# 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 $\alpha$-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. Copyright 2021 by the author(s).

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 $\alpha$-optimal strategy, with $\Delta$ the difference between the optimal and best suboptimal extreme point solution of the LP, and where $\alpha$ 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 $\pi$, 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 $\pi$ 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 $\pi$. For any instance $I$ and time horizon $T$, the expected reward can be expressed as follows:

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

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

**Approximate Regret.** Let $\pi^*$ 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 *$\alpha$-regret* of an algorithm $\pi$ is the difference between $\alpha$ times the expected reward of an optimal online policy[2] and the reward collected by $\pi$, 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 $\alpha$-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:

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

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 $\alpha$-regret by assuming that the optimal online algorithm is clairvoyant.

cursively computed as follows:

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

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 $\alpha$-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 (**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 (**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 (**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 $t = 1, 2, \ldots$ 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 $\alpha$-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 $\Delta_Z$ 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 $\beta_\tau$ 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 $\{0, 1, \ldots, d_i - 1\}$, 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 $\tau$ 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 $\tau \geq (t - M_t + d_i)$ 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 $\alpha$-regret of* UCB-CBB*, for $\alpha = \frac{d_{\max}}{2d_{\max}-1}$ and $M = \Theta(\log T + d_{\max})$, we have*

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

**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 $t \geq T_c$, 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 $t \geq T_c$ (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 $l \geq 1$, we define the TP group $\mathcal{Z}_{i,j,l} = \{Z \in \mathcal{Z} \mid 2^{-l} < z^Z_{i,j} \leq 2^{-l+1}\}$, where $\{\mathcal{Z}_{i,j,l}\}_{l\geq 1}$ 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 $T_{i,j}(t) \geq \frac{1}{3}2^{-l}N_{i,j,l}(t)$ 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 $T_{i,j}(t) \geq O(\frac{2^{-l}}{d_i}N_{i,j,l}(t))$ 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 $\mathcal{O}(2^l\log(t)) \leq s \leq t-1$, we have:*

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

**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 $\alpha$-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:*

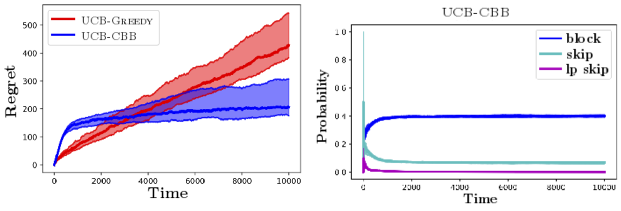
$$\lim_{T\to+\infty}\sup_{\pi}\rho^{\pi}(T) \leq \frac{d_{\max}}{2d_{\max}-1}.$$

**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 $\alpha$-regret of UCB-CBB is the best possible. Unfortunately, there exists no known framework in the literature for lower bounding the $\alpha$-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 $\alpha$-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 $\alpha$-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 $\alpha$-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 $\alpha$-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 $\alpha$-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 $\alpha$-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 $\alpha$-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.

# Supplementary Material

## A TECHNICAL NOTATION

For any event $\mathcal{E}$, we denote by $\mathbb{I}(\mathcal{E}) \in \{0, 1\}$ the indicator variable that takes the value of 1 if $\mathcal{E}$ occurs and 0, otherwise. For any number $n \in \mathbb{N}$, we define $[n] = \{1, 2, \dots, n\}$ and for any integer $r \in \mathbb{Z}$, we define $[r]^+ := \max\{r, 0\}$. Moreover, we use the notation $t \in [a, b]$ (for $a \leq b$) for some time index $t$, in lieu of $t \in [T] \cap \{a, a+1. \dots, b-1, b\}$. Unless otherwise noted, we use the indices $i$ or $i'$ to refer to arms, $j$ or $j'$ to refer to contexts and $t$, $t'$ or $\tau$ to refer to time. We use $\log(\cdot)$ for the logarithm of base 2 and $\ln(\cdot)$ for the natural logarithm. Let $A^{\pi}_t \in \mathcal{A} \cup \{\emptyset\}$ be the arm played by some algorithm $\pi$ at time $t$ and let $F^{\pi}_{i,t}$ be the event that arm $i$ is *free* (i.e. not blocked) at time $t$ for some algorithm $\pi$. We denote by $C_t \in \mathcal{C}$ (or simply $j_t \in \mathcal{C}$) the observed context of round $t$. For a given instance $I$, let $d_{\max} = \max_{i \in \mathcal{A}}\{d_i\}$ be the maximum delay of an arm. In this reading, expectations can be taken over the randomness of the nature, including the sampling of contexts (denoted by $\mathcal{R}_C$) and the arm rewards (denoted by $\mathcal{R}_X$), as well as the random bits of the corresponding algorithm (denoted by $\mathcal{R}_\pi$ for an algorithm $\pi$). We denote by $\mathcal{R}_{N,\pi}$ the randomness generated by the combination of the aforementioned factors.

## B OMITTED PSEUDOCODES

### B.1 Pseudocode of Algorithm fi-cbb

**Algorithm 2:** FI-CBB

```
Compute an optimal solution $\{z^*_{i,j}\}_{\forall i,j}$ to (LP).
Initialize the non-skipping probabilities: $q_{i,1} \leftarrow 1$, $\beta_{i,1} \leftarrow \frac{d_i}{2d_i-1}$, $\forall i \in \mathcal{A}$.
for $t = 1, 2, \dots$ do
    Observe context $j_t \in \mathcal{C}$.
    Generate $u, v \sim U[0,1]$.
    Sample arm $i_t$ such that $u \in \left[\sum_{i'=1}^{i-1} \frac{z^*_{i',j_t}}{f_{j_t}}, \sum_{i'=1}^{i} \frac{z^*_{i',j_t}}{f_{j_t}}\right)$ (assuming a fixed arm order).
    if $i_t \neq \emptyset$ and $i_t$ is available and $v \leq \beta_{i_t,t}$ then
        Play arm $i_t$.
    else
        Skip the round without playing any arm.

    for $i \in \mathcal{A}$ such that $d_i \geq 2$ do
        $q_{i,t+1} \leftarrow q_{i,t}\left(1 - \beta_{i,t}\sum_{j \in \mathcal{C}} z^*_{i,j}\right) + \mathbb{I}(t \geq d_i)\, q_{i,t-d_i+1}\beta_{i,t-d_i+1}\sum_{j \in \mathcal{C}} z^*_{i,j}$.
        $\beta_{i,t+1} \leftarrow \min\{1, \frac{d_i}{2d_i-1}\frac{1}{q_{i,t+1}}\}$.
```

### B.2 Computation of the Conditional Non-skipping Probability, compq $(i, t, H_{t-M_t})$

**Algorithm 3:** COMPQ$(i, t, H_{t-M_t})$

```
if $(i, t)$ in Cache then
    return Cache$[(i, t)]$ // Global Cache
Let $Z(t')$ be the solution of $(\mathbf{LP})(t')$ $\forall t' \in [T]$ and $Z(0) = Z(\tau)$ $\forall \tau \leq 0$ be an initial solution.
Set $t_0 \leftarrow$ the first time on or after $\max\{1, t - M_t\}$, when arm $i$ becomes available.
Set $q_{i,t_0} \leftarrow 1$.
for $t' = t_0, \dots, t-1$ do
    $t'' \leftarrow t' - d_i + 1$.
    $\beta_{i,\tau} \leftarrow \min\{1, \frac{d_i}{2d_i - 1} \frac{1}{\text{COMPQ } (i, \tau, H_{\tau - M_\tau})}\}$ for $\tau \in \{t', t''\}$.
    $q_{i,t'+1} \leftarrow q_{i,t'} \left(1 - \beta_{i,t'} \sum_{j \in \mathcal{C}} z_{i,j}(t' - M_{t'})\right) + \mathbb{I}\left(t'' \geq t_0\right) q_{i,t''} \beta_{i,t''} \sum_{j \in \mathcal{C}} z_{i,j}(t'' - M_{t''})$.
Cache$[(i, t)] = q_{i,t}$. // Memorization
Remove all $(i, t')$ s.t. $t' < t - M_t$ from Cache. //Garbage Collection
return $q_{i,t}$.
```

## C DISCUSSIONS

### C.1 Optimizing over (LP) using Combinatorial Methods

The linear formulation $(\mathbf{LP})$ contains $k \cdot m$ variables and $k \cdot m + k + m$ constraints (including the non-negativity constraints).

From a practical perspective, an optimal extreme point solution to $(\mathbf{LP})$ can be computed efficiently using fast combinatorial methods. Indeed, every instance of the $(\mathbf{LP})$ can be transformed into an instance of the well-studied MAXIMUM WEIGHTED FLOW problem and solved by standard techniques such as cycle canceling [Goldberg and Tarjan, 1989] or fast implementations of the dual simplex method for network polytopes [Orlin et al., 1993].

