives from observing that $\gamma_i(t_{i,j}) \leq \gamma_i(j)$ for Assumption 1, line (76) follows (72) and a decomposition of the pulls over the cliques of $\bar{\mathbf{G}}^U$, line (77) from the inequality $\min\{1,x\} \leq \min\{1,x\}^z \leq x^z$ for $z \in [0,1/2]$, line (78) is obtained from Hölder's inequality with exponents $\frac{1}{z} \geq 1$ and $\frac{1}{1-z} \geq 1$ respectively, line (79) is an application of Lemma 12 to independently to both inner summations, line (80) derives from bounding the harmonic sum, i.e., $\sum_{\lfloor 2\epsilon \rfloor}^{\lfloor \epsilon(N_{i,T}-1) \rfloor} \frac{1}{j} \leq 1 + \log(\epsilon(N_{i,T}-1)) \leq 1 + \log(\epsilon T)$, line (81) follows from Jensen's inequality and by bounding $\widetilde{N}^U_{C^U_m,T} \leq T$, line (82) is obtained from Lemma 13. By recalling $q = \frac{z}{1-z} \in [0,1]$, we obtain:

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

□

Theorem 7 (Regret Upper Bound for `R-□-UCB` in General Matrices)**.** *Let* $(\boldsymbol{\nu}, \mathbf{G})$ *be an instance of the* GTRB *problem, 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, Algorithm 3 suffers an expected regret bounded as:*

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

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 11 instead of Lemma 10 to bound cumulative estimator's bias.

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:

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

We have to analyze the following expression:

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

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

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

where we exploited $\mu_{i_t^*}(t) - \mu_{I_t}(t) \leq 1$ in the inequality. Now, we bound the second summation, recalling that $\alpha > 2$:

$$\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\{\neg\mathcal{E}_t\} \right] = \sum_{t=1}^{T} \mathbb{P}\left(\neg\mathcal{E}_t \right) \leq 1 + \sum_{t=2}^{T} \mathbb{P}\left(\neg \bigcap_{i \in [k]} \mathcal{E}_{i,t} \right) = 1 + \sum_{t=2}^{T} \mathbb{P}\left(\bigcup_{i \in [k]} \neg\mathcal{E}_{i,t} \right) \leq 1 + \sum_{i \in [k]} \sum_{t=2}^{T} \mathbb{P}\left(\neg\mathcal{E}_{i,t} \right),$$

where the first inequality is obtained with $\mathbb{P}(\neg\mathcal{E}_1) \leq 1$ and the second with 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 to get:

$$\sum_{i \in [k]} \sum_{t=2}^{T} \mathbb{P}\left(\neg\mathcal{E}_{i,t} \right) \leq \sum_{i \in [k]} \sum_{t=2}^{T} 2t^{1-\alpha} \leq 2k \int_{x=1}^{+\infty} x^{1-\alpha} dx = \frac{2k}{\alpha - 2}.$$

From now on, we will proceed the analysis 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 $t = 1$, 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:

$$\begin{aligned} B_{I_t}(t) - \mu_{I_t}(t) &= \widehat{\mu}_{I_t}^{h_{I_t,t}}(t) + \beta_{I_t}^{h_{I_t,t}}(t) - \mu_{I_t}(t) && (83) \\ &\leq \underbrace{\widetilde{\mu}_{I_t}^{h_{I_t,t}}(t) - \mu_{I_t}(t)}_{\text{(a)}} + \underbrace{2\beta_{I_t}^{h_{I_t,t}}(t)}_{\text{(b)}}, && (84) \end{aligned}$$

where line (83) follows from the definition of $B_i(t)$ and line (84) from the good event $\mathcal{E}_t$. We make use of Lemma 11 and Lemma 14 to bound the summations over t of (a) and (b), respectively.

Putting all together, we obtain:

$$\begin{aligned} & R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{R-}\square\texttt{-UCB}) \leq \\ & \leq 1 + \frac{2k}{\alpha - 2} + 5k + \frac{k}{\epsilon} + \frac{3k}{\epsilon} (2\sigma T)^{\frac{2}{3}} \left(10\alpha \log T \right)^{\frac{1}{3}} + \end{aligned}$$

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

□

C. Technical Lemmas for Stochastic Rising Bandits

In this appendix, we report some useful technical lemmas from the literature of stochastic rising bandits that will play a role in the results of Section B.

