Contextual Blocking Bandits

Soumya Basu[*,‡] UT Austin
Orestis Papadigenopoulos[*] UT Austin
Constantine Caramanis UT Austin
Sanjay Shakkottai UT Austin

Abstract

We study a novel variant of the multi-armed bandit problem, where at each time step, the player observes an independently sampled context that determines the arms' mean rewards. However, playing an arm blocks it (across all contexts) for a fixed number of future time steps. The above contextual setting captures important scenarios such as recommendation systems or ad placement with diverse users. This problem has been recently studied [Dickerson et al., 2018] in the full-information setting (i.e., assuming knowledge of the mean context-dependent arm rewards), where competitive ratio bounds have been derived. We focus on the bandit setting, where these means are initially unknown; we propose a UCB-based variant of the full-information algorithm that guarantees a $\mathcal{O}(\log T)$-regret w.r.t. an α-optimal strategy in T time steps, matching the $\Omega(\log(T))$ regret lower bound in this setting. Due to the time correlations caused by blocking, existing techniques for upper bounding regret fail. For proving our regret bounds, we introduce the novel concepts of delayed exploitation and opportunistic subsampling and combine them with ideas from combinatorial bandits and non-stationary Markov chains coupling.

1 INTRODUCTION

There has been much interest in variants of the stochastic *multi-armed bandit* (MAB) problem to model the phenomenon of *local* performance loss, where after each play, an arm either becomes unavailable for several subsequent rounds [Basu et al., 2019], or its mean reward temporarily decreases [Kleinberg and Immorlica, 2018, Cella and Cesa-Bianchi, 2019]. These studies provide state-of-the-art finite time regret guarantees. However, many practical applications of bandit algorithms are contextual in nature (e.g., in recommendation systems, task allocations), and these studies do not capture such scenarios where the rewards depend on a task-dependent context.

Our paper focuses on the following *contextual* blocking bandit problem: We consider a set of *arms* such that, once an arm is pulled, it cannot be played again (i.e., is blocked) for a fixed number of consecutive rounds. At each round, a unique *context* is sampled according to some fixed distribution over a finite set of contexts and the player observes this context before playing an arm. The reward of each arm is drawn independently from a different distribution, depending on the context of the round under which the arm is played. The objective of the player is to maximize the expected cumulative reward over an unknown time horizon.

Applications of the above model include scheduling in data-centers, task assignment in online or physical service systems, and more generally, settings where the contextual nature as well as transient unavailability are important. As an example, consider a group of agents (arms) with different expertise on a (monetized) question-answering platform (e.g., JustAnswers, Chegg, Quora). When a question is presented, the platform assigns it to one of the agents, who answers the question after a fixed amount of research time (a.k.a. blocked time). If the answer is satisfactory, the reward is '1', else it is '0'. The probability that the answer is satisfactory varies across agents based on their individual expertise. Here, the context is the question type, and the context-dependent mean reward is the probability of a satisfactory answer. The goal of the

[*] These authors have equal contribution.
[‡] Part of the work was done after the author joined Google, Mountain View, USA.
✉: `basusoumya@utexas.edu`, `papadig@cs.utexas.edu`, `constantine@utexas.edu`, `sanjay.shakkottai@utexas.edu`

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

platform is to match questions to agents who are both available and have relevant expertise. At a high level, settings such as that above forms the main motivation of our model.

1.1 Key Technical Challenges

We introduce and study the problem of contextual blocking bandits (CBB). In this setting, greedy approaches that play the best available arm fail. Instead, for adapting to unknown future contexts, a combination of randomized arm selection and selective round skipping (i.e., not playing any arm in some rounds) is required for achieving optimal competitive guarantees. This technique, that ensures sufficient future arm availability, has been noted in [Dickerson et al., 2018] and [Chawla et al., 2010, Alaei et al., 2012].

Prior work in the full-information case where the mean rewards are known [Dickerson et al., 2018], devises a randomized LP rounding algorithm that is based on round skipping. Critically, the round skipping probabilities are time-dependent and computed offline given the LP solution (see Section 3). These skipping probabilities, however, cannot be precomputed in a bandit setting, thus requiring some form of online learning.

To address the challenges of a bandit setting, a natural idea is to use a (dynamic) LP. This LP would use upper confidence bound (UCB) values (that vary over time) in place of the true mean values that would be available in the full information setting (as in [Agrawal and Devanur, 2014, Sankararaman and Slivkins, 2018]).

This strategy, however, creates a significant technical hurdle: the LP is now a function of the trajectory, and the availability state of the system depends on the dynamically changing LP solution several steps into the future. This correlates past and future decisions and thus, prior techniques for analyzing the impact of skipping rounds cannot be applied.

The LP using UCB values has a further challenge: An action derived from the LP might not be available in a particular round (due to blocking); thus no action would be taken leading to no new sample of reward, and thus, no evolution of the information state (maintained by the bandit to learn the environment).

1.2 Our Contributions

(i) We develop an efficient time-oblivious bandit algorithm for k arms and m contexts, that achieves $\mathcal{O}\left(\frac{km(k+m)\log(T)}{\Delta}\right)$ regret bound w.r.t. an α-optimal strategy, with Δ the difference between the optimal and best suboptimal extreme point solution of the LP, and where α is the best possible competitive guarantee. This requires two key technical innovations:

(a) *Delayed Exploitation.* At each time t, our algorithm uses the UCB from the (past) time $(t-M_t)$, where $M_t = \Theta(\log(t))$, for computing a new solution to the LP. Introducing this delay is crucial – it ensures that the dynamics of the underlying Markov chain over the interval $[t-M_t, t]$ have mixed, and decorrelates the UCB from each arm's availability at time t. We believe that this technique might be of independent interest.

(b) *LP Convergence under Blocking.* We leverage techniques from combinatorial bandits [Chen et al., 2016, Wang and Chen, 2017] and combine them with an *opportunistic subsampling* scheme, in order to ensure a sufficient rate of new samples associated with suboptimal LP solutions.

We validate our theory with simulations on synthetic instances in Section 6 and Appendix H.

(ii) For the full-information case, we prove an unconditional hardness of $\alpha = \frac{d_{\max}}{2d_{\max}-1}$, where $d_{\max}$ is the maximum blocking time, establishing that our algorithm (and the one in [Dickerson et al., 2018]) achieves the optimal competitive guarantee. This improves on the 0.823-hardness result of [Dickerson et al., 2018].

(iii) As a byproduct of our work, we improve on [Dickerson et al., 2018], in the special case where the blocking times are deterministic and time-independent. Specifically, our algorithm (a) does not require knowledge of the time horizon T, (b) involves a (smaller) LP that can be optimized via fast combinatorial methods, and (c) leads to a slightly improved competitive guarantee (asymptotically) for finite blocking times.

