Recurrent Submodular Welfare and Matroid Blocking Semi-Bandits

Orestis Papadigenopoulos
Department of Computer Science
The University of Texas at Austin
papadig@cs.utexas.edu

Constantine Caramanis
Electrical and Computer Engineering
The University of Texas at Austin
constantine@utexas.edu

Abstract

A recent line of research focuses on the study of stochastic multi-armed bandits (MAB), in the case where temporal correlations of specific structure are imposed between the player's actions and the reward distributions of the arms. These correlations lead to (sub-)optimal solutions that exhibit interesting dynamical patterns – a phenomenon that yields new challenges both from an algorithmic as well as a learning perspective. In this work, we extend the above direction to a combinatorial semi-bandit setting and study a variant of stochastic MAB, where arms are subject to matroid constraints and each arm becomes unavailable (blocked) for a fixed number of rounds after each play. A natural common generalization of the state-of-the-art for blocking bandits, and that for matroid bandits, only guarantees a $1/2$-approximation for general matroids. In this paper we develop the novel technique of correlated (interleaved) scheduling, which allows us to obtain a polynomial-time $(1-1/e)$-approximation algorithm (asymptotically and in expectation) for any matroid. Along the way, we discover an interesting connection to a variant of Submodular Welfare Maximization, for which we provide (asymptotically) matching upper and lower approximability bounds. In the case where the mean arm rewards are unknown, our technique naturally decouples the scheduling from the learning problem, and thus allows to control the $(1-1/e)$-approximate regret of a UCB-based adaptation of our online algorithm.

1 Introduction

Despite the large number of variants of the stochastic *multi-armed bandits* (MAB) model [46, 33] that have been introduced [8, 34], the majority of the results comply with the common assumption that playing an action does not alter the environment, namely, the reward distributions of the subsequent rounds (with notable exceptions discussed below). Only recently, researchers have focused their attention on settings where temporal dependencies of specific structure are imposed between the player's actions and the reward distributions [27, 10, 7, 39, 6]. In [27], Kleinberg and Immorlica consider the setting of *recharging bandits*, where the expected reward of each arm is a concave and weakly increasing function of the time passed since its last play, modeling in that way scenarios of local performance loss. In a similar spirit, Basu et al. [7] consider the problem of *blocking bandits*, in which case once an arm is played at some round, it cannot be played again (i.e., it becomes blocked) for a fixed number of consecutive rounds. Notice that all the aforementioned examples are variations of the stochastic MAB setting, where the decision maker plays (at most) one arm per time step.

When combinatorial constraints and time dynamics come together, the result is a much richer and more challenging setting, precisely because their interplay creates a complex dynamical structure. Indeed, in the standard combinatorial bandits setting [11], the optimal solution in hindsight is to consistently play the feasible subset of arms of maximum expected reward. However, in the presence

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

of local temporal constraints on the arms, an optimal (or even suboptimal) solution cannot be trivially characterized– a fact that significantly complicates the analysis, both from the algorithmic as well as from the learning perspective. In this work, we study the following bandit setting– a common generalization of *matroid bandits*, introduced by Kveton et al. [30], and blocking bandits [7]:

Problem 1.1 (Matroid Blocking Semi-Bandits (MBB))**.** *We consider a set $\mathcal{A}$ of k* arms*, a* matroid *$\mathcal{M} = (\mathcal{A}, \mathcal{I})$, and an unknown time horizon of T rounds. Each arm $i \in \mathcal{A}$ is associated with an unknown bounded* reward *distribution of mean μ_i, and with a known deterministic* delay *d_i, such that whenever an action i is played at some round, it cannot be played again for the next $d_i - 1$ rounds. At each time step, the* player *pulls a subset of the available (i.e., not blocked) arms restricted to be an independent set of $\mathcal{M}$. Subsequently, she observes the reward realization of each arm played (*semi-bandit *feedback) and collects their sum as the reward for this round. The goal of the player is to maximize her expected cumulative reward over T rounds.*

The above model captures a number of applications, varying from team formation to ad placement, when arms represent actions that cannot be played repeatedly without restriction. As a concrete example, consider a recommendation system that repeatedly suggests a variety of products (e.g., songs, movies, books) to a user. The need for diversity on the collection of suggested products (arms), to capture different aspects of user's preferences, can be modeled as a linear matroid. Further, the blocking constraints preclude the incessant recommendation of the same product (which can be detrimental, as the product might be perceived as a "spam"), while the maximum rate of recommendation (controlled by the delay) might depend on factors such as popularity, promotion and more. Finally, the expected reward of each product is the probability of purchasing (or clicking).

From a technical viewpoint, the MBB problem is already NP-hard for the simple case of a uniform rank-1 matroid (see Theorem 2.1 in [43]), even in the *full-information* setting, where the reward distributions are known to the player a priori. The natural common generalization of the algorithms in [7, 30], computes and plays, at each time step, an independent set of maximum mean reward consisting of the available elements. While the above strategy is a $(1 - 1/e)$-approximation asymptotically (that is, for $T \to \infty$) for partition matroids, unfortunately, it only guarantees a $1/2$-approximation for general matroids [1] and this guarantee is tight (see Appendix E for an example). A natural question that arises is whether a $(1 - 1/e)$-approximation is possible for any matroid.

The main result of this paper shows that this is indeed possible. Along the way, we identify that the key insight (and also the weak point of the naive $1/2$-approximation) is the underlying *diminishing returns* property hidden in the matroid structure. In particular, we discover an interesting connection of MBB to the following problem of interest in its own right:

Problem 1.2 (Recurrent Submodular Welfare (RSW))**.** *We consider a monotone (non-decreasing) submodular function $f : 2^{\mathcal{A}} \to \mathbb{R}_{\geq 0}$ over a universe $\mathcal{A}$ and a time horizon T. At each round $t \in [T]$ we choose a subset $\mathcal{A}_t \subseteq \mathcal{A}$ and collect a reward $f(\mathcal{A}_t)$. However, using an element $i \in \mathcal{A}$ at some round $t \in [T]$ makes it unavailable (i.e., blocked) for a fixed and known number of $d_i - 1$ subsequent rounds, namely, during the interval $[t, t + d_i - 1]$. The objective is to maximize $\sum_{t \in [T]} f(\mathcal{A}_t)$, subject to the blocking constraints, within a (potentially unknown) time horizon T.*

For the above model, which can be thought of as a variant of *Submodular Welfare Maximization* [47], we provide an efficient randomized $(1 - 1/e)$-approximation (asymptotically), accompanied by a matching hardness result. Note that the RSW problem is a very natural model, capturing applications of submodular maximization in repeating scenarios, where the elements cannot be constantly used without restriction. As an example, consider the process of renting goods to a stream of customers with identical submodular utility functions modeling their satisfaction.

As we show, our approach for the RSW problem immediately implies an algorithm of the same approximation guarantee for the full-information case of MBB and, additionally, it has important implications for the *bandit* setting, where the reward distributions are initially unknown. The standard goal in this case is to provide a (sublinear in the time horizon) upper bound on the *regret*, namely, the difference between the expected reward of a bandit algorithm and a (near-)optimal algorithm, due to the initial lack of knowledge of the former[1].

[1]In fact, we upper bound the $(1 - 1/e)$-(approximate) regret, defined as the difference between $(1 - 1/e)\,\mathrm{OPT}(T)$ and the expected reward collected by a bandit algorithm. The notion of α-regret is widely used in the combinatorial bandits literature [15, 48] for combinatorial problems where an efficient algorithm

1.1 Related Work

A recent line of research focuses on non-stationary models in the case where each reward distribution is a special function of the player's actions [10, 39, 6]. In [7], Basu et al. provide a greedy $(1 - 1/e)$-approximation for the full-information case of the blocking bandits problem (a special case of the MBB model for a uniform rank-1 matroid). As we have already mentioned, generalizing their strategy to the MBB problem fails to provide the same guarantee for general matroids. In the bandit setting, where the reward distributions are initially unknown, the authors have to overcome the burden of characterizing a (sub)optimal solution, where the rate of mean collected reward exhibits significant fluctuations over time. The key insight is to observe that every time the full-information algorithm plays an arm, its bandit variant, which relies on estimations of the mean rewards, has at least one chance of playing the same arm. However, this key coupling argument, that enables sublinear regret bounds, becomes significantly more involved in the presence of matroid constraints.

In [27], Kleinberg and Immorlica study the case of recharging bandits. Their approach first computes the "optimal" playing frequency $1/x_i$ of each arm i via a mathematical formulation. In order to play each arm with this frequency, they develop the technique of *interleaved rounding*, where they associate each arm i with a sequence of real numbers $\{(\alpha_i+k)/x_i\}_{k\in\mathbb{N}}$, with $\alpha_i \sim U[0,1]$. Then, the arms are played sequentially in the same order they appear on the real line. This novel rounding technique exhibits reduced variance and, thus, an improved approximation guarantee comparing to other natural approaches such as independent randomized rounding.

The MBB model is also related to the literature on *periodic scheduling* [5, 4]. In [43], Sgall et al. consider the problem of periodically scheduling jobs on a set of machines. Each job is associated with a *processing time*, during which it occupies the machine it is executed on, a *vacation time*, namely, a minimum time required after its completion in order to be rescheduled, and a *reward*. It is not hard to see that the case of unit processing times is a special case of MBB with a uniform matroid of rank equal to the number of machines, under the objective of maximizing the total reward. Further, it is known [7] that the rank-1 case of MBB generalizes the *Pinwheel Scheduling* problem [24]: Given k colors associated a set of integers $\{d_i\}_{i\in[k]}$, such that $\sum_{i\in[k]} 1/d_i = 1$, decide whether there is a coloring of the natural numbers $\nu : \mathbb{N} \to [k]$ such that every color $i \in [k]$ appears at least once every d_i numbers. As it is proved in [25], the above problem does not admit a pseudopolynomial time algorithm unless SAT can be solved by a randomized algorithm in expected quasi-polynomial time.

In a concurrent work [1], the authors consider the blocking bandit model in a generic combinatorial setting and under stochastic delays. As they show, the greedy algorithm that plays at each time the maximum feasible subset of available arms is a $\mathcal{O}(1)$-approximation for downward-closed set systems. However, when specialized to matroids, they cannot do better than a $1/2$-approximation. We need new ideas to reach a $(1 - 1/e)$-approximation algorithm and associated regret guarantees for the rich class of matroid bandits.

Our work is related to the line of research regarding bandit optimization of submodular functions (see [13, 20] and references therein). We refer the reader to Appendix A for additional related work on *non-stationary* bandits, *combinatorial* bandits, and *submodular welfare maximization*.

1.2 Our Contributions

Reducing full-information MBB to RSW. We first focus on the full-information variant of MBB, where the mean rewards of the arms are known to the player a priori. We assume that the player has access to the matroid $\mathcal{M}$ via an independence oracle and knowledge of the arms' fixed delays, yet she is oblivious to the time horizon T. In this sense, she plays *online*. An interesting aspect of dynamics, as illustrated in [27, 7, 6], is that one needs to guarantee, via *scheduling*, that each arm is roughly played at a frequency close to its "optimal" rate. This is particularly important in the presence of "hard" blocking constraints, where no reward can be obtained by a blocked arm.

In order to address the above scheduling problem, we propose a particular "decoupled" *two-phase strategy*. We refer to each phase as (cooperative) Player A and Player B. Initially, Player A decides on a schedule that determines arm availability, namely, a subset of rounds where each arm is allowed to be played. Subsequently, Player B chooses a subset of available arms that maximizes the total

does not exist, and, thus, any efficient algorithm would inevitably suffer linear regret in standard definition (where $\alpha = 1$).

expected reward, subject to the matroid constraints. In order to completely decouple the two phases, the availability schedule produced by Player A is never affected by which arms are eventually chosen by Player B (that is, it is impossible for Player B to violate the blocking constraints).

In the case where Player B knows the expected rewards of the arms and due to the above decoupling property, his optimal strategy (given any availability schedule) can be easily characterized: Since the arms of each round are subject to matroid constraints, Player B achieves his goal by playing a maximum expected reward independent set among the available arms of each round, which can be computed efficiently using the greedy algorithm for matroids. Thus, the role of Player A becomes to choose an availability schedule that maximizes the total reward, knowing that Player B will behave exactly as described above. The key observation is that the solution computed by Player B at each round, corresponds to the *weighted rank function* of the matroid evaluated on the set of available arms of the round. More importantly, it can be proved that this function is monotone submodular and, hence, Player A's task is a special case of the RSW problem.

Optimal approximation for RSW. Focusing our attention on the RSW problem, any "good" solution should guarantee that each element $i \in \mathcal{A}$ is selected a fraction of the time close to $1/d_i$ (the maximum possible), where d_i is the delay. However, a naive randomized approach that selects (if available) each element i with probability $1/d_i$ independently at each round, can be as bad as a $(1 - e^{-1/2}) \approx 0.393$-approximation (see Appendix E for an example). Instead, motivated by the rounding technique of Kleinberg and Immorlica [27], we develop a (time-)correlated sampling strategy, which we call *interleaved scheduling*. While our technique is based on the same principle of transforming (randomly interleaved) sequences of real numbers into a feasible schedule, our implementation is very different. Indeed, as opposed to [27], we additionally face the issue of scheduling more than one arms per round, subject to matroid constraints, and the fact that our "hard" blocking constraints are particularly sensitive to the variance of the produced schedule. Using our technique, we construct a polynomial-time randomized algorithm, named INTERLEAVED-SUBMODULAR (IS), that achieves the following guarantee for RSW:

Theorem 1.3. *The expected reward collected by* INTERLEAVED-SUBMODULAR *over T rounds, $\mathcal{R}^{IS}(T)$, is at least $(1 - 1/e)\,\mathrm{OPT}(T) - \mathcal{O}(d_{\max} f(\mathcal{A}))$, where $\mathrm{OPT}(T)$ is the optimal reward of RSW for T rounds and $d_{\max} = \max_{i \in \mathcal{A}} d_i$ is the maximum delay of the instance.*

The proof of the above guarantee relies on the construction of a *convex program* (CP), based on the *concave closure* of f (see below), that yields an (approximate up to an additive term) upper bound on the optimal reward. Although our algorithm never computes an optimal solution to this convex program, it allows us to compare its expected collected reward with the optimal solution of CP, leveraging known results on the *correlation gap* of submodular functions. As we show via a reduction from the SWM problem with identical utilities, the $(1 - 1/e)$ term in the above guarantee is asymptotically the best possible, unless P = NP; further, the additive term results from the fact that our algorithm is oblivious to the time horizon T.