Lemma 12 (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:*

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

Proof. We simply observe that the minimum value of $\lfloor \beta j \rfloor$ is $\lfloor 3\beta \rfloor$ and its maximum value is $\lfloor \beta M \rfloor$. Each element $\lfloor \beta j \rfloor$ changes value at least one time every $\left\lceil \frac{1}{\beta} \right\rceil$ times. □

Lemma 13 (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:*

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

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:

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

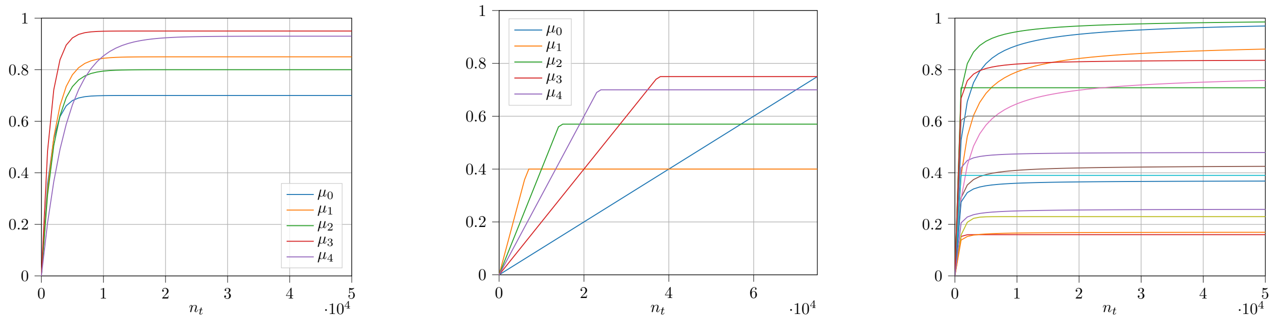
Proof Sketch. Using a Doob's *optional skipping* argument (Doob, 1953; Bubeck et al., 2008), and noting that, at round t, $t_{i,l}^I$ is a stopping time for every arm $i \in [k]$ and pull number $l \in \{1, \ldots, N_{i,t-1}\}$ w.r.t. the filtration $\mathcal{F}_{\tau-1} = \sigma(I_1, X_1, \ldots, I_{\tau-1},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

Lemma 14 (Bound on Estimator's Variance, Metelli et al. (2022), Theorem 4.4)**.** *Let $(I_t)_{t \in [T]}$ be a sequence of actions such that*

$$
\left| \widehat{\mu}_{I_t}^{h_{I_t,t}}(t) - \widetilde{\mu}_{I_t}^{h_{I_t,t}}(t) \right| \leq \beta_{I_t}^{h_{I_t,t}}(t, t^{-\alpha}), \ \forall t \in [T], \tag{85}
$$

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

$$
\sum_{t=1}^{T} \min \left\{ 1, 2\beta_{I_t}^{h_{I_t,t}}(t, t^{-\alpha}) \right\} \leq k \left(3 + \frac{1}{\epsilon} \right) + \frac{3k}{\epsilon} (2\sigma T)^{\frac{2}{3}} (10\alpha \log T)^{\frac{1}{3}}. \tag{86}
$$



(a) $\boldsymbol{\nu}_1, \quad \mu_i = m_i(1 - e^{-\kappa_i n_t})$.

(b) $\boldsymbol{\nu}_2, \quad \mu_i = \min\{\kappa_i n_t, m_i\}$.

(c) $\boldsymbol{\nu}_3, \quad \mu_i = m_i(1 - e^{-\kappa_i n_t})$.

Figure 2. Sets of functions used in the experimental campaign over deterministic settings with block-diagonal adjacency matrices. Under each figure, we report the family of analytical functions used for the instance construction, where $\kappa_i > 0$ and $m_i \in [0, 1]$.

D. Experiments

In this appendix, we provide an experimental campaign to validate the proposed algorithmic solutions from an empirical perspective.

