Combinatorial Blocking Bandits with Stochastic Delays

Alexia Atsidakou [*1] **Orestis Papadigenopoulos** [*2] **Soumya Basu** [3] **Constantine Caramanis** [1] **Sanjay Shakkottai** [1]

Abstract

Recent work has considered natural variations of the *multi-armed bandit* problem, where the reward distribution of each arm is a special function of the time passed since its last pulling. In this direction, a simple (yet widely applicable) model is that of *blocking bandits*, where an arm becomes unavailable for a deterministic number of rounds after each play. In this work, we extend the above model in two directions: (i) We consider the general combinatorial setting where more than one arms can be played at each round, subject to feasibility constraints. (ii) We allow the blocking time of each arm to be stochastic. We first study the computational/unconditional hardness of the above setting and identify the necessary conditions for the problem to become tractable (even in an approximate sense). Based on these conditions, we provide a tight analysis of the approximation guarantee of a natural greedy heuristic that always plays the maximum expected reward feasible subset among the available (non-blocked) arms. When the arms' expected rewards are unknown, we adapt the above heuristic into a bandit algorithm, based on UCB, for which we provide sublinear (approximate) regret guarantees, matching the theoretical lower bounds in the limiting case of absence of delays.

1. Introduction

It is only recently that researchers have focused their attention on variants of the stochastic *multi-armed bandit* (MAB) problem where the mean reward of each arm is a specific function of the time passed since its last pulling (Kleinberg & Immorlica, 2018; Pike-Burke & Grünewälder, 2019; Cella & Cesa-Bianchi, 2019). These variants capture applications where the mean outcome of an action temporarily decreases or fluctuates after each use. A simple yet very expressive model in this direction is that of *blocking bandits* (Basu et al., 2019), where each arm becomes blocked (i.e., it cannot be played again) for a deterministic number of subsequent rounds after each play, known as the *delay*.

In this paper, we generalize blocking bandits to model two important and well-motivated properties: (i) *Stochastic* blocking, where the delay of each arm is randomly sampled from a distribution after each pull, and (ii) *Combinatorial* actions, where more than one arm can be pulled at each round, subject to general feasibility constraints.

Our model can accommodate a variety of settings where arm pulling is subject to combinatorial constraints, while the repeated selection of a certain arm is undesirable or even infeasible. For example, in ride sharing platforms, where the users are matched with rides, the allocation of resources imposes *matching* constraints, while the transit times of rides give rise to *stochastic blocking* of resources (Dickerson et al., 2018). Matching constraints are also natural in task assignment in the cloud, and combinatorial network optimization where the processing/serving times of tasks can be modeled as stochastic blocking. Further, *knapsack* constraints emerge inherently in ad placement (multiple ad slots in a webpage) (Chen et al., 2016b), and movie/song recommendation (multiple recommendation slots) (Kveton et al., 2015b); while repetitive ads and recommendations should be avoided for user satisfaction by introducing (possibly) random delays. Whereas these varied applications naturally are within the scope of our results, to the best of our knowledge, they are beyond the reach of prior works, and in particular (Basu et al., 2019; 2021; Papadigenopoulos & Caramanis, 2021; Bishop et al., 2020).

1.1. Central Challenges and Main Contributions

Model. A *player* is given a set of *arms*, each corresponding to an unknown *reward-delay* distribution. At each round, the player plays a subset of the available arms, subject to feasibility constraints, and collects, for each arm played, the realized reward of the round. Additionally, each of the selected arms becomes unavailable for a random duration equal to the realized delay of the round. The player has

[*]Equal contribution [1]Department of Electrical and Computer Engineering, The University of Texas at Austin, USA [2]Department of Computer Science, The University of Texas at Austin, USA [3]Google, Mountain View, USA. Correspondence to: Alexia Atsidakou <atsidakou@utexas.edu>, Orestis Papadigenopoulos <papadig@cs.utexas.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

access to an oracle that takes as input a subset of arms and a weight vector, and returns an approximately/probably maximum weight feasible subset. The high-level goal is to maximize cumulative reward that is acquired over time.

There are three central challenges towards the player's goal described above. We outline these challenges, and use these as an organizing principle to describe our main contributions.

Challenge: Inapproximability. In the *full-information setting*, where we assume prior knowledge of the arm mean rewards, the blocking bandits problem is already NP-hard, even in the simple case where the player can pull at most one arm per round and the arm delays are deterministic (Sgall et al., 2009). In the generic combinatorial bandit setting, the resulting scheduling problem may be *NP-hard to even approximate*, even if the underlying feasible set of arms is "easy" to optimize. Thus, a fundamental challenge is to identify sufficient conditions on the combinatorial structure of the feasibility constraints under which our problem accepts efficient $\mathcal{O}(1)$-approximation algorithms.

Contribution 1: We show that if the family of feasible sets satisfies the *hereditary property* (subsets of feasible sets are feasible), then our problem accepts $\mathcal{O}(1)$-approximations. This is a natural property that appears in many practical applications (e.g., knapsack and matching constraints). Specifically, using a gap-preserving reduction from the *Edge-Disjoint Paths* (EDP) problem, we prove that, without the hereditary property, there exists no polynomial-time $\Omega(k^{-\frac{1}{2}+\epsilon})$-approximation algorithm for any $\epsilon > 0$, where k is the number of arms, unless P = NP. Interestingly, this hardness result holds even when the underlying combinatorial set accepts a polynomial-time algorithm for linear maximization. Finally, we show that even assuming feasible sets that satisfy the hereditary property and admit efficient linear program solution, no efficient algorithm can achieve an approximation ratio greater than $1 - \frac{1}{e} + \epsilon$, for any $\epsilon > 0$, unless P = NP. We prove this result using a reduction from the *Max-k-Cover* problem. Hence, constant approximations are the best one can hope for.

Challenge: Analyzing the Greedy Algorithm. A natural heuristic for the full-information setting can be obtained by greedily playing the maximum expected reward feasible subset among the available arms of each round. Though this greedy step may itself be hard, we follow the paradigm of the combinatorial bandits literature (see (Wang & Chen, 2017) and references therein) and assume access to a black-box (approximate/randomized) oracle for this (static) greedy problem. Though natural, the performance of such a greedy heuristic in our setting remains unknown. In part, the technical challenge stems from the black-box oracle assumption. Because we have only oracle-access to the greedy algorithm, there is no easy way to characterize the solution returned by the oracle at each round. Further, as opposed to the case of deterministic delays considered in prior work, in our case, any optimal algorithm in hindsight is inherently online and adaptive to random delay realizations.