1.3 Related Work

From the advent of stochastic MAB [Thompson, 1933] and later [Lai and Robbins, 1985], decades of research in stochastic MAB have culminated in a rich body of results (c.f. [Bubeck and Cesa-Bianchi, 2012, Lattimore and Szepesvári, 2018]). Focusing on directions which are relevant to ours, we first note that our problem differs from *contextual bandits* as in [Langford and Zhang, 2008, Beygelzimer et al., 2011, Agarwal et al., 2014]. Although these works face the challenge of arbitrarily many contexts, they do not handle blocking.

Our problem lies in the space of stochastic *non-stationary* bandits, where the reward distributions (states) of the arms can change over time. Two important threads in this area are: *rested bandits* [Gittins, 1979, Tekin and Liu, 2012, Cortes et al., 2017], where the arm state (hence, reward distribution) changes only when the arm is played, and *restless bandits* [Whittle, 1988, Tekin and Liu, 2012], where the state changes at each time, independently of when the arm is pulled. Our problem differs from these settings (and from *sleeping bandits* [Kleinberg et al., 2010]), as our reward distributions change in a very special manner, both during arm playing (becoming blocked) and not playing (i.i.d. context and becoming available). Our problem also falls into the class of *controlled MDPs* [Altman, 1999] with unknown parameters. However, the exponentially large state space (i.e., $\mathcal{O}(d_{\max}^k)$) makes this approach highly space and time consuming, and the finite time regret of known algorithms [Auer and Ortner, 2007, Tewari and Bartlett, 2008, Gajane et al., 2019] non-admissible.

In recent works [Kleinberg and Immorlica, 2018, Basu et al., 2019, Cella and Cesa-Bianchi, 2019, Pike-Burke and Grünewälder, 2019], the reward distribution changes are determined by some fixed special functions. Our setting belongs to this line of work, as blocking can be translated w.l.o.g. into deterministically zero reward. However, our problem differs from the above, as the optimal algorithm in hindsight must adapt to random context realizations. The models in [György et al., 2007, Pike-Burke and Grünewälder, 2019] also assume stochastic side information and arm delays, but consider different notions of regret, comparing to our work.

From an algorithmic side, the full-information case of our problem has been studied in [Dickerson et al., 2018], in the context of *online bipartite matching* with stochastic arrivals and reusable nodes (see also [Johari et al., 2017] for an interesting, yet unrelated to ours, combination of matching and learning). In addition, the non-contextual case [Basu et al., 2019] is related to the literature on *periodic scheduling* [Holte et al., 1989, Bar-Noy et al., 1998, Sgall et al., 2009].

The idea of combining UCB [Auer et al., 2002] and LP formulations also appears in *bandits with knapsacks* [Badanidiyuru et al., 2018, Sankararaman and Slivkins, 2018, Agrawal and Devanur, 2014, Agrawal et al., 2016]. Our problem differs from this model (and from *bandits with budgets* [Slivkins, 2013, Combes et al., 2015a]), as we assume both resource consumption and budget renewal (i.e., arm availability) that depend on the player's actions. Finally, due to blocking, our problem differs from *combinatorial bandits* and *semi-bandits* [Combes et al., 2015b, Chen et al., 2013, 2016, Kveton et al., 2014, 2015]. However, we draw from the techniques in [Wang and Chen, 2017] for analyzing the regret of our LP-based algorithm.

2 PROBLEM DEFINITION

Model. Let $\mathcal{A}$ be a set of k *arms* (or *actions*), $\mathcal{C}$ be a set of m *contexts* and $T \in \mathbb{N}$ be the time horizon of our problem. At every round $t \in \{1, 2, \dots, T\}$, a context $j \in \mathcal{C}$ is sampled by *nature* with probability f_j (such that $\sum_{j \in \mathcal{C}} f_j = 1$). The *player* observes the realization of each context at the beginning of the corresponding round, before making any decision on the next action. When arm $i \in \mathcal{A}$ is pulled at round t under context $j \in \mathcal{C}$, the player receives a *reward* $X_{i,j,t}, \forall t \in \{1, 2, \dots, T\}$. We assume that the (context and arm dependent) rewards $\{X_{i,j,t}\}_{t \in [T]}$ are i.i.d. random variables with mean $\mu_{i,j}$ and bounded support in $[0, 1]$. In the *blocking* bandits setting, each arm is in addition associated with a *delay* $d_i \in \mathbb{N}_{\geq 1}$, indicating the fact that, once arm i is played at some round t, the arm becomes unavailable for the next $d_i - 1$ rounds (in addition to round t), namely, in the interval $\{t, \dots, t+d_i-1\}$. The player is unaware of the time horizon, but we assume that she has prior knowledge of the context distribution $\{f_j\}_{j \in \mathcal{C}}$ and arm delays. As we explain in Section 7, it is straightforward to relax the above technical assumption, since these attributes are independent of the player's actions and, thus, can be efficiently learned by the algorithm.

A specific problem instance I is defined by the tuple $(\mathcal{A}, \mathcal{C}, \{d_i\}_{\forall i \in \mathcal{A}}, \{f_j\}_{\forall j \in \mathcal{C}}, \{X_{i,j,t}\}_{\forall i,j,t})$, with each element as defined above. We refer the reader to Appendix A for additional technical notation.

Online Algorithms. In our setting, an *online algorithm* is a strategy according to which, at every round t, the player observes the context of the round, and chooses to play one of the available arms (or skip the round). Specifically, the decisions of an online algorithm depend only on the observed context of each round and the availability state of the system. We are interested in constructing an online algorithm π, that maximizes the *expected cumulative reward* over the randomness of the nature and of the algorithm itself, in the case of a randomized algorithm. Let $A_t^\pi \in \mathcal{A} \cup \emptyset$ be the arm played by algorithm π at time t, C_t be the context of the round, and $\mathcal{R}_{N,\pi}$ be the randomness due to the contexts/rewards realizations and the possible random bits of π. For any instance I and time horizon T, the expected reward can be expressed as follows:

$$\mathbf{Rew}_I^\pi(T) = \mathop{\mathbb{E}}_{\mathcal{R}_{N,\pi}} \left[\sum_{t \in [T]} \sum_{j \in \mathcal{C}} \sum_{i \in \mathcal{A}} X_{i,j,t} \, \mathbb{I}\left(A_t^\pi = i, C_t = j\right) \right].$$

Oracle. In order to characterize an optimal online algorithm, one way is to formulate it as a Markov Decision Process (MDP) on a state space of size $\mathcal{O}(d_{\max}^k)$, which is exponential in the number of arms. Instead, we take a different route by comparing our algorithms with an offline *oracle*, i.e., an optimal (offline) algorithm that has a priori knowledge of the context realizations of all rounds and infinite computational power (a.k.a. *optimal clairvoyant algorithm*). Clearly, the expected reward of the *oracle*, denoted by $\mathbf{Rew}^*_I(T)$, upper bounds the reward of any online algorithm.