We now describe the reduction: We consider a node $i$ for every arm $i \in \mathcal{A}$ and a node $j$ for every context $j \in \mathcal{C}$. We define two additional nodes: a source node $s$ and a sink node $t$. For each variable $z_{i,j}$, we associate an edge $(i, j)$ of capacity $c_{i,j} = +\infty$ and weight $w_{i,j} = \mu_{i,j}$. In addition, for each node $i \in \mathcal{A}$, we consider an edge $(s, i)$ of weight $w_{s,i} = 0$ and capacity $c_{s,i} = 1/d_i$, while for each node $j \in \mathcal{C}$, we consider an edge $(j, t)$ of weight $w_{j,t} = 0$ and capacity $c_{j,t} = f_j$. It is not hard to verify that the optimal solution to $(\mathbf{LP})$ coincides with a flow of maximum weight in the aforementioned network.

### C.2 Suboptimality Gaps

In general, the suboptimality gaps, $\Delta^{i,j}_{\text{min}}$, of the LP are complex functions of the means, $\{\mu_{i,j}\}_{i,j}$, arm delays, $\{d_i\}_i$, and context distribution, $\{f_j\}_j$. This fact should not be surprising– it is the combination of all these parameters that determines how an optimal (or near-optimal) solution must behave.

Interestingly, when applied to the standard MAB[4] problem [Lai and Robbins, 1985] (i.e., single context and unit delays), the gap $\Delta^{i,j}_{\text{min}}$ for $i > 1$, matches the standard notion of gap $\Delta_i = \mu_{1,j} - \mu_{i,j}$, where $i = 1$ is the arm of highest mean reward (and $j$ the unique context).

As another example of suboptimality gaps, consider the following structured instance: Let $k > 2$ arms and $m = k$ contexts. All the arms have equal delay $d_i = k, \forall i \in \mathcal{A}$ and all contexts appear with equal probability $f_j = \frac{1}{k}, \forall j \in \mathcal{C}$. We assume that $\mu_{i,j} = \Delta > 0$, if $i = j$, and $\mu_{i,j} = 0$, otherwise. In the above instance, it is not hard to verify that the variables $\{z_{i,j}\}_{i,j}$ in any extreme point solution of $(\mathbf{LP})$ take values in $\{0, 1/k\}$. Moreover, the support of the optimal extreme point solution corresponds to a maximum bipartite matching (w.r.t. the edge weights $\{\mu_{i,j}\}_{i,j}$) in the underlying bipartite graph consisting of arm (left) and context (right) nodes.

Let $M \subset [k] \times [k]$ be the maximum matching in the above bipartite graph with respect to the mean values. Moreover, we define $M_{i,j} \subset [k] \times [k]$ for any $i \neq j$ to be a maximal matching in the above graph that necessarily

[4] The standard MAB regret lower bound is $\mathcal{O}(k \cdot \frac{\log(T)}{\Delta})$, where $\Delta$ is the minimum gap between two arms.

contains the edge $(i,j)$ of $\mu_{i,j} = 0$ (which corresponds to a matching of $k-2$ edges). In addition, we define $M_{i,i} = M \setminus (i,i)$, namely, the maximum matching with the edge $(i,i)$ removed. Using the above definitions, we can see that the optimal solution to (**LP**) can be expressed as $\sum_{(i,j)\in M} \Delta z^*_{i,j} = k\Delta\frac{1}{k} = \Delta$. It is not hard to verify that the suboptimality gap of any pair $(i,j)$ with $i \neq j$ can be expressed as

$$\Delta^{i,j}_{\min} = \Delta - \sum_{(i',j')\in M_{i,j}} \Delta\frac{1}{k} = \Delta - \frac{k-2}{k}\Delta = \frac{2}{k}\Delta.$$

Finally, for the suboptimality gap of any pair $(i,i)$, we have

$$\Delta^{i,i}_{\min} = \Delta - \sum_{(i',j')\in M_{i,i}} \Delta\frac{1}{k} = \Delta - \frac{k-1}{k}\Delta = \frac{1}{k}\Delta.$$

### C.3 Difference in $\alpha$-regret Definition

We note that in Definition 5 in [Chen et al., 2016], a super-arm (which is analogous to an extreme point of (**LP**) in our paper) is defined as *bad*, if the reward from this super arm is less than $\alpha$ times the reward of an optimal super arm. However, in our case an extreme point is *bad* if its reward is less than 1 times (not $\frac{d_{\max}}{2d_{\max}-1}$ times) the optimal solution of the LP (**LP**). This difference is present in our paper, as we require solving the LP (**LP**) *optimally with probability* 1 at each time slot, in order to ensure a $\frac{d_{\max}}{2d_{\max}-1}$-approximation algorithm. This is in contrast with the combinatorial bandits literature [Wang and Chen, 2017, Chen et al., 2016], where in each time slot the oracle provides an $\alpha$-approximate solution to the combinatorial problem with probability at least $\beta$, for $\alpha, \beta \in (0,1]$. Our approximation loss comes from the online rounding, rather than from the LP solution at each time slot.

## D CONCENTRATION INEQUALITIES

In this section, we outline the standard concentration results that we use in our proofs.

**Theorem 4** (Hoeffding's Inequality)**.** [5] *Let $X_1, \dots, X_n$ be independent identically distributed random variables with common support in $[0,1]$ and mean $\mu$. Let $Y = X_1 + \dots + X_n$. Then for all $\delta \geq 0$,*

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

**Theorem 5** (Multiplicative Chernoff Bound)**.** [6] *Let $X_1, \dots, X_n$ be Bernoulli random variables taking values from $\{0,1\}$, and $\mathbb{E}\left[X_t | X_{t-1}, \dots, X_1\right] \geq \mu$ for every $t \leq n$. Let $Y = X_1 + \dots + X_n$. Then, for all $0 < \delta < 1$,*

$$\mathbb{P}\left(Y \leq (1-\delta)n\mu\right) \leq e^{-\frac{\delta^2 n\mu}{2}}.$$

## E FULL-INFORMATION PROBLEM AND COMPETITIVE ANALYSIS: OMITTED PROOFS

### E.1 Proof of Theorem 1

We now prove a lower bound on the competitive guarantee of FI-CBB, against any optimal clairvoyant algorithm. The proofs of the lemmas we use in the proof of the following theorem are also contained in this section of the Appendix.

**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 $d_{\max} = \max_{i\in\mathcal{A}} d_i$.*

[5] This is a standard concentration result and the statement can be found, e.g., in [Lattimore and Szepesvári, 2018]

[6] The result is a combination of Theorem 4.5 and Exercise 4.7 in [Mitzenmacher and Upfal, 2017], in the case where the $\{X_i\}_{i\in[n]}$ are independent. The authors in [Wang and Chen, 2017, Chen et al., 2016] describe a slight modification that directly proves the statement.

*Proof.* The first step in our analysis is to show that the optimal solution of (**LP**), denoted by $\mathbf{Rew}_I^{LP}$ yields a $\left(1-\frac{d_{\max}-1}{d_{\max}-1+T}\right)$-approximate upper bound to the maximum (average) expected reward collected by any (clairvoyant) algorithm, denoted by $\mathbf{Rew}_I^*(T)$. Note that, since $\mathbf{Rew}_I^{LP}$ represents an upper bound on the average collected reward, we multiply it with $T$, in order to compare it with $\mathbf{Rew}_I^*(T)$. Finally, we emphasize that the multiplicative approximation of the upper bound asymptotically goes to 1 as $T$ increases.

**Lemma 5.** *For any time horizon $T$, we have*

$$T \cdot \boldsymbol{Rew}_I^{LP} \geq \left(1 - \frac{d_{\max}-1}{d_{\max}-1+T}\right) \boldsymbol{Rew}_I^*(T).$$

We denote by $F_{i,t}^\pi$ the event that arm $i$ is available in time $t$, and by $A_t^\pi$ the arm played at time $t$, where $A_t^\pi \in \{\mathcal{A} \cup \emptyset\}$. Moreover, we denote the event of playing arm $i$ at context $j$ at time $t$ as $\mathbb{I}(A_t^\pi = i, C_t = j)$ for all $i \in \mathcal{A}$ and $j \in \mathcal{C}$. We fix a time horizon $T$, for the purpose of the analysis.