Bandit algorithm and regret guarantees. We now turn our attention to the *bandit setting* of MBB, where the mean rewards are initially unknown. Our interleaved scheduling method exhibits an additional property: *It does not rely on the monotone submodular function itself*, a fact that is particularly important for the bandit setting. Indeed, in the full-information setting Player B computes a maximum expected reward independent set at each round, for any availability schedule provided by Player A. In the bandit setting, however, the reward distributions are not a priori known and, thus, must be learned. Nevertheless, we do not need to wait to learn these distributions to find a good availability schedule. This allows us to make a natural coupling between the strategy of Player B in the bandit and in the full-information case and, thus, to compare the expected reward collected "pointwise", assuming a fixed common availability schedule. We remark that the above coupling is very different than the one in [7], as ours is independent of the trajectory of the observed rewards.

The above analysis allows us to develop a bandit algorithm for MBB based on the UCB method, called INTERLEAVED-UCB (IB). Specifically, given any availability schedule provided by Player A (independently of the rewards) and in increasing order of rounds, Player B greedily computes a maximal independent set consisting the available arms of each round, based on estimates (known as UCB indices) of the mean rewards. In order to analyze the regret, we use the independence of the availability schedule in combination with the *strong basis exchange* property of matroids. This allows us to decompose the overall regret of our algorithm into contributions from each individual arm.

Once we have established this regret decomposition, we can bound the individual regret attributed to each arm using more standard UCB type arguments [30], leading to the following guarantee:

Theorem 1.4. *The expected reward collected by* INTERLEAVED-UCB *in T rounds, $\mathcal{R}^{IB}(T)$, for k arms, a matroid of rank $r = \mathrm{rk}(\mathcal{M})$ and maximum delay $d_{\max}$ is at least*

$$\left(1 - \frac{1}{e}\right) \mathrm{OPT}(T) - \mathcal{O}\left(k\sqrt{T\ln(T)} + k^2 + d_{\max} r\right).$$

In the above bound, the additive loss corresponds to the regret with respect to $\left(1 - \frac{1}{e}\right) \mathrm{OPT}(T)$. Interestingly, our regret bound is very close (even in constant factors) to the information-theoretically optimal bound provided in [30] for the non-blocking setting. In fact, except for the small additive $\mathcal{O}(d_{\max} r)$ term, the regret bound in [30] is the same as ours, if we replace the number of arms k with $\sqrt{k \cdot r}$. Intuitively, this is due to the fact that our algorithm must learn the complete order of mean rewards, as opposed to the non-blocking setting where learning the maximum expected reward independent set in hindsight is sufficient for eliminating the regret.

All the omitted proofs of our results have been moved to the Appendix.

2 Preliminaries on Matroids and Submodular Functions

Continuous extensions and correlation gap of submodular functions. Consider any set function $f : 2^{\mathcal{A}} \to \mathbb{R}_{\geq 0}$ over a ground set $\mathcal{A}$. Recall that f is submodular, if $\forall S, T \subseteq \mathcal{A}$ we have $f(S \cup T) + f(S \cap T) \leq f(S) + f(T)$. For any point $\mathbf{x} \in [0,1]^k$, we denote by $S \sim \mathbf{x}$ the random set $S \subseteq \mathcal{A}$, such that $\mathbb{P}(i \in S) = x_i$. We consider two canonical continuous extensions of a set function:

Definition 2.1 (Continuous extensions)**.** *For any set function f the* multi-linear extension *is*

$$F(\mathbf{x}) = \mathop{\mathbb{E}}_{S \sim \mathbf{x}}[f(S)] = \sum_{S \subseteq \mathcal{A}} f(S) \prod_{i \in S} x_i \prod_{i \notin S} (1 - x_i).$$

Moreover, the concave closure *is defined as*

$$f^+(\mathbf{x}) = \max_{\alpha} \{ \sum_{S \subseteq \mathcal{A}} \alpha_S f(S) \mid \sum_{S \subseteq \mathcal{A}} \alpha_S \mathbf{1}_S = \mathbf{x}, \sum_{S \subseteq \mathcal{A}} \alpha_S = 1, \alpha \succeq 0 \},$$

where $\mathbf{1}_S \in \{0,1\}^k$ is an indicator vector such that $(\mathbf{1}_S)_i = 1$, if $i \in S$, and $(\mathbf{1}_S)_i = 0$, otherwise.

Lemma 2.2 (Correlation gap [9])**.** *Let $f : 2^k \to \mathbb{R}_{\geq 0}$ be a monotone (non-decreasing) submodular function. Then for any point $\mathbf{x} \in [0,1]^k$, we have*

$$F(\mathbf{x}) \leq f^+(\mathbf{x}) \leq (1 - 1/e)^{-1} F(\mathbf{x}).$$

Matroids and weighted rank functions. Consider a matroid $\mathcal{M} = (\mathcal{A}, \mathcal{I})$, where $\mathcal{A}$ is the *ground set* and $\mathcal{I}$ is the family of *independent sets*. Recall that in any matroid, the family $\mathcal{I}$ satisfies the following two properties: (i) Every subset of an independent set (including the empty set) is an independent set, namely, if $S' \subset S \subseteq \mathcal{A}$ and $S \in \mathcal{I}$, then $S' \in \mathcal{I}$ (*hereditary property*). (ii) Let $S, S' \subseteq \mathcal{A}$ be two independent sets with $|S| < |S'|$, then there exists some $e \in S' \backslash S$ such that $S \cup \{e\} \in \mathcal{I}$ (*augmentation property*). See [42, 37] for more details on matroids.

We assume that access to $\mathcal{M}$ is given through an *independence oracle* [22, 40], namely, a black-box routine that, given a set $S \subseteq \mathcal{A}$, answers whether S is an independent set of $\mathcal{M}$. For any set $R \subset \mathcal{A}$ we define the *restriction* of $\mathcal{M}$ to R, denoted by $\mathcal{M}\,|R$, to be the matroid $\mathcal{M}\,|R = (R, \{I \in \mathcal{I} \mid I \subseteq R\})$.

Given any non-negative linear *weight* vector $\mathbf{w} \in \mathbb{R}^k_{\geq 0}$, the problem of computing a maximum weight independent set can be solved optimally by the standard greedy algorithm: Starting from the empty set $S = \emptyset$, add each ground element $e \in \mathcal{A}$ to the set S in a non-increasing order of weights, as long as the set $S \cup \{e\}$ does not contain a circuit. Given a matroid $\mathcal{M} = (\mathcal{A}, \mathcal{I})$ and a weight vector $\mathbf{w}$, the function $f_{\mathcal{M},\mathbf{w}}(S) = \max_{I \in \mathcal{I}, I \subseteq S} \{\mathbf{w}(I)\}$ is called the *weighted rank function* of $\mathcal{M}$ and returns the weight of the maximum independent set of the restriction $\mathcal{M}\,|S$.

Lemma 2.3 (Weighted rank function [9])**.** *For any matroid $\mathcal{M}$ and non-negative weight vector $\mathbf{w}$, the function $f_{\mathcal{M},\mathbf{w}}(S) = \max_{I \in \mathcal{I}, I \subseteq S} \{\mathbf{w}(I)\}$ is monotone (non-decreasing) submodular.*

Technical notation. For any event $\mathcal{E}$, we denote by $\mathcal{X}(\mathcal{E}) \in \{0,1\}$ the indicator variable such that $\mathcal{X}(\mathcal{E}) = 1$, if $\mathcal{E}$ occurs, and $\mathcal{X}(\mathcal{E}) = 0$, otherwise. For any non-negative integer $n \in \mathbb{N}$, we define $[n] = \{1, 2, \ldots, n\}$. For any vector $\mu \in \mathbb{R}^k$ and set $S \subseteq [k]$, we define $\mu(S) = \sum_{i \in S} \mu_i$. Moreover, we use the notation $t \in [a, b]$ (for $a \leq b$) for some time index t, in place of $t \in [T] \cap [a, \ldots, b\}$. Unless otherwise noted, we use the indices i, j or i' to refer to arms and t, t' or τ to refer to time. Let $\mathcal{A}_t^\pi \in \mathcal{I}$ be the set of arms played by some algorithm $\pi \in \{IS, IG, IB\}$ (defined in Sections 3 and 4) at time t. Unless otherwise noted, all expectations are taken over the randomness of the offsets $\{r_i\}_{i \in [k]}$ (see Section 3) and the reward realizations.

3 Recurrent Submodular Welfare

Let $f(S) : 2^{\mathcal{A}} \to \mathbb{R}_{\geq 0}$ be a monotone submodular function over a universe $\mathcal{A}$ of k elements, such that $f(\emptyset) = 0$. In the *blocking* setting, each element $i \in \mathcal{A}$ is associated with a known deterministic *delay* $d_i \in \mathbb{N}_{>0}$, such that once the arm is played at some round t, it becomes unavailable for the next $d_i - 1$ rounds, namely, in the interval $\{t, \ldots, t + d_i - 1\}$. At each round $t \in [T]$, the player chooses a subset $\mathcal{A}_t$ of available (i.e., non-blocked) elements and collects a reward $f(\mathcal{A}_t)$. The goal is to maximize the total reward collected, i.e., $\sum_{t \in [T]} f(\mathcal{A}_t)$, within an unknown time horizon T.

We provide an efficient randomized $(1 - 1/e)$-approximation algorithm for RSW. Informally, the algorithm starts by considering, for each element $i \in \mathcal{A}$, a sequence of rational numbers of the form $\{t \cdot \frac{1}{d_i}\}_{t \in [T]}$. Then, these sequences are *interleaved* by randomly adding an *offset* r_i, drawn uniformly at random from $[0, 1]$, for each $i \in \mathcal{A}$ to the corresponding sequence. At every round $t \in [T]$, the algorithm chooses a set $\mathcal{A}_t$, consisting only of elements for which the (perturbed) interval $L_{i,t} = [t \cdot \frac{1}{d_i} + r_i, (t+1) \cdot \frac{1}{d_i} + r_i)$ contains an integer.

Algorithm 3.1 (INTERLEAVED-SUBMODULAR (IS))**.** *For each element $i \in \mathcal{A}$, let $r_i \sim U[0, 1]$ be a random* offset *drawn uniformly from $[0, 1]$. At every round $t = 1, 2, \ldots$, let $\mathcal{A}_t \subseteq \mathcal{A}$ be the subset of elements such that for any $i \in \mathcal{A}_t$, the interval $L_{i,t} = [t \cdot \frac{1}{d_i} + r_i, (t+1) \cdot \frac{1}{d_i} + r_i)$ contains an integer. Choose the elements $\mathcal{A}_t$ and collect the reward $f(\mathcal{A}_t)$.*

3.1 Correctness and approximation guarantee.

We first show the algorithm is correct, namely, that the elements chosen at each round respect the blocking constraints. The correctness is established by the following simple observation:

Fact 3.2. *At any $t \in [T]$, all the elements in $\mathcal{A}_t$ are available (i.e., not blocked).*

In order to prove the competitive guarantee of our algorithm, we first construct a convex programming (CP)-based (approximate) upper bound on the optimal reward. Although our algorithm never computes an optimal solution to this CP, this step allows us to prove our guarantee, leveraging results on the correlation gap of submodular functions. For $\boldsymbol{d}^{-1} \in \mathbb{R}^k$ such that $(\boldsymbol{d}^{-1})_i = 1/d_i, \forall i \in [k]$, consider the following formulation based on the concave closure f^+ of f:

$$\underset{\mathbf{z} \in \mathbb{R}^k}{\textbf{maximize}}: \; T \cdot f^+(\mathbf{z}) \;\; \textbf{s.t.} \;\; \mathbf{0} \preceq \mathbf{z} \preceq \boldsymbol{d}^{-1}. \tag{\textbf{CP}}$$

In (**CP**), each variable z_i can be thought of as the fraction of rounds where element $i \in \mathcal{A}$ is chosen. Intuitively, the constraints indicate the fact that, due to the blocking, each element $i \in \mathcal{A}$ can be played at most once every d_i steps. In order to derive (**CP**), we start from a non-convex integer program (IP) with 0-1 variables $\{x_{i,t}\}_{i \in \mathcal{A}, t \in [T]}$, each indicating whether element $i \in \mathcal{A}$ is used at round $t \in [T]$. The objective is to maximize $\sum_{t \in [T]} \sum_{S \subseteq \mathcal{A}} f(S) \prod_{i \in S} x_{i,t} \prod_{i \notin S} (1 - x_{i,t})$ subject to natural blocking constraints. For integral solutions, the above objective is equivalent to $\sum_{t \in [T]} f^+(\mathbf{x}_t)$ (where $(\mathbf{x}_t)_i = x_{i,t}$) and, thus, the above relaxation is simply the result of averaging over time the variables and constraints of this IP. By using the concavity of f^+, we are able to show that (**CP**) yields an (approximate) upper bound on the optimal solution of RSW, while the approximation becomes exact as T increases.