Competitive Ratio. The *competitive ratio*, $\rho^\pi(T)$, of an algorithm π for T time steps is defined as the (worst case over the problem instance) ratio between the expected reward collected by π and the expected reward of the oracle, and is a standard notion in the field of online algorithms [1]. An algorithm π is called α-*competitive* if there exists some $\alpha \in (0, 1]$ such that $\rho^\pi(T) \geq \alpha, \forall T \in \mathbb{N}_+$. Thus, an α-competitive algorithm achieves at least $\alpha \cdot \mathbf{Rew}^*_I(T)$ expected reward.

Approximate Regret. Let π^* be the oracle. Note that, for any finite T, and due to the finiteness of the number of contexts and actions, such an algorithm is well-defined. The α-*regret* of an algorithm π is the difference between α times the expected reward of an optimal online policy[2] and the reward collected by π, for $\alpha \in (0, 1]$, i.e.,

$$\alpha\, \mathbf{Reg}^\pi_I(T) = \alpha \cdot \mathbf{Rew}^*_I(T) - \mathbf{Rew}^\pi_I(T).$$

The notion of α-regret is widely accepted in the combinatorial bandits literature [Chen et al., 2016, Wang and Chen, 2017] for problems where an efficient algorithm does not exist, even for the case where the mean rewards $\{\mu_{i,j}\}_{\forall i,j}$ are known a priori (thus leading inevitably to linear regret in the standard definition).

3 FULL-INFORMATION SETTING

We begin by considering the *full-information* (non-bandit) variant of the problem, where the mean rewards $\{\mu_{i,j}\}_{i\in\mathcal{A},j\in\mathcal{C}}$ are known to the player a priori. Note that in both variants, we assume that the distribution of contexts $\{f_j\}_{j\in\mathcal{C}}$ and the delays $\{d_i\}_{i\in\mathcal{A}}$ are known to the player (see Section 7), but the time horizon is unknown. This case of our problem has been also studied in [Dickerson et al., 2018], in the setting where the delays can be stochastic and time-dependent, but the time horizon is known.

LP Upper Bound. Our first step is to upper bound the reward of an optimal clairvoyant policy, $\mathbf{Rew}^*_I(T)$, which uses an optimal schedule of arms for each context realization. Consider the following LP:

$$\begin{aligned}
\textbf{maximize: } & \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j} z_{i,j} && (\mathbf{LP})\\
\textbf{s.t.: } & \sum_{j\in\mathcal{C}} z_{i,j} \leq \frac{1}{d_i}, \forall i \in \mathcal{A} && (\mathbf{C1})\\
& \sum_{i\in\mathcal{A}} z_{i,j} \leq f_j, \forall j \in \mathcal{C} && (\mathbf{C2})\\
& z_{i,j} \geq 0, \forall i \in \mathcal{A}, \forall j \in \mathcal{C}\,.
\end{aligned}$$

In (**LP**), each variable $z_{i,j}$ can be thought of as the (fluidized) average rate of playing arm i under context j. Intuitively, constraints (**C1**) indicate the fact that each arm $i \in \mathcal{A}$ can be pulled at most once every d_i steps, due to the blocking constraints. Similarly, constraints (**C2**) suggest that playing (any arm) under context $j \in \mathcal{C}$ happens with probability at most f_j. As we show in the proof of Theorem 1, (**LP**) provides an (approximate) upper bound to the expected reward collected by an optimal clairvoyant policy (when we multiply its objective value by T), and this approximation becomes tighter as T increases. Finally, we remark that, as opposed to the LP used in [Dickerson et al., 2018]: (a) We do not require knowledge of the time horizon T in order to compute an optimal solution to (**LP**), and (b) its structural simplicity allows the efficient computation of an optimal extreme point solution, using fast combinatorial methods (see Appendix C.1).

Online Randomized Rounding. Our algorithm, FI-CBB, rounds an optimal solution to (**LP**) in an online randomized manner (as in [Dickerson et al., 2018], but for a different LP), and serves as a basis for the bandit algorithm we design in the next section (see Appendix B.1 for a pseudocode):

FI-CBB: The algorithm initially computes an optimal solution, $\{z^*_{i,j}\}_{i,j}$, to (**LP**). At any round t, and after observing the context $j_t \in \mathcal{C}$ of the round, the algorithm *samples* an arm, based on the marginal distribution $\{z^*_{i,j_t}/f_{j_t}\}_{i\in\mathcal{A}}$. At this phase, any arm can be sampled, independently of its availability state. If no arm is sampled (because $\sum_{i\in\mathcal{A}} z^*_{i,j_t}/f_{j_t} < 1$), the round is skipped and no arm is played. Let $i_t \in \mathcal{A}$ be the sampled arm of this phase. If the arm i_t is available, the algorithm plays the arm with probability $\beta_{i_t,t}$ (formally defined shortly)– otherwise, the round is skipped.

For any arm $i \in \mathcal{A}$ and round t, we set $\beta_{i,t} = \min\{1, \frac{d_i}{2d_i-1}\frac{1}{q_{i,t}}\}$, where $q_{i,t}$ is the a priori probability of i being available at time t (i.e., before observing any context realization). The value of $q_{i,t}$, can be re-

[1] Formally, $\rho^\pi(T) = \inf_I \frac{\mathbf{Rew}^\pi_I(T)}{\mathbf{Rew}^*_I(T)}$.

[2] In fact, we use a stronger notion of α-regret by assuming that the optimal online algorithm is clairvoyant.

cursively computed as follows:

$$q_{i,1} = 1 \text{ and } q_{i,t+1} = q_{i,t}(1 - \beta_{i,t} \sum_{j \in \mathcal{C}} z^*_{i,j}) + \mathbb{I}\,(t \geq d_i)\, q_{i,t-d_i+1}\beta_{i,t-d_i+1} \sum_{j \in \mathcal{C}} z^*_{i,j}. \qquad (1)$$

In the above algorithm, the arm sampling at the beginning of each round ensures that, on average, each arm-context pair, (i, j), is selected a $z^*_{i,j}$-fraction of time. Moreover, $\{\beta_{i,t}\}_{\forall i,t}$ correspond to the *non-skipping* probabilities– their role is to ensure a constant rate of arm availability over time. The technique of precomputing these probabilities as a function of the expected arm availability has been proven useful for achieving optimal competitive guarantees in various online optimization settings (see, e.g., [Dickerson et al., 2018, Chawla et al., 2010, Alaei et al., 2012]), where other approaches (such as greedy LP rounding) fail.

In the following theorem, we provide the competitive guarantee of our algorithm FI-CBB. Due to space constraints and the partial overlapping with [Dickerson et al., 2018], its proof has been moved to Appendix E.