The FI-CBB algorithm at each time $t$ plays an arm $i$ if it is *(i)* sampled, *(ii)* available and *(iii)* not skipped. The sampling of arm $i$ under context $j \in \mathcal{C}$ happens with probability $z_{i,j}^*$ and the arm is not skipped with probability $\beta_{i,t}$, independently. Finally, the arm is played if it is available, which happens independently of sampling and skipping, with probability $\mathbb{P}\left(F_{i,t}^\pi\right)$. The above analysis leads to a recursive characterization of $\mathbb{P}\left(F_{i,t}^\pi\right)$. Upon inspection, this is the same characterization as for $q_{i,t}$ given in Eq. (1). We formally summarize the above in the following lemma:

**Lemma 6.** *At every round $t \in [T]$ and for any arm $i \in \mathcal{A}$ and context $j \in \mathcal{C}$, it is the case that $\mathbb{P}(A_t^\pi = i, C_t = j) = z_{i,j}^* \beta_{i,t} \mathbb{P}\left(F_{i,t}^\pi\right)$. Moreover, we have $q_{i,t} = \mathbb{P}\left(F_{i,t}^\pi\right)$, $\forall i \in \mathcal{A}, t \in [T]$.*

We observe that by design of the skipping mechanism $\beta_{i,t}$, the quantity $\beta_{i,t} \cdot \mathbb{P}\left(F_{i,t}^\pi\right)$ never exceeds $\frac{d_i}{2d_i-1}$. Leveraging this observation, we show that at every time $t \in [T]$, it is the case that $\mathbb{P}\left(F_{i,t}^\pi\right) \geq \frac{d_i}{2d_i-1}$. This allows us to completely characterize the behavior of the algorithm as it is shown in the following lemma:

**Lemma 7.** *At every round $t \in [T]$, the probability that* FI-CBB *plays an arm $i \in \mathcal{A}$ under context $j \in \mathcal{C}$ is exactly $\mathbb{P}(A_t^\pi = i, C_t = j) = \frac{d_i}{2d_i-1} z_{i,j}^*$.*

In order to complete the proof of the theorem, the expected cumulative reward collected by FI-CBB in $T$ time steps can be expressed as

$$\begin{aligned}
\mathbf{Rew}_I^\pi(T) &= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\pi}}\left[\sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} X_{i,j,t}\,\mathbb{I}(A_t^\pi = i, C_t = j)\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{R}_C\mathcal{R}_\pi}\left[\sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mathop{\mathbb{E}}_{\mathcal{R}_X}[X_{i,j,t}]\,\mathbb{I}(A_t^\pi = i, C_t = j)\right] &(3)\\
&= \mathop{\mathbb{E}}_{\mathcal{R}_C\mathcal{R}_\pi}\left[\sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\,\mathbb{I}(A_t^\pi = i, C_t = j)\right] \\
&= \sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j} \mathop{\mathbb{E}}_{\mathcal{R}_C\mathcal{R}_\pi}[\mathbb{I}(A_t^\pi = i, C_t = j)] \\
&= T\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\frac{d_i}{2d_i-1} z_{i,j}^* &(4)\\
&\geq \frac{d_{\max}}{2d_{\max}-1} T \cdot \mathbf{Rew}_I^{LP} \\
&\geq \frac{d_{\max}}{2d_{\max}-1}\left(1-\frac{d_{\max}-1}{d_{\max}-1+T}\right)\mathbf{Rew}_I^*(T), &(5)
\end{aligned}$$

where (3) follows by independence of $\{X_{i,j,t}\}_{i,j,t}$, (4) follows by Lemma 7 and (5) follows by Lemma 5. □

### E.2 Proof of Lemma 5

**Lemma 5.** *For any time horizon $T$, we have*

$$T \cdot \boldsymbol{Rew}_I^{LP} \geq \left(1 - \frac{d_{\max} - 1}{d_{\max} - 1 + T}\right) \boldsymbol{Rew}_I^*(T).$$

*Proof.* We denote by $\Sigma : [T] \to \mathcal{C}$ a fixed sequence of context realizations over $T$ rounds, where, at each time step $t \in [T]$, context $j \in \mathcal{C}$ appears independently with probability $f_j$. Let $\mathcal{S}$ be the family of all possible sequences. Given that the context of each round is sampled independently according to the fixed probabilities $\{f_j\}_{j \in \mathcal{C}}$, the probability of each sequence is given by $\mathbb{P}(\Sigma) = \prod_{t \in [T]} f_{\Sigma(t)}$. Note that we overload the notation and denote by $\Sigma$ the event that the sequence is realized.

Consider the optimal clairvoyant algorithm that first observes the full context realization and, then, chooses a fixed feasible arm-pulling sequence that yields the maximum expected reward for this realization. Let $\mathbb{I}(A_t^* = i, C_t = j \mid \Sigma)$ be the indicator of the event that under the realization $\Sigma$, the optimal algorithm plays arm $i$ on time $t$ under context $j$. We emphasize the fact that the event $C_t = j$ is deterministic conditioned on the realization $\Sigma$. Finally, notice that we can assume w.l.o.g. that there exists an optimal clairvoyant policy maximizing the expected reward that ignores the realizations of the collected rewards.

We fix any realization $\Sigma \in \mathcal{S}$. In any feasible solution and for any arm $i \in \mathcal{A}$, we have

$$\sum_{t' \in [t, t+d_i-1]} \sum_{j \in \mathcal{C}} \mathbb{I}(A_t^* = i, C_t = j \mid \Sigma) \leq 1, \qquad \forall t \in [T],$$

as the arm can be played at most once during any $d_i$ consecutive time steps. By summing the above inequalities over all $t \in [T]$, for any arm $i \in \mathcal{A}$, we get

$$\begin{aligned} &\sum_{t \in [1, d_i-1]} t \sum_{j \in \mathcal{C}} \mathbb{I}(A_t^* = i, C_t = j \mid \Sigma) + \sum_{t \in [d_i, T]} d_i \sum_{j \in \mathcal{C}} \mathbb{I}(A_t^* = i, C_t = j \mid \Sigma) \leq T \\ \Leftrightarrow &\sum_{t \in [T]} d_i \sum_{j \in \mathcal{C}} \mathbb{I}(A_t^* = i, C_t = j \mid \Sigma) \leq T + \sum_{t \in [1, d_i-1]} (d_i - t) \sum_{j \in \mathcal{C}} \mathbb{I}(A_t^* = i, C_t = j \mid \Sigma). \end{aligned}$$

By feasibility of (**LP**), we have that $\sum_{t \in [1, d_i-1]} (d_i - t) \sum_{j \in \mathcal{C}} \mathbb{I}(A_t^* = i, C_t = j \mid \Sigma) \leq d_i - 1$. Therefore, by dividing the above inequality by $d_i \cdot T$, we get

$$\frac{1}{T} \sum_{t \in [T]} \sum_{j \in \mathcal{C}} \mathbb{I}(A_t^* = i, C_t = j \mid \Sigma) \leq \frac{1}{d_i} \left(1 + \frac{d_i - 1}{T}\right), \forall i \in \mathcal{A}.$$

Now, by multiplying the above inequality with the probability of each context realization $\Sigma$ and taking the sum over all $\Sigma \in \mathcal{S}$, we get

$$\sum_{j \in \mathcal{C}} \sum_{\Sigma \in \mathcal{S}} \frac{1}{T} \sum_{t \in [T]} \mathbb{I}(A_t^* = i, C_t = j \mid \Sigma) \mathbb{P}(\Sigma) \leq \frac{1}{d_i} \left(1 + \frac{d_i - 1}{T}\right), \forall i \in \mathcal{A}. \tag{6}$$

For each context $j \in \mathcal{C}$ and any time $t \in [T]$, we have

$$\sum_{i \in \mathcal{A}} \mathbb{I}(A_t^* = i, C_t = j \mid \Sigma) \leq \mathbb{I}(C_t = j \mid \Sigma),$$

where the inequality follows by the fact that at most one arm is played at each time in any feasible solution. By taking the expectation in the above expression over the context realization, we get

$$\mathop{\mathbb{E}}_{\mathcal{R}_C} \left[\sum_{i \in \mathcal{A}} \mathbb{I}(A_t^* = i, C_t = j \mid \Sigma)\right] \leq \mathop{\mathbb{E}}_{\mathcal{R}_C} [\mathbb{I}(C_t = j \mid \Sigma)] = \sum_{\Sigma \in \mathcal{S}} \mathbb{I}(C_t = j \mid \Sigma) \mathbb{P}(\Sigma) = f_j,$$

where the last equality follows by the fact that the probability that any context realization sequence satisfies $C_t = j$ is exactly $f_j$. Finally, by taking the sum of the above inequality over all $t \in [T]$ and dividing by $T$ yields

$$\sum_{i\in\mathcal{A}}\sum_{\Sigma\in\mathcal{S}}\frac{1}{T}\sum_{t\in[T]}\mathbb{I}\left(A^*_t = i, C_t = j \mid \Sigma\right)\mathbb{P}\left(\Sigma\right) \leq f_j. \tag{7}$$

For the expected cumulative reward of the above optimal clairvoyant policy, we have:

$$\begin{aligned}
\mathbf{Rew}^*_I(T) &= \mathop{\mathbb{E}}_{\mathcal{R}_X,\mathcal{R}_C}\left[\max_{\text{feasible}\{A^*_t\}_{t\in[T]}}\left\{\sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}X_{i,j,t}\,\mathbb{I}\left(A^*_t = i, C_t = j\right)\right\}\right]\\
&= \mathop{\mathbb{E}}_{\mathcal{R}_X}\left[\sum_{\Sigma\in\mathcal{S}}\sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}X_{i,j,t}\,\mathbb{I}\left(A^*_t = i, C_t = j \mid \Sigma\right)\mathbb{P}\left(\Sigma\right)\right]\\
&= \sum_{\Sigma\in\mathcal{S}}\sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mathop{\mathbb{E}}_{\mathcal{R}_X}\left[X_{i,j,t}\right]\mathbb{I}\left(A^*_t = i, C_t = j \mid \Sigma\right)\mathbb{P}\left(\Sigma\right)\\
&= \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\sum_{\Sigma\in\mathcal{S}}\sum_{t\in[T]}\mu_{i,j}\,\mathbb{I}\left(A^*_t = i, C_t = j \mid \Sigma\right)\mathbb{P}\left(\Sigma\right),
\end{aligned} \tag{8}$$

where the second and third equalities follow by the fact that the optimal clairvoyant policy plays a fixed arm-pulling solution for any observed context realization sequence and that this solution is independent of the observed reward realizations.