Lemma 3.3. *Let $\mathcal{R}^{CP}(T)$ be the optimal solution to* (**CP**) *and* $\mathrm{OPT}(T)$ *be the optimal solution over T rounds. We have $\mathcal{R}^{CP}(T) \geq \mathrm{OPT}(T) - \mathcal{O}(d_{\max} f(\mathcal{A}))$, where $d_{\max} = \max_{i \in \mathcal{A}} \{d_i\}$.*

Before we complete the proof of our first main result, we first compute the probability that $i \in \mathcal{A}_t$, i.e., an element $i \in \mathcal{A}$ is sampled at round $t \in [T]$:

Fact 3.4. *For any $i \in \mathcal{A}$ and $t \in [T]$, we have $\mathbb{P}(i \in \mathcal{A}_t) = \mathbb{P}(L_{i,t} \cap \mathbb{N} \neq \emptyset) = 1/d_i$.*

Proof of Theorem 1.3. Let us denote by $S \sim \mathbf{p}$ with $\mathbf{p} \in [0,1]^k$ the random set $S \subseteq \mathcal{A}$, where each element i participates in S independently with probability equal to p_i. By Fact 3.4 and due to the randomness of the offsets $\{r_i\}_{i \in \mathcal{A}}$, we have that $\mathcal{A}_t \sim \mathbf{d}^{-1}$ for each $t \in [T]$. Let $\mathbf{z}^*$ be an optimal solution to (**CP**). By monotonicity of f and the fact that $\mathbf{z}^* \preceq \boldsymbol{d}^{-1}$, for the expected value of $f(\mathcal{A}_t)$ at any round $t \in [T]$, we know that $\mathop{\mathbb{E}}_{\mathcal{A}_t \sim \boldsymbol{d}^{-1}}[f(\mathcal{A}_t)] \geq \mathop{\mathbb{E}}_{\mathcal{A}_t \sim \mathbf{z}^*}[f(\mathcal{A}_t)]$. Moreover, by definition of the multi-linear extension, we have that $\mathop{\mathbb{E}}_{\mathcal{A}_t \sim \mathbf{z}^*}[f(\mathcal{A}_t)] = F(\mathbf{z}^*)$, while by Lemma 2.2 (the correlation gap of submodular functions), we have that, $F(\mathbf{z}) \geq \left(1 - \frac{1}{e}\right) f^+(\mathbf{z})$ for any vector $\mathbf{z} \in [0,1]^k$. By combining the above facts, we can see that

$$\mathcal{R}^{IS}(T) = \sum_{t \in [T]} \mathop{\mathbb{E}}_{\mathcal{A}_t \sim \boldsymbol{d}^{-1}}[f(\mathcal{A}_t)] \geq \sum_{t \in [T]} F(\mathbf{z}^*) \geq \left(1 - \frac{1}{e}\right) T \cdot f^+(\mathbf{z}^*) = \left(1 - \frac{1}{e}\right) \mathcal{R}^{CP}(T).$$

Therefore, by Lemma 3.3, we can conclude that $\mathcal{R}^{IS}(T) \geq \left(1 - \frac{1}{e}\right) \mathrm{OPT}(T) - \mathcal{O}(d_{\max} f(\mathcal{A}))$. $\square$

In Appendix C.2, we provide a $(1 - 1/e)$-hardness result for RSW, thus proving that the guarantee of Theorem 1.3 is asymptotically tight. This result, which holds even for the special case where $d_{\max} = o(T)$ (that is when the delays are significantly smaller than the time horizon), is proved via a reduction from the SWM problem with identical utilities, in a way that the constructed RSW instance accepts w.l.o.g. solutions of a simple periodic structure.

Theorem 3.5. *For any $\epsilon > 0$, there exists no polynomial-time $\left(1 - \frac{1}{e} + \epsilon\right)$-approximation algorithm for the RSW problem, unless $\mathbf{P} = \mathbf{NP}$, even in the special case where $d_{\max} = o(T)$.*

4 Matroid Blocking Semi-Bandits

Let $\mathcal{A}$ be a set of k arms and T be an unknown time horizon. At any round $t \in [T]$ and for each $i \in \mathcal{A}$ a reward $X_{i,t}$ is drawn independently from an unknown distribution of mean μ_i and bounded support in $[0,1]$. Let $d_i \in \mathbb{N}_{>0}$ be the known determinisitc delay of each arm $i \in \mathcal{A}$, and $d_{\max} = \max_{i \in \mathcal{A}}\{d_i\}$. At any round $t \in [T]$, the player pulls any subset $\mathcal{A}_t$ of the available (i.e., non-blocked) arms, as long as it forms an independent set of a given matroid $\mathcal{M} = (\mathcal{A}, \mathcal{I})$. The player only observes the realized reward of each arm she plays and collects their sum. The goal is to maximize the *expected cumulative reward* collected within T rounds, denoted by $\mathcal{R}^{IG}(T) = \mathbb{E}\left[\sum_{t \in [T]} \sum_{i \in \mathcal{A}} X_{i,t}\, \mathcal{X}(i \in \mathcal{A}_t)\right]$.

The full-information setting The following algorithm is the implementation of IS in the special case of the full-information MBB setting, where the mean rewards $\{\mu_i\}_{i \in \mathcal{A}}$ are known a priori:

Algorithm 4.1 (INTERLEAVED-GREEDY (IG))**.** *For each arm $i \in \mathcal{A}$, let $r_i \sim U[0,1]$ be a random* offset *drawn uniformly from $[0,1]$. At every round $t = 1, 2, \ldots$, let $\mathcal{G}_t \subseteq \mathcal{A}$ be the subset of arms $i \in \mathcal{A}$, such that the interval $L_{i,t} = [t \cdot \frac{1}{d_i} + r_i, (t+1) \cdot \frac{1}{d_i} + r_i)$ contains an integer. Greedily compute a maximum independent set $\mathcal{A}_t$ of $\mathcal{M} \,|\, \mathcal{G}_t$ with respect to $\{\mu_i\}_{i \in \mathcal{G}_t}$ and play these arms.*

The following result is an immediate corollary of Theorem 1.3, given that the value of the greedily computed maximum independent set in $\mathcal{M} \,|\, \mathcal{G}_t$ corresponds to the weighted rank function $f_{\mathcal{M},\mu}(\mathcal{G}_t)$ which, by Lemma 2.3, is monotone submodular:

Theorem 4.2. *The expected reward collected by* INTERLEAVED-GREEDY *for T rounds, $\mathcal{R}^{IG}(T)$, is at least $\left(1 - \frac{1}{e}\right) \mathrm{OPT}(T) - \mathcal{O}(d_{\max} \operatorname{rk}(\mathcal{M}))$, where $\mathrm{OPT}(T)$ is the optimal expected reward.*

Remark 4.3. *The analysis of IG is tight for rank-1 matroids. Indeed, consider k arms, each of delay k and deterministic reward equal to 1. For $T \to \infty$, the optimal average reward is equal to 1, simply by playing the arms in a round-robin manner. However, the probability that at least one arm is sampled at some round t is $\sum_{i=1}^{k} \binom{k}{i} \left(\frac{1}{k}\right)^i \left(1 - \frac{1}{k}\right)^{k-i} = 1 - \left(1 - \frac{1}{k}\right)^k \to 1 - \frac{1}{e}$ as $k \to \infty$.*

The bandit setting and regret analysis In the setting where the mean rewards are initially unknown, we develop a UCB-based bandit algorithm, INTERLEAVED-UCB (IB). The algorithm is identical to IG, except for the greedy computation of the maximum independent set over the sampled

arms, which is now performed using estimates. Specifically, the algorithm maintains for every $i \in \mathcal{A}$, $t \in [T]$ the following upper estimate of μ_i:

$$\bar{\mu}_{i,t} = \hat{\mu}_{i,T_i(t)} + c_{i,t} \text{ with } c_{i,t} = \sqrt{\frac{2\ln(t)}{T_i(t)}},$$

where $T_i(t)$ denotes the number of times arm i has been played at the beginning of round t and $\hat{\mu}_{i,T_i(t)}$ denotes the empirical average of the $T_i(t)$ i.i.d. samples from its reward distribution. The term $c_{i,t}$ is the *confidence length* around $\hat{\mu}_{i,T_i(t)}$ that guarantees $\bar{\mu}_{i,t}$ lies in $[\mu_i, \mu_i + 2c_{i,t}]$ with high probability. Note that all the above quantities are random variables depending on the random offsets and the observed reward realizations.

We are interested in upper bounding the α-regret, for $\alpha = 1 - \frac{1}{e}$, namely, the difference between $\alpha\text{OPT}(T)$ and the expected reward collected by IB. Due to the complex time dynamics, characterizing the optimal expected reward as a function of the instance is hard. However, using Theorem 4.2 we can upper bound $\alpha\text{OPT}(T)$ by the expected reward collected by IG, thus giving:

$$\alpha\text{OPT}(T) - \mathcal{R}^{UCB}(T) \leq \mathcal{R}^{IG}(T) - \mathcal{R}^{UCB}(T) + \mathcal{O}(d_{\max} \cdot \text{rk}(\mathcal{M})). \tag{1}$$

By the above inequality, it becomes clear that in order to upper bound the regret, it suffices to bound the difference between the expected reward collected by IG and IB. This difference not only depends on the reward realizations (through the UCB estimates), but also on the trajectory of sampled arms in each algorithm, which is itself a function of the random offsets. However, by construction of our interleaved scheduling scheme, these offsets are sampled at the initialization phase of each algorithm and are identically distributed. Thus, the trajectories of sampled arms in the two algorithms exhibit a coupled evolution. This allows us to analyse the regret "pointwise", under the assumption that the sequences of sampled arms are identical throughout the time horizon. To make this idea precise, let $\mathbf{r}^\pi \in [0,1]^k$ be the random offsets used and let $\{\mathcal{G}_t^\pi(\mathbf{r}^\pi)\}_{t\in[T]}$ be the sequence of sampled arms by algorithm $\pi \in \{IG, IB\}$. Using (henceforth) $\mathcal{Q}$ to denote the randomness due to the reward realizations of the arms, the next lemma gives our pointwise regret bound.

Lemma 4.4. *Let $\bar{\mu}_t(S) = \sum_{i\in S} \bar{\mu}_{i,t}$ and $\mu(S) = \sum_{i\in S} \mu_i$. We have*

$$\mathcal{R}^{IG}(T) - \mathcal{R}^{IB}(T) = \mathop{\mathbb{E}}_{\mathbf{r}\sim U[0,1]^k, \mathcal{Q}} \left[\sum_{t\in[T]} \left(\max_{S\subseteq\mathcal{G}_t(\mathbf{r}), S\in\mathcal{I}} \{\mu(S)\} - \mu\left(\arg\max_{S\subseteq\mathcal{G}_t(\mathbf{r}), S\in\mathcal{I}} \{\bar{\mu}_t(S)\} \right) \right) \right].$$

Thus w.l.o.g., we focus on the case where the sequences of sampled arms are identical. Let $\mathcal{E}_\mathbf{r}$ denote the event that both algorithms, IG and IB, sample the same offset vector $\mathbf{r}$, namely, $\mathbf{r}^{IG} = \mathbf{r}^{IB} = \mathbf{r}$. Assuming that $\mathcal{E}_\mathbf{r}$ holds for some $\mathbf{r} \in [0,1]^k$, let $\{\mathcal{G}_t\}_{t\in[T]} = \{\mathcal{G}_t(\mathbf{r})\}_{t\in[T]}$ be the sequence of sampled arms, common in both algorithms. Clearly, IB accumulates regret only when it plays independent sets of arms that are suboptimal w.r.t. the true means, i.e., when $\mu(\mathcal{A}_t^{IB}) < \mu(\mathcal{A}_t^{IG})$ for some $t \in [T]$. We assume w.l.o.g. that the arms are indexed in decreasing order of mean rewards and that these mean rewards are distinct. We now formally define the gaps related to our analysis:

Definition 4.5 (Gaps)**.** *For any subset $S \subseteq \mathcal{A}$ and reward vector $\nu \in \mathbb{R}^k$, we define*

$$\Delta_S(\nu) = \max_{I\in\mathcal{I}, I\subseteq S} \{\mu(I)\} - \mu\left(\arg\max_{B\in\mathcal{I}, B\subseteq S} \{\nu(B)\} \right).$$

Moreover, let $\Delta_{i,j} = \mu_i - \mu_j$ be the standard suboptimality gap between two arms $i, j \in \mathcal{A}$.

By Lemma 4.4 and assuming that the event $\mathcal{E}_\mathbf{r}$ holds for some $\mathbf{r}$, we are interested in bounding the expectation of $\sum_{t\in[T]} \Delta_{\mathcal{G}_t(\mathbf{r})}(\bar{\mu}_t)$ w.r.t. the reward realizations. The next step is to decompose the suboptimality of IB by noticing that both algorithms play, at each round $t \in [T]$, a basis of $\mathcal{M} \,|\, \mathcal{G}_t$ and thus $|\,\mathcal{A}_t^{IG}\,| = |\,\mathcal{A}_t^{IB}\,|$. We use the following fundamental property of matroids:

Theorem 4.6 (Strong Basis Exchange, Corollary 39.12a in [42])**.** *Let $\mathcal{M} = (\mathcal{A}, \mathcal{I})$ be a matroid and $I_1, I_2 \in \mathcal{I}$ be two independent sets such that $|I_1| = |I_2|$. Then, there exists a bijection $\sigma : I_1 \to I_2$, such that for any $i \in I_1$ the set $I_1 - i + \sigma(i)$ is an independent set of $\mathcal{M}$.*

Let $\sigma_t : \mathcal{A}_t^{IB} \to \mathcal{A}_t^{IG}$ for each $t \in [T]$ be the bijection described in Theorem 4.6 with respect to the sets $\mathcal{A}_t^{IB}$ and $\mathcal{A}_t^{IG}$ and let σ_t^{-1} be its inverse mapping. Note that in any bijection σ_t and any $i \in \mathcal{A}_t^{IB} \cap \mathcal{A}_t^{IG}$ we can assume w.l.o.g. that $\sigma_t(i) = i$. Notice, further, that under the event $\mathcal{E}_\mathbf{r}$, the bijections $\{\sigma_t\}_{t\in[T]}$ are still random variables that depend on the observed realizations.