We start with the deterministic setting: in Appendix D.1 we evaluate DR-BG-UB in 15 GTRB instances, varying both the functions and the adjacency matrices; in Appendix D.2 we evaluate DR-G-UB in 3 GTRB instances, but varying the sub-matrix used in the Algorithm 2 routine. Finally, we evaluate R-□-UCB in 10 stochastic GTRB instances, varying both the functions and the adjacency matrices, and comparing its performances to a baseline from the literature, Sliding Window UCB (Garivier & Moulines, 2011). We decided not to evaluate R-□-UCB under general adjacency matrices since there is no feasible way to compute a clairvoyant and no reasonable sensitivity analysis can be conducted on the algorithm's inputs as we did in the deterministic setting studying the impact of the specified sub-matrix.

D.1. Deterministic Setting with Block-Diagonal Matrices

This section assesses the empirical performances of DR-BG-UB in a synthetic environment. To do so, we propose a total of 15 different instances of Graph-Triggered Rising Bandits with block-diagonal matrices and adding no noise in the rewards generation process.

Setting. In Figure 2, we report three different set of functions satisfying Assumption 1, $\{\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\nu}_3\}$, of 5, 5 and 15 arms, respectively.

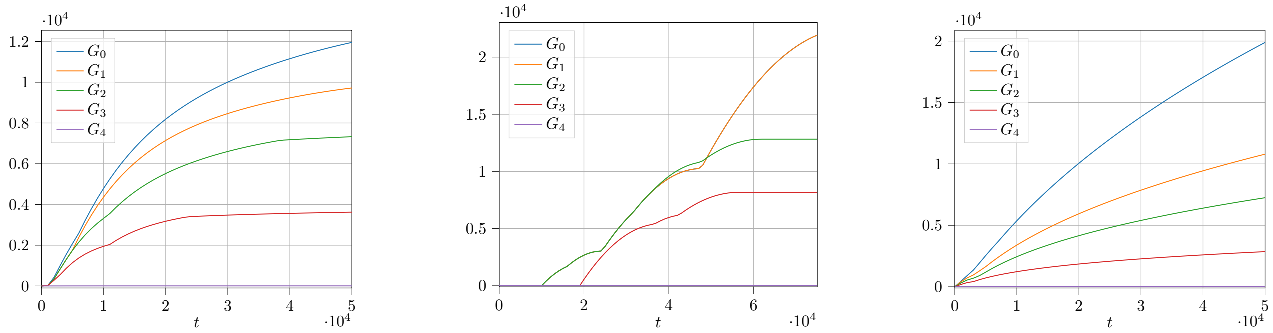
Some remarks are in order. The total increment assumes different behaviors depending on the set of functions, indeed $\Upsilon_{\boldsymbol{\nu}_1} = \mathcal{O}(\log T)$ and $\Upsilon_{\boldsymbol{\nu}_3} = \mathcal{O}(\log T)$, while $\Upsilon_{\boldsymbol{\nu}_2} = \mathcal{O}(T)$. This has been done voluntarily to stress the algorithm towards these two corner cases and assess its performance on both. Moreover, in $\mathcal{F}_1$ we can see one function, namely μ_3, dominating all the others. This is a corner case in which, whatever the underlying graph, all the optimal policies coincide. Instead, in $\mathcal{F}_2$, we observe that the optimal policy in the restless scenario would include pulling 4 different actions across the trial. Instead, $\mathcal{F}_3$ is aimed at assessing Algorithm 1 performance when the action space is larger.

We now introduce a compact notation for block-diagonal matrices. Let $\mathbf{G} \in \{0, 1\}^{k \times k}$ a block-diagonal matrix with $\widetilde{k}$ distinct blocks. Then, we indicate the matrix by means of its block sizes as $\mathbf{G} := \{b_1, \ldots, b_{\widetilde{k}}\}$, with the convention that the first block is the one on the upper-left corner, and so on. Note that $\sum_{i \in [\widetilde{k}]} b_i = k$.