Consider now a (candidate) solution of (**LP**), such that:

$$z_{i,j} = \left(1 + \frac{d_{\max}-1}{T}\right)^{-1}\sum_{\Sigma\in\mathcal{S}}\frac{1}{T}\sum_{t\in[T]}\mathbb{I}\left(A^*_t = i, C_t = j \mid \Sigma\right)\mathbb{P}\left(\Sigma\right), \forall i\in\mathcal{A}, j\in\mathcal{C}\,.$$

It is not hard to verify that, for this assignment, constraints (**C1**) and (**C2**) are satisfied by making use of (6) and (7), respectively. Moreover, for the objective of (**LP**), using (8), we have:

$$\begin{aligned}
T\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mu_{i,j}z_{i,j} &= \left(1 + \frac{d_{\max}-1}{T}\right)^{-1}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mu_{i,j}\sum_{\Sigma\in\mathcal{S}}\sum_{t\in[T]}\mathbb{I}\left(A^*_t = i, C_t = j \mid \Sigma\right)\mathbb{P}\left(\Sigma\right)\\
&= \left(1 + \frac{d_{\max}-1}{T}\right)^{-1}\mathbf{Rew}^*_I(T)\\
&= \left(1 - \frac{d_{\max}-1}{d_{\max}-1+T}\right)\mathbf{Rew}^*_I(T),
\end{aligned}$$

where in the last equality follows by the fact that $\frac{1}{1+\delta} = 1 - \frac{\delta}{1+\delta}$ for any $\delta\in\mathbb{R}$. Therefore, by exhibiting a feasible solution to (**LP**) of value $\left(1 - \frac{d_{\max}-1}{d_{\max}-1+T}\right)\mathbf{Rew}^*_I(T)$, we can conclude that $\mathbf{Rew}^{LP}_I \geq \left(1 - \frac{d_{\max}-1}{d_{\max}-1+T}\right)\mathbf{Rew}^*_I(T)$. □

### E.3 Proof of Lemma 6

**Lemma 6.** *At every round $t\in[T]$ and for any arm $i\in\mathcal{A}$ and context $j\in\mathcal{C}$, it is the case that $\mathbb{P}\left(A^\pi_t = i, C_t = j\right) = z^*_{i,j}\beta_{i,t}\,\mathbb{P}\left(F^\pi_{i,t}\right)$. Moreover, we have $q_{i,t} = \mathbb{P}\left(F^\pi_{i,t}\right)$, $\forall i\in\mathcal{A}, t\in[T]$.*

*Proof.* Although our algorithm fi-cbb computes and uses an optimal extreme point solution to (**LP**), the analysis that follows holds for any feasible solution $\{z_{i,j}\}_{i,j}$. We denote by $S^\pi_{i,t}$ the event that arm $i$ is sampled by fi-cbb at round $t$ (with probability $\mathbb{P}\left(S^\pi_{i,t}\right) = \frac{z_{i,j_t}}{f_{j_t}}$ for a sampled context $j_t$) and by $B^\pi_{i,t}$ the event that arm $i$ is not skipped at round $t$. Finally, we denote by $F^\pi_{i,t}$ the event that arm $i$ is available at the beginning of round $t$.

In order to prove the first part of the claim, we first notice that the event $\{A^\pi_t = i\}$ is equivalent to $\{S^\pi_{i,t}, B^\pi_{i,t}, F^\pi_{i,t}\}$, namely, in order for an arm $i$ to be played during $t$, the arm needs to be sampled, not skipped and available.

For any fixed $i \in \mathcal{A}$, $j \in \mathcal{C}$ and $t \in [T]$, we have:

$$\begin{aligned}
\mathbb{P}\left(A^\pi_t = i, C_t = j\right) &= \mathbb{P}\left(A^\pi_t = i | C_t = j\right) \mathbb{P}\left(C_t = j\right) \\
&= f_j \, \mathbb{P}\left(A^\pi_t = i | C_t = j\right) && (9) \\
&= f_j \, \mathbb{P}\left(S^\pi_{i,t}, F^\pi_{i,t}, B^\pi_{i,t} | C_t = j\right) \\
&= f_j \, \mathbb{P}\left(S^\pi_{i,t} | C_t = j\right) \mathbb{P}\left(B^\pi_{i,t} | C_t = j\right) \mathbb{P}\left(F^\pi_{i,t} | C_t = j\right) && (10) \\
&= f_j \, \mathbb{P}\left(S^\pi_{i,t} | C_t = j\right) \mathbb{P}\left(B^\pi_{i,t}\right) \mathbb{P}\left(F^\pi_{i,t}\right) && (11) \\
&= f_j \beta_{i,t} \, \mathbb{P}\left(S^\pi_{i,t} | C_t = j\right) \mathbb{P}\left(F^\pi_{i,t}\right) && (12) \\
&= f_j \frac{z_{i,j}}{f_j} \beta_{i,t} \, \mathbb{P}\left(F^\pi_{i,t}\right) && (13) \\
&= z_{i,j} \beta_{i,t} \, \mathbb{P}\left(F^\pi_{i,t}\right).
\end{aligned}$$

In the above analysis, equality (9) follows by the fact that $\mathbb{P}\left(C_t = j\right) = f_j$, while in (10) we use the fact that the events $S^\pi_{i,t}$, $B^\pi_{i,t}$ and $F^\pi_{i,t}$ are mutually independent, by construction of our algorithm. Moreover, in (11) we use that the events $B^\pi_{i,t}$ and $F^\pi_{i,t}$ are independent of the observed type $j \in \mathcal{C}$. Finally, in (12) and (13), we use the fact that $\mathbb{P}\left(S^\pi_{i,t} | C_t = j\right) = \frac{z_{i,j}}{f_j}$ and $\mathbb{P}\left(B^\pi_{i,t}\right) = \beta_{i,t}$, by construction of our algorithm.

We now prove the second part of the statement, namely, that the computed probabilities, $\{q_{i,t}\}_{\forall i,t}$ (by the recursive formula (1)), indeed match the actual a priori probabilities of the events $\{F^\pi_{i,t}\}_{\forall i,t}$. The main idea behind the computation of $q_{i,t}$ is that an arm is available at some round $t$, if it is available but not played at time $t-1$, or if it is played at time $t - d_i$.

For any fixed arm $i \in \mathcal{A}$, we prove the statement by induction on the number of rounds. Note that we only consider arms such that $d_i \geq 2$, since, otherwise, we trivially have that $q_{i,t} = \mathbb{P}\left(F^\pi_{i,t}\right) = 1, \forall t \in [T]$. Clearly, for $t = 1$ the computed probabilities are correct, since $\mathbb{P}\left(F^\pi_{i,1}\right) = 1$. We assume that up to round $t-1$, the computed probabilities are correct, namely, $q_{i,t'} = \mathbb{P}\left(F^\pi_{i,t'}\right)$, $\forall t' \in [t-1]$. Considering the event $F^\pi_{i,t}$, we have:

$$\begin{aligned}
\mathbb{I}\left(F^\pi_{i,t}\right) &= \mathbb{I}\left(F^\pi_{i,t}, F^\pi_{i,t-1}\right) + \mathbb{I}\left(F^\pi_{i,t}, \neg F^\pi_{i,t-1}\right) \\
&= \mathbb{I}\left(F^\pi_{i,t}, F^\pi_{i,t-1}\right) + \mathbb{I}\left(F^\pi_{i,t}, \neg F^\pi_{i,t-1}, t \geq d_i + 1\right) + \mathbb{I}\left(F^\pi_{i,t}, \neg F^\pi_{i,t-1}, t \leq d_i\right) \\
&= \mathbb{I}\left(F^\pi_{i,t}, F^\pi_{i,t-1}\right) + \mathbb{I}\left(F^\pi_{i,t}, \neg F^\pi_{i,t-1}, t \geq d_i + 1\right) && (14) \\
&= \mathbb{I}\left(F^\pi_{i,t}, F^\pi_{i,t-1}\right) + \mathbb{I}\left(F^\pi_{i,t}, \neg F^\pi_{i,t-1}, F^\pi_{i,t-d_i}, t \geq d_i + 1\right) \\
&\qquad\qquad + \mathbb{I}\left(F^\pi_{i,t}, \neg F^\pi_{i,t-1}, \neg F^\pi_{i,t-d_i}, t \geq d_i + 1\right) \\
&= \mathbb{I}\left(F^\pi_{i,t}, F^\pi_{i,t-1}\right) + \mathbb{I}\left(F^\pi_{i,t}, \neg F^\pi_{i,t-1}, F^\pi_{i,t-d_i}, t \geq d_i + 1\right). && (15)
\end{aligned}$$

In equality (14), we use the fact that the event $\{F^\pi_{i,t}, \neg F^\pi_{i,t-1}, t \leq d_i\}$ is empty. This follows by noticing that for $t \leq d_i$, if an arm is not available at round $t-1$, then it has to be pulled during some round $t' \in [t-1] \subseteq [d_i - 1]$ and, thus, cannot be available on round $t$. Similarly, in (15), we use the fact that the event $\{F^\pi_{i,t}, \neg F^\pi_{i,t-1}, \neg F^\pi_{i,t-d_i}, t \geq d_i + 1\}$ is empty. The reason is that, if the arm is not available at round $t - d_i$ and neither at round $t-1$, this implies that the arm is pulled during some round $t' \in [t - d_i + 1, t - 2]$. However, if the arm is played at any such $t'$, then it cannot be available at round $t$.