Contribution 2: We provide a *tight approximation analysis* of the greedy heuristic when the feasible sets of arms satisfy the hereditary property. Specifically, assuming access to an oracle such that, given a weight vector and a set of arms, it returns an α-approximation of the maximum weight feasible subset with probability β, we prove that the greedy heuristic yields an $\frac{\alpha\beta}{1+\alpha\beta}$-approximation (asymptotically) for the full-information setting. The key in proving the above bound is to consider the *expected pulling rate* of each arm in an optimal solution. Then, by the hereditary property and the fact that the optimal solution is not a priori aware of the delay realizations, we show properties of these rates and, thus, provide the above guarantee.

Challenge: Bandit algorithm. In the case where distributional knowledge of rewards and delays is not assumed, our problem seems significantly harder than standard combinatorial bandits. Even in the absence of blocking, the analysis of bandit algorithms based on the technique of *Upper Confidence Bound* (UCB) (Auer et al., 2002) becomes highly non-trivial in the combinatorial setting and it required a considerably long line of research to achieve optimal regret guarantees (Wang & Chen, 2017; Gai et al., 2012; Chen et al., 2013; Combes et al., 2015; Kveton et al., 2015a; Chen et al., 2016b). These guarantees are obtained by measuring the loss compared to the optimal feasible subset, which is *fixed* throughout the time horizon. In the presence of stochastic blocking, these results do not seem to apply, as the optimal solution in hindsight changes dynamically over time according to the set of available arms.

Contribution 3: We develop a UCB-based variant of the greedy heuristic for the bandit case of our problem, where the reward and delay distributions are initially unknown. As already mentioned, the optimal arm-pulling schedule in our case is not fixed, but depends on the randomness of the delays and the player's actions. Our key insight is to control regret by comparing to a static fractional (hence fictitious) solution, rather than to the dynamics of the optimal solution. By leveraging and extending the framework of (Wang & Chen, 2017) in a way to capture dynamically changing sets of available arms, we are able to provide logarithmic $\frac{\alpha\beta}{1+\alpha\beta}$-approximate regret guarantees. As we note, these guarantees are optimal in the sense that they match the theoretically proven lower bound presented in (Kveton et al., 2015a) in the limiting case of absence of delays.

1.2. Related Work

Following its introduction (Thompson, 1933; Lai & Robbins, 1985), several variants of the stochastic MAB frame-

work have been thoroughly studied (see (Lattimore & Szepesvári, 2020; Slivkins, 2019) for an overview). We focus on the case of *stochastic combinatorial bandits* that is mostly related to our setting. In this direction, a long line of work (Chen et al., 2013; Combes et al., 2015; Kveton et al., 2015a;b; Chen et al., 2016a;b; Wang & Chen, 2017) focuses on variations of the MAB problem, where more than one arm can be played at each round. The above model has been studied in a high level of generality, including arbitrary feasible sets of arms where linear optimization is achieved via (approximate/randomized) oracles, non-linear reward functions of bounded smoothness, probabilistically triggered arms and more. Starting with the work of Gai et al. (2012), the case of arbitrary feasible sets and linear reward functions has been particularly studied. Given k arms and a time horizon T, the state-of-the-art (Wang & Chen, 2017; Kveton et al., 2015a) in that case is a UCB-based bandit policy of $\mathcal{O}(\frac{k \cdot r}{\Delta} \log(T))$ regret against the best possible solution that is achievable in polynomial time, where r is the maximum cardinality of a feasible set and Δ is the gap in expected reward between the optimal and the best suboptimal feasible set. As proved in (Kveton et al., 2015a), the above regret bound matches the theoretical lower bound for this setting.

Our model falls into the area of *stochastic non-stationary bandits*. Important lines in this area include *restless bandits*, where the arms' mean rewards change at every round (Whittle, 1988; Guha et al., 2010), and *rested bandits*, where the means can change only when the arm is played (Gittins, 1979; Tekin & Liu, 2012). The above classes appear to contain notoriously hard problems from a computational viewpoint. In fact, even approximating the optimal solution for the class of restless bandits is PSPACE-hard (Papadimitriou & Tsitsiklis, 1999). Moreover, our model also belongs to the class of *Markov Decision Processes* (MDPs) (Puterman, 1994). However, modeling our problem as an MDP would be inefficient, as it would require a state space that grows exponentially in the number of arms.

Recently, there has been much interest in several variants of the non-stationary model where reward distributions exhibit special temporal correlations with the player's past actions (Kleinberg & Immorlica, 2018; Pike-Burke & Grünewälder, 2019; Cella & Cesa-Bianchi, 2019; Leqi et al., 2020). In (Basu et al., 2019), the authors first study the blocking bandits problem, where each arm becomes blocked for a deterministic number of time steps after it has been played. The above problem has also been studied in an adversarial setting (Bishop et al., 2020) and in a contextual setting (Basu et al., 2021), where the mean rewards of the arms depend on a stochastic context that is observed by the player at the beginning of each round. Very recently, (Papadigenopoulos & Caramanis, 2021) generalize the problem to the setting where more than one arms can be played at each round, subject to matroid constraints. However, none of these studies handle either stochastic delays, or a general class of independence systems, both of which are important in practice (e.g., knapsack problems cannot be handled by any prior studies). Furthermore, in terms of techniques, the analysis of (Papadigenopoulos & Caramanis, 2021) (the only one above addressing combinatorial constraints) heavily relies on the *submodularity* of the *matroid rank function* and the fact that the delays are known and deterministic. We require neither of these to hold in our setting.

2. Preliminaries

We consider a set of k arms, denoted by A, and an *unknown* time horizon $T \in \mathbb{N}$. Each arm $i \in \mathrm{A}$ is associated with a reward distribution $\mathcal{X}_i$ that is bounded w.l.o.g. in $[0,1]$. We denote by μ_i the mean reward of arm i. In the blocking setting, whenever an arm i is pulled at some round t, it cannot be played again for $D_{i,t}-1$ subsequent rounds, namely, within the interval $\{t, ..., t+D_{i,t}-1\}$ (i.e., $D_{i,t}=1$ implies that the arm is not blocked). For each arm $i \in \mathrm{A}$, the value $D_{i,t} \in \mathbb{N}_{\geq 1}$, which we refer to as the *delay*, is a random variable drawn from some arm-dependent distribution $\mathcal{D}_i$ of mean $d_i = \mathbb{E}[D_{i,t}], \forall t \in [T]$ and bounded support in $[1, d_i^{\max}]$. At each round t and for each arm $i \in A$, the reward and delay realization, $X_{i,t}$ and $D_{i,t}$, respectively, are drawn independently from the joint distribution $\mathcal{X}_i, \mathcal{D}_i$ and, thus, they are allowed to be correlated. We denote by $d_{\max} = \max_{i \in \mathrm{A}}\{d_i^{\max}\}$ the maximum delay in a given instance.