Together with those, we define two sets of block-diagonal matrices, namely $\mathcal{B}_1$ and $\mathcal{B}_2$, composed of 5 matrices each:

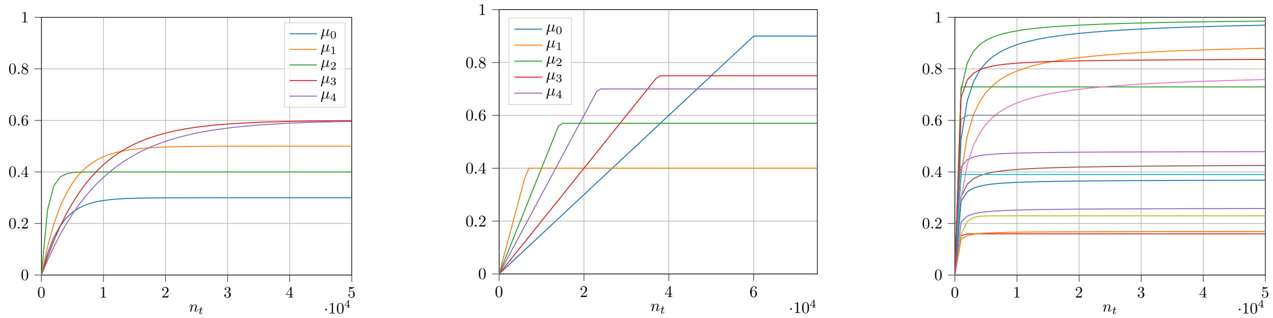
$$\mathcal{B}_1 = \{\mathbf{I}_5, \{2, 1, 1, 1\}, \{2, 1, 2\}, \{3, 2\}, \mathbf{1}_{5 \times 5}\},$$
$$\mathcal{B}_2 = \{\mathbf{I}_{15}, \{3, 3, 3, 3, 3\}, \{5, 5, 5\}, \{5, 10\}, \mathbf{1}_{15 \times 15}\}.$$

Note that the definition of sub-matrix partially orders matrices increasingly, *i.e.*, every matrix is a sub-matrix of the following one. When referring to a set of "increasing" matrices, we will indicate with $\mathbf{G}_0$ the minimum (*e.g.*, $\mathbf{G}_0 = \mathbf{I}_5$), and so on until the maximum (*e.g.*, $\mathbf{G}_4 = \mathbf{1}_{5 \times 5}$). Set $\mathcal{B}_1$ is composed of sequentially nested matrices, while set $\mathcal{B}_2$ is composed of matrices. Both have a decreasing number of blocks.



(a) $\boldsymbol{\nu}_1,\ \mathbf{G}_i \in \mathcal{B}_1,\ T = 5 \cdot 10^4$. (b) $\boldsymbol{\nu}_2,\ \ \mathbf{G}_i \in \mathcal{B}_1,\ T = 7.5 \cdot 10^4$. (c) $\boldsymbol{\nu}_3,\ \ \mathbf{G}_i \in \mathcal{B}_2,\ T = 5 \cdot 10^4$.

Figure 3. Cumulative regrets obtained by `DR-BG-UB`. The algorithm faces every set of functions 5 times under a different adjacency matrix, from the sparser ($\mathbf{G}_0 = \mathbf{I}$) to the complete matrix ($\mathbf{G}_4 = \mathbf{1}$).



(a) $\boldsymbol{\nu}_4,\ \ \mu_i = m_i(1 - e^{-\kappa_i n_t})$. (b) $\boldsymbol{\nu}_5,\ \ \mu_i = \min\{\kappa_i n_t, m_i\}$. (c) $\boldsymbol{\nu}_6,\ \ \mu_i = m_i(1 - e^{-\kappa_i n_t})$.

Figure 4. Sets of functions used in the experimental campaign over stochastic settings with non-block-diagonal adjacency matrices. Under each figure, we report the family of analytical functions used for the instance construction, where $\kappa_i > 0$ and $m_i \in [0, 1]$.

Together with the reward functions, the matrices will form the 15 instances as follows: $(\boldsymbol{\nu}_1, \mathbf{G})_{\mathbf{G}\in\mathcal{B}_1}$, $(\boldsymbol{\nu}_2, \mathbf{G})_{\mathbf{G}\in\mathcal{B}_1}$, $(\boldsymbol{\nu}_3, \mathbf{G})_{\mathbf{G}\in\mathcal{B}_2}$, where $\mathbf{G}_0$ Thus, every set of functions will be evaluated on 5 different block-diagonal matrices.