Lemma 4.7. *Under the event $\mathcal{E}_{\mathbf{r}}$ and at any time $t \in [T]$, we have $\Delta_{\mathcal{G}_t}(\bar{\mu}_t) = \sum_{i \in \mathcal{A}_t^{IG}} \Delta_{i,\sigma_t^{-1}(i)}$.*

Conditioned on the fact that both algorithms operate on the same sequence $\{\mathcal{G}_t\}_{t\in[T]}$ of sampled arms, Lemma 4.7 allows us to decompose the suboptimality gap $\Delta_{\mathcal{G}_t}(\bar{\mu}_t)$ of each round $t \in [T]$, into simpler gaps of the form $\Delta_{i,j}$ between any arms $i \in \mathcal{A}_t^{IG}$ and $j \in \mathcal{A}_t^{IB}$ that are perfectly matched according to the bijection σ_t, namely, $\sigma_t(j) = i$. Assuming that the event $\{\sigma_t(j) = i\}$ directly implies that $i \in A_t^{IG}$ and $j \in A_t^{IB}$, we can further upper bound the regret as

$$\sum_{t\in[T]} \Delta_{\mathcal{G}_t}(\bar{\mu}_t) = \sum_{t\in[T]} \sum_{i\in\mathcal{A}_t^{IG}} \Delta_{i,\sigma_t^{-1}(i)} \leq \sum_{t\in[T]} \sum_{i\in\mathcal{A}_t^{IG}} \sum_{j\in\mathcal{A},\Delta_{i,j}>0} \Delta_{i,j}\,\mathcal{X}\left(\sigma_t(j) = i\right).$$

The above inequality allows us to study the regret attributed to each arm independently, using more standard arguments for UCB-based algorithms in combination with Theorem 4.6. Specifically, for every pair of arms $i, j \in \mathcal{A}$ with $i < j$ (thus, $\Delta_{i,j} > 0$), we define a threshold $\ell_{i,j}$ with the following key-property: After IB "exchanges" arm j for arm $i = \sigma_t(j)$ more than $\ell_{i,j}$ times, due to insufficient exploration, then it has collected enough samples to infer that $\mu_j < \mu_i$ with high probability.

Lemma 4.8. *Let $\ell_{i,j} = \left\lfloor \frac{8\ln(T)}{\Delta_{i,j}^2} \right\rfloor$ for any $i < j$. Under event $\mathcal{E}_{\mathbf{r}}$ and for any arm $j > 1$, we have*

$$\sum_{t\in[T]} \sum_{i<j} \Delta_{i,j}\,\mathcal{X}\left(\sigma_t(j) = i, T_j(t) \leq \ell_{i,j}\right) \leq \frac{16}{\Delta_{j-1,j}} \ln(T) \qquad \textit{(Under-sampled regret)} \quad (2)$$

$$\mathop{\mathbb{E}}_{\mathcal{Q}}\left[\sum_{t\in[T]} \sum_{i<j} \Delta_{i,j}\,\mathcal{X}\left(\sigma_t(j) = i, T_j(t) > \ell_{i,j}\right)\right] \leq \frac{\pi^2}{3} \sum_{i=1}^{j-1} \Delta_{i,j} \qquad \textit{(Sufficiently sampled regret)} \quad (3)$$

Proof sketch of Theorem 1.4. By inequality (1) and Lemma 4.4, in order to bound the regret of IB, it suffices to upper bound the difference between $\mathcal{R}^{IG}(T)$ and $\mathcal{R}^{IB}(T)$, conditioned on the fact that both algorithms use exactly the same offset vector $\mathbf{r}$ and, thus, they operate on the exact same sequence of sampled arms, denoted by $\{\mathcal{G}_t\}_{t\in[T]}$. By construction, IG plays at any round $t \in [T]$ a basis of $\mathcal{M} \mid \mathcal{G}_t$ of maximum expected reward, while IB plays a basis of $\mathcal{M} \mid \mathcal{G}_t$ that is maximum with respect to the estimates $\{\bar{\mu}_{i,t}\}_{i\in\mathcal{A}}$. By Theorem 4.6, we can consider a perfect matching between exchangeable arms of $\mathcal{A}_t^{IG}$ and $\mathcal{A}_t^{IB}$ and, thus, to decompose the regret into suboptimality gaps between individual arms. Then, using Lemma 4.8, we can upper bound on the expected regret due to the fact that IB erroneously plays arm j instead of arm i, when $\Delta_{i,j} > 0$. The above analysis culminates in the following *gap-dependent* regret upper bound:

$$\sum_{j>1} \frac{16}{\Delta_{j-1,j}} \ln(T) + \frac{\pi^2}{3} \sum_{j>1} \sum_{i=1}^{j-1} \Delta_{i,j} + \mathcal{O}(d_{\max} \cdot \mathrm{rk}(\mathcal{M})) \quad \text{(gap-dependent regret)}.$$

In order to derive a gap-independent regret bound, we partition the gaps into "small" and "large" and notice that any pair of arms $i, j \in \mathcal{A}$ with $\Delta_{i,j} < \Theta(\sqrt{\frac{\ln(T)}{T}})$ cannot contribute more than $\sqrt{T\ln(T)}$ loss in the regret. □

Conclusion and Further Directions

We explore the effect of action-reward dependencies in the combinatorial MAB setting by introducing and studying the MBB problem. After relating the problem to RSW, we provide a $(1 - 1/e)$-approximation for its full-information case, based on the technique of interleaved scheduling. Importantly, our technique is oblivious to the reward distributions of the arms– a fact that allows us to provide regret bounds of optimal dependence in T, when these distributions are initially unknown. We believe that this idea could be further applied to different classes of (combinatorial) non-stationary bandits, other than blocking bandits.

Our work leaves behind numerous interesting questions. By exhaustive search over $\mathcal{O}(1)$-periodic schedules, one can construct a PTAS for the (asymptotic) MBB problem, assuming *constant* $\mathrm{rk}(\mathcal{M})$ and $\{d_i\}_{i\in[k]}$. It remains an open question, however, whether the $(1 - 1/e)$-approximation is the

best possible in general. We remark that the hardness of MBB cannot solely rely on an argument similar to Theorem 3.5, since the welfare maximization problem for the class of *gross substitutes*, which includes weighted matroid rank functions, is easy [35]. Finally, it is easy to show that our algorithm gives a $\mathcal{O}(1)$-approximation for the case of stochastic delays. Whether we can recover a $(1 - 1/e)$-approximation in this case is an interesting open question.

Acknowledgements

The authors would like to thank an anonymous reviewer of a previous version of this work for an unusually thoughtful and helpful review, which aided us in improving the document — in particular, for pointing out the idea of correlated rounding. Further, the authors would like to thank Jannik Matuschke for noticing that the weighted matroid rank function falls into the class of gross substitutes.

Funding Transparency Statement

This research was partially supported by NSF Grant 2019844. In addition, this research was sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-19-2-0333. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] A. Atsidakou, O. Papadigenopoulos, S. Basu, C. Caramanis, and S. Shakkottai. Combinatorial blocking bandits with stochastic delays. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 404–413. PMLR, 18–24 Jul 2021.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2–3):235–256, May 2002.

[3] A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with knapsacks. *J. ACM*, 65(3), March 2018.

[4] A. Bar-Noy, R. Bhatia, J. Naor, and B. Schieber. Minimizing service and operation costs of periodic scheduling. *Mathematics of Operations Research*, 27(3):518–544, 2002.

[5] A. Bar-Noy, R. E. Ladner, and T. Tamir. Windows scheduling as a restricted version of bin packing. *ACM Trans. Algorithms*, 3(3):28–es, August 2007.

[6] S. Basu, O. Papadigenopoulos, C. Caramanis, and S. Shakkottai. Contextual blocking bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 271–279. PMLR, 2021.

[7] S. Basu, R. Sen, S. Sanghavi, and S. Shakkottai. Blocking bandits. In *Advances in Neural Information Processing Systems (NeurIPS) 32*, pages 4785–4794. Curran Associates, Inc., 2019.

[8] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Machine Learning*, 5(1):1–122, 2012.

[9] G. Calinescu, C. Chekuri, M. Pál, and J. Vondrák. Maximizing a submodular set function subject to a matroid constraint (extended abstract). In Matteo Fischetti and David P. Williamson, editors, *Integer Programming and Combinatorial Optimization*, pages 182–196, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[10] L. Cella and N. Cesa-Bianchi. Stochastic bandits with delay-dependent payoffs. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1168–1177, Online, 26–28 Aug 2020. PMLR.

[11] N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404 – 1422, 2012. JCSS Special Issue: Cloud Computing 2011.

[12] L. Chen, A. Gupta, and J. Li. Pure exploration of multi-armed bandit under matroid constraints. *Proceeding of the 29th Annual Conference on Learning Theory (COLT 2016)*, 2016.

[13] L. Chen, A. Krause, and A. Karbasi. Interactive submodular bandit. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[14] W. Chen, W. Hu, F. Li, J. Li, Y. Liu, and P. Lu. Combinatorial multi-armed bandit with general reward functions. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 1659–1667, Red Hook, NY, USA, 2016. Curran Associates Inc.

[15] W. Chen, Y. Wang, Y. Yuan, and Q. Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *J. Mach. Learn. Res.*, 17(1):1746–1778, January 2016.

[16] R. Combes, C. Jiang, and R. Srikant. Bandits with budgets: Regret lower bounds and optimal algorithms. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '15, page 245–257, New York, NY, USA, 2015. Association for Computing Machinery.

[17] R. Combes, M. S. Talebi, A. Proutiere, and M. Lelarge. Combinatorial bandits revisited. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2116–2124, Cambridge, MA, USA, 2015. MIT Press.

[18] U. Feige and J. Vondrák. The submodular welfare problem with demand queries. *Theory Comput.*, 6:247–290, 2010.

[19] J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, pages 148–177, 1979.

[20] D. Golovin and A. Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *J. Artificial Intelligence Research*, 2011.

[21] S. Guha, K. Munagala, and P. Shi. Approximation algorithms for restless bandit problems. *J. ACM*, 58(1), December 2010.

[22] D. Hausmann and B. Korte. *Algorithmic versus axiomatic definitions of matroids*, pages 98–111. Springer Berlin Heidelberg, Berlin, Heidelberg, 1981.

[23] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[24] R. Holte, A. Mok, L. Rosier, I. Tulchinsky, and D Varvel. Pinwheel: a real-time scheduling problem. volume 2, pages 693 – 702 vol.2, 02 1989.

[25] T. Jacobs and S. Longo. A new perspective on the windows scheduling problem. *CoRR*, abs/1410.7237, 2014.

[26] S. Khot, R. J. Lipton, E. Markakis, and A. Mehta. Inapproximability results for combinatorial auctions with submodular utility functions. In *Proceedings of the First International Conference on Internet and Network Economics*, WINE'05, page 92–101, Berlin, Heidelberg, 2005. Springer-Verlag.

[27] R. Kleinberg and N. Immorlica. Recharging bandits. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 309–319, 2018.

[28] R. Kleinberg, A. Niculescu-Mizil, and Y. Sharma. Regret bounds for sleeping experts and bandits. *Mach. Learn.*, 80(2–3):245–272, September 2010.

[29] N. Korula, V. Mirrokni, and M. Zadimoghaddam. Online submodular welfare maximization: Greedy beats 1/2 in random order. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 889–898, New York, NY, USA, 2015. Association for Computing Machinery.

[30] B. Kveton, Z. Wen, A. Ashkan, H. Eydgahi, and B. Eriksson. Matroid bandits: Fast combinatorial optimization with learning. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, page 420–429, Arlington, Virginia, USA, 2014. AUAI Press.

[31] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvári. Combinatorial cascading bandits. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1450–1458, Cambridge, MA, USA, 2015. MIT Press.

[32] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari. Tight Regret Bounds for Stochastic Combinatorial Semi-Bandits. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 535–543, San Diego, California, USA, 09–12 May 2015. PMLR.

[33] T.L Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6(1):4–22, March 1985.

[34] T. Lattimore and C. Szepesvári. Bandit algorithms. *preprint*, page 28, 2018.

[35] R. P. Leme. Gross substitutability: An algorithmic survey. *Games Econ. Behav.*, 106:294–316, 2017.

[36] V. Mirrokni, M. Schapira, and J. Vondrak. Tight information-theoretic lower bounds for welfare maximization in combinatorial auctions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, EC '08, page 70–77, New York, NY, USA, 2008. Association for Computing Machinery.

[37] J. G. Oxley. *Matroid Theory (Oxford Graduate Texts in Mathematics)*. Oxford University Press, Inc., New York, NY, USA, 2006.

[38] C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999.

[39] C. Pike-Burke and S. Grunewalder. Recovering bandits. In *Advances in Neural Information Processing Systems 32*, pages 14122–14131. 2019.

[40] G. C. Robinson and D. J. A. Welsh. The computational complexity of matroid properties. *Mathematical Proceedings of the Cambridge Philosophical Society*, 87(1):29–45, 1980.

[41] K. A. Sankararaman and A. Slivkins. Combinatorial semi-bandits with knapsacks. In *AISTATS*, 2018.

[42] A. Schrijver. *Combinatorial Optimization - Polyhedra and Efficiency*. Springer, 2003.

[43] J. Sgall, H. Shachnai, and T. Tamir. Periodic scheduling with obligatory vacations. *Theoretical Computer Science*, 410(47):5112 – 5121, 2009.

[44] A. Slivkins. Dynamic ad allocation: Bandits with budgets. *ArXiv*, abs/1306.0155, 2013.

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

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

[47] J. Vondrak. Optimal approximation for the submodular welfare problem in the value oracle model. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, page 67–74, New York, NY, USA, 2008. Association for Computing Machinery.

[48] Q. Wang and W. Chen. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. In *Advances in Neural Information Processing Systems*, pages 1161–1171, 2017.

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

A Further Related Work

The MBB model belongs to the family of stochastic *non-stationary* bandits, given that the reward distributions of the arms can change over time. Significant members of this family are *restless bandits* [49, 21], where the reward distribution of each arm changes at each time step, and *rested bandits* [19, 45], where the distribution changes only when the arm is played. For the setting of restless bandits and without further assumptions on the transition functions, it is PSPACE-hard to even approximate the optimal solution [38]. Our model differs from the above cases as we consider a transition function of special form and the transitions can occur both during playing and not playing an arm. In addition, the MBB model falls into the category of Markov Decision Processes (MDPs) with deterministic transitions and stochastic rewards, but requires an exponential (in the size of the arms) state space, which makes this approach inefficient in practice.