We consider the setting of combinatorial bandits, where more than one arm can be played at each round, subject to feasibility constraints. Let $\mathcal{I} \subseteq \{0,1\}^k$ be the family of *feasible* subsets of A. For any subset of arms $S \subseteq \mathrm{A}$, we denote by $\mathcal{I}(S) = \{S' \in \mathcal{I} \mid S' \subseteq S\}$ the subset of feasible sets that only contain arms from $S \subseteq \mathrm{A}$. Let r be the maximum cardinality of a set in $\mathcal{I}$, namely, $r = \max_{S \in \mathcal{I}}\{|S|\}$. Since the problem of maximizing a linear function over a feasible family $\mathcal{I}$ can be NP-hard, following the paradigm of combinatorial bandits as in (Wang & Chen, 2017; Chen et al., 2016b), access to the feasible set is given through an oracle. Given a non-negative weight vector $\mu \in \mathbb{R}^k_{\geq 0}$ and a set $S \subseteq \mathrm{A}$, let $\mathrm{OPT}_\mu(S)$ be the maximum weight feasible set in $\mathcal{I}(S)$ w.r.t. μ. For any $S \subseteq \mathrm{A}$ and vector $\mu \in \mathbb{R}^k_{\geq 0}$, we use the notation $\mu(S) = \sum_{i \in S} \mu_i$. For $\alpha, \beta \in (0,1]$, we assume access to a polynomial-time (α, β)-approximation oracle, $f_\mu^{\alpha,\beta}$, for the underlying combinatorial problem, defined as follows:

Definition 1 ((α,β)-approximation oracle)**.** *Given a weight vector $\mu \in \mathbb{R}^k$ and a subset $S \subseteq \mathrm{A}$, an (α,β)-approximation oracle for $\alpha, \beta \in (0,1]$ outputs a set $f_\mu^{\alpha,\beta}(S) \in \mathcal{I}(S)$, such that*

$$\mathbb{P}\left(\mu(f_\mu^{\alpha,\beta}(S)) \geq \alpha \cdot \mu(\mathit{OPT}_\mu(S))\right) \geq \beta.$$

We are now ready to describe the setting: At each round t, after observing the set of available arms (that is, the arms that are not blocked by some previous pulling), the *player* plays any feasible subset A_t of these arms, namely, $\mathrm{A}_t \in \mathcal{I}$ and collects the realized rewards. At that point, each arm $i \in \mathrm{A}_t$ becomes blocked with delay $D_{i,t}$. We emphasize that the player is *initially unaware of the delay distributions* and can only infer each delay realization $D_{i,t}$ by observing the time where arm i becomes available again. The player seeks to maximize her *total expected reward* in T rounds, formally defined as:

$$\mathrm{Rew}^{\pi}(T) = \mathbb{E}\left[\sum_{t\in[T]}\sum_{i\in\mathrm{A}} X_{i,t}\,\mathbb{I}\left(i \in A_t^{\pi}\right)\right],$$

where A_t^{π} denotes the subset of arms played by an algorithm π at round t. We refer to the above setting as *Combinatorial Blocking Bandits with Stochastic Delays* (CBBSD).

In the bandit setting, we compare the performance of our algorithm with the expected reward of an optimal algorithm that has distributional knowledge of the arm rewards and delays, is aware of the time horizon T and has infinite computational power. Let $\mathrm{Rew}^{*}(T)$ be the optimal expected reward. To evaluate our algorithm we compute its approximate regret compared to the optimal. The following definition of *ρ-approximate regret* (or ρ-regret) is standard in the field of combinatorial bandits for underlying feasibile sets where linear maximization is NP-hard:

$$\mathrm{Reg}_{\rho}^{\pi}(T) := \rho\,\mathrm{Rew}^{*}(T) - \mathrm{Rew}^{\pi}(T).$$

We present two important conditions and investigate their necessity for proving the approximation guarantee of our proposed algorithm. In the standard MAB framework, it is assumed that the optimal algorithm only knows the distributions of the rewards, but not the realizations. The following condition states that the optimal algorithm is also unaware of the delay realizations before pulling an arm:

Condition 1. *The optimal online algorithm has knowledge of the delay distributions, but is not a priori aware of the delay realizations.*

A family $\mathcal{I}$ of feasible subsets is called an *independence system*, if it satisfies the following property:

Definition 2 (Hereditary Property)**.** *Every subset of a feasible set in $\mathcal{I}$ is a also a feasible set, that is, $S' \subset S \subseteq \mathrm{A}$ and $S \in \mathcal{I}$ implies that $S' \in \mathcal{I}$.*

We introduce the following condition on the family of feasible sets of our problem:

Condition 2. *We assume that the family of feasible set $\mathcal{I}$ in every instance of our problem is an independence system.*

The above two conditions are critical. Indeed, as we show in Section 3, dropping any of these makes the problem intractable, even in an approximate sense.

3. Computational-Unconditional Hardness

In this section, we study the hardness of the CBBSD problem. We emphasize that all the results of this section hold even in the simple case where the arm rewards are deterministic and known to the player.

We first show that even in the simple case of a single arm that is always feasible to play (if available), we cannot compete withing a $\omega(1/d_{\max})$-factor against an optimal solution that knows the delay realizations.

Theorem 1. *In the case where Condition 1 does not hold (that is, the optimal algorithm knows the delay realizations of all rounds), there exists no algorithm (not even one of infinite computational power) for the CBBSD problem that achieves an approximation ratio of $\omega(\frac{1}{d_{\max}})$.*

Proof sketch. We consider the case of an infinite time horizon and a single arm ($k = 1$) of deterministic reward, equal to 1. The delay of the arm is either $d > 1$ or 1, each with probability $\frac{1}{2}$. The proof of the theorem follows by modeling the optimal policy in the above example as a Markov Chain, and comparing its expected average reward with that of a possibly sub-optimal policy, which is a priori aware of the delay realizations and plays the arm only when the realized delay is 1. □

We now focus on the hardness of the problem, restricting our attention to the simpler case where the arm rewards and delays are deterministic and known. In the next result, we show that if the underlying feasible set $\mathcal{I}$ does not satisfy Condition 2, then, for any $\epsilon > 0$, there cannot exist any polynomial-time algorithm of approximation guarantee better than $\Omega(k^{-\frac{1}{2}+\epsilon})$, where k is the number of arms, unless P = NP. Interestingly, the above result holds even if linear maximization over the family of feasible sets $\mathcal{I}$ can be done efficiently at every round (i.e., having access to a $(1, 1)$-approximation oracle). We prove the above claim by a gap-preserving reduction from the following problem:

Definition 3 (Edge-disjoint paths (EDP))**.** *Given a directed graph $\mathcal{G} = (V, E)$, where V is the set of vertices and E is the set of k' edges, and m pairs of vertices $\mathcal{T} = \{(s_i, t_i) \mid s_i, t_i \in V, i \in [m]\}$, compute the maximum number of (s_i, t_i) pairs that can be connected using edge-disjoint paths.*

As we show, the full-information case of our bandit problem captures the hardness of EDP, if we do not assume that the underlying feasible set of arms in $\mathcal{I}$ is an independence system (that is, satisfying Condition 2). The following result is known for the EDP problem.