Theorem 1. *For any T, the competitive ratio of* FI-CBB *against any optimal clairvoyant algorithm is at least $\frac{d_{\max}}{2d_{\max}-1}\left(1 - \frac{d_{\max}-1}{d_{\max}-1+T}\right)$, where $d_{\max} = \max_{i \in \mathcal{A}} d_i$.*

4 BANDIT SETTING

In the bandit setting of our problem, where the mean rewards $\{\mu_{i,j}\}_{\forall i,j}$ are initially unknown, we design a bandit variant of FI-CBB, that attempts to learn the mean values of the distributions $\{X_{i,j,t}\}_{\forall t}$ for all $i \in \mathcal{A}, j \in \mathcal{C}$, while collecting the maximum possible reward. Our objective is to achieve an α-regret bound growing as $\mathcal{O}(\log(T))$, for $\alpha = \frac{d_{\max}}{2d_{\max}-1}$. Due to space constraints, the proofs are deferred to Appendix F.

4.1 The Bandit Algorithm: ucb-cbb

Our algorithm, named UCB-CBB, maintains UCB indices for all arm-context pairs, and uses them (in place of the actual means) to compute a new optimal solution to (**LP**) at each round. Given this solution, the algorithm samples an arm in a similar way as FI-CBB. We expect that, as the time progresses, the LP solution computed using the UCB estimates will converge to the optimal solution of (**LP**) and, thus, the two algorithms will gradually operate in an similar manner.

However, as the UCB indices are intrinsically linked with arm sampling, the future arm availability and, thus, the sequence of LP solutions become correlated across time. This makes the precomputation of non-skipping probabilities, $\{\beta_{i,t}\}_{i,t}$, as before, no longer possible. In order to disentangle these dependencies, we introduce the novel technique of *delayed exploitation*, where at each round, UCB-CBB uses UCB estimates from relatively far in the past. This ensures that the extreme points used in the meantime are fixed and unaffected by the online rounding and reward realizations in the entire duration. Using this fixed sequence of extreme points, we *adaptively* compute non-skipping probabilities that strike the right balance between skipping and availability.

We now outline the new elements of UCB-CBB (which we denote by $\tilde{\pi}$), comparing to FI-CBB.

Dynamic LP. As opposed to the case of FI-CBB, where the mean rewards are initially unknown, our bandit algorithm solves at each time $t \in [T]$ a linear program $(\mathbf{LP})(t)$. This LP has the same constraints as (**LP**), but uses UCB estimates, $\{\bar{\mu}_{i,j}(t)\}_{i,j}$, in place of the actual means in the objective. Following the standard UCB paradigm, for every $i \in \mathcal{A}$ and $j \in \mathcal{C}$, this estimate is defined as

$$\bar{\mu}_{i,j}(t) = \min\left\{\hat{\mu}_{i,j,T_{i,j}(t)} + \sqrt{\frac{3\ln(t)}{2T_{i,j}(t)}}, 1\right\}. \qquad (2)$$

In the above formula, $T_{i,j}(t)$ denotes the number of times arm i is played under context j up to (and excluding) time t, and $\hat{\mu}_{i,j,T_{i,j}(t)}$ denotes the empirical estimate of $\mu_{i,j}$, using $T_{i,j}(t)$ i.i.d. samples.

Delayed Exploitation. In order to decouple the UCB estimates and, thus, the extreme point choices, from the arm availability state of the system, our algorithm, at any round t, uses the UCB indices from several rounds in the past. For any $t \in [T]$, let $Z(t) = \{z_{i,j}(t)\}_{i,j}$ be the optimal extreme point solution to $(\mathbf{LP})(t)$, i.e., using the indices $\{\bar{\mu}_{i,j}(t)\}_{i,j}$ in place of the actual mean rewards. Moreover, let $Z(0)$ be an arbitrary extreme point of (**LP**). For any $t \in [T]$, we fix $M_t = \Theta(\log t)$, in a way that there is a unique integer $T_c \geq 1$, such that $t \geq M_t + 1$ if and only if $t \geq T_c$ (see Appendix F.1).

At any $t \in [T]$, and after observing the context $j_t \in \mathcal{C}$ of the round, our algorithm samples arms according to the marginal distribution $\{z_{i,j_t}(t - M_t)/f_{j_t}\}_{i \in \mathcal{A}}$, namely, using the solution of $(\mathbf{LP})(t - M_t)$. In the case where $t - M_t \leq 0$, the algorithm samples arms according to the marginal distribution $\{z_{i,j_t}(0)/f_{j_t}\}_{i \in \mathcal{A}}$, based on the initial extreme point $Z(0)$.

Conditional Skipping. In UCB-CBB the non-skipping probabilities of each round $t \in [T]$, $\{\beta_{i,t}\}_{\forall i}$, now depend on the sequence of solutions of (**LP**) up to time t, that are used for sampling arms. We define by H_t the history up to time t for any $t \geq 1$, which includes all context realizations, pulling of arms, and reward realizations of played arms. For every arm $i \in \mathcal{A}$ and time t, the non-skipping probability is defined as $\beta_{i,t} = \min\{1, \frac{d_i}{2d_i-1}\frac{1}{q_{i,t}(H_{t-M_t})}\}$, where $q_{i,t}(H_{t-M_t})$ now corresponds to the probability of i being available at time t, conditioned on the history up to time H_{t-M_t}.

For $t < T_c$, where the extreme point $Z(0)$ is used at every round until t, the probability $q_{i,t}(H_0)$, for any $i \in \mathcal{A}$ can be recursively computed similarly as in the full-information case (using the recursive equation (1), where every $z^*_{i,j}$ is replaced with $z_{i,j}(0)$ for any $i \in \mathcal{A}, j \in \mathcal{C}$).

For $t \geq T_c$, the value $q_{i,t}(H_{t-M_t})$ is the probability that arm i is available at time t, conditioned on H_{t-M_t}. By definition of T_c, for any $\tau \in [t-M_t, t]$, it is the case that $\tau - M_\tau \leq t - M_t$ and, thus, $H_{\tau-M_\tau} \subseteq H_{t-M_t}$. This implies that all the extreme points in the trajectory of $(\mathbf{LP})(\tau - M_\tau)$ for $\tau \in [t-M_t, t]$, as well as the involved non-skipping probabilities $\{\beta_{i,\tau}\}_{i\in\mathcal{A}}$ are deterministic and, thus, computable, conditioned on H_{t-M_t}. The computation of $q_{i,t}(H_{t-M_t})$ can be done recursively similarly to (1). However, the extreme point solutions depend on arm mean estimates that vary over time, thus requiring a more involved recursion (see Appendix F.2 for more details). Our choice of $M_t = \Theta(\log t)$ is large enough to guarantee sufficient decorrelation of the extreme point choices and the future arm availability, but also small enough to incur a small additive loss in the regret bound.

The above changes are summarized in Algorithm 1. In Appendix B.2, we provide a routine, called COMPQ(i,t,H), for the computation of $q_{i,t}(H_{t-M_t})$.