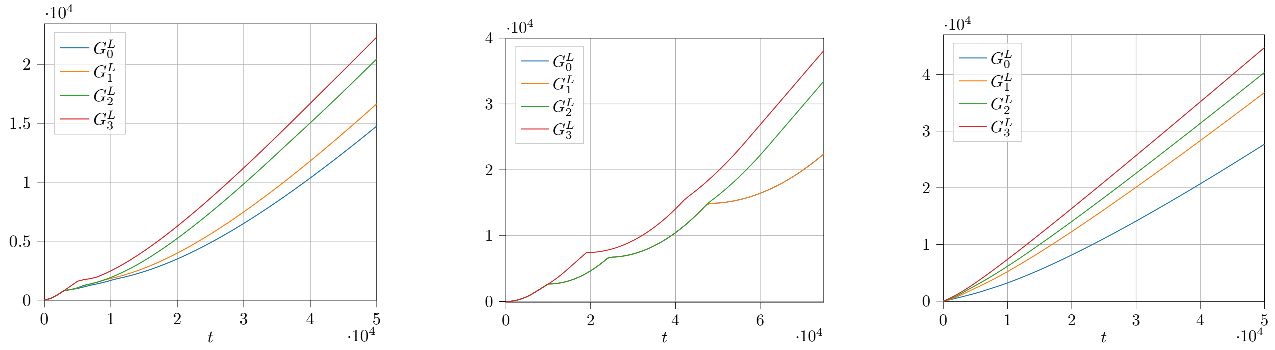
Results. In Figure 3, we report the cumulative regrets obtained by `DR-BG-UB` on the three instances previously described, setting T to 50.000, 75.000 and 50.000, respectively. Since `DR-BG-UB` is an anytime algorithm, for every time $t \in [T]$ we computed the cumulative reward achieved by the optimal policy for that specific time horizon and then tracked the algorithm's cumulative regret at every time.

Some comments are in order. The instances corresponding to purely restless settings (*i.e.*, $\mathbf{G}_4$) are the ones in which `DR-BG-UB` achieves the best performances. This is expected since the restless contribution to the regret's upper bound is sensibly lower than the contribution given by rested arms (Theorem 3). In general, this phenomenon is even more evident when looking at the progression of regret when the number of blocks increases. The higher cumulative regret is always observed when the matrix is the identity (*i.e.*, $\mathbf{G}_0$). However, the cumulative regret always assumes a sub-linear shape, thus validating the theoretical findings of Theorem 3. Finally, the cumulative regret's shape assumes a non-concave behavior (see, *e.g.*, Figure 3b): this is expected since the clairvoyant is computed for every possible t and the optimal policy may drastically change from one time to the subsequent.

D.2. Deterministic Setting with General Matrices

This section assesses the empirical performances of `DR-G-UB` in a synthetic environment. To do so, we propose a total of 3 different instances of Graph-Triggered Rising Bandits with non-block-diagonal matrices and adding no noise in the rewards generation process. Moreover, we analyze the behavior of `DR-G-UB` under different choices of the employed sub-matrix, also deviating from the maximal sub-matrix choice prescribed by the pseudo-code in Algorithm 2.

Setting. In Figure 4, we report three different set of functions satisfying Assumption 1, $\{\boldsymbol{\nu}_4, \boldsymbol{\nu}_5, \boldsymbol{\nu}_6\}$, of 5, 5 and 15



(a) $\boldsymbol{\nu}_4,\ \mathbf{G}_a,\ T = 5 \cdot 10^4$.

(b) $\boldsymbol{\nu}_5,\ \mathbf{G}_a,\ T = 7.5 \cdot 10^4$.

(c) $\boldsymbol{\nu}_6,\ \mathbf{G}_b,\ T = 5 \cdot 10^4$.

Figure 5. Cumulative rewards obtained by DR−G−UB. The algorithm faces every set of functions 5 times under the same different adjacency matrix, however the used sub-matrix G_i^L changes, from the sparser ($\mathbf{G}_0^L = \mathbf{I}$) to the maximal sub-matrix ($\mathbf{G}_3^L$).

functions, respectively.

These set of functions are very similar to the ones in the previous section. Thus, all the remarks done before still apply.

We define two non-block-diagonal matrices, namely $\mathbf{G}_a \in \{0,1\}^{5\times 5}$ and $\mathbf{G}_b \in \{0,1\}^{15\times 15}$:

$$\mathbf{G}_a = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}, \quad \mathbf{G}_b = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Together with the set of functions, these will constitute the 3 instances as follows: $(\nu_4, \mathbf{G}_a)$, $(\nu_5, \mathbf{G}_a)$, and $(\nu_6, \mathbf{G}_b)$.