Theorem 2 ((Guruswami et al., 1999)). *For any $\epsilon > 0$, given a directed graph $\mathcal{G} = (V, E)$ with $k' = |E|$ and a set of m pairs of vertices $\mathcal{T} = \{(s_i, t_i) \mid s_i, t_i \in V, i \in [m]\}$, it is NP-hard to distinguish whether all pairs in $\mathcal{T}$ or at most a fraction of $\frac{1}{k'^{1/2-\epsilon}}$ of the pairs can be connected by edge-disjoint paths.*

We now prove the following theorem:

Theorem 3. *Unless P = NP and for any $\epsilon > 0$, there exists no polynomial-time $\Omega(k^{-\frac{1}{2}+\epsilon})$-approximation algorithm for CBBSD problem, in the case of feasible sets that are not independence systems. The above holds even for feasible sets where linear optimization can be performed efficiently.*

Proof sketch. The key idea is to construct an instance of CBBSD, where each arm represents an edge in a given directed graph. In this instance, a subset of arms is feasible only if it includes a unique path between some (s_i, t_i) pair. Our instance is constructed in a way that any feasible solution yields a reward equal to 1 (using linear rewards on the arms). Further, finding a feasible solution in a given subset of arms/edges can be achieved in polynomial-time via breadth/depth first search. As we show, the existence of a polynomial-time $\Omega(k^{-\frac{1}{2}+\epsilon})$-approximation algorithm for CBBSD problem would allow us to resolve the gap problem of Theorem 2. □

Finally, we focus on the hardness of the CBBSD problem, in the case of feasible sets that are independence systems (i.e., satisfying Condition 2). As we show, although under this assumption the problem is easier, yet it still exhibits $\Omega(1)$-hardness of approximation. Specifically, we prove the hardness of CBBSD in that case, using a reduction from Max-k-Cover:

Definition 4 (Max-k-Cover). *Given a universe $U = \{e_1, \ldots, e_k\}$ of k of elements and a collection of m subsets $S_1, \ldots, S_m \subseteq U$, choose l sets that maximize the number of covered elements.*

The following hardness result is known for Max-k-Cover:

Theorem 4 ((Feige, 1998)). *Unless P = NP, there is no polynomial-time $1 - \frac{1}{e} + \epsilon$-approximation algorithm for Max-k-Cover, for any $\epsilon > 0$.*

Using the above result, we are able to show the following hardness result for our problem:

Theorem 5. *Unless P = NP and for any $\epsilon > 0$, there exists no polynomial-time $(1 - \frac{1}{e} + \epsilon)$-approximation algorithm for the full-information CBBSD problem, even for feasible sets that satisfy the hereditary property. The above holds even for feasible sets where linear optimization can be done efficiently.*

Proof sketch. For any instance of Max-k-Cover, we can construct in polynomial-time an instance of the CBBSD problem by considering an arm of reward 1 and delay l for each element of U, where l is the number of sets that can be chosen in Max-k-Cover. Further, a subset of arms S is feasible, only if $S \subseteq S_j$ for at least one given subset $S_j \subseteq U$ (notice that this construction satisfies the hereditary property). As we show, any ρ-approximation algorithm for the above instance of CBBSD, would imply a ρ-approximation for Max-k-Cover, which, in turn, implies the hardness of the former. □

4. Full-Information Setting

We first study the full-information setting of our problem, where we assume that the player has *prior knowledge of the arm mean rewards*. In the case where Conditions 1 and 2 hold, we analyze the approximation guarantee of the following greedy heuristic:

GREEDY-HEURISTIC: At each time $t = 1, 2, \ldots$, observe the set $\mathrm{F}_t \subseteq \mathrm{A}$ of available (not blocked) arms. Let $\mathrm{A}_t = f_\mu^{(\alpha,\beta)}(\mathrm{F}_t)$ be the subset returned by the (α, β)-oracle for support F_t and weight function μ. Play the arms of A_t and collect the associated rewards.

We remark there are two sources of randomness: arm rewards and stochastic delays. In the full information setting, we assume that the player knows the mean rewards of the arms, beforehand. We drop this assumption in the bandit setting, studied in Section 5. Moreover, we note that the algorithm provided above does not make use, and thus does not require knowledge of the delay distributions (or realizations). As we prove in the rest of this section, the above algorithm is $\rho(\alpha, \beta)$-competitive in expectation, against an optimal online algorithm, for $\rho(\alpha, \beta) = \frac{\alpha\beta}{1+\alpha\beta}$.

Theorem 6. *Given any (α, β)-oracle and assuming Condition 1 and Condition 2, the algorithm* GREEDY-HEURISTIC *is a $\frac{\alpha\beta}{1+\alpha\beta}$-approximation (asymptotically) for the full-information case of CBBSD.*

The rest of this section is dedicated to proving the above result. For an algorithm π, we denote as A_t^π the set of arms played by π at time t and $\mathrm{F}_t^\pi \subseteq \mathrm{A}$ (resp., $\mathrm{B}_t^\pi \subseteq \mathrm{A}$) the set of *available* (resp., *blocked*) arms at the beginning of the round. Finally, for any $S \subseteq \mathrm{A}$ and vector $\mu \in \mathbb{R}^k_{\geq 0}$, we use the notation $\mu(S) = \sum_{i \in S} \mu_i$. Let A_t^* be the set of arms played by an optimal algorithm at time $t \in [T]$. For proving the performance guarantee of our algorithm in the presence of stochastic delays, the first step is to consider the *expected pulling rate* of each arm $i \in \mathrm{A}$ by the optimal algorithm, defined as: $z_i = \mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} \mathbb{I}\left(i \in \mathrm{A}_t^*\right)\right]$.