Notice, that the event $\{F^\pi_{i,t}, F^\pi_{i,t-1}\}$ occurs with probability $(1 - \beta_{i,t-1} \sum_{j \in \mathcal{C}} z_{i,j}) \, \mathbb{P}\left(F^\pi_{i,t-1}\right)$, since the arm is, either not selected on round $t-1$, i.e., $\mathbb{I}\left(S^\pi_{i,t-1}\right) = 0$, or skipped, i.e., $\mathbb{I}\left(B^\pi_{i,t-1}\right) = 0$. Moreover, the event $\{F^\pi_{i,t}, \neg F^\pi_{i,t-1}, F^\pi_{i,t-d_i}, t \geq d_i + 1\}$, for $t \geq d_i + 1$ is equivalent to the event $\{A^\pi_{t-d_i} = i\}$, since the arm has to be played at time $t - d_i$, in order to be available at round $t$ for the first time after $t - d_i$. By taking expectations in (15) and combining the above facts, we have:

$$\begin{aligned}
\mathbb{P}\left(F^{\pi}_{i,t}\right) &= \mathbb{P}\left(F^{\pi}_{i,t}, F^{\pi}_{i,t-1}\right) + \mathbb{I}\left(t \geq d_i + 1\right) \mathbb{P}\left(F^{\pi}_{i,t}, \neg F^{\pi}_{i,t-1}, F^{\pi}_{i,t-d_i}\right) \\
&= \left(1 - \mathbb{P}\left(S^{\pi}_{i,t-1}, B^{\pi}_{i,t-1}\right)\right) \mathbb{P}\left(F^{\pi}_{i,t-1}\right) + \mathbb{I}\left(t \geq d_i + 1\right) \mathbb{P}\left(A^{\pi}_{t-d_i} = i\right) \\
&= \left(1 - \mathbb{P}\left(S^{\pi}_{i,t-1}, B^{\pi}_{i,t-1}\right)\right) \mathbb{P}\left(F^{\pi}_{i,t-1}\right) + \mathbb{I}\left(t \geq d_i + 1\right) \mathbb{P}\left(S^{\pi}_{i,t-d_i}, B^{\pi}_{i,t-d_i}, F^{\pi}_{i,t-d_i}\right) \\
&= \left(1 - \sum_{j \in \mathcal{C}} \mathbb{P}\left(S^{\pi}_{i,t-1}, B^{\pi}_{i,t-1}, C_{t-1} = j\right)\right) \mathbb{P}\left(F^{\pi}_{i,t-1}\right) \\
&\qquad\qquad + \mathbb{I}\left(t \geq d_i + 1\right) \sum_{j \in \mathcal{C}} \mathbb{P}\left(S^{\pi}_{i,t-d_i}, B^{\pi}_{i,t-d_i}, F^{\pi}_{i,t-d_i}, C_{t-d_i} = j\right) \\
&= \left(1 - \beta_{i,t-1} \sum_{j \in \mathcal{C}} z_{i,j}\right) \mathbb{P}\left(F^{\pi}_{i,t-1}\right) + \mathbb{I}\left(t \geq d_i + 1\right) \beta_{i,t-d_i} \left(\sum_{j \in \mathcal{C}} z_{i,j}\right) \mathbb{P}\left(F^{\pi}_{i,t-d_i}\right), \qquad (16)
\end{aligned}$$

where (16) follows by the analysis of the first part of this proof. By setting $t+1$ instead of $t$ in the above relation and setting $z_{i,j} = z^*_{i,j}, \forall i \in \mathcal{A}, j \in \mathcal{C}$, we can easily verify that the formula that computes these probabilities in formula (1) and Algorithm 2 is correct, which concludes the proof of this lemma. □

### E.4 Proof of Lemma 7

**Lemma 7.** *At every round $t \in [T]$, the probability that* FI-CBB *plays an arm $i \in \mathcal{A}$ under context $j \in \mathcal{C}$ is exactly $\mathbb{P}\left(A^{\pi}_t = i, C_t = j\right) = \frac{d_i}{2d_i - 1} z^*_{i,j}$.*

*Proof.* Similarly to the proof of Lemma 6, the analysis of this proof holds true for any feasible solution $\{z_{i,j}\}_{\forall i,j}$ of (**LP**), including the optimal extreme point solution. Recall that by Lemma 6, the probability of each event $F^{\pi}_{i,t}$ is equal to the actual probability of the event, namely, $q_{i,t} = \mathbb{P}\left(F^{\pi}_{i,t}\right), \forall i \in \mathcal{A}, t \in [T]$. Moreover, by the same lemma, we have:

$$\mathbb{P}\left(A^{\pi}_t = i, C_t = j\right) = z^*_{i,j} \beta_{i,t} \mathbb{P}\left(F^{\pi}_{i,t}\right) \qquad (17)$$

Recall that $\beta_{i,t} = \min\{1, \frac{d_i}{2d_i-1} \frac{1}{q_{i,t}}\}$. We now prove by induction that for every fixed arm $i \in \mathcal{A}$ and for every time $t \in [T]$, it is the case that: $\mathbb{P}\left(A^{\pi}_t = i, C_t = j\right) = \frac{d_i}{2d_i-1} z^*_{i,j}$. Clearly, for $t = 1$, we have $\mathbb{P}\left(F^{\pi}_{i,1}\right) = 1 = q_{i,1}$ (by initialization) and, thus, $\beta_{i,1} = \frac{d_i}{2d_i-1}$, implying that $\mathbb{P}\left(A^{\pi}_1 = i, C_1 = j\right) = \frac{d_i}{2d_i-1} z^*_{i,j}$. Suppose the argument is true for any $\tau \in [t-1]$. For time $t$, we distinguish between two cases:

**Case (a).** Suppose $\beta_{i,t} < 1$. Then, by construction, it has to be that $\beta_{i,t} = \frac{d_i}{2d_i-1} \frac{1}{q_{i,t}}$, while by Lemma 6, we have that $q_{i,t} = \mathbb{P}\left(F^{\pi}_{i,t}\right)$. By (17), this immediately implies that $\mathbb{P}\left(A^{\pi}_t = i, C_t = j\right) = \frac{d_i}{2d_i-1} z^*_{i,j}$.

**Case (b).** Suppose $\beta_{i,t} = 1$. Then, by definition of $\beta_{i,t}$, it has to be that $q_{i,t} \leq \frac{d_i}{2d_i-1}$, which in turn implies that $\mathbb{P}\left(F^{\pi}_{i,t}\right) \leq \frac{d_i}{2d_i-1}$. Therefore, we can upper bound the probability of interest as: $\mathbb{P}\left(A_t = i, C_t = j\right) = z^*_{i,j} \beta_{i,t} \mathbb{P}\left(F^{\pi}_{i,t}\right) \leq \frac{d_i}{2d_i-1} z^*_{i,j}$. In order to complete the induction step, it suffices to also show that $\mathbb{P}\left(A^{\pi}_t = i, C_t = j\right) \geq \frac{d_i}{2d_i-1} z^*_{i,j}$. By a simple union bound, we can lower bound the probability of arm $i$ being available using the probabilities that the arm has been played within time $[t - d_i + 1, t - 1]$:

$$\begin{aligned}
\mathbb{P}\left(A^{\pi}_t = i, C_t = j\right) &= z^*_{i,j} \mathbb{P}\left(F^{\pi}_{i,t}\right) \\
&= z^*_{i,j} \left(1 - \mathbb{P}\left(\neg F^{\pi}_{i,t}\right)\right) \\
&\geq z^*_{i,j} \left(1 - \sum_{t' \in [t-d_i+1, t-1]} \sum_{j' \in \mathcal{C}} \mathbb{P}\left(A^{\pi}_{t'} = i, C_{t'} = j'\right)\right).
\end{aligned}$$