The goal of this section is mainly to study the behaviour of DR−G−UB under different choices of the sub-matrix, and to validate that the maximal sub-matrix is empirically the best choice. Thus, we will propose 4 possible sub-matrices of $\mathbf{G}_a$ and $\mathbf{G}_b$, respectively, and run Algorithm 2 for every possible choice.

We define two sets of sub-matrices, namely $\mathcal{B}_a^L$ and $\mathcal{B}_b^L$, s.t.

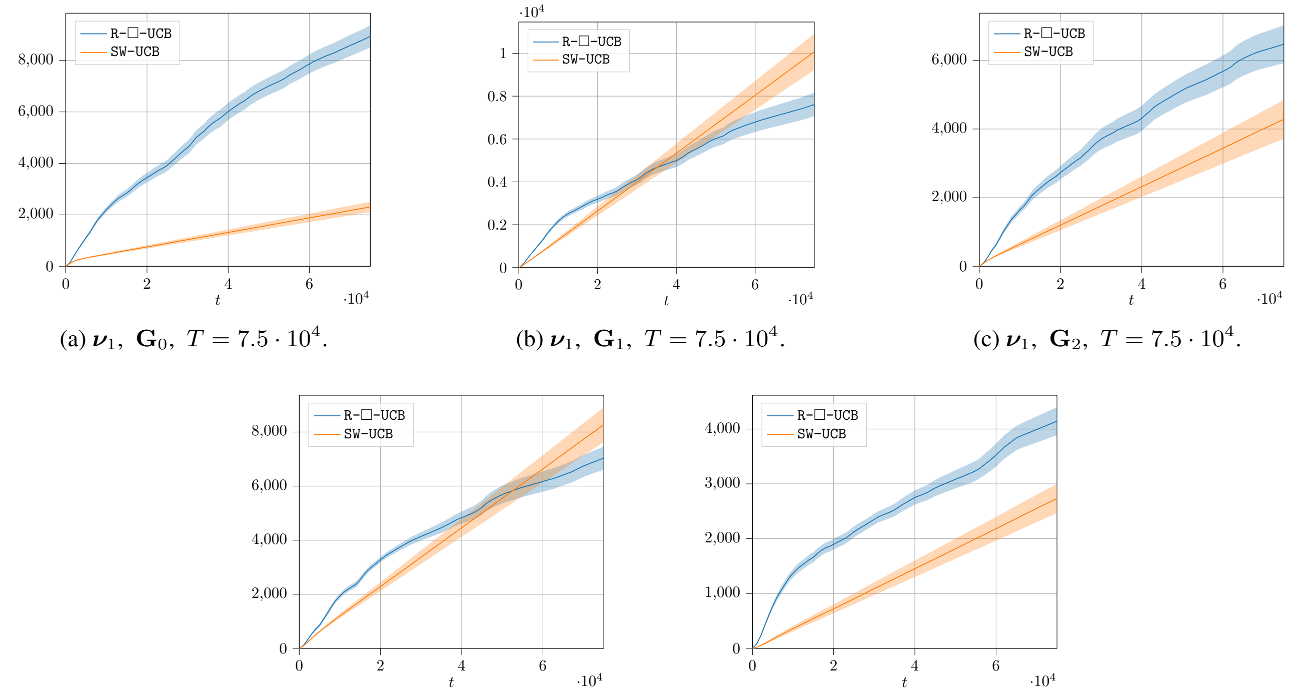
$$\mathcal{B}_a^L = \{\mathbf{I}_5, \{2,1,1,1\}, \{2,1,2\}, \{3,2\}\}, \tag{87}$$

$$\mathcal{B}_b^L = \{\mathbf{I}_{15}, \{3,3,3,3,3\}, \{5,5,5\}, \{5,10\}\}. \tag{88}$$

We will indicate with $\mathbf{G}_0^L$ the minimum sub-matrix (*e.g.*, $\mathbf{G}_0 = \mathbf{I}_5$), and so on until the maximal sub-matrix (*e.g.*, $\mathbf{G}_4^L = \{3,2\}$), that is here uniquely defined.

Results. When the matrix is non-block-diagonal, computing the clairvoyant is NP-hard. Thus, in Figure 5, we report the cumulative rewards obtained by DR−G−UB on the three instances defined above when varying the sub-matrix that is used, setting T to 50.000, 75.000 and 50.000, respectively.