Given the fact that the optimal algorithm is not aware of $\{X_{i,t}\}_{i\in\mathrm{A}}$ before deciding which arms to pull at time t, its

expected cumulative reward is:

$$\mathbb{E}\left[\sum_{t\in[T]}\sum_{i\in \mathrm{A}} X_{i,t}\,\mathbb{I}\left(i\in \mathrm{A}_t^*\right)\right] = \sum_{t\in[T]}\sum_{i\in \mathrm{A}}\mu_i\,\mathbb{E}\left[\mathbb{I}\left(i\in \mathrm{A}_t^*\right)\right],$$

where the equality follows by the fact that $X_{i,t}$ and $\mathbb{I}\left(i\in \mathrm{A}_t^*\right)$ are independent. Thus, we obtain that:

$$\mathrm{Rew}^*(T) = T\sum_{i\in A}\mu_i z_i. \tag{1}$$

We now prove two important properties of the expected pulling-rates $\{z_i\}_{i\in \mathrm{A}}$, each following from Condition 1 and Condition 2, respectively. In the following lemma, we provide an upper bound on z_i, for each arm $i\in \mathrm{A}$.

Lemma 1. *Let $i\in \mathrm{A}$ be an arm of expected delay $d_i = E[D_{i,t}]$ and maximum delay $d_i^{\max}$. We have:*

$$z_i \le \frac{1}{d_i} + \mathcal{O}\Big(\frac{d_i^{\max}}{T}\Big).$$

For proving the above claim, we make crucial use of Condition 1. Indeed, it is not hard to see that in the construction of Theorem 1, the above inequality does not hold.

Lemma 2. *For any subset $S\subseteq \mathrm{A}$ and reward vector μ, assuming that $\mathcal{I}$ is an independence system, we have:*

$$\mu(OPT_\mu(S)) \ge \sum_{i\in S}\mu_i z_i.$$

Proof. Let A_t^* be the set of arms played by the optimal algorithm at time t. Since for any set S and time t the intersection $S\cap \mathrm{A}_t^*$ is strictly contained in S and A_t^*, by Condition 2 we have that $S\cap \mathrm{A}_t^*\in\mathcal{I}(S)$. Therefore, for the expected reward of the optimal solution in $\mathcal{I}(S)$, we have:

$$\mu(\mathrm{OPT}_\mu(S)) \ge \mu(S\cap \mathrm{A}_t^*) = \sum_{i\in S}\mu_i\mathbb{I}\left(i\in \mathrm{A}_t^*\right). \tag{2}$$

By averaging over time and taking the expectation over the randomness of the delays and the possible random bits of the optimal algorithm, we can conclude that:

$$\mu(\mathrm{OPT}_\mu(S)) \ge \frac{1}{T}\sum_{t\in[T]}\sum_{i\in S}\mu_i\,\mathbb{P}\left(i\in \mathrm{A}_t^*\right) = \sum_{i\in S}\mu_i z_i.$$

□

Equipped with the above lemmas, we can now complete the approximation analysis of GREEDY-HEURISTIC.

Proof of Theorem 6. Using the fact that the choices of GREEDY-HEURISTIC do not depend on the reward realizations, for each round $t\in[T]$ we have:

$$\sum_{i\in \mathrm{A}}\mathbb{E}\left[X_{i,t}\,\mathbb{I}\left(i\in \mathrm{A}_t\right)\right] = \sum_{i\in \mathrm{A}}\mathbb{E}\left[\mu_i\,\mathbb{I}\left(i\in \mathrm{A}_t\right)\right] = \mathbb{E}\left[\mu(\mathrm{A}_t)\right].$$

We denote by F_t^π and B_t^π the set of available and blocked arms at the beginning of round t, respectively. Let $\mathcal{Q}_t$ denote the event that the oracle succeeds in returning an α-approximate solution at time t. By taking expectation over the randomness of the oracle and the delay realizations, for every $t\in[T]$, we have that:

$$\begin{aligned}\mathbb{E}\left[\mu(\mathrm{A}_t)\right] &\ge \mathbb{E}\left[\alpha\cdot\mu(\mathrm{OPT}_\mu(\mathrm{F}_t^\pi))\,\mathbb{I}\left(\mathcal{Q}_t\right)\right]\\ &\ge \alpha\cdot\beta\cdot\mathbb{E}\left[\mu(\mathrm{OPT}_\mu(\mathrm{F}_t^\pi)]\right)\\ &\ge \alpha\cdot\beta\cdot\mathbb{E}\left[\sum_{i\in \mathrm{F}_t^\pi}\mu_i z_i\right],\end{aligned}$$

where the first inequality follows by definition of the oracle, the second by the fact that $\mathcal{Q}_t$ is independent of the set F_t^π and that $\mathbb{P}\left(\mathcal{Q}_t\right)\ge\beta$, and the third by Lemma 2. Thus, using Equation (1), we have:

$$\begin{aligned}&\sum_{t\in[T]}\mathbb{E}\left[\sum_{i\in \mathrm{A}} X_{i,t}\,\mathbb{I}\left(i\in \mathrm{A}_t\right)\right]\\ &\ge \alpha\cdot\beta\sum_{t\in[T]}\mathbb{E}\left[\sum_{i\in \mathrm{F}_t^\pi}\mu_i z_i\right]\\ &= \alpha\cdot\beta\cdot T\sum_{i\in \mathrm{A}}\mu_i z_i - \alpha\cdot\beta\sum_{t\in[T]}\mathbb{E}\left[\sum_{i\in \mathrm{B}_t^\pi}\mu_i z_i\right]\\ &= \alpha\cdot\beta\,\mathrm{Rew}^*(T) - \alpha\cdot\beta\sum_{t\in[T]}\mathbb{E}\left[\sum_{i\in \mathrm{B}_t^\pi}\mu_i z_i\right],\end{aligned} \tag{3}$$

where, for the first equality, we use that at any round $t\in[T]$, we have $\mathrm{F}_t^\pi\cup\mathrm{B}_t^\pi = \mathrm{A}_t^\pi$ and $\mathrm{F}_t^\pi\cap\mathrm{B}_t^\pi = \emptyset$.