Algorithm 1: UCB-CBB

```
Set $\bar{\mu}_{i,j}(0) \leftarrow 1$ for all $i \in \mathcal{A}, j \in \mathcal{C}$ and compute an initial solution $Z(0)$ to $(\mathbf{LP})$.
for $t = 1, 2, \dots$ do
    Set $M \leftarrow \lfloor 2\log_{c_1}(t) \rfloor + 2 \cdot d_{\max} + 8$, where $c_1 = e^2/(e^2-1)$.
    if $t \leq M$ then Set $M = t$.
    Compute solution $Z(t-M) = \{z_{i,j}\}_{i\in\mathcal{A}, j\in\mathcal{C}}$ to $(\mathbf{LP})(t-M)$.
    Observe context $j_t \in \mathcal{C}$ and sample arm $i_t \in \mathcal{A}$ with probability $z_{i_t,j_t}/f_{j_t}$.
    if $i_t \neq \emptyset$ and $i_t$ is available then
        Set $q_{i_t,t}(H_{t-M}) \leftarrow$ COMPQ$(i_t, t, H_{t-M})$ and $\beta_{i_t,t} \leftarrow \min\{1, \frac{d_i}{2d_i-1}\frac{1}{q_{i_t,t}(H_{t-M})}\}$.
        if $u \leq \beta_{i_t,t}$, for $u \sim U[0,1]$ then Play $i_t$.
    Update the UCB indices according to Eq. (2).
```

4.2 Analysis of the α-regret

We define the family of extreme point solutions $Z = \{z^Z_{i,j}\}_{i,j}$ of $(\mathbf{LP})$ as $\mathcal{Z}$. We note that, as $(\mathbf{LP})(t)$ varies from $(\mathbf{LP})$ only in the objective, the *family* of extreme points remains fixed and known to the player. We denote by $Z^* = \{z^*_{i,j}\}_{\forall i,j}$ any optimal extreme point of $(\mathbf{LP})$ with respect to the mean values $\{\mu_{i,t}\}_{\forall i,t}$, and we denote by $\mathcal{Z}^{\mathrm{S}}$ the set of suboptimal extreme points. We now define the relevant gaps of our problem by specializing the corresponding definitions of [Wang and Chen, 2017], in the case where the family of feasible solutions coincides with the extreme points solutions of $(\mathbf{LP})$. As we discuss in Appendix C.2, the following suboptimality gaps are complex functions of the means, $\{\mu_{i,j}\}_{i\in\mathcal{A}, j\in\mathcal{C}}$, arm delays, $\{d_i\}_{i\in\mathcal{A}}$, and context distribution, $\{f_j\}_{j\in\mathcal{C}}$.

Definition 1 (Gaps [Wang and Chen, 2017])**.** *For any extreme point $Z \in \mathcal{Z}^{\mathrm{S}}$ the suboptimality gap is $\Delta_Z = \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}(z^*_{i,j} - z^Z_{i,j})$ and $\Delta_{\max} = \sup_{Z\in\mathcal{Z}} \Delta_Z$. For any arm-context pair (i,j), we define $\Delta^{i,j}_{\min} = \inf_{Z\in\mathcal{Z}^{\mathrm{S}}, z^Z_{i,j}>0} \Delta_Z$, i.e., the minimum Δ_Z over all $Z \in \mathcal{Z}^{\mathrm{S}}$ such that $z^Z_{i,j} > 0$.*

The first step of our analysis is to show that delayed exploitation, indeed ensures that the dynamics of the underlying Markov Chain (MC) over the interval $[t - M_t, t]$ have mixed. This weakens the dependence between online rounding and extreme point choices and, thus, decorrelates the UCB from arm availability at time t. Let $F^{\tilde{\pi}}_{i,t}$ be the event that arm i is available at time t. Using techniques from *non-homogeneous MC coupling*, we prove the above weakening formally in the following lemma.

Lemma 1. *For any arm $i \in \mathcal{A}$ and rounds $t, t' \in [T]$ such that $0 < t - t' < d_i$ and $t \geq T_c$, we have:*

$$\frac{\mathbb{P}\left(F^{\tilde{\pi}}_{i,t'} | H_{t-M_t}\right)}{\mathbb{P}\left(F^{\tilde{\pi}}_{i,t'} | H_{t'-M_{t'}}\right)} \leq 1 + c_0 \cdot c_1^{-M_t},$$

for $c_0 = e\left(\frac{e^2}{e^2-1}\right)^{2d_{\max}}$ and $c_1 = \frac{e^2}{e^2-1}$.

Proof sketch. The key idea of the proof is to link the quantities $\mathbb{P}\left(F^{\tilde{\pi}}_{i,t'} | H_{t-M_t}\right)$ and $\mathbb{P}\left(F^{\tilde{\pi}}_{i,t'} | H_{t'-M_{t'}}\right)$ to the evolution of a fast-mixing non-homogeneous MC. Let us fix an arbitrary run of UCB-CBB upto time $t - M_t$, which fixes the sequence of extreme points in the run as $z_{ij}(\tau - M_\tau)$, and the skipping probabilities as β_τ for $1 \leq \tau \leq t$ (see Appendix F.2 for details). For this run and any fixed arm i, we construct the non-homogeneous MC with state space $\{0, 1, \dots, d_i - 1\}$, where each state represents the number of remaining rounds until the arm becomes available. At time $\tau \geq 1$, the MC transitions from state 0 to state $(d_i - 1)$

w.p. $\beta_{i,\tau}\sum_j z_{i,j}(\tau - M_\tau)$, and from state $d > 0$ to state $(d-1)$ w.p. 1. Let $\nu(\tau)$, be the first time on or after τ when arm i becomes available. We show that $\mathbb{P}\left(F^{\tilde\pi}_{i,t'}|H_{s-M_s}\right)$ equals the probability that an independent copy of the above MC which starts from state 0 (available) at time $\nu(s-M_s)$, named $\mathcal{X}_s$, is available at time t'. As the two independent MCs $\{\mathcal{X}_s, s = t, t'\}$ evolve on the same non-homogeneous MC, we show using coupling ideas that at time t' the $L1$ distance between their distributions decays exponentially with M_t. Specifically, we construct a Doeblin coupling [Lindvall, 2002] of the two MCs, where at each time $\tau \ge (t - M_t + d_i)$ w.p. at least $1/e^2$, the two MCs meet at state 0, thus coupling exponentially fast. ■

As we show below, Lemma 1 allows us to relate $\alpha\,\mathbf{Reg}^{\tilde\pi}_I(T)$ to the suboptimality gaps of the sequence of LP solutions used by UCB-CBB. This comes with an additive $\Theta(\log(T)\Delta_{\max})$ cost in the regret.