A rich body of research on combinatorial bandits [17, 15, 14, 32, 31, 48] focuses on bandit optimization problems over general combinatorial structures. In [30], Kveton et al. consider the problem of stochastic combinatorial bandits where the underlying feasible set is a matroid defined over the ground set of arms. At each round, the player pulls an independent subset of arms and collects their realized rewards, assuming *semi-bandit* feedback (as opposed to the pure exploration *full-feedback* variant studied in [12]). The authors develop a greedy algorithm based on the Upper Confidence Bound (UCB) method [2], while they exploit well-known exchange properties of matroids for achieving optimal regret bounds. Their approach relies on the fact that the optimal solution in hindsight is fixed throughout the time horizon– a fact that is no longer true in the presence of blocking constraints. Additional lines of research that are related to, yet incompatible with, our problem are *bandits with knapsacks* [3, 41] or *with budgets* [16, 44], and *sleeping bandits* [28].

The RSW problem is closely related to the problem of *Submodular Welfare Maximization* (SWM) [47, 36, 26, 18]: Given k items and m players, each associated with a monotone submodular utility function $u_i : 2^{[k]} \to \mathbb{R}_{\geq 0}$, the goal is to partition the elements into m sets $S_1, \ldots, S_m$, one for each player, such that to maximize $\sum_{i \in [m]} u_i(S_i)$. Specifically, RSW can be thought of as a version of the SWM problem, when the items are distributed to a (possibly infinite) stream of players with identical utilities, and each item can be reused after some fixed time period (note that this is different than the online setting in [29]). Interestingly, as noted in [47], the SWM problem with identical utilities is *approximation resistant* in the sense that allocating the items to the players uniformly at random achieves the optimal approximation guarantee of $\left(1 - \frac{1}{e}\right)$ for this setting.

B Concentration inequalities

Theorem B.1 (Hoeffding's Inequality [23])**.** *Let $X_1, \ldots, X_n$ be independent identically distributed random variables with common support in $[0, 1]$ and mean μ. Let $Y = X_1 + \cdots + X_n$. Then for any $\delta \geq 0$,*

$$\mathbb{P}\left(Y - n\mu \geq \delta\right) \leq e^{-2\delta^2/n} \text{ and } \mathbb{P}\left(Y - n\mu \leq -\delta\right) \leq e^{-2\delta^2/n}.$$

C Recurrent Submodular Welfare: Omitted Proofs

C.1 Correctness and approximation guarantee

Fact 3.2. *At any $t \in [T]$, all the elements in $\mathcal{A}_t$ are available (i.e., not blocked).*

Proof. Recall that at any round $t \in [T]$, the algorithm only chooses a subset $\mathcal{A}_t$ of the elements. Consider any element $i \in \mathcal{A}$ such that $i \in \mathcal{A}_t$ for some $t \in [T]$. By definition of $\mathcal{A}_t$, the interval $L_{i,t} = [t \cdot \frac{1}{d_i} + r_i, (t+1) \cdot \frac{1}{d_i} + r_i)$ contains an integer. It is not hard to see that, in that case, none of the intervals $L_{i,t'}$ for $t' \in [t - d_i + 1, d_i - 1]$ can contain an integer. Therefore, the last time element i has been chosen must be before $t - d_i$, which implies feasibility with respect to the blocking constraints. □

Fact 3.4. *For any $i \in \mathcal{A}$ and $t \in [T]$, we have $\mathbb{P}\left(i \in \mathcal{A}_t\right) = \mathbb{P}\left(L_{i,t} \cap \mathbb{N} \neq \emptyset\right) = 1/d_i$.*

Proof. For any fixed $i \in \mathcal{A}$ and $t \in [T]$, because of the fact that $\frac{1}{d_i} \leq 1$ and $r_i \in [0,1]$, the interval $L_{i,t} = [t \cdot \frac{1}{d_i} + r_i, (t+1) \cdot \frac{1}{d_i} + r_i)$ clearly contains at most one integral point. The event that $\{[t \cdot \frac{1}{d_i} + r_i, (t+1) \cdot \frac{1}{d_i} + r_i) \cap \mathbb{N} \neq \emptyset\}$ is equivalent to the event that a continuous window of size equal to $\frac{1}{d_i}$ starting from the (real) point $t \cdot \frac{1}{d_i} + r_i$ contains an integer. For r_i ranging in $[0,1]$, the starting point of the interval lies between $t \cdot \frac{1}{d_i}$ and $t \cdot \frac{1}{d_i} + 1$. It is not hard to see that fraction of possible realizations of r_i such that the window contains an integer equals its size. The fact follows since for any $i \in \mathcal{A}$, the window has size $\frac{1}{d_i}$ and the offset r_i is sampled uniformly at random from $[0,1]$. □

Lemma 3.3. *Let $\mathcal{R}^{CP}(T)$ be the optimal solution to* (**CP**) *and* $\text{OPT}(T)$ *be the optimal solution over T rounds. We have $\mathcal{R}^{CP}(T) \geq \text{OPT}(T) - \mathcal{O}(d_{\max} f(\mathcal{A}))$, where $d_{\max} = \max_{i \in \mathcal{A}}\{d_i\}$.*

Proof. In order to prove the lemma, we first construct an (non-convex) IP upper bound on the optimal expected reward over T rounds, based on the multi-linear extension of f.

$$\textbf{maximize:} \sum_{i \in [T]} \sum_{S \subseteq \mathcal{A}} f(S) \prod_{i \in S} x_{i,t} \prod_{i \notin S} (1 - x_{i,t}) \tag{MP}$$

$$\textbf{s.t.} \sum_{t' \in [t, t+d_i-1]} x_{i,t'} \leq 1, \forall i \in \mathcal{A}, \forall t \in [T] \tag{4}$$

$$\mathbf{x}_t \in \{0,1\}^k, \forall t \in [T]$$

In the formulation (**MP**), each variable $x_{i,t}$ can be thought of as the 0-1 indicator of playing arm $i \in \mathcal{A}$ at time $t \in [T]$. Intuitively, constraints (4) of (**MP**) indicate the fact that, due to blocking constraints, each arm $i \in \mathcal{A}$ can be played at most once every d_i steps. Clearly, any optimal solution to RSW can be mapped onto the above formulation and, thus, the optimal solution of (**MP**) provides an upper bound on $\text{OPT}(T)$.

Let $\mathbf{x}_t \in \{0,1\}^k$ for each $t \in [T]$ be a vector such that $(\mathbf{x}_t)_i = x_{i,t}$. Notice that for any integral $\mathbf{x} \in \{0,1\}^k$, the multi-linear extension is equal to the concave closure of any set function f, that is, $f^+(\mathbf{x}) = F(\mathbf{x})$. Therefore, (**MP**) remains an upper bound, even if we replace its objective function with $g(\mathbf{x}_1, \ldots, \mathbf{x}_T) = \sum_{t \in [T]} f^+(\mathbf{x}_t)$.

We now fix any optimal solution $\{x^*_{i,t}\}_{i \in \mathcal{A}, t \in [T]}$ to (**MP**) under the objective $g(\mathbf{x}_1, \ldots, \mathbf{x}_T) = \sum_{t \in [T]} f^+(\mathbf{x}_t)$. Let us define the variables $\{z'_i\}_{i \in \mathcal{A}}$, such that

$$z'_i = \frac{1}{T} \sum_{t \in [T]} x^*_{i,t} \geq 0, \quad \forall i \in \mathcal{A}\,.$$

In the above definition, each z'_i is the fraction of time an element $i \in \mathcal{A}$ is chosen in an optimal solution. Let $\mathbf{z}' \in [0,1]^k$, such that $(\mathbf{z}')_i = z'_i \; \forall i \in \mathcal{A}$.

By concavity of f^+, we have

$$g(\mathbf{x}^*_1, \ldots, \mathbf{x}^*_T) = \sum_{t \in [T]} f^+(\mathbf{x}^*_t) = T \sum_{t \in [T]} \frac{1}{T} f^+(\mathbf{x}^*_t) \leq T f^+(\frac{1}{T} \sum_{t \in [T]} \mathbf{x}^*_t) = T f^+(\mathbf{z}'),$$

where the inequality follows by the fact that $\mathbf{z}'$ can be thought of as a convex combination of $\{\mathbf{x}^*_1, \ldots, \mathbf{x}^*_T\}$.

Moreover, for each $i \in \mathcal{A}$ and by averaging constraints (4) of (**MP**) over all $t \in [T]$, we can see that

$$\frac{1}{T} \sum_{t \in [1, d_i - 1]} t x^*_{i,t} + \frac{1}{T} \sum_{t \in [d_i, T]} d_i x^*_{i,t} \leq 1 \Leftrightarrow \frac{1}{T} \sum_{t \in [T]} d_i x^*_{i,t} \leq 1 + \frac{1}{T} \sum_{t \in [1, d_i - 1]} (d_i - t) x^*_{i,t}.$$

Given the fact that $\sum_{t \in [1, d_i - 1]} x^*_{i,j} \leq 1$, the above inequality immediately implies that

$$z'_i \leq \frac{1}{d_i} \left(1 + \frac{d_i - 1}{T}\right) \quad \forall i \in \mathcal{A}\,.$$

Consider now the assignment $z_i = \left(1 + \frac{d_{\max}-1}{T}\right)^{-1} z'_i, \forall i \in \mathcal{A}$. For this assignment, we can easily verify that the constraints of (**CP**) are trivially satisfied, since $0 \leq z_i \leq \frac{1}{d_i}, \forall i \in \mathcal{A}$.

Let $\mathbf{z} \in [0,1]^k$, such that $(\mathbf{z})_i = z_i \; \forall i \in \mathcal{A}$. By the above analysis, we can see that

$$\mathbf{z} = \mathbf{z}' - \frac{d_{\max}-1}{T + d_{\max} - 1}\, \mathbf{z}',$$

where we use the fact that $\frac{1}{1+\beta} = 1 - \frac{\beta}{1+\beta}$ for any $\beta \in \mathbb{R}$. Finally, by concavity of f^+ we have

$$\begin{aligned} f^+(\mathbf{z}) &= f^+\left(\left(1 - \frac{d_{\max}-1}{T + d_{\max} - 1}\right)\mathbf{z}' + \frac{d_{\max}-1}{T + d_{\max} - 1}\mathbf{0}\right) \\ &\geq \left(1 - \frac{d_{\max}-1}{T + d_{\max} - 1}\right) f^+(\mathbf{z}') + \frac{d_{\max}-1}{T + d_{\max} - 1} f^+(\mathbf{0}) \\ &\geq f^+(\mathbf{z}') - \frac{d_{\max}-1}{T + d_{\max} - 1} f(\mathcal{A}), \end{aligned}$$

where the last inequality follows by the facts that $f^+(\mathbf{0}) = f(\mathbf{0}) = 0$ and $f^+(\mathbf{z}') \leq f^+(\mathbf{1}) = f(\mathcal{A})$, since f is monotone.

Therefore, by exhibiting a feasible solution $\mathbf{z}$ of (**CP**) such that

$$Tf^+(\mathbf{z}) \geq Tf^+(\mathbf{z}') - \mathcal{O}(d_{\max} f(\mathcal{A})) \geq g(\mathbf{x}_1^*, \ldots, \mathbf{x}_T^*) - \mathcal{O}(d_{\max} f(\mathcal{A})) \geq \mathrm{OPT}(T) - \mathcal{O}(d_{\max} f(\mathcal{A})),$$

the proof is completed. □

C.2 Hardness of approximation

The goal of this section is to show that the $\left(1 - \frac{1}{e}\right)$-multiplicative factor in the approximation guarantee of Theorem 1.3 cannot be improved, unless $\mathbf{P} = \mathbf{NP}$. Specifically, we prove the following result:

Theorem 3.5. *For any $\epsilon > 0$, there exists no polynomial-time $\left(1 - \frac{1}{e} + \epsilon\right)$-approximation algorithm for the RSW problem, unless $\mathbf{P} = \mathbf{NP}$, even in the special case where $d_{\max} = o(T)$.*

In order show the above hardness result, we study for simplicity the average version of RSW, where the objective is to maximize the average reward over T time steps, namely, $\frac{1}{T}\left(\sum_{t\in[T]} f(\mathcal{A}_t)\right)$, where $\mathcal{A}_t$ is the set of elements used at time $t \in [T]$. Notice that in the average case, the additive term in the approximation guarantee of INTERLEAVED-GREEDY, as presented in Theorem 1.3, vanishes as $T \to \infty$. Let OPT be the average reward collected by any optimal algorithm for RSW.

Our proof relies on a reduction from the Submodular Welfare (SW) problem [47], in the special case where the players have identical utility functions. The problem can be formally defined as follows:

Definition C.1 (Submodular Welfare with Identical Utilities (SWIU))**.** *We consider a set of k items and m players, each associated with the same monotone submodular utility function $u : 2^{[k]} \to \mathbb{R}_{\geq 0}$ over the items. The goal is to partition the k items into m subsets $S_1, \ldots, S_m$, such that to maximize $\sum_{i\in[m]} u(S_i)$.*

As noted in [47], the hardness result presented in [26] for the SW problem also holds for SWIU, namely, the special case of SW where all the players have the same utility function. Note, also that the RSW problem is defined in the *value oracle* model, as we are only allowed to make queries of the function value for any input set.