Note that for any time $t\in[T]$ and arm $i\in\mathrm{B}_t^\pi$, we have:

$$\mathbb{I}\left(i\in B_t^\pi\right) = \sum_{t'<t}\mathbb{I}\left(i\in \mathrm{A}_{t'}^\pi\right)\mathbb{I}\left(D_{i,t'} > t-t'\right).$$

Indeed, in order for an arm to be blocked at time t, it must be played at some earlier round $t'<t$, and the realized delay at t' must be greater than $t-t'$. Using this fact, we

have:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in[T]}\sum_{i\in \mathrm{B}_t^\pi}\mu_i z_i\right]\\
&=\mathbb{E}\left[\sum_{t\in[T]}\sum_{i\in \mathrm{A}}\mu_i z_i\sum_{t'<t}\mathbb{I}\left(i\in \mathrm{A}_{t'}^\pi\right)\mathbb{I}\left(D_{i,t'}>t-t'\right)\right]\\
&=\sum_{t\in[T]}\sum_{i\in \mathrm{A}}\mu_i z_i\sum_{t'<t}\mathbb{P}\left(i\in \mathrm{A}_{t'}^\pi\right)\mathbb{P}\left(D_{i,t'}>t-t'\right)\\
&=\sum_{t'\in[T]}\sum_{i\in \mathrm{A}}\mathbb{P}\left(i\in \mathrm{A}_{t'}^\pi\right)\mu_i z_i\sum_{t>t'}\mathbb{P}\left(D_i>t-t'\right)\\
&\leq\sum_{t'\in[T]}\sum_{i\in \mathrm{A}}\mathbb{P}\left(i\in \mathrm{A}_{t'}^\pi\right)\mu_i z_i d_i\\
&\leq \mathrm{Rew}^\pi(T)+\mathcal{O}\left(d_{\max}\cdot k\right), \qquad (4)
\end{aligned}$$

where the second equality holds since the realization $D_{i,t'}$ is independent of the event $i\in A_{t'}^\pi$. The first inequality is due to the fact that $D_{i,t'}$ is a non-negative integer random variable, thus, $d_i=\mathbb{E}\left[D_{i,t'}\right]=\sum_{\tau=1}^{\infty}\mathbb{P}\left(D_{i,t'}\geq\tau\right)$. Finally, the last inequality is due to Lemma 1. By combining Inequalities (3) and (4), we get:

$$\mathrm{Rew}^\pi(T)\geq\frac{\alpha\cdot\beta}{1+\alpha\cdot\beta}\,\mathrm{Rew}^*(T)-\mathcal{O}\left(d_{\max}\cdot k\right). \quad \square$$

The above analysis is tight, since there exists an instance of the CBBSD problem where GREEDY-HEURISTIC collects an exact $\frac{\alpha\cdot\beta}{1+\alpha\cdot\beta}$-fraction of the optimal reward (see Appendix E in (Papadigenopoulos & Caramanis, 2021)).

5. Bandit Setting

We now turn our attention to the bandit setting where the player is initially unaware of the mean rewards. Given an (α,β)-oracle for the feasible set of arms and assuming Conditions 1 and 2, we develop a UCB-based variant of GREEDY-HEURISTIC, for which we prove $\rho(\alpha,\beta)$-regret guarantees.

Bandit algorithm. Let us denote by $T_{i,t}$ the number of times arm i has been played up to and including round $t\in[T]$, and by $\hat{\mu}_{i,t}$ the empirical average of $T_{i,t}$ independent samples from the distribution $\mathcal{X}_i$. We design a UCB-based variant of GREEDY-HEURISTIC, which we call CBBSD-UCB, that maintains at each time t the following estimates (UCB-indices) of the mean rewards:

$$\bar{\mu}_{i,t}=\min\left\{\hat{\mu}_{i,t-1}+\sqrt{\frac{3\ln t}{2T_{i,t-1}}},1\right\},\forall t\in \mathrm{A}, t\in[T].$$

In the following, we use $\tilde{\pi}$ for any reference to CBBSD-UCB.

Algorithm 1 CBBSD-UCB

Input: Set of arms A and oracle $f^{\alpha,\beta}$.
For every arm $i\in \mathrm{A}$, set $T_i\leftarrow 0$ and $\hat{\mu}_i\leftarrow 1$.
for t=1,2,... **do**
  For every arm $i\in \mathrm{A}$, $\bar{\mu}_i\leftarrow\min\{\hat{\mu}_i+\sqrt{\frac{3\ln t}{2T_i}},1\}$
  Play the set $\mathrm{A}^{\tilde{\pi}}\leftarrow f_{\bar{\mu}}^{\alpha,\beta}(\mathrm{F}^{\tilde{\pi}})$.
  Observe the realization X_i of $\mathcal{X}_i$, for $i\in \mathrm{A}^{\tilde{\pi}}$.
  Set $T_i\leftarrow T_i+1$ and $\hat{\mu}_i\leftarrow\hat{\mu}_i\frac{T_i-1}{T_i}+\frac{X_i}{T_i}$, for $i\in \mathrm{A}^{\tilde{\pi}}$.
end for

The CBBSD-UCB algorithm (Algorithm 1) works as follows: Let $\mathrm{F}_t^{\tilde{\pi}}$ denote the set of available arms. At each time $t\in[T]$ the algorithm observes $\mathrm{F}_t^{\tilde{\pi}}$, and computes and plays the feasible set of arms $f_{\bar{\mu}_t}^{\alpha,\beta}(\mathrm{F}_t^{\tilde{\pi}})$ returned by the oracle, where $\bar{\mu}_t\in\mathbb{R}_{\geq 0}^k$ is the vector of the UCB-indices, such that $(\bar{\mu}_t)_i=\bar{\mu}_{i,t}$. Then, it observes and collects the reward realizations of the played arms and updates the UCB-indices.

Regret analysis. We now provide $\rho(\alpha,\beta)$-regret guarantees for algorithm CBBSD-UCB, with $\rho=\rho(\alpha,\beta)=\frac{\alpha\beta}{1+\alpha\beta}$. As opposed to the non-blocking combinatorial bandits setting, where the optimal solution in hindsight is to repeatedly play the maximum expected reward feasible subset of arms, the mean reward collected at each round by an optimal solution in the blocking setting might exhibit significant fluctuations over time due to blocking. Standard regret analysis typically uses the optimal solution's value as a baseline; but these fluctuations require a different idea.

Instead, we use as a baseline the expected reward collected by an optimal full-information algorithm that can play arms *fractionally* yet *consistently* over time. Indeed, the (optimal) expected pulling rates, $\{z_i\}_{i\in \mathrm{A}}$, as defined in Section 4, precisely characterize such a baseline. These rates can be used to mirror the analysis of Theorem 6, this time for CBBSD-UCB, keeping track of the extra loss due to the lack of information. In order to quantify the above loss, which is due to the fact that CBBSD-UCB calls the (α,β)-oracle using the vector of UCB-indices $\bar{\mu}_t$ at each time t, we use the notion of *instantaneous regret*:

Definition 5 (Instantaneous Regret)**.** *At every time $t\in[T]$, the* instantaneous regret *of $\tilde{\pi}$ is defined as:*

$$\Gamma_t^{\tilde{\pi}}=\alpha\cdot\beta\cdot\mu(\mathit{OPT}_\mu(\mathrm{F}_t^{\tilde{\pi}}))-\mu(\mathrm{A}_t^{\tilde{\pi}}).$$

Informally, $\Gamma_t^{\tilde{\pi}}$ measures the difference between an $\alpha\beta$-fraction of the expected reward of the optimal feasible subset of the available arms, and the expected reward collected by $\tilde{\pi}$ at time t.