Lemma 2. *For the α-regret of* UCB-CBB, *for $\alpha = \frac{d_{\max}}{2d_{\max}-1}$ and $M = \Theta(\log T + d_{\max})$, we have*

$$\alpha\,\boldsymbol{Reg}^{\tilde\pi}_I(T) \le \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde\pi}}\left[\sum_{t=1}^{T-M}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mu_{i,j}\left(z^*_{i,j} - z_{i,j}(t)\right)\right] + \frac{1}{3}\ln(T)\Delta_{\max} + 6d_{\max} + 71.$$

Proof sketch. Starting from the definition of $\alpha\,\mathbf{Reg}^{\tilde\pi}_I(T)$: We upper bound $\alpha\cdot\mathbf{Rew}^*_I(T)$ using Theorem 1, while we incur regret in four distinct ways. (a) We incorporate the $\left(1-\Theta(T^{-1})\right)$-multiplicative loss as a $\Theta(d_{\max})$ additive term in the regret. (b) We upper bound the total regret during time 1 to T_c by $(\max_{ij}\mu_{ij})T_c = \Theta(d_{\max})$. (c) We separate the rounds $t \ge T_c$, when M_t is increased (and, thus, the same UCB values are used more than once). This happens $\Theta(\log(T))$ times, adding another $\Theta(\log(T))\Delta_{\max}$ term to the regret. (d) For the rest of the "synchronized" rounds $t \ge T_c$ (i.e., where each one uses strictly updated UCB estimates), using Lemma 1, we show that $i \in \mathcal{A}$ is played under $j \in \mathcal{C}$ with probability "close" to $\frac{d_i}{2d_i-1}z_{i,j}(t-M_t)$, where the total approximation loss leads to an additive $\Theta(1)$ term in the regret. ■

By Lemma 2, we see that UCB-CBB accumulates only constant regret in expectation, once all the extreme points of $\mathcal{Z}^{\mathrm{S}}$ are eliminated with high probability. For this to happen, we need enough samples from each of the arm-context pairs in the support of any $Z \in \mathcal{Z}^{\mathrm{S}}$ (i.e., $\mathrm{supp}(Z) = \{(i,j)\,|z^Z_{i,j} > 0\}$). Once the algorithm computes a point $Z \in \mathcal{Z}^{\mathrm{S}}$ (as a solution of $(\mathbf{LP})(t)$), each pair $(i,j) \in \mathrm{supp}(Z)$ is played with probability $z^Z_{i,j} > 0$, assuming there is no blocking or skipping. Leveraging this observation, we draw from techniques in combinatorial bandits with *probabilistically triggered arms* [Chen et al., 2016, Wang and Chen, 2017][3]. In this direction, following the paradigm of [Wang and Chen, 2017], we define the following subfamilies of extreme points called *triggering probability* (TP) groups:

Definition 2 (TP groups [Wang and Chen, 2017])**.** *For any pair $(i,j) \in \mathcal{A}\times\mathcal{C}$ and integer $l \ge 1$, we define the TP group $\mathcal{Z}_{i,j,l} = \{Z \in \mathcal{Z} \mid 2^{-l} < z^Z_{i,j} \le 2^{-l+1}\}$, where $\{\mathcal{Z}_{i,j,l}\}_{l\ge1}$ forms a partition of $\{Z \in \mathcal{Z} \mid z^Z_{i,j} > 0\}$.*

The regret analysis relies on the following counting argument (known in literature as suboptimality charging) – now standard in the combinatorial bandits literature [Kveton et al., 2015, Chen et al., 2016, Wang and Chen, 2017]: For each TP group $\mathcal{Z}_{i,j,l}$, we associate a counter $N_{i,j,l}$. The counters are all initialized to 0 and are updated as follows: At every round t, where the algorithm computes the extreme point solution $Z(t)$, we increase by one every counter $N_{i,j,l}$, such that $Z(t) \in \mathcal{Z}_{i,j,l}$. We denote by $N_{i,j,l}(t)$ the value of the counter at the beginning of round t.

Opportunistic Subsampling. In the absence of blocking, it can be shown [Wang and Chen, 2017] that at any time t and TP group $\mathcal{Z}_{i,j,l}$, we have $T_{i,j}(t) \ge \frac{1}{3}2^{-l}N_{i,j,l}(t)$ with probability $1 - O(1/t^3)$. This guarantees that by sampling arm-context pairs frequently enough, the algorithm learns to avoid all the points in $\mathcal{Z}^{\mathrm{S}}$ with high probability. However, no such conclusion can be drawn in our situation, where arm blocking can potentially preclude information gain. Specifically, the naive approach of subsampling the counter increases every d_i rounds, can only guarantee that $T_{i,j}(t) \ge O(\frac{2^{-l}}{d_i}N_{i,j,l}(t))$ with high probability, thus, leading to a $\Theta(\sqrt{d_{\max}})$ multiplicative loss in the regret. We address the above issue via a novel *opportunistic subsampling* scheme, which guarantees that, even in the presence of strong local temporal correlations, we still obtain a constant fraction (independent of d_i) of independent samples with high probability.

Lemma 3. *For any time $t \in [T]$, TP group $\mathcal{Z}_{i,j,l}$ and $\mathcal{O}(2^l\log(t)) \le s \le t-1$, we have:*

$$\mathbb{P}\left(N_{i,j,l}(t) = s, T_{i,j}(t) \le \tfrac{1}{24e}2^{-l}N_{i,j,l}(t)\right) \le \tfrac{1}{t^3}.$$

Proof sketch. Due to blocking, there is no uniform lower bound for playing a pair (i,j) each time $N_{i,j,l}(t)$ is increased. Therefore, we subsample the increases of $N_{i,j,l}(t)$ in a way that: (a) the subsampled instances of increases are at least d_i rounds apart, and (b) the subsampled sequence captures a constant fraction (independent of d_i) of non-skipped rounds of the original sequence. The two properties ensure that, in the subsampled sequence, the number of times a pair (i,j)

[3] The papers [Chen et al., 2016, Wang and Chen, 2017] capture a more general setting, which we omit for brevity.

is played concentrates around its mean. For a TP group $\mathcal{Z}_{i,j,l}$, we consider blocks of $(2d_i - 1)$ contiguous counter increases. From each block we obtain one sample in the first d_i counter increases, opportunistically picking a non-skipped round if there is one. By construction, the samples remain d_i rounds apart, ensuring property (a). Also, we show there is at least one non-skipped round per block with probability at least $\frac{(2d_i-1)}{8e}2^{-l}$, ensuring property (b). ■

As we observe, the small size of (**LP**) implies that all its extreme points are *sparse*. This makes it less sensitive to the error in the estimates; which, in turn, leads to tighter regret bounds (see Theorem 2).

Lemma 4. *For any $Z \in \mathcal{Z}$, $|supp(Z)| \leq k + m$.*

By combining Lemmas 2, 3 and 4, along with suboptimality charging arguments of [Wang and Chen, 2017] (as described above), we provide our final regret upper bound in the following theorem.