However, by induction hypothesis we know that $\forall t' \in [t - d_i + 1, t - 1]$ and $\forall j' \in \mathcal{C}$, it is the case that $\mathbb{P}\left(A^{\pi}_{t'} = i, C_{t'} = j'\right) = \frac{d_i}{2d_i-1} z^*_{i,j'}$. Moreover, by constraints (**C1**) of (**LP**), we know that

$\sum_{t' \in [t-d_i+1, t-1]} \sum_{j' \in \mathcal{C}} z^*_{i,j'} \leq \frac{d_i - 1}{d_i}$. Combining the above facts, we have:

$$\begin{aligned}
\mathbb{P}\left(A^\pi_t = i, C_t = j\right) &\geq z^*_{i,j} \left(1 - \frac{d_i}{2d_i - 1} \sum_{t' \in [t-d_i+1, t-1]} \sum_{j' \in \mathcal{C}} z^*_{i,j'}\right) \\
&\geq z^*_{i,j} \left(1 - \frac{d_i}{2d_i - 1} \frac{d_i - 1}{d_i}\right) \\
&\geq \frac{d_i}{2d_i - 1} z^*_{i,j},
\end{aligned}$$

which completes our induction step, since the combination of two inequalities implies that $\mathbb{P}\left(A^\pi_t = i, C_t = j\right) = \frac{d_i}{2d_i - 1} z^*_{i,j}$. □

# F BANDIT PROBLEM AND REGRET ANALYSIS: OMITTED PROOFS

## F.1 Properties of $M_t$ and the Critical Time $T_c$

We consider the delayed exploitation parameter, $M_t$, specifically defined as

$$M_t = \lfloor 2\log_{c_1}(t) \rfloor + \lceil \log_{c_1}(c_0) \rceil + 1 = \lfloor 2\log_{c_1}(t) \rfloor + 2d_{\max} + 8,$$

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

We define the *critical* round $T_c$, as the smallest integer such that $T_c - M_{T_c} \geq 1$. It is not hard to verify that, by definition of $M_t$, this implies that $t - M_t \geq 1$ for all $t \geq T_c$ (see the next paragraph), and $t - M_t \leq 0$ for all $t \leq T_c - 1$ (by definition of $T_c$). By definition of the algorithm, at each round $t \geq T_c$ and in order to sample the next arm to be played, UCB-CBB uses an extreme point solution computed with respect to the UCB estimates before exactly $M_t$ time steps (i.e., at round $t - M_t$). For $t \leq T_c$, where $t - M_t \leq 0$, the algorithm uses an initially computed extreme point solution $Z(0) = \{z_{i,j}(0)\}_{i,j}$. In this section, we study several useful properties of $M_t$ and $T_C$.

**Bounded Increases of $M_t$.** We now show that for $t \geq T_c$, the value of $M_t$ increases by at most one unit per round. This fact significantly simplifies our proofs and results in the analysis of the $\alpha$-regret.

Let us first compute the condition that must be satisfied for any $t \in [T]$, such that the value $t - M_t$ is strictly positive. Formally

$$t - M_t \geq 1 \iff t - \lfloor 2\log_{c_1}(t) \rfloor \geq 2d_{\max} + 9.$$

By noticing that $d_{\max} \geq 1$, we can easily verify that by the time $t - \lfloor 2\log_{c_1}(t) \rfloor \geq 2d_{\max} + 9$, it also holds that $t - \lfloor 2\log_{c_1}(t) \rfloor \geq 11$, which, in turn, implies that $t \geq 69$.

We are now looking for the smallest $t$, after which $M_t$ increases by at most one unit per round. Consider the (fractional) *breakpoints* of the form $c_1^{i/2}$ for any positive integer $i$. These breakpoints, corresponds to the points such that the value of $M_t$ increases, when time $t$ passes them. We consider intervals of the form $C_i = [c_1^{i/2}, c_1^{(i+1)/2})$. The first step is to find the smallest $i$, such that there is at least one integral point in $[c_1^{i/2}, c_1^{(i+1)/2})$. Notice that the above condition is true if

$$c_1^{(i+1)/2} - c_1^{i/2} \geq 1 \iff c_1^{i/2} \geq \frac{1}{\sqrt{c_1} - 1} \approx 13.25.$$

Therefore, for any $t \geq 14$, the value of $M_t$ increases by at most one unit per time step.

We can conclude that, for any $t \in [T]$ such that $t - M_t \geq 1$ (in other words $t \geq T_c$), the value of $M_t$ changes by at most one unit per time step, since in that case $t \geq 69 \geq 14$.

**Fact 1.** *For any $t \geq T_c$, the value of $M_t$ increases by at most one unit per round, namely, $M_t \leq M_{t-1}+1, \forall t \geq T_c$.*

Notice that the above fact implies that for $t \geq T_c$, the value $t - M_t$ is nondecreasing.

**Upper and Lower Bounds on $T_c$.** We would like to compute some non-trivial upper and lower bounds on the value of $T_c$. The lower bound is used in the proof of Lemma 1, while the upper bound is used in Lemma 2.

We first compute an upper bound on $T_c$. Recall, that, by definition, $T_c$ is the smallest positive integer such that $T_c - \lfloor 2\log_{c_1}(T_c) \rfloor \geq 2d_{\max} + 9$. Therefore, for $T_c - 1$ it has to be the case that

$$T_c - 1 \leq 2d_{\max} + 8 + \lfloor 2\log_{c_1}(T_c - 1) \rfloor \leq 2d_{\max} + 8 + 2\log_{c_1}(T_c)$$

We can get an upper bound to $T_c$, by noticing that for any $t \geq 1$, it is the case that $2\log_{c_1}(t) \leq t/3 + 38$. Using this, we can see that

$$T_c \leq 2d_{\max} + T_c/3 + 47.$$

By the above, we conclude that $T_c \leq \frac{3}{2}(2d_{\max} + 47) \leq 3d_{\max} + 71$.

We are now looking for a lower bound on $T_c$. Since $T_c$ satisfies $T_c - M_{T_c} \geq 1$, by the analysis of the previous paragraph (on the boundedness of $M_t$), it has to be that $T_c \geq 69$. Using that, we have

$$T_c \geq 2d_{\max} + 9 + \lfloor 2\log_{c_1}(T_c) \rfloor \geq 2d_{\max} + 9 + \lfloor 2\log_{c_1}(69) \rfloor \geq 2d_{\max} + 67.$$

**Fact 2.** *We can bound $T_c$ as $2d_{\max} + 67 \leq T_c \leq 3d_{\max} + 71$.*

Consider now any $t \geq T_c$ and any $t' \in [t - d_{\max}, t-1]$. We have:

$$t' \geq t - d_{\max} \geq T_c - d_{\max} \geq 2d_{\max} + 67 - d_{\max} \geq d_{\max} + 67,$$

where in the second inequality we use Fact 2. Therefore, since $t' \geq 67 \geq 14$, then by the above paragraph (on the bounded increases of $M_t$), we have that for any $\tau \in [t', t]$, the value of $M_t$ is increased by at most one.

**Fact 3.** *For any $t \geq T_c \geq d_{\max}$ and $t' \in [t - d_{\max}, t-1]$, then for any $\tau \in [t', t]$ we have that $M_\tau \leq M_{\tau-1} + 1$. This also implies that $t' - M_{t'} \leq \tau - M_\tau \leq t - M_t$ for any $\tau \in [t', t]$.*

**Correctness of Delayed Exploitation.** Finally, we present one additional property that is proved useful in proving the correctness of the routine COMPQ$(i, t, H_{t-M_t})$ and the overall correctness of our algorithm. Specifically, we would like to prove the following inequality for any $t \in [T]$, $t' \in [t - d_{\max}, t-1]$ and $\tau \in [t' - M_{t'}, t'-1]$:

$$\max\{\tau - M_\tau, 0\} \leq \max\{t' - M_{t'}, 0\} \leq \max\{t - M_t, 0\}.$$

Consider any fixed $t, t'$ and $\tau$ that satisfy $t' \in [t - d_{\max}, t-1]$ and $\tau \in [t' - M_{t'}, t'-1]$. We first notice that if $\tau - M_\tau \leq 0$, then we trivially have that $\max\{\tau - M_\tau, 0\} \leq \max\{t' - M_{t'}, 0\}$ and $\max\{\tau - M_\tau, 0\} \leq \max\{t - M_t, 0\}$. We focus on the case where $\tau - M_\tau \geq 1$. By the above analysis, we can see that for any $\tau$ such that $\tau - M_\tau \geq 1$, it has to be the case that $\tau \geq 69 \geq 14$ and, thus, for any time step $\tau'$ in the interval $\tau' \in [\tau, t-1]$ the value of $M_{\tau'}$ increases by at most one unit. This immediately guarantees that $\tau - M_\tau \leq t' - M_{t'} \leq t - M_t$.