Using the properties of the expected pulling rates in Equation (1) and Lemmas 1 and 2, and following Theorem 6, we provide the following upper bound to the regret:

Lemma 3. *The ρ-approximate regret of our algorithm, where $\rho = \frac{\alpha\cdot\beta}{1+\alpha\cdot\beta}$ can be upper bounded as:*

$$\rho\, Reg^{\tilde{\pi}}(T) \leq \frac{1}{1+\alpha\cdot\beta}\,\mathbb{E}\left[\sum_{t\in[T]}\Gamma_t^{\tilde{\pi}}\right] + \mathcal{O}(d_{\max}\cdot k).$$

Thus in order to upper bound the ρ-regret, we need to control the instantaneous regret over time. While this task now resembles the regret analysis of standard (non-blocking) combinatorial bandits, our setting poses an additional technical challenge: the instantaneous regret $\Gamma_t^{\tilde{\pi}}$ depends on the history of arm pulling and reward/delay realizations, not only via the state of the UCB-indices at time t, but also through the set of available arms $\mathrm{F}_t^{\tilde{\pi}}$. As we show, any potential issue due to these additional correlations can be avoided by carefully defining the suboptimality gaps of our problem and adapting the techniques in (Wang & Chen, 2017) in a way to capture the constantly changing availability state.

Definition 6 (Bad Feasible Set)**.** *We refer to a feasible set $S \in \mathcal{I}(F)$ as* bad *w.r.t. a availability set F, when $\mu(S) < \alpha\cdot\mu(OPT_\mu(F))$. Moreover, the family of bad feasible sets w.r.t. the availability set F is defined as:*

$$\mathcal{S}_B(F) = \{S \in \mathcal{I}(F) \mid \mu(S) < \alpha\cdot\mu(OPT_\mu(F))\}.$$

Finally, we denote by $\mathcal{S}_{i,B}(F)$ the family of feasible sets in $\mathcal{S}_B(F)$ that contain arm $i \in \mathrm{A}$.

As opposed to (Chen et al., 2013; Wang & Chen, 2017), the following notion of suboptimality gap is now a function of both the availability set F, which depends on the choices of the algorithm, and the played set $S \in \mathcal{I}(\mathrm{F})$.

Definition 7 (Suboptimality Gap)**.** *The* suboptimality gap *of a bad feasible set $S \in \mathcal{I}(\mathrm{F})$ w.r.t. an availability set F is defined as:*

$$\Delta(S,F) = \alpha\cdot\mu(OPT_\mu(F)) - \mu(S).$$

Focusing on the bad feasible sets that contain arm i, we can further define the minimum possible gap of any such set as: $\Delta_{min}^i = \min_{F\subseteq\mathrm{A},\ S\in\mathcal{S}_{i,B}(F)}\Delta(S,F)$. Further, we can define maximum gap of any bad feasible set that can be played as: $\Delta_{max} = \max_{F\subseteq\mathrm{A},\ i\in\mathrm{A},\ S\in\mathcal{S}_{i,B}(F)}\Delta(S,F)$. We remark that the above choice of Δ_{min}^i captures the following crucial aspect of our problem: the algorithm needs to be able to distinguish between optimal and suboptimal choices over any possible availability set F in the worst case.

For the rest of our analysis, we adapt the ideas introduced in (Wang & Chen, 2017) to accommodate stochastic blocking. The key technical challenge we circumvent stems from the dynamic (un)availability of the arms, which impacts our ability to obtain accurate estimates of μ. This becomes important in our proofs of Lemma 5 and Theorem 7.

Definition 8 (Nice sampling)**.** *At each the beginning of each round t, we say that* CBBSD-UCB *has a* nice sampling *if $|\hat{\mu}_{i,t-1} - \mu_i| \leq \sqrt{\frac{3\ln(t)}{2T_{i,t-1}}}$ for each arm $i \in \mathrm{A}$.*

Let $\mathcal{N}_t$ denote the event that the algorithm has a nice sampling at time t. We can bound the probability of the event $\mathcal{N}_t$ as follows:

Lemma 4. *The event $\mathcal{N}_t$ holds for* CBBSD-UCB *at round t with probability at least $1 - \frac{2k}{t^2}$.*

Intuitively, the event $\mathcal{N}_t$ implies that the UCB-indices of the arms at time t are a good approximation of the actual mean rewards. At each round t, there are three reasons why our algorithm might play a suboptimal set of arms: (i) Failure of the oracle, denoted by $\neg\mathcal{Q}_t$, (ii) Non-representative collection of samples, denoted by $\neg\mathcal{N}_t$, and (iii) Insufficient number of samples for distinguishing the gaps. The probabilities of $\neg\mathcal{Q}_t$ and $\neg\mathcal{N}_t$ at each round t can be upper bounded by $1-\beta$, using the definition of the oracle, and by $\frac{2k}{t^2}$, using Lemma 4, respectively. Thus, we focus on the rounds where $\mathcal{Q}_t$ and $\mathcal{N}_t$ hold.

Definition 9 (Sampling Threshold)**.** *We define the* sampling threshold *as:*

$$\ell_T(\Delta) = \frac{24r^2\ln(T)}{\Delta^2},$$

where $r = \max_{S\in\mathcal{I}}\{|S|\}$. We also define the following function:

$$\kappa_T(\Delta, s) = \begin{cases} 2, & \text{if } s = 0,\\ \sqrt{\frac{24\ln(T)}{s}}, & \text{if } 1 \leq s \leq \ell_T(\Delta),\\ 0, & \text{if } s > \ell_T(\Delta).\end{cases}$$

In the above definitions, which appear in the framework of (Wang & Chen, 2017), the domain of Δ is over any possible suboptimality gap given by a combination of availability set F and feasible set $S \in \mathcal{I}(\mathrm{F})$. Specifically, the sampling threshold ℓ_T has the following property: When the number of times an arm i has been played exceeds $\ell_T(\Delta_{\min}^i)$, then the bad feasible sets containing this arm cannot further contribute to the regret (assuming that $\mathcal{Q}_t$ and $\mathcal{N}_t$ hold). This idea is depicted in the following result:

Lemma 5. *For any $t \in [T]$, if the event $\{\mathcal{Q}_t, \mathcal{N}_t, \mathrm{A}_t^{\tilde{\pi}} \in \mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}})\}$ holds, then*

$$\Delta(\mathrm{A}_t^{\tilde{\pi}}, \mathrm{F}_t^{\tilde{\pi}}) \leq \sum_{i\in\mathrm{A}_t^{\tilde{\pi}}}\kappa_T(\Delta_{\min}^i, T_{i,t-1}).$$