Theorem 2. *The α-regret of* UCB-CBB *for* $\alpha = \frac{d_{\max}}{2d_{\max}-1}$ *and a universal constant* $C > 0$ *satisfies*

$$\alpha\, \boldsymbol{Reg}_I^{\tilde{\pi}}(T) \leq \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \frac{C\,(k+m)\log(T)}{\Delta_{\min}^{i,j}} + \frac{\pi^2}{6}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \log\left(\frac{2\,(k+m)}{\Delta_{\min}^{i,j}}\right)\Delta_{\max} + 6\cdot d_{\max}.$$

5 HARDNESS RESULTS

Unconditional hardness. The NP-hardness of the full-information CBB problem follows by [Sgall et al., 2009, Basu et al., 2019], even in the non-contextual (offline) setting [Basu et al., 2019]. In the following theorem, we provide unconditional hardness for the contextual case of our problem (see Appendix G for the proof). This result implies that the competitive guarantee of FI-CBB is (asymptotically) optimal, even for the single arm case. Moreover, since the construction in our proof involves deterministic rewards, the theorem also implies the optimality of the algorithm in [Dickerson et al., 2018], thus, improving on the 0.823-hardness presented in that work.

Theorem 3. *For the (asymptotic) competitive ratio of the full-information CBB problem, it holds:*

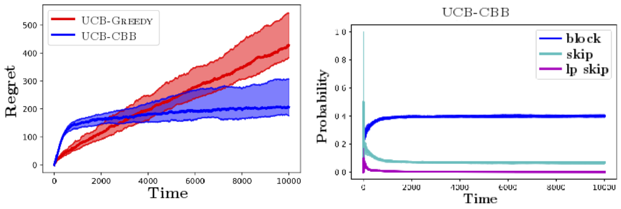
$$\lim_{T\to+\infty}\sup_{\pi}\rho^{\pi}(T) \leq \frac{d_{\max}}{2d_{\max}-1}.$$

Tightness of the regret bound. It is an intriguing open question whether the $\mathcal{O}(\frac{k\cdot m\cdot(k+m)}{\Delta}\log(T))$ dependence in the α-regret of UCB-CBB is the best possible. Unfortunately, there exists no known framework in the literature for lower bounding the α-regret of a bandit algorithm for $\alpha < 1$. In part, this is because, for instances that are "hard" to learn, the considered family of algorithms must *strictly* collect (in expectation) an α-fraction of the optimal expected reward, in the full-information setting. In the opposite case, these algorithms can exhibit *negative regret* [Basu et al., 2019], thus, invalidating any attempt to construct lower bounds.

Nevertheless, there is evidence to support that the dependence of our α-regret bound is not far from optimal. Indeed, consider the following *easy instance* where lower-bounding regret for $\alpha = 1$ is possible: assume k arms, each with delay m, and m contexts each with frequency $\frac{1}{m}$. Let the contexts arrive in a deterministic *round-robin* manner. In this setting, the optimal algorithm (i.e., $\alpha = 1$) plays a specific arm for a specific context obtained by solving the max-weight bipartite matching problem between arms and contexts, with $\mu_{i,j}$ being the weight between arm i and context j. By leveraging existing works (e.g., [Merlis and Mannor, 2020]), we can provide a lower bound of $\Omega((km\min\{k,m\}/\Delta_{\min})\log(\frac{T}{m}))$, where $\Delta_{\min}$ is the mean difference between the reward of the max-weight and second-max-weight matching, nearly matching our dependence in k and m. Converting this to a full proof for our setting (with stochastic contextual arrivals) remains open.

6 NUMERICAL SIMULATIONS

We compare the cumulative regret of our algorithm, UCB-CBB, with a natural greedy heuristic, called UCB-GREEDY, that plays the arm of highest UCB index for the observed context, among the available arms using numerical simulations.



(a) Cumulative Regret (b) Round Skipping Rates

Figure 1: We present the empirical α-regret of UCB-CBB/UCB-GREEDY and skipping rates for UCB-CBB. We consider 10 arms and 10 contexts (i.e., effectively 100 reward distributions). The arm delays are selected uniformly from $\{8, 9\}$ and the context distribution is sampled uniformly from a simplex. The best arm per context has mean 0.9, whereas all other arm-context pairs have means chosen uniformly in $[0, 0.3]$.

We simulate the UCB-CBB and UCB-GREEDY algorithms, for 60 sample paths and $10k$ iterations, and report the mean, 25% and 75% trajectories of (a) cumulative α-regret. Additionally, for the UCB-CBB algorithm we report (b) the empirical probabilities of (i) *LP skip*: skipping due to the sampling according to the LP solution (i.e., when $\sum_{i \in \mathcal{A}} \frac{z_{i,j}}{f_j} < 1$ for some context j), (ii) *skip*: skipping due to our adaptive skipping technique (with probability $1 - \beta_{i,t}$ for a sampled arm i), and (iii) *block*: skipping because the sampled arm is already blocked.

As we observe in Figure 1, the UCB-CBB algorithm using adaptive skipping balances the instantaneous reward and the future availability to achieve a logarithmic α-regret, whereas the UCB-GREEDY algorithm suffers a linear regret. See Appendix H for extended definitions of the above metrics, the UCB-GREEDY algorithm, and additional simulations.

7 EXTENSIONS

In this section, we discuss possible extensions of our model and techniques.

Delayed feedback. In practice, it is natural to assume that the reward of an action is realized at the end of the blocking period (see, for example, the question-answering application provided in Section 1). Clearly, this is already captured by our model in the full-information setting where the actual reward realizations do not matter. Further, due to our technique of delayed exploitation, our bandit algorithm, UCB-CBB, already incorporates the above characteristic, since, by construction, it only uses knowledge on an outcome of an action at least $d_{\max}$ rounds after its playing.

Unknown Context Frequencies and Delays. Our technical assumption of known context frequencies can be relaxed by using empirical estimates of the frequencies in constraints (**C2**) of the (**LP**), instead of the actual frequencies. As the context realizations are independent of actions, the above estimation does not suffer from explore-exploit tradeoffs and, thus, our proofs can be extended to provide a sublinear $\frac{d_{\max}}{2d_{\max}-1}$-regret bound. Further, deterministic delays can be estimated trivially by playing each arm once.

Context Dependent Delays. A generalization of our model that would extend the range of applications captured is that of context dependent delays, namely, the case where each arm $i \in \mathcal{A}$ can have a different delay $d_{i,j}$, when played under context $j \in \mathcal{C}$. On the technical side, it can be proved that our algorithm maintains its $\alpha = \frac{d_{\max}}{2d_{\max}-1}$-competitive ratio, simply by replacing constraints (**C1**) with $\sum_{j \in \mathcal{C}} d_{i,j} z_{i,j} \leq 1, \forall i \in \mathcal{A}$ and by adjusting the recursive computation of the non-skipping probabilities as $\beta_{i,t} = \min\{1, \frac{d_{\max}}{(2d_{\max}-1)q_{i,t}}\}$, where $q_{i,t} = \mathbb{P}\left(\exists j \in \mathcal{C}, \exists t' \in [t - d_{i,j} + 1, t-1] : A_t = i, C_i = j\right)$. In the bandit case, our technique of delayed exploitation together with the coupling arguments suffices to provide a logarithmic α-regret bound. A caveat in the above extension is that the aforementioned guarantees hold against a *non-clairvoyant* optimal solution.