Theorem C.2 ([26])**.** *For any $\epsilon > 0$, there exists no polynomial-time $\left(1 - \frac{1}{e} + \epsilon\right)$-approximation algorithm for the SWIU problem in the value oracle model, unless $\mathbf{P} = \mathbf{NP}$.*

We start from a simple construction for the non-average case of RSW in order to show how our problem is directly associated with SWIU: Consider an instance of SWIU of k items and m players. Let $u : 2^{[k]} \to \mathbb{R}_{\geq 0}$ be the monotone submodular utility function which is commonly used by all players. Given the above instance, we can construct in polynomial time an instance of RSW as follows: Let $\mathcal{A}$ be the set of k elements, each corresponding to an item, and let $f : 2^{\mathcal{A}} \to \mathbb{R}_{\geq 0}$ be our

function, chosen such that $f \equiv u$. We set the delay of each element $i \in \mathcal{A}$ as well as the time horizon to be equal to the number of players, namely, $d_i = T = m$ for each $i \in \mathcal{A}$.

Clearly, in the above construction where the delays are all equal to the time horizon, each element can be chosen at most once by any algorithm for RSW. Therefore, the above constructed instance of RSW exactly corresponds to SWIU, given that any solution to latter immediately translates into a solution of RSW of the same total reward, and the opposite.

The above construction immediately relates the two problems in the case where the delays can be of the same order as the time horizon. However, it does not rule out the possibility that the RSW problem might become easier in the special case where $d_{\max} = o(T)$. Indeed, one could argue that for small enough delays, exploiting the possible periodicity of the RSW solutions might lead to improved approximation guarantees. Notice, further, that the approximation guarantee we provide in Theorem 1.3 for IS becomes meaningless in the above scenario, since the additive loss for $d_{\max} = T$ becomes $\mathcal{O}(T \cdot f(\mathcal{A}))$.

In order to overcome the above technical issue and show that the multiplicative factor of $\left(1 - \frac{1}{e}\right)$ in Theorem 1.3 cannot be improved, we map any instance of SWIU onto an instance of RSW such that $T \gg d_{\max}$. Given any instance of SWIU, we can construct in polynomial time an instance of RSW as follows: We define $\mathcal{A}$ to be the set of k items, $f \equiv u$ to be the monotone submodular function and $d_i = m \; \forall i \in \mathcal{A}$ to be the delay of all elements. In this case, we consider a time horizon $T = m \cdot \lceil \text{poly}(k, m) \rceil$, where by $\text{poly}(k, m)$ we denote some polynomial function in k and m.

We first show that, without loss of generality, we can focus our attention on solutions to the average case of RSW that exhibit a periodic structure of period m.

Lemma C.3. *Let $\nu : [T] \to 2^{\mathcal{A}}$ be any feasible assignment to the above instance of RSW of average reward R. We can construct in polynomial time a feasible assignment $\nu' : [T] \to 2^{\mathcal{A}}$ of average reward at least $R' \geq R$, such that $\nu'(t) = \nu(t + m) \; \forall t \in \mathbb{N}$, namely, ν' is a periodic assignment of period m.*

Proof. Given that the average reward of the assignment ν is R, there must exist a continuous subsequence of rounds of length m, that is, $\{t, \ldots, t + m - 1\}$ for some $t \in [T - m]$, such that

$$\frac{1}{m} \sum_{\tau=t}^{t+m-1} f(\nu(t)) \geq R.$$

In the opposite case, we immediately get a contradiction to the fact that the average reward is at least R.

Let L with $|L| = m$ be such a sequence. We now construct the periodic assignment ν' by repeating the assignment of the subinterval L, as follows:

$$\nu'(t) = \nu(L(t \mod m)) \in 2^{\mathcal{A}} \; \forall t \in [T].$$

It is not hard to verify that since $d_i = m$ for each $i \in \mathcal{A}$ and since L is a subsequence of a feasible assignment of length m, the assignment ν' never violates the blocking constraints. Moreover, the average reward of ν' equals the average reward of the interval L which is at least R. Finally, notice that the subsequence L can be found in polynomial time, given the fact that the time horizon T is defined to be polynomial in k and m. □

We can now complete the proof of our hardness result.

Proof of Theorem 3.5. We prove the result via a reduction from the SWIU problem to the average version of the RSW. Clearly, the average and non-average version of RSW share the same approximability status, as the two problems are essentially identical up to a scaling of the objective function.

Given an instance I of SWIU, we can construct in polynomial time an instance I' of the average version of RSW, as described above. Let $\text{OPT}_{SWIU}(I)$ and $\text{OPT}_{RSW}(I')$ be the optimal solution of SWIU and RSW on the corresponding instance, respectively.

We first show that when $\text{OPT}_{SWIU}(I) \geq R$ for some reward R, then we necessarily have that $\text{OPT}_{RSW}(I') \geq \frac{R}{m}$. Indeed, let $L : [m] \to 2^{[k]}$ be an allocation that achieves a reward $R' =$

$\text{OPT}_{SWIU}(I) \geq R$ for the instance I of SWIU. As indicated in proof of Lemma C.3, we can construct in polynomial time a periodic assignment for the RSW problem of average reward exactly $\frac{R'}{m}$, which implies that $\text{OPT}_{RSW}(I') \geq \frac{R'}{m} \geq \frac{R}{m}$.

Now, we would like to show that if $\text{OPT}_{SWIU}(I) \leq \alpha R$ for some reward R and $\alpha \in (0,1)$, then it has to be that $\text{OPT}_{RSW}(I') \leq \alpha\frac{R}{m}$. We prove the statement via its contrapositive, assuming that $\text{OPT}_{RSW}(I') > \alpha\frac{R}{m}$ for some reward R and $\alpha \in (0,1)$. Let $\frac{R'}{m} > \alpha\frac{R}{m}$ be the optimal average reward of RSW. By Lemma C.3, we can assume w.l.o.g. that the assignment $\text{OPT}_{RSW}(I')$, that achieves an average reward of $\frac{R'}{m}$, is a periodic assignment of period m. However, given that all the delays are equal to m in the instance I' of RSW, it is easy to see that in any period of m consecutive rounds, each element is played at most once. Moreover, the average reward of each period is exactly $\frac{R'}{m}$. Therefore, any continuous subsequence of length m in the solution of the RSW naturally induces a solution to the instance I of SWIU of total reward exactly R'. This, in turn, implies that $\text{OPT}_{SWIU}(I) \geq R' \geq \alpha R$.

By the above discussion, we have completed the proof of a reduction from SWIU to RSW. Therefore, any polynomial-time $\left(1 - \frac{1}{e} + \epsilon\right)$-approximation algorithm for RSW, for some $\epsilon > 0$, would imply a $\left(1 - \frac{1}{e} + \epsilon\right)$-approximation algorithm for SWIU. However, by Theorem C.2 this is not possible, unless $\mathbf{P} = \mathbf{NP}$. □

We believe that, through a similar reduction as above, we can prove information-theoretic hardness of the RSW problem by leveraging the results in [36]. We leave this as future work.

D Matroid Blocking Semi-Bandits: Omitted Proofs

Theorem 4.2. *The expected reward collected by* INTERLEAVED-GREEDY *for T rounds, $\mathcal{R}^{IG}(T)$, is at least $\left(1 - \frac{1}{e}\right)\text{OPT}(T) - \mathcal{O}(d_{\max}\,\text{rk}(\mathcal{M}))$, where $\text{OPT}(T)$ is the optimal expected reward.*

Proof. Fix any algorithm for the MBB problem and let $\mathcal{A}_t$ be the set of arms played at round t. Notice that the sets $\{\mathcal{A}_t\}_{t\in[T]}$ are independent of the reward realizations, since the selection of arms pulled at each round is made before observing their actual rewards. Thus, the expected reward collected (over the randomness of the reward realizations) can be expressed as

$$\mathbb{E}\left[\sum_{t\in[T]}\sum_{i\in\mathcal{A}_t} X_{i,t}\right] = \sum_{t\in[T]}\sum_{i\in\mathcal{A}_t}\mathbb{E}\left[X_{i,t}\right] = \sum_{t\in[T]}\sum_{i\in\mathcal{A}_t}\mu_i.$$

Therefore, INTERLEAVED-GREEDY can be thought of as an instance of INTERLEAVED-SUBMODULAR for the weighted rank function of the given matroid, that is, for $f_{\mathcal{M},\mu}(S) = \max_{I\in\mathcal{I}, I\subseteq S}\{\mu(I)\}$. By Lemma 2.3, this function is monotone submodular and, also, $f_{\mathcal{M},\mu}(\mathcal{A}) \leq \text{rk}(\mathcal{M})$, given that the distribution of rewards is bounded in $[0,1]$.

Thus, by applying Theorem 1.3, we can conclude that

$$\mathcal{R}^{IG}(T) \geq \left(1 - \frac{1}{e}\right)\mathcal{R}^{LP}(T) \geq \left(1 - \frac{1}{e}\right)\text{OPT}(T) - \mathcal{O}(d_{\max}\,\text{rk}(\mathcal{M})).$$

□

Lemma 4.4. *Let $\bar{\mu}_t(S) = \sum_{i\in S}\bar{\mu}_{i,t}$ and $\mu(S) = \sum_{i\in S}\mu_i$. We have*

$$\mathcal{R}^{IG}(T) - \mathcal{R}^{IB}(T) = \mathop{\mathbb{E}}_{\mathbf{r}\sim U[0,1]^k,\mathcal{Q}}\left[\sum_{t\in[T]}\left(\max_{S\subseteq\mathcal{G}_t(\mathbf{r}),S\in\mathcal{I}}\{\mu\left(S\right)\} - \mu\left(\arg\max_{S\subseteq\mathcal{G}_t(\mathbf{r}),S\in\mathcal{I}}\{\bar{\mu}_t(S)\}\right)\right)\right].$$

Proof. Let $\{\mathcal{G}_t(\mathbf{r})\}_{t\in[T]}$ be the sequence of sampled arms over T rounds as a function of the sampled offsets $\mathbf{r} \in [0,1]^k$. Moreover, let $X_t(S)$ be the realized rewards of a subset $S \subseteq \mathcal{A}$ of arms at round $t \in [T]$. We denote by $\mathcal{A}_t^\pi$ the arms played at round $t \in [T]$ and by $H_t^\pi = \{\mathcal{A}_1^\pi, X_1(\mathcal{A}_1^\pi), \ldots, \mathcal{A}_t^\pi, X_t(\mathcal{A}_t^\pi)\}$ the *history* of arm playing and observed realizations up to (and

including) time t by algorithm $\pi \in \{IG, IB\}$. Recall that we denote by $\mathcal{Q}$ the randomness due to the reward realizations of the arms.

Notice that in the case of IB and for fixed offsets, the player's actions only depend on the previous realized rewards of the arms. Thus, for any fixed offset vector $\mathbf{r}^{IB}$, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{Q}}\left[\sum_{i\in\mathcal{A}} X_{i,t}\,\mathcal{X}\left(i \in \arg\max_{S\subseteq\mathcal{G}_t(\mathbf{r}^{IB}),S\in\mathcal{I}}\{\bar{\mu}_t(S)\}\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{Q}}\left[\sum_{i\in\mathcal{A}} \mathop{\mathbb{E}}_{\mathcal{Q}}\left[X_{i,t}\,\mathcal{X}\left(i \in \arg\max_{S\subseteq\mathcal{G}_t(\mathbf{r}^{IB}),S\in\mathcal{I}}\{\bar{\mu}_t(S)\}\right) \mid H_{t-1}^{IB}\right]\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{Q}}\left[\sum_{i\in\mathcal{A}} \mathop{\mathbb{E}}_{\mathcal{Q}}\left[X_{i,t} \mid H_{t-1}^{IB}\right]\mathcal{X}\left(i \in \arg\max_{S\subseteq\mathcal{G}_t(\mathbf{r}^{IB}),S\in\mathcal{I}}\{\bar{\mu}_t(S)\}\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{Q}}\left[\sum_{i\in\mathcal{A}} \mu_i\,\mathcal{X}\left(i \in \arg\max_{S\subseteq\mathcal{G}_t(\mathbf{r}^{IB}),S\in\mathcal{I}}\{\bar{\mu}_t(S)\}\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{Q}}\left[\mu\left(\arg\max_{S\subseteq\mathcal{G}_t(\mathbf{r}^{IB}),S\in\mathcal{I}}\{\bar{\mu}_t(S)\}\right)\right].
\end{aligned}$$

Similarly, notice that the algorithm IG is oblivious to the realized rewards. Therefore, for any fixed offset vector $\mathbf{r}^{IG}$ and at any time $t \in [T]$, we get

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{Q}}\left[\sum_{i\in\mathcal{A}} X_{i,t}\,\mathcal{X}\left(i \in \arg\max_{S\subseteq\mathcal{G}_t(\mathbf{r}^{IG}),S\in\mathcal{I}}\{\mu(S)\}\right)\right] &= \mathop{\mathbb{E}}_{\mathcal{Q}}\left[\sum_{i\in\mathcal{A}} \mu_i\,\mathcal{X}\left(i \in \arg\max_{S\subseteq\mathcal{G}_t(\mathbf{r}^{IG}),S\in\mathcal{I}}\{\mu(S)\}\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{Q}}\left[\max_{S\subseteq\mathcal{G}_t(\mathbf{r}^{IG}),S\in\mathcal{I}}\{\mu(S)\}\right].
\end{aligned}$$

The lemma follows by observing that the offsets $\mathbf{r}^{IG}$ and $\mathbf{r}^{IB}$ of the two algorithms follow exactly the same distribution. Therefore, we have

$$\begin{aligned}
&\mathcal{R}^{IG}(T) - \mathcal{R}^{IB}(T) \\
&= \mathop{\mathbb{E}}_{\mathbf{r}^{IG}\sim[0,1]^k,\mathcal{Q}}\left[\sum_{t\in[T]} \max_{S\subseteq\mathcal{G}_t(\mathbf{r}^{IG}),S\in\mathcal{I}}\{\mu(S)\}\right] - \mathop{\mathbb{E}}_{\mathbf{r}^{IB}\sim[0,1]^k,\mathcal{Q}}\left[\sum_{t\in[T]} \mu\left(\arg\max_{S\subseteq\mathcal{G}_t(\mathbf{r}^{IB}),S\in\mathcal{I}}\{\bar{\mu}_t(S)\}\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathbf{r}\sim[0,1]^k,\mathcal{Q}}\left[\sum_{t\in[T]}\left(\max_{S\subseteq\mathcal{G}_t(\mathbf{r}),S\in\mathcal{I}}\{\mu(S)\} - \mu\left(\arg\max_{S\subseteq\mathcal{G}_t(\mathbf{r}),S\in\mathcal{I}}\{\bar{\mu}_t(S)\}\right)\right)\right].
\end{aligned}$$