By combining Lemmas 3 to 5 with the techniques of (Wang & Chen, 2017), we obtain the following regret bounds:

Theorem 7. *For the $\frac{\alpha\cdot\beta}{1+\alpha\cdot\beta}$-approximate regret of our algorithm, we provide the following* distribution-dependent

bound:

$$\frac{48}{1+\alpha\beta}\sum_{i\in \mathrm{A}}\frac{r\cdot\ln T}{\Delta^i_{\min}}+k\cdot(2+\frac{\pi^2}{3}\Delta_{\max})+\mathcal{O}(d_{\max}\cdot k),$$

and the following distribution-independent *bound:*

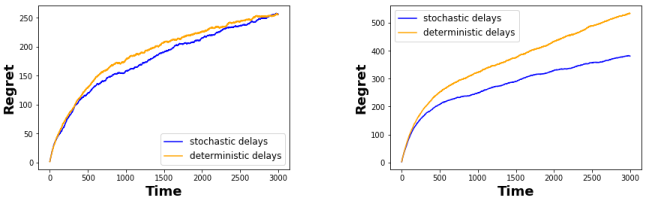
$$\frac{14\sqrt{k\cdot r\cdot T\ln T}}{1+\alpha\beta}+k\cdot(2+\frac{\pi^2}{3}\Delta_{\max})+\mathcal{O}(d_{\max}\cdot k),$$

where $r=\max_{S\in\mathcal{I}}\{|S|\}$.

Note that except for the additive $\mathcal{O}(d_{\max}\cdot k)$ term, the above regret bounds match the theoretical lower bounds for combinatorial bandits setting presented in (Kveton et al., 2015a) in the absence of delays. Indeed, given that our problem strictly generalizes the setting of standard combinatorial bandits with linear rewards and without blocking constraints (in which case our regret is exact and the $\mathcal{O}(d_{\max}\cdot k)$ term disappears), our dependence on k, r, T and $\Delta_{\min}$ is unimprovable.

6. Experimental Evaluation

We evaluate our results on synthetic experiments, where the feasible set is defined by knapsack and matching constraints. We compare the performance of our greedy heuristic for deterministic and stochastic delays of the same mean. In both experimental settings we used $k = 50$ arms, exact oracles for the underlying feasibility problem ($\alpha = \beta = 1$) and averaged the results over 50 trials. The empirical ratio in both settings is computed against an LP relaxation (see appendix D for a detailed description of the experimental settings). In the case of stochastic delays, the empirical ratio is slightly worse but regret is smaller (Fig. 1).



arms	apx. ratio fixed d	apx. ratio rand. d	arms	apx. ratio fixed d	apx. ratio rand. d
10	0.5730	0.565113	10	0.9727	0.9332
20	0.8997	0.8803	20	0.9693	0.9383
30	0.9463	0.9208	30	0.8797	0.8454
40	0.9628	0.9331	40	0.9063	0.8609
50	0.9760	0.9494	50	0.9094	0.8686

Figure 1. Empirical α-regret (for 50 arms) and approximation ratio ("apx. ratio") for the cases of: i) *1-D knapsack constraints* (left) and ii) *matching constraints* (right), for deterministic ("fixed d") and stochastic ("rand. d") delays.

Acknowledgements

This research was partially supported by NSF Grants 1704778, 1826320, 2019844, ARO grant W911NF-17-1-0359, and the Wireless Networking and Communications Group Industrial Affiliates Program.

References

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2–3):235–256, May 2002. ISSN 0885-6125.

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

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

Bishop, N., Chan, H., Mandal, D., and Tran-Thanh, L. Adversarial blocking bandits. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Cella, L. and Cesa-Bianchi, N. Stochastic bandits with delay-dependent payoffs, 2019.

Chen, W., Wang, Y., and Yuan, Y. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pp. 151–159, 2013.

Chen, W., Hu, W., Li, F., Li, J., Liu, Y., and Lu, P. Combinatorial multi-armed bandit with general reward functions. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29, pp. 1659–1667. Curran Associates, Inc., 2016a.

Chen, W., Wang, Y., Yuan, Y., and Wang, Q. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *J. Mach. Learn. Res.*, 17(1):1746–1778, January 2016b. ISSN 1532-4435.

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

Dickerson, J. P., Sankararaman, K. A., Srinivasan, A., and Xu, P. Allocation problems in ride-sharing platforms: Online matching with offline reusable resources. In *AAAI*, 2018.

Feige, U. A threshold of ln n for approximating set cover. *J. ACM*, 45(4):634–652, July 1998. ISSN 0004-5411.

Gai, Y., Krishnamachari, B., and Jain, R. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Trans. Netw.*, 20(5):1466–1478, October 2012. ISSN 1063-6692.

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

Guha, S., Munagala, K., and Shi, P. Approximation algorithms for restless bandit problems. *J. ACM*, 58(1), December 2010. ISSN 0004-5411.

Guruswami, V., Khanna, S., Rajaraman, R., Shepherd, B., and Yannakakis, M. Near-optimal hardness results and approximation algorithms for edge-disjoint paths and related problems. In *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing*, STOC '99, pp. 19–28, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581130678.

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

Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. Tight Regret Bounds for Stochastic Combinatorial Semi-Bandits. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 535–543, San Diego, California, USA, 09–12 May 2015a. PMLR.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. Combinatorial cascading bandits. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28, pp. 1450–1458. Curran Associates, Inc., 2015b.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1): 4–22, 1985.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Leqi, L., Kilinc-Karzan, F., Lipton, Z. C., and Montgomery, A. L. Rebounding bandits for modeling satiation effects, 2020.

Mitzenmacher, M. and Upfal, E. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, USA, 2005. ISBN 0521835402.

Papadigenopoulos, O. and Caramanis, C. Recurrent submodular welfare and matroid blocking bandits. *arXiv preprint arXiv:2102.00321*, 2021.

Papadimitriou, C. and Tsitsiklis, J. The complexity of optimal queuing network control. *Math. Oper. Res.*, 24: 293–305, 1999.

Pike-Burke, C. and Grünewälder, S. Recovering bandits. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 14122–14131, 2019.

Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994. ISBN 0471619779.

Sgall, J., Shachnai, H., and Tamir, T. Periodic scheduling with obligatory vacations. *Theor. Comput. Sci.*, 410 (47–49):5112–5121, November 2009. ISSN 0304-3975.

Slivkins, A. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019. ISSN 1935-8237. doi: 10.1561/2200000068.

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

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. ISSN 00063444.

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

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