The cumulative regret is always sub-linear, independent of the choice of the used sub-matrix. However, the best performances



(a) $\boldsymbol{\nu}_1,\ \mathbf{G}_0,\ T = 7.5 \cdot 10^4$.

(b) $\boldsymbol{\nu}_1,\ \mathbf{G}_1,\ T = 7.5 \cdot 10^4$.

(c) $\boldsymbol{\nu}_1,\ \mathbf{G}_2,\ T = 7.5 \cdot 10^4$.

(d) $\boldsymbol{\nu}_1,\ \mathbf{G}_3,\ T = 7.5 \cdot 10^4$.

(e) $\boldsymbol{\nu}_1,\ \mathbf{G}_4,\ T = 7.5 \cdot 10^4$.

Figure 6. Cumulative regrets obtained by R−□−UCB and SW−UCB. The algorithms face the same set of functions ν_1 for 5 times under different adjacency matrices. G_i moves from the sparser ($\mathbf{G}_0 = \mathbf{I}_5$) to the complete matrix ($\mathbf{G}_4 = \mathbf{1}_{5\times 5}$). For each instance, we performed 20 trials and reported mean ± std.

are obtained using the maximal sub-matrix, and this is an expected consequence of Theorem 4. When the provided sub-matrix becomes smaller in the number of blocks, the cumulative reward improves despite keeping the true matrix underlying the process fixed. This outcome validates the design choice in Algorithm 2 to use the maximal sub-matrix.

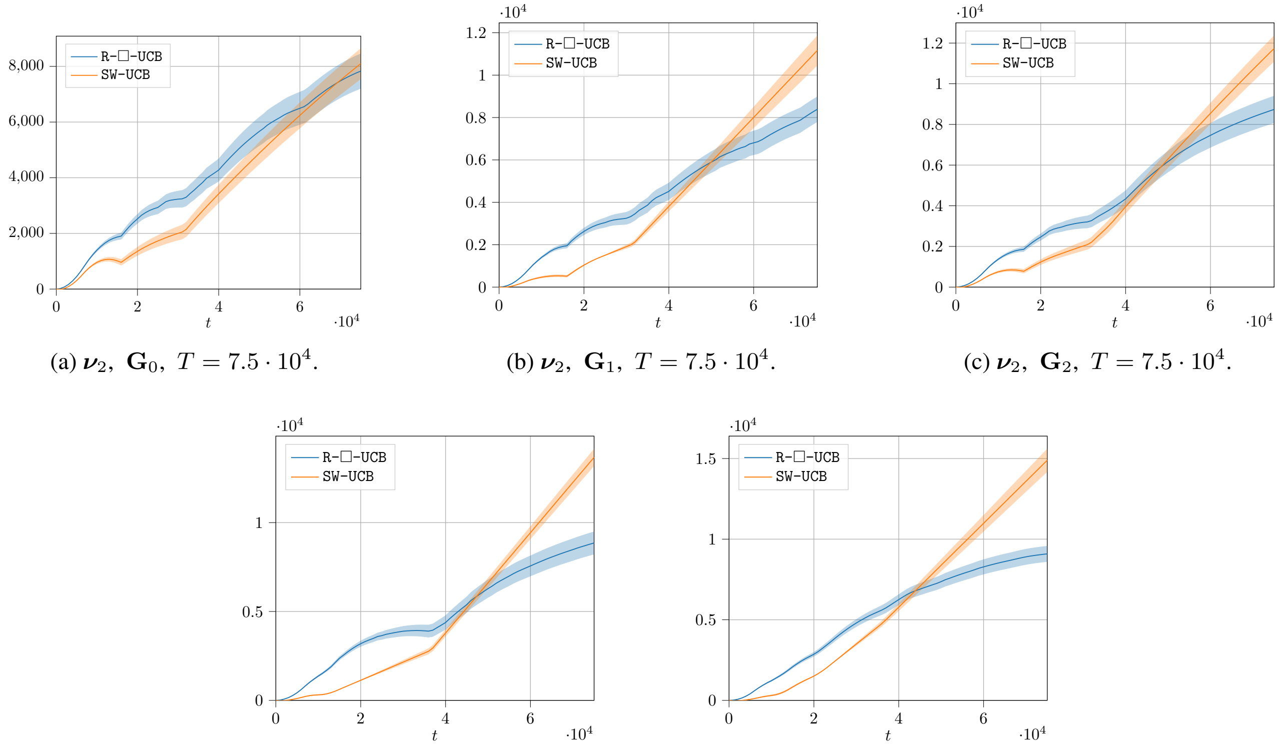
D.3. Stochastic Setting with Block-Diagonal Matrices, and comparison with Sliding-Window UCB