□

Lemma 4.7. *Under the event $\mathcal{E}_{\mathbf{r}}$ and at any time $t \in [T]$, we have $\Delta_{\mathcal{G}_t}(\bar{\mu}_t) = \sum_{i\in\mathcal{A}_t^{IG}} \Delta_{i,\sigma_t^{-1}(i)}$.*

Proof. Recall that under the event $\mathcal{E}_{\mathbf{r}}$, both algorithms IG and IB use the same offset vector $\mathbf{r}$ and, thus, they operate on same sequence of sampled arms over time. Let $\mathcal{G}_t = \mathcal{G}_t(\mathbf{r})$ be the common set of sampled arms and let $\mathcal{A}_t^{IG}$ and $\mathcal{A}_t^{IB}$ be the maximal independent sets computed by IG and IB, respectively, at any round $t \in [T]$. Notice that for any $t \in [T]$ both $\mathcal{A}_t^{IG}$ and $\mathcal{A}_t^{IB}$ are bases of the restricted matroid $\mathcal{M} \mid \mathcal{G}_t$ and, thus, correspond to independent sets of $\mathcal{I}$ of equal cardinality. Let σ_t be the bijection between $\mathcal{A}_t^{IG}$ and $\mathcal{A}_t^{IB}$ described by Theorem 4.6. For any $t \in [T]$, we have that

$$\Delta_{\mathcal{G}_t}(\bar{\mu}) = \mu(\mathcal{A}_t^{IG}) - \mu(\mathcal{A}_t^{IB}) = \sum_{i\in\mathcal{A}_t^{IG}} \mu_i - \sum_{j\in\mathcal{A}_t^{IB}} \mu_j = \sum_{i\in\mathcal{A}_t^{IG}} \left(\mu_i - \mu_{\sigma_t^{-1}(i)}\right) = \sum_{i\in\mathcal{A}_t^{IG}} \Delta_{i,\sigma_t^{-1}(i)}.$$

□

Lemma 4.8. *Let $\ell_{i,j} = \left\lfloor \frac{8\ln(T)}{\Delta_{i,j}^2} \right\rfloor$ for any $i < j$. Under event $\mathcal{E}_{\mathbf{r}}$ and for any arm $j > 1$, we have*

$$\sum_{t\in[T]} \sum_{i<j} \Delta_{i,j}\, \mathcal{X}\left(\sigma_t(j) = i, T_j(t) \leq \ell_{i,j}\right) \leq \frac{16}{\Delta_{j-1,j}} \ln(T) \qquad \textit{(Under-sampled regret)} \quad (2)$$

$$\mathop{\mathbb{E}}_{\mathcal{Q}}\left[\sum_{t\in[T]} \sum_{i<j} \Delta_{i,j}\, \mathcal{X}\left(\sigma_t(j) = i, T_j(t) > \ell_{i,j}\right)\right] \leq \frac{\pi^2}{3} \sum_{i=1}^{j-1} \Delta_{i,j} \qquad \textit{(Sufficiently sampled regret)} \quad (3)$$

Proof. We first focus on proving inequality (2), that is, the part of the regret attributed to an arm $j > 1$ when not enough samples have been collected. Notice that the algorithm IB never accumulates regret when it plays the arm $j = 1$ of highest mean reward. Recall that for any fixed $j \in \mathcal{A}$, we have $\Delta_{1,j} > \Delta_{2,j} > \cdots > \Delta_{j,j} = 0$, since we assume w.l.o.g. that the arms have distinct mean rewards. By construction of our algorithm, if the number of samples from arm $j \in \mathcal{A}$ is increased at some round t, it is because there exists exactly one arm $i \in \mathcal{A}$ with $\Delta_{i,j} > 0$, such that $\sigma_t(j) = i$. The above is implied by Theorem 4.6, given the fact that each bijection σ_t for all $t \in [T]$ maps each arm played by IB in $\mathcal{A}_t^{IB}$ to a single arm played by IG in $\mathcal{A}_t^{IG}$. On the other hand, as the number of obtained samples $T_j(t)$ from arm $j \in \mathcal{A}$ by time $t \in [T]$ increases, the maximum suboptimality gap $\Delta_{i,j}$ that can be charged in the under-sampled part of the regret is that of the maximum reward $i \in \mathcal{A}$ that satisfies $T_j(t) \leq \ell_{i,j}$. By the above analysis, for any $j > 1$, we get that

$$\begin{aligned}\sum_{t\in[T]} \sum_{i=1}^{j-1} \Delta_{i,j}\, \mathcal{X}\left(\sigma_t(j) = i, T_j(t) \leq \ell_{i,j}\right) &\leq \sum_{i=1}^{j-1} \left(\Delta_{i,j} - \Delta_{i+1,j}\right) \ell_{i,j} \\ &\leq \sum_{i=1}^{j-1} \left(\Delta_{i,j} - \Delta_{i+1,j}\right) \frac{8\ln(T)}{\Delta_{i,j}^2}, \end{aligned} \quad (5)$$

where the last inequality follows by definition of $\ell_{i,j}$.

The rest of the claim follows by simple algebra. Indeed,

$$\begin{aligned}(5) &\leq \left(\sum_{i=1}^{j-1} \frac{\Delta_{i,j} - \Delta_{i+1,j}}{\Delta_{i,j}^2}\right) 8\ln(T) \\ &\leq \left(\frac{1}{\Delta_{j-1,j}} + \sum_{i=1}^{j-2} \frac{\Delta_{i,j} - \Delta_{i+1,j}}{\Delta_{i,j}^2}\right) 8\ln(T) \\ &\leq \left(\frac{1}{\Delta_{j-1,j}} + \sum_{i=1}^{j-2} \frac{\Delta_{i,j} - \Delta_{i+1,j}}{\Delta_{i,j}\Delta_{i+1,j}}\right) 8\ln(T) \\ &= \left(\frac{1}{\Delta_{j-1,j}} + \sum_{i=1}^{j-2} \left(\frac{1}{\Delta_{i+1,j}} - \frac{1}{\Delta_{i,j}}\right)\right) 8\ln(T) \\ &= \left(\frac{2}{\Delta_{j-1,j}} - \frac{1}{\Delta_{1,j}}\right) 8\ln(T) \\ &\leq \frac{16}{\Delta_{j-1,j}} \ln(T). \end{aligned}$$

We now focus on proving inequality (3), that is, the regret accumulated after a sufficient number of samples has been collected from an arm $j > 1$. Notice, that given the event $\mathcal{E}_{\mathbf{r}}$, the expectation in the LHS of inequality (3) is taken only over the randomness of the realized rewards that are observed by IB.

For proving the upper bound, we fix any arm $j > 1$ and focus on each arm $i \in \mathcal{A}$ such that $i < j$ and, thus, $\Delta_{i,j} > 0$. Let us fix any such arm $i \in \mathcal{A}$. For any $t \in [T]$, the event $\{\sigma_t(j) = i\}$ implies that $\{\mu_i > \mu_j, \bar{\mu}_{i,t} \leq \bar{\mu}_{j,t}\}$, namely, the order of the UCB-indices at time $t \in [T]$ of i and j is inconsistent with the order of their true mean rewards. In the opposite case, the algorithm IB would

have chosen the set $\mathcal{A}_t^{IB} - j + i$, which, as suggested by Theorem 4.6, is an independent set of $\mathcal{M}$. Therefore, for any arm $i < j$, we have

$$\{\sigma_t(j) = i, T_j(t) > \ell_{i,j}\} \subseteq \{\bar{\mu}_{i,t} \leq \bar{\mu}_{j,t}, \mu_i > \mu_j, T_j(t) > \ell_{i,j}\}. \tag{6}$$

Note that the inclusion in the above expression is because the inconsistency in the order of UCB-indices does not necessarily imply that $\sigma_t(j) = i$ (i.e., that IB actually exchanges j for i at time $t \in [T]$).

By definition of the UCB-indices, the event $\bar{\mu}_{i,t} \leq \bar{\mu}_{j,t}$ at time $t \in [T]$ implies that

$$\hat{\mu}_{i,T_i(t)} + \sqrt{\frac{2\ln(t)}{T_i(t)}} \leq \hat{\mu}_{j,T_j(t)} + \sqrt{\frac{2\ln(t)}{T_j(t)}}. \tag{7}$$

We fix $s_i = T_i(t)$ and $s_j = T_j(t) > \ell_{i,j}$ to be the number of samples obtained from arm i and j, respectively, by time $t \in [T]$. Notice that in order for (7) to hold, at least one of the following events must be true:

(i) $\left\{\hat{\mu}_{i,s_i} + \sqrt{\frac{2\ln(t)}{s_i}} \leq \mu_i\right\}$, **(ii)** $\left\{\hat{\mu}_{j,s_j} \geq \mu_j + \sqrt{\frac{2\ln(t)}{s_j}}\right\}$, **(iii)** $\left\{\mu_i < \mu_j + 2\sqrt{\frac{2\ln(t)}{s_j}}\right\}$.

Indeed, it can be easily verified that the simultaneous negation of the above three events contradicts (7) for any fixed number of samples s_i, s_j.

By our choice of $\ell_{i,j} = \left\lfloor \frac{8\ln(T)}{\Delta_{i,j}^2} \right\rfloor$ and the fact that $s_j \geq \ell_{i,j} + 1 \geq \frac{8\ln(T)}{\Delta_{i,j}^2}$, we can see that event **(iii)** cannot be true, since in that case, we have

$$\mu_j + 2\sqrt{\frac{2\ln(t)}{s_j}} \leq \mu_j + 2\sqrt{\frac{2\Delta_{i,j}^2\ln(t)}{8\ln(T)}} \leq \mu_j + \Delta_{i,j} = \mu_i.$$

Moreover, by Hoeffding's inequality, for the probabilities of the events **(i)** and **(ii)**, we have that

$$\mathbb{P}\left(\hat{\mu}_{i,s_i} + \sqrt{\frac{2\ln(t)}{s_i}} \leq \mu_i\right) \leq e^{-4\ln(t)} = t^{-4} \text{ and } \mathbb{P}\left(\hat{\mu}_{j,s_j} \geq \mu_j + \sqrt{\frac{2\ln(t)}{s_j}}\right) \leq e^{-4\ln(t)} = t^{-4},$$

where the probability is taken over the randomness of the reward realizations.

Therefore, for any numbers of samples $s_i = T_i(t)$ and $s_j = T_j(t) > \ell_{i,j}$, we have

$$\begin{aligned}\mathbb{P}\left(\bar{\mu}_{i,t} \leq \bar{\mu}_{j,t}, \mu_i > \mu_j, T_j(t) = s_j, T_i(t) = s_i\right) &\leq \mathbb{P}\left(\hat{\mu}_{i,s_i} + \sqrt{\frac{2\ln(t)}{s_i}} \leq \mu_i\right) + \mathbb{P}\left(\hat{\mu}_{j,s_j} \geq \mu_j + \sqrt{\frac{2\ln(t)}{s_j}}\right) \\ &\leq 2 \cdot t^{-4}. \end{aligned} \tag{8}$$

Finally, by union bound over the possible number of samples, s_i and s_j, and using the aforementioned results, for any $j > 1$ and time $t \in [T]$, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{Q}}\left[\sum_{t\in[T]}\sum_{i=1}^{j-1}\Delta_{i,j}\,\mathcal{X}\left(\sigma_t(j) = i, T_j(t) > \ell_{i,j}\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{Q}}\left[\sum_{t\in[T]}\sum_{i=1}^{j-1}\sum_{s_i=0}^{t-1}\sum_{s_j=\ell_{i,j}+1}^{t-1}\Delta_{i,j}\,\mathcal{X}\left(\sigma_t(j) = i, T_j(t) = s_j, T_i(t) = s_i\right)\right] &(9)\\
&\leq \mathop{\mathbb{E}}_{\mathcal{Q}}\left[\sum_{t\in[T]}\sum_{i=1}^{j-1}\sum_{s_i=0}^{t-1}\sum_{s_j=\ell_{i,j}+1}^{t-1}\Delta_{i,j}\,\mathcal{X}\left(\bar{\mu}_{i,t} \leq \bar{\mu}_{j,t}, \mu_i > \mu_j, T_j(t) = s_j, T_i(t) = s_i\right)\right] &(10)\\
&= \sum_{t\in[T]}\sum_{i=1}^{j-1}\sum_{s_i=0}^{t-1}\sum_{s_j=\ell_{i,j}+1}^{t-1}\Delta_{i,j}\,\mathbb{P}\left(\bar{\mu}_{i,t} \leq \bar{\mu}_{j,t}, \mu_i > \mu_j, T_j(t) = s_j, T_i(t) = s_i\right) \\
&\leq \sum_{t\in[T]}\sum_{i=1}^{j-1}\Delta_{i,j}2t(t-1)t^{-4}, &(11)
\end{aligned}$$

where in (9) we consider any possible number of samples by time t for each arm. Moreover, inequality (10) follows by (6) and (11) follows by (8). The proof of inequality (3) follows by the fact that

$$\sum_{t\in[T]} t(t-1)t^{-4} \leq \sum_{t\in[T]} t^{-2} \leq \sum_{t=1}^{+\infty} t^{-2} = \frac{\pi^2}{6}.$$

□

D.1 Proof of Theorem 1.4

Theorem 1.4. *The expected reward collected by* INTERLEAVED-UCB *in T rounds, $\mathcal{R}^{IB}(T)$, for k arms, a matroid of rank $r = \mathrm{rk}(\mathcal{M})$ and maximum delay $d_{\max}$ is at least*

$$\left(1 - \frac{1}{e}\right) \mathrm{OPT}(T) - \mathcal{O}\left(k\sqrt{T\ln(T)} + k^2 + d_{\max} r\right).$$

Proof. By inequality (1), Lemma 4.4 and Definition 4.5, we can upper bound the α-regret, for $\alpha = 1 - \frac{1}{e}$, as

$$\alpha\mathrm{OPT}(T) - \mathcal{R}^{IB}(T) \leq \mathop{\mathbb{E}}_{\mathbf{r}\sim[0,1]^k,\mathcal{Q}} \left[\sum_{t\in[T]} \Delta_{\mathcal{G}_t(\mathbf{r})}(\bar{\mu}_t)\right] + \mathcal{O}(d_{\max}\cdot \mathrm{rk}(\mathcal{M})), \tag{12}$$

where the expectation is taken over the randomness of the offset vector $\mathbf{r}$ and the reward realizations.