Consider now the remaining case, where $\tau - M_\tau \leq 0$, thus, $\max\{\tau - M_\tau, 0\} \leq \max\{t' - M_{t'}, 0\}$ and $\max\{\tau - M_\tau, 0\} \leq \max\{t - M_t, 0\}$. We still have to verify that $\max\{t' - M_{t'}, 0\} \leq \max\{t - M_t, 0\}$. Following the same reasoning, if $t' - M_{t'} \leq 0$, then the inequality is trivially satisfied. On the other hand, if $t' - M_{t'} \geq 1$, then $t' \geq 69 \geq 14$ and, thus, the value of $M_{\tau'}$ for any round $\tau' \in [t', t-1]$ can be increased by at most one unit. This suffices to conclude that $t' - M_{t'} \leq t - M_t$.

**Fact 4.** *For any $t \in [T]$, $t' \in [t - d_{\max}, t-1]$ and $\tau \in [t' - M_{t'}, t'-1]$, we have*

$$\max\{\tau - M_\tau, 0\} \leq \max\{t' - M_{t'}, 0\} \leq \max\{t - M_t, 0\}.$$

### F.2 Computing the Probability $q_{i,t}(H_{t-M_t})$

In this section, we show that during a run of UCB-CBB, each value of the form $q_{i,t}(H_{t-M_t})$, as computed by COMPQ$(i, t, H_{t-M_t})$ (Algorithm 3), is equal to the probability of arm $i$ being available at round $t$, conditioned on $H_{t-M_t}$, that is, $q_{i,t}(H_{t-M_t}) = \mathbb{P}\left(F^{\tilde{\pi}}_{i,t} | \, H_{t-M_t}\right), \forall i \in \mathcal{A}, t \in [T]$ (assuming that $t - M_t \geq 1$). For any integer $t$, we define $[t]^+ := \max\{t, 0\}$. Therefore, for any $t \in [T]$ such that $t - M_t \leq 0$, we have $H_{[t-M_t]^+} = H_0$ (recall that, in this case, UCB-CBB samples arms according to an initial extreme point solution $Z(0) = \{z_{i,j}(0)\}_{i,j}$). In the following, we fix any arm $i \in \mathcal{A}$ and any point in time $t \in [T]$. Recall that $T_c$ is defined as the smallest $t \in [T]$ such that $t - M_t \geq 1$.

We first consider the case where $t < T_c$ (and, thus, $t - M_t \leq 0$ and $H_{[t-M_t]^+} = H_0$). In that case, for every round $\tau \in [t]$, the algorithm uses the initially computed extreme point $Z(0)$ in order to sample arms. Following the same reasoning as used in Lemma 6 for the full-information case of our problem, we can see that $q_{i,t}(H_0)$ (and, thus, the conditional probability $\mathbb{P}\left(F^{\tilde{\pi}}_{i,t} | \, H_0\right)$) can be computed by the following recursive formula: We set $q_{i,1}(H_0) = 1$ and

$$q_{i,t'+1}(H_0) = q_{i,t'}(H_0)\left(1 - \beta_{i,t'}\sum_{j \in \mathcal{C}} z_{i,j}(0)\right) + \mathbb{I}\left(t' \geq d_i\right) q_{i,t'-d_i+1}(H_0)\beta_{i,t'-d_i+1}\sum_{j \in \mathcal{C}} z_{i,j}(0),$$

where each $\beta_{i,\tau}$ is by construction equal to $\min\{1, \frac{d_i}{2d_i-1}\frac{1}{q_{i,t'}(H_{[\tau-M_\tau]^+})}\}$.

It is not hard to verify that in the above recursive formula, $q_{i,t}(H_0) = q_{i,t}(H_{[t-M_t]^+})$ for $t < T_c$ is indeed equal to $\mathbb{P}\left(F^{\tilde{\pi}}_{i,t} | \; H_{[t-M_t]^+}\right)$, and that $\textsc{compq}(i,t,H_{t-M_t})$ computes exactly this value. The correctness of this computation follows by the fact that for any $t' \leq t$, we have that $q_{i,t'}(H_{[t'-M_{t'}]^+}) = q_{i,t'}(H_{[t-M_t]^+})$, since for all rounds $t' \leq t < T_c$, we have $H_{[t'-M_{t'}]^+} = H_{[t-M_t]^+} = H_0$. Therefore, all the non-skipping probabilities $\beta_{i,t'}$ for $t' < t$ are deterministic and, thus, computable, conditioned on $H_{[t-M_t]^+}$ (thus, the algorithm can simulate them recursively at time $t$).

We now consider the case, where $t \geq T_c$ (and, thus, $t - M_t \geq 1$)). In this case, the algorithm uses the extreme point $Z(t-M_t)$ for sampling arms. Recall that the skipping probability of each round $t'$, is defined given the value of $q_{i,t'}$, as computed, conditioned on $H_{[t'-M_{t'}]^+}$, namely, $\beta_{i,t'} = \min\{1, \frac{d_i}{2d_i-1}\frac{1}{q_{i,t'}(H_{[t'-M_{t'}]^+})}\}$. Therefore, for being able to compute (i.e., simulate) $\beta_{i,t'}$, while being at some round $t > t'$, it suffices to show that $[t'-M_{t'}]^+ \leq [t-M_t]^+$.

In the case where $t' < T_c \leq t$, the extreme point solution used for sampling arms at time $t'$ is $Z(0)$ and, thus, is computable conditioned on $H_{t-M_t}$. The same holds for the non-skipping probability, $\beta_{i,t'}$, used at time $t'$. On the other hand, consider the case where $T_c \leq t' \leq t$. By the analysis in Appendix F.1 (see Fact 1), since $t' \geq T_c$, we know that for any $\tau$ in the interval $\tau \in [t', t]$, the value of $M_\tau$ can increase by at most one unit per round, namely, $M_{\tau+1} \leq M_\tau + 1$. By using this argument, we can directly show by induction, that $t' - M_{t'} \leq t - M_t$ and, thus, $H_{t'-M_{t'}} \subseteq H_{t-M_t}$. Therefore, both the extreme point $Z(t'-M_{t'})$ and the non-skipping probability $\beta_{i,t'}$ used at time $t'$ can be computed (recursively) by the algorithm at time $t$.

The above discussion leads to the following recursive computation of $q_{i,t}(H_{t-M_t})$ for any arm $i$ and time $t \geq T_c$. Let $\nu(i, t-M_t, H_{t-M_t})$ be the first time $\tau \geq t - M_t$ that arm $i$ is deterministically available, conditioned on the history $H_{t-M_t}$. We set $q_{i,\nu(i,t-M_t,H_{t-M_t})}(H_{t-M_t}) = 1$ and for any $t' \geq \nu(i, t-M_t, H_{t-M_t})$, we set

$$q_{i,t'+1}(H_{t-M_t}) = q_{i,t'}(H_{t-M_t}) \left(1 - \beta_{i,t'} \sum_{j \in \mathcal{C}} z_{i,j}\left([t' - M_t']^+\right)\right)$$
$$+ \mathbb{I}\left(t' - d_i + 1 \geq \nu(i, t-M_t, H_{t-M_t})\right) q_{i,t'-d_i+1}(H_{t-M_t}) \beta_{i,t'-d_i+1} \sum_{j \in \mathcal{C}} z_{i,j}([t' - d_i + 1 - M_{t'-d_i+1}]^+).$$

It is easy to verify that $\textsc{compq}(i,t,H_{t-M_t})$ produces exactly the same result as the above recursive formula for $t \geq T_c$.

Given the above analysis, we have now established the correctness of $\textsc{compq}(i,t,H_{t-M_t})$. We remark that in the pseudocode provided in Algorithm 3, the recursive computation of the non-skipping probabilities is implemented efficiently by caching and reusing past values.

### F.3 Proof of Lemma 1

**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.* Recall from Section F.2 that for any fixed arm $i$, the quantity $q_{i,s}(H_{t-M_t})$ in Algorithm 3 equals $\mathbb{P}\left(F^{\tilde{\pi}}_{i,s} | H_{t-M_t}\right)$ for $t - d_i \leq s \leq t$. Therefore, we are interested in the ratio $\frac{q_{i,t'}(H_{t-M_t})}{q_{i,t'}(H_{t'-M_{t'}})}$. In the rest of this proof and for simplicity of notation, we assume that $H_\tau = H_0$ and $\{z_{i,j}(\tau)\}_{i,j} = \{z_{i,j}(0)\}_{i,j}$, for any $\tau \leq 0$.

Let us fix any run of the $\textsc{ucb-cbb}$ algorithm upto time $t$ as $h_t$. The sequence of random variables $\{Z(\tau), \beta_{i,\tau} : 1 \leq \tau \leq t - M_t\}$ is computable at time $t$ given the history $H_{t-M_t}$ (see Fact 4 in Appendix F.1). Therefore, fixing a run of the $\textsc{ucb-cbb}$ algorithm up to any time $t$ (in terms of sampling and non-skipping probabilities), corresponds to fixing $H_{t-M_t} = h_{t-M_t}$, which, in turn, fixes the sequence $\{Z(\tau - M_\tau), \beta_{i,\tau} : 1 \leq \tau \leq t\}$. This follows from the computability of $\beta_\tau$ as discussed in Appendix F.2.