In this section, we validate the need for *ad-hoc* algorithmic solutions to solve the Graph-Triggered Rising Bandits problem. In particular, we evaluate the performance of Algorithm 3 in a stochastic setting with block-diagonal adjacency matrices and compare it to the one of Sliding Window UCB (shortly, SW−UCB, Garivier & Moulines (2011)). To the authors' knowledge, no existing algorithm from the literature deals appropriately with the GTRB setting. So, as a comparison baseline, we decided to use SW−UCB since it is one of the most robust and known algorithms from the non-stationary bandit literature.

Setting. We evaluate the two algorithm on a total of 10 instances, corresponding to the first 10 instances of the experimental campaign in Appendix D.1 (rescaled), *i.e.*, $(\boldsymbol{\nu}_1, \mathbf{G})_{\mathbf{G}\in\mathcal{B}_1}$ and $(\boldsymbol{\nu}_2, \mathbf{G})_{\mathbf{G}\in\mathcal{B}_1}$. The hyper-parameters of SW−UCB have been set according to the original paper (Garivier & Moulines, 2011) and then optimized to get the smaller regret upper bound. Instead, the hyper-parameters of R−□−UCB have been fixed for all experiments, using $\epsilon = 0.1$ and $\alpha = 3$. The time horizon was set to $T = 75.000$, and the optimal policy's cumulative reward has been computed for every time $t \in [T]$. We perform 20 trials for each setting, varying the seed to the additive, zero-mean, Gaussian noise generator, where $\sigma = 0.1$.

Results. In Figure 6, we report the average cumulative regrets obtained by the two algorithms in the first 5 instances and their standard deviations. We can observe that, in most of these, the performance of the two algorithms is comparable, even if SW−UCB tends to achieve a lower regret. However, the shape of SW−UCB cumulative regret has a linear behavior. We remark that the set of functions $\boldsymbol{\nu}_1$ contains a dominant function, μ_3, thus the optimal policy prescribes to always play such arm, independently from the matrix. In this kind of setting, standard algorithms for non-stationary bandits can still achieve satisfactory performance in practice.

However, as we can observe in Figure 7, which contains the regrets for the second half of the instances, the performance of SW−UCB quickly deteriorates w.r.t. R−□−UCB as the optimal policy becomes less trivial. Indeed, when "increasing" the



(a) $\boldsymbol{\nu}_2,\ \mathbf{G}_0,\ T = 7.5 \cdot 10^4$.

(b) $\boldsymbol{\nu}_2,\ \mathbf{G}_1,\ T = 7.5 \cdot 10^4$.

(c) $\boldsymbol{\nu}_2,\ \mathbf{G}_2,\ T = 7.5 \cdot 10^4$.

(d) $\boldsymbol{\nu}_2,\ \mathbf{G}_3,\ T = 7.5 \cdot 10^4$.

(e) $\boldsymbol{\nu}_2,\ \mathbf{G}_4,\ T = 7.5 \cdot 10^4$.

Figure 7. Cumulative regrets obtained by R-☐-UCB and SW-UCB. The algorithms face the same set of functions ν_2 for 5 times under different adjacency matrices. G_i moves from the sparser ($\mathbf{G}_0 = \mathbf{I}_5$) to the complete matrix ($\mathbf{G}_4 = \mathbf{1}_{5\times 5}$). For each instance, we performed 20 trials and reported mean ± std.

adjacency matrix, the optimal policy pulls a larger set of different arms, and standard techniques for non-stationary bandits fail to model this kind of interaction among the arms. The results of this section highlight the need for a specific algorithm to deal with GTRB problems, as standard algorithms from the non-stationary bandits literature may perform well in simpler instances but will severely deteriorate in harder ones. Finally, this section also assesses the good performance of R-☐-UCB in stochastic settings. Indeed, all the cumulative regrets assume a sub-linear shape.