Under the event $\mathcal{E}_{\mathbf{r}}$, that is, where both IG and IB use the same offsets $\mathbf{r}$, let $\{\sigma_t\}_{t\in[T]}$ be the sequence of bijections between $\mathcal{A}_t^{IB}$ and $\mathcal{A}_t^{IG}$ over all rounds $t \in [T]$, as described in Theorem 4.6. Using Lemma 4.7, we have that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{r}\sim[0,1]^k,\mathcal{Q}} \left[\sum_{t\in[T]} \Delta_{\mathcal{G}_t(\mathbf{r})}(\mu_t)\right] &= \mathop{\mathbb{E}}_{\mathbf{r}\sim[0,1]^k,\mathcal{Q}} \left[\sum_{t\in[T]} \sum_{i\in\mathcal{A}_t^{IG}} \Delta_{i,\sigma_t^{-1}(i)}\right] \\
&= \mathop{\mathbb{E}}_{\mathbf{r}\sim[0,1]^k,\mathcal{Q}} \left[\sum_{t\in[T]} \sum_{i\in\mathcal{A}_t^{IG}} \sum_{j\in\mathcal{A}} \Delta_{i,j}\,\mathcal{X}\left(\sigma_t(j) = i\right)\right] \\
&\leq \mathop{\mathbb{E}}_{\mathbf{r}\sim[0,1]^k,\mathcal{Q}} \left[\sum_{t\in[T]} \sum_{j\in\mathcal{A}} \sum_{i<j} \Delta_{i,j}\,\mathcal{X}\left(\sigma_t(j) = i\right)\right],
\end{aligned} \tag{13}$$

where in the last inequality we restrict ourselves to arms $i < j$, where $\Delta_{i,j} > 0$.

Now using the results of Lemma 4.8, we can further upper bound (13) as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{r}\sim[0,1]^k,\mathcal{Q}} \left[\sum_{t\in[T]} \sum_{j\in\mathcal{A}} \sum_{i<j} \Delta_{i,j}\,\mathcal{X}\left(\sigma_t(j) = i\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathbf{r}\sim[0,1]^k,\mathcal{Q}} \left[\sum_{t\in[T]} \sum_{j\in\mathcal{A}} \sum_{i<j} \Delta_{i,j}\,\mathcal{X}\left(\sigma_t(j) = i, T_j(t) \leq \ell_{i,j}\right)\right] \\
&\qquad + \mathop{\mathbb{E}}_{\mathbf{r}\sim[0,1]^k} \left[\mathop{\mathbb{E}}_{\mathcal{Q}} \left[\sum_{t\in[T]} \sum_{j\in\mathcal{A}} \sum_{i<j} \Delta_{i,j}\,\mathcal{X}\left(\sigma_t(j) = i, T_j(t) > \ell_{i,j}\right)\right]\right] \\
&\leq \sum_{j>1} \frac{16}{\Delta_{j-1,j}} \ln(T) + \frac{\pi^2}{3} \sum_{j>1} \sum_{i=1}^{j-1} \Delta_{i,j}.
\end{aligned} \tag{14}$$

By combining inequalities (12), (13) and (14), we can upper bound the regret as a function of the gaps as follows:

$$\alpha\text{OPT}(T) - \mathcal{R}^{IB}(T)$$
$$\leq \sum_{j>1} \frac{16}{\Delta_{j-1,j}} \ln(T) + \frac{\pi^2}{3} \sum_{j>1} \sum_{i=1}^{j-1} \Delta_{i,j} + \mathcal{O}(d_{\max} \cdot \text{rk}(\mathcal{M})) \quad \text{(gap-dependent regret)}.$$

In order to conclude the proof of the theorem, we would like to construct a regret bound that is independent of the gaps. The standard method is to partition the suboptimality gaps into "small" and "large" and, then, separately study their contribution to the regret. Specifically, for each $j \in \mathcal{A}$ and fixed $\epsilon > 0$, we define:

$$S_j = \{i < j \mid \Delta_{i,j} \leq \epsilon\} \text{ and } L_j = \{i < j \mid \Delta_{i,j} > \epsilon\}.$$

Starting again from (13) and noticing that the total regret due to small gaps can be at most $\epsilon \cdot T$ per arm, we have

$$\mathop{\mathbb{E}}_{\mathbf{r}\sim[0,1]^k,\mathcal{Q}} \left[\sum_{t\in[T]} \sum_{j\in\mathcal{A}} \sum_{i<j} \Delta_{i,j} \, \mathcal{X}\left(\sigma_t(j) = i\right) \right]$$
$$= \mathop{\mathbb{E}}_{\mathbf{r}\sim[0,1]^k,\mathcal{Q}} \left[\sum_{t\in[T]} \sum_{j\in\mathcal{A}} \sum_{i\in S_j} \Delta_{i,j} \, \mathcal{X}\left(\sigma_t(j) = i\right) \right] + \mathop{\mathbb{E}}_{\mathbf{r}\sim[0,1]^k,\mathcal{Q}} \left[\sum_{t\in[T]} \sum_{j\in\mathcal{A}} \sum_{i\in L_j} \Delta_{i,j} \, \mathcal{X}\left(\sigma_t(j) = i\right) \right]$$
$$\leq \epsilon k T + \mathop{\mathbb{E}}_{\mathbf{r}\sim[0,1]^k,\mathcal{Q}} \left[\sum_{t\in[T]} \sum_{j\in\mathcal{A}} \sum_{i\in L_j} \Delta_{i,j} \, \mathcal{X}\left(\sigma_t(j) = i\right) \right]. \tag{15}$$

We now focus only on the regret due to the large gaps, namely, the pairs i, j such that $j \in \mathcal{A}$ and $i \in L_j$, which implies that $\Delta_{i,j} > \epsilon$. By exactly the same analysis as in the gap-dependent case, we can reach inequality (14), in the restricted case where the summations only include pairs of arms such that $\Delta_{i,j} > \epsilon$ (notice that we can apply Lemma 4.8 considering only the set L_j of arms for each $j > 1$). In addition, using the fact that $\Delta_{i,j} \leq 1$ for any $i, j \in \mathcal{A}$, we have

$$\mathop{\mathbb{E}}_{\mathbf{r}\sim[0,1]^k,\mathcal{Q}} \left[\sum_{t\in[T]} \sum_{j\in\mathcal{A}} \sum_{i\in L_j} \Delta_{i,j} \, \mathcal{X}\left(\sigma_t(j) = i\right) \right] \leq \sum_{j>1} \frac{16}{\epsilon} \ln(T) + \frac{\pi^2}{6} k(k-1). \tag{16}$$

By combining inequalities (15) and (16) with (12) and (13), we have

$$\alpha\text{OPT}(T) - \mathcal{R}^{IB}(T) \leq \epsilon k T + \frac{16k}{\epsilon} \ln(T) + \frac{\pi^2}{6} k(k-1) + \mathcal{O}(d_{\max} \cdot \text{rk}(\mathcal{M})).$$

Finally, by setting $\epsilon = 4\sqrt{\frac{\ln(T)}{T}}$, we get that

$$\alpha\text{OPT}(T) - \mathcal{R}^{IB}(T) \leq 8k\sqrt{T\ln(T)} + \frac{\pi^2}{6} k(k-1) + \mathcal{O}(d_{\max} \cdot \text{rk}(\mathcal{M})) \quad \text{(gap-independent regret)}.$$

Therefore, we can conclude that the expected reward collected by IB in T rounds is at least

$$\left(1 - \frac{1}{e}\right) \text{OPT}(T) - \mathcal{O}\left(k\sqrt{T\ln(T)} + k^2 + d_{\max} \cdot \text{rk}(\mathcal{M})\right).$$

□

E Tight examples for natural approaches

Tight example for the naive greedy algorithm for MBB.

Lemma E.1. *For any $d \geq 2$, there exists an instance of the full-information variant of the MBB problem (where the mean rewards are known a priori) such that the greedy strategy that plays a maximum mean reward independent set among the available arms collects a $\left(\frac{1}{2} + \frac{1}{2d}\right)$-fraction of the optimal expected reward.*

Proof. We consider an infinite time horizon and a graphic matroid based on the graph $G_d = (V_d, E_d)$, which is recursively defined as follows: Let $G_1 = (V_1, E_1)$ with $V_1 = \{u, v\}$, $E_1 = \{\{u, v\}\}$ and assume that the arm associated with edge $\{u, v\}$ has delay 1 and mean reward $1 - \epsilon$, for some $\epsilon > 0$. For the graph $G_d = (V_d, E_d)$, we have $V_d = V_{d-1} \cup \{u_d\}$ and $E_d = E_{d-1} \cup \{\{u, u_d\}, \forall u \in V_{d-1}\}$ (namely, G_d is essentially the result of the join operation between G_{d-1} and a single vertex graph). The arms that are associated with the edges of $E_d \setminus E_{d-1}$ all have delay equal to d and mean reward equal to $1 - \frac{\epsilon}{d}$. The above recursive construction is illustrated in Figure 1.

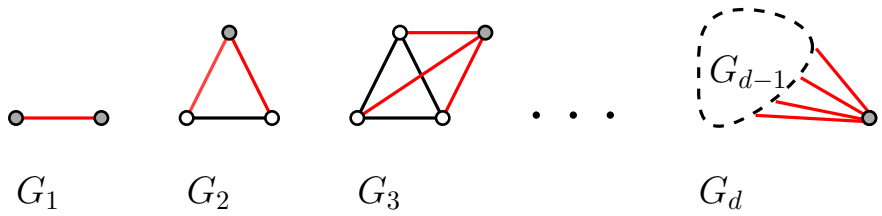

Figure 1: Recursive definition of G_d.

Consider now the arm-pulling schedule constructed by the greedy strategy. Let $T_p = E_p \setminus E_{p-1}$ be the new edges added at each step $p \in [d]$ in the recursive definition of G_d (assuming that $E_0 = \emptyset$). Notice that for any integers $d \geq p_1 > p_2 \geq 1$ the edges of T_{p_1} correspond to arms of higher mean reward than the edges of T_{p_2}. Therefore, the algorithm produces a periodic schedule of period d as follows: Initially, the algorithm plays the d arms of group T_d, collecting reward $d\left(1 - \frac{\epsilon}{d}\right) = d - \epsilon$. Notice that, by construction, these edges form a spanning tree in G_d and, thus, no additional arm can be played at the same time step. In the second time step of the period, the arms of T_d are blocked and the algorithm plays the arms of T_{d-1} collecting $d - 1 - \epsilon$ reward. Again, this is the maximum reward independent set of G_d among the available arms. The algorithm proceeds similarly in the following steps and collects an average reward of

$$\frac{\sum_{p=1}^{d}(p - \epsilon)}{d} = \frac{d \cdot (d+1)/2 - d\epsilon}{d} = \frac{d+1}{2} - \epsilon.$$

In the above example, the optimal arm-pulling sequence is to play at each time $t \in [T]$, one arm of each group T_p for $p \in [d]$. Notice that by construction of the delays and at each time step, there always exists at least one arm per group that is available. Moreover, by definition of the graph G_d, any such selection of arms never contains a circuit and, thus, it is an independent set of the graphic matroid. The expected reward collected by the optimal algorithm at each step is $d - \epsilon \sum_{p \in [d]} \frac{1}{p} = d - \epsilon H(d)$, were $H(d) = \sum_{p \in [d]} \frac{1}{p}$.

In the above example, the ratio between the average reward collected by the greedy strategy and the optimal reward for $\epsilon \to 0$ becomes

$$\lim_{\epsilon \to 0} \frac{\frac{d+1}{2} - \epsilon}{d - \epsilon H(d)} = \frac{1}{2} + \frac{1}{2d}.$$

Therefore, by choosing large enough d, we can bring the approximation ratio of the above example arbitrarily close to $\frac{1}{2}$. □

Tight example for the naive greedy algorithm for RSW.

Remark E.2. *The greedy approach of choosing $\mathcal{A}_t$ to be the set of all available elements at round $t \in [T]$ can be as bad as a $\frac{1}{k}$-approximation. In order to see that, consider the monotone (budget-additive) submodular function $f(S) = \min\{|S|, 1\}$. Let k be the number of elements with delay $d_i = k$ for each $i \in \mathcal{A}$. Assuming an infinite time horizon, the optimal strategy collects an average reward of 1, simply by choosing one element at a time in a round-robin manner. However, the average reward of the greedy approach in this case is $\frac{1}{k}$.*

Tight example for independent sampling for RSW.

Remark E.3. *The independent randomized sampling approach of adding each arm i to $\mathcal{A}_t$ independently with probability $\frac{1}{d_i}$, if available, can be as bad as a $(1 - \frac{1}{\sqrt{e}})$-approximation. Consider the*

same setting as in Remark E.2, where for $T \to \infty$ the optimal average reward is 1. However, the average expected reward of the independent randomized sampling strategy is $1-(1-p)^k$, where $p = \frac{1}{2k-1}$ is the probability that each element is selected at each round (in stationarity). For $k \to \infty$, we have that $1-(1-p)^k \to 1 - e^{-\frac{1}{2}} \approx 0.393$.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Conclusion and Further Directions.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This is a theoretical work. Negative (or positive) societal impact depends on the application.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 1.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Sections 3 and 4 and Supplementary Material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]