Stochastic Delays. An interesting open direction is the case where the arm delays are stochastic and their distributions are initially unknown. On the positive side, our techniques can be extended in the case where the delay distribution is known, simply by replacing the constant d_i with $\mathbb{E}[d_i]$ in constraints (**C1**) of (**LP**) and adjusting the computation of non-skipping probabilities. However, this computation now relies on the complete knowledge of the distribution, which, in the bandit case, can only be learned empirically using samples. It would be interesting to explore whether our techniques suffice for maintaining a sublinear α-regret bound, under this additional online learning aspect.

Conclusion

In this work, we consider a variant of the blocking bandits problem [Basu et al., 2019], where a stochastic context is observed at the beginning of each round that determines the arm mean rewards. Using the novel techniques of delayed exploitation and opportunistic subsampling, we have developed a bandit algorithm with logarithmic (approximate) regret guarantee. We believe that these techniques could potentially serve as building blocks for approaching similar problems.

Acknowledgements

This research was partially supported by NSF Grants 1826320, 1609279, 1646522, 1934932 and 1704778, ARO grant W911NF-17-1-0359, and the Wireless Networking and Communications Group Industrial Affiliates Program.

References

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646, 2014.

Shipra Agrawal and Nikhil R. Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, page 989–1006, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450325653.

Shipra Agrawal, Nikhil R. Devanur, and Lihong Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 4–18, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

Saeed Alaei, MohammadTaghi Hajiaghayi, and Vahid Liaghat. Online prophet-inequality matching with applications to ad allocation. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, page 18–35, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450314152.

Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 49–56, 2007.

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

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *J. ACM*, 65(3):13:1–13:55, 2018.

Amotz Bar-Noy, Randeep Bhatia, Joseph (Seffi) Naor, and Baruch Schieber. Minimizing service and operation costs of periodic scheduling. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '98, page 11–20, USA, 1998. Society for Industrial and Applied Mathematics. ISBN 0898714109.

Soumya Basu, Rajat Sen, Sujay Sanghavi, and Sanjay Shakkottai. Blocking bandits. In *Advances in Neural Information Processing Systems (NeurIPS) 32*, pages 4785–4794. Curran Associates, Inc., 2019.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26, 2011.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Machine Learning*, 5(1): 1–122, 2012.

Leonardo Cella and Nicolò Cesa-Bianchi. Stochastic bandits with delay-dependent payoffs, 2019.

Shuchi Chawla, Jason D. Hartline, David L. Malec, and Balasubramanian Sivan. Multi-parameter mechanism design and sequential posted pricing. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, STOC '10, page 311–320, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450300506.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pages 151–159, 2013.

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

Richard Combes, Chong Jiang, and Rayadurgam 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, Portland, OR, USA, June 15-19, 2015*, pages 245–257. ACM, 2015a.

Richard Combes, M. Sadegh Talebi, Alexandre Proutiere, and Marc 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, 2015b. MIT Press.

Corinna Cortes, Giulia DeSalvo, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. Discrepancy-based algorithms for non-stationary rested bandits. *arXiv preprint arXiv:1710.10657*, 2017.

John P. Dickerson, Karthik Abinav Sankararaman, Aravind Srinivasan, and Pan Xu. Allocation problems in ride-sharing platforms: Online matching with offline reusable resources. In *AAAI*, 2018.

Pratik Gajane, Ronald Ortner, and Peter Auer. Variational regret bounds for reinforcement learning. *arXiv preprint arXiv:1905.05857*, 2019.

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

Andrew V. Goldberg and Robert E. Tarjan. Finding minimum-cost circulations by canceling negative cycles. *J. ACM*, 36(4):873–886, October 1989. ISSN 0004-5411.

András György, Levente Kocsis, Ivett Szabó, and Csaba Szepesvári. Continuous time associative bandit problems. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, page 830–835, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

Robert Holte, Aloysius Mok, Louis Rosier, Igor Tulchinsky, and Donald Varvel. Pinwheel: a real-time scheduling problem. In *Proceedings of the Hawaii International Conference on System Science*, volume 2, pages 693 – 702 vol.2, 02 1989. ISBN 0-8186-1912-0.

Ramesh Johari, Vijay Kamble, and Yash Kanoria. Matching while learning. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, EC '17, page 119, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450345279.

Robert Kleinberg and Nicole 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.

Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. *Machine learning*, 80(2-3):245–272, 2010.

Branislav Kveton, Zheng Wen, Azin Ashkan, Hoda Eydgahi, and Brian Eriksson. Matroid bandits: Fast combinatorial optimization with learning. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence, UAI 2014, Quebec City, Quebec, Canada, July 23-27, 2014*, pages 420–429. AUAI Press, 2014.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. 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*, pages 535–543, San Diego, California, USA, 09–12 May 2015. PMLR.

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

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, page 28, 2018.

Torgny Lindvall. *Lectures on the coupling method.* Courier Corporation, 2002.

Nadav Merlis and Shie Mannor. Tight lower bounds for combinatorial multi-armed bandits. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2830–2857. PMLR, 09–12 Jul 2020.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis.* Cambridge University Press, USA, 2nd edition, 2017. ISBN 110715488X.

James B. Orlin, Serge A. Plotkin, and Éva Tardos. Polynomial dual network simplex algorithms. *Math. Program.*, 60(1-3):255–276, June 1993. ISSN 0025-5610.

Ciara Pike-Burke and Steffen Grünewälder. 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*, pages 14122–14131, 2019.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014.

Karthik Abinav Sankararaman and Aleksandrs Slivkins. Combinatorial semi-bandits with knapsacks. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, volume 84 of *Proceedings of Machine Learning Research*, pages 1760–1770. PMLR, 2018.

Jirí Sgall, Hadas Shachnai, and Tami Tamir. Periodic scheduling with obligatory vacations. *Theor. Comput. Sci.*, 410(47-49):5112–5121, 2009.

Aleksandrs Slivkins. Dynamic ad allocation: Bandits with budgets. *ArXiv*, abs/1306.0155, 2013.

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

Ambuj Tewari and Peter L Bartlett. Optimistic linear programming gives logarithmic regret for irreducible mdps. In *Advances in Neural Information Processing Systems*, pages 1505–1512, 2008.

William 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. ISSN 00063444.

Qinshi Wang and Wei 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.

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