For a particular run $h_{t-M_t}$ upto time $t - M_t$ and a specific arm $i$, the computation of $q_{i,\tau}(H_{t-M_t})$ for any $\tau \leq t$ corresponds to simulating a specific Markov chain as detailed next. We consider the time-nonhomogeneous Markov transition probability matrices (TPM) $\mathcal{M} = \{\mathbf{P}_\tau : 1 \leq \tau \leq t\}$, that at any time $\tau \leq t$ makes transitions as follows. If it is in state 0 it moves to state $d_i$ w.p. $\beta_{i,\tau} \sum_{j \in \mathcal{C}} z_{i,j}([\tau - M_\tau]^+)$, otherwise it stays in state 0. In the case the Markov chain is in state $d > 0$, then it moves to state $(d-1)$ w.p. 1. Here, we denote the TPM at time $\tau$ as $\mathbf{P}_\tau$.

Let us also denote the first time on or after time $\tau$ where the arm $i$ becomes available as $\nu(\tau)$ (which is fixed for a run $h_t$, as it is computable using $H_t$). Using this definition, we denote by $\mathcal{X}_{t'}$ (resp. $\mathcal{X}_t$) the Markov chain that lies in state 0 (w.p. 1) at time $\nu(t' - M_{t'})$ (resp. $\nu(t - M_t)$), and moves following the TPM $\mathcal{M}$. We emphasize the fact that both $\mathcal{X}_t$ and $\mathcal{X}_{t'}$ have the same transition probabilities for all time steps between $\max\{\nu(t - M_t), \nu(t' - M_{t'})\}$ and $t'$ (see Fact 4 in Appendix F.1).

We claim that the probability that the Markov chain $\mathcal{X}_{t'}$ is in state 0 at time $t'$ equals $q_{i,t'}(H_{t'-M_{t'}})$, namely, $\mathbb{P}(\mathcal{X}_{t'}(t') = 0) = q_{i,t'}(H_{t'-M_{t'}})$. This follows induction on $\tau$ for the statement

$$\mathbb{P}(\mathcal{X}_{t'}(\tau) = 0) = q_{i,\tau},$$

where $q_{i,\tau}$ is as given in Algorithm 3. As the base case, at $t_0 = \nu(t' - M_{t'})$ we have by construction that $\mathbb{P}(\mathcal{X}_{t'}(t_0) = 0) = q_{i,t_0} = 1$. Let us assume that the argument is true for all time up to $\tau$. Then we have,

$$\begin{aligned}
&\mathbb{P}(\mathcal{X}_{t'}(\tau+1) = 0) \\
&= \mathbb{P}(\mathcal{X}_{t'}(\tau) = 0) \left(1 - \beta_{i,\tau} \sum_{j \in \mathcal{C}} z_{i,j}(\tau - M_\tau)\right) \\
&+ \mathbb{I}(\tau - d_i + 1 \geq t_0)\, \mathbb{P}(\mathcal{X}_{t'}(\tau - d_i + 1) = 0)\, \beta_{i,\tau-d_i+1} \sum_{j \in \mathcal{C}} z_{i,j}(\tau - d_i + 1 - M_{\tau-d_i+1}). \\
&= q_{i,\tau} \left(1 - \beta_{i,\tau} \sum_{j \in \mathcal{C}} z_{i,j}(\tau - M_\tau)\right) \\
&+ \mathbb{I}(\tau - d_i + 1 \geq t_0)\, q_{i,\tau-d_i+1} \beta_{i,\tau-d_i+1} \sum_{j \in \mathcal{C}} z_{i,j}(\tau - d_i + 1 - M_{\tau-d_i+1}) \\
&= q_{i,\tau+1},
\end{aligned}$$

which proves our claim. Using similar arguments we have that $\mathbb{P}(\mathcal{X}_t(t') = 0) = q_{i,t'}(H_{t-M_t})$.

The rest of the proof relies on showing that $\mathbb{P}(\mathcal{X}_t(t') = 0) \approx \mathbb{P}(\mathcal{X}_t(t') = 0)$ for large enough time (specifically, for time $\min\{t' - \nu(t' - M_{t'}), t' - \nu(t - M_t)\}$). We accomplish that by the use of a Doeblin type coupling argument for the two Markov chains $\mathcal{X}_t$ and $\mathcal{X}_{t'}$.

**Doeblin Coupling of two Markov Chains.** The argument of the rest of the proof relies on a Doeblin type coupling of the above two MCs. Let $\mathcal{X}_t(\tau)$ and $\mathcal{X}_{t'}(\tau)$ be the states of the MC $\mathcal{X}_t$ and $\mathcal{X}_{t'}$ at time $\tau$, respectively. Recall that $\mathcal{X}_t$ starts from state 0 at time $\nu(t - M_t)$, and $\mathcal{X}_{t'}$ starts from state 0 at time $\nu(t' - M_{t'})$. Given the fact that the transition functions are common in both MCs, the two chains evolve independently up until the point they meet for the first moment. Afterwards, they get coupled and evolve together.

We consider the evolution of the bi-variate Markov chain $\{(\tilde{\mathcal{X}}_t(\tau), \tilde{\mathcal{X}}_{t'}(\tau))\}$, where, for $\tau \geq \nu_{\max} := \max\{\nu(t - M_t), \nu(t' - M_{t'})\}$, we have the following evolution of the two Markov chains,

$$\begin{aligned}
&\mathbb{P}\left(\tilde{\mathcal{X}}_t(\tau+1) = s_1, \tilde{\mathcal{X}}_{t'}(\tau+1) = s_2 \mid \tilde{\mathcal{X}}_t(\tau) = s_1', \tilde{\mathcal{X}}_{t'}(\tau) = s_2'\right) \\
&\qquad = \begin{cases} \mathbf{P}_\tau(s_1', s_1)\mathbf{P}_\tau(s_2', s_2), & \text{if } s_1' \neq s_2', \\ \mathbf{P}_\tau(s_1', s_1), & \text{if } s_1' = s_2' \wedge s_1 = s_2, \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}$$

It is easy to check that the bi-variate MC has the property $\tilde{\mathcal{X}}_t(\tau) \overset{d}{=} \mathcal{X}_t(\tau)$ and $\tilde{\mathcal{X}}_{t'}(\tau) \overset{d}{=} \mathcal{X}_{t'}(\tau)$ for all integers $\tau \geq \nu_{\max}$ (here, $\overset{d}{=}$ indicates equality in distribution).

Let the random variable $R_c = \inf\{r \geq \nu_{\max} \mid \mathcal{X}_t(\tau) = \mathcal{X}_{t'}(\tau)\}$ denote the first time after $\nu_{\max}$, when the two chains $\mathcal{X}_t$ and $\mathcal{X}_{t'}$ become coupled. From standard arguments in Doeblin coupling [Lindvall, 2002], we have $|\,\mathbb{P}\left(\mathcal{X}_{t'}(t') = 0\right) - \mathbb{P}\left(\mathcal{X}_t(t') = 0\right)| \leq \mathbb{P}\left(\mathcal{X}_{t'}(t') \neq \mathcal{X}_t(t')\right) \leq \mathbb{P}\left(R_c > t'\right).$

We now make a claim that under the Markov TPM $\mathcal{M}$ at any time $\tau \geq \nu(t - M_t)$ we have $\mathbb{P}\left(\mathcal{X}_t(\tau) = 0\right) \geq \frac{1}{e}$. The claim follows by noticing that arm $i$ is sampled by UCB-CBB with probability at most $1/d_i$ at each time, and it is available, if not sampled in the last $(d_i - 1)$ time slots. Formally, for all $\tau \geq \nu(t - M_t)$ we consider the event $E = \{\mathcal{X}_t(\tau') \neq 0, \forall \tau' \in [\tau - d_i, \tau - 1]\}$, and derive the following

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{X}_t(\tau) = 0\right) = \mathbb{P}\left(E \wedge \mathcal{X}_t(\tau) = 0\right) + \mathbb{P}\left(E^c \wedge \mathcal{X}_t(\tau) = 0\right)\\
&\overset{(i)}{=} \mathbb{P}\left(E\right) + \sum_{\tau'=\tau-d_i}^{\tau-1} \mathbb{P}\left(\mathcal{X}_t(\tau') = 0\right) \mathbb{P}\left(\mathcal{X}_t(\tau) = 0 | \mathcal{X}_t(\tau') = 0\right)\\
&\overset{(ii)}{\geq} \mathbb{P}\left(E\right) + \sum_{\tau'=\tau-d_i}^{\tau-1} \mathbb{P}\left(\mathcal{X}_t(\tau') = 0\right) \prod_{\tau'' \in [\tau', \tau-1]} \left(1 - \beta_{\tau''} \sum_{j \in \mathcal{C}} z_{i,j}(\tau'' - M_{\tau''})\right)\\
&\overset{(iii)}{\geq} \mathbb{P}\left(E\right) + \left(\sum_{\tau'=\tau-d_i}^{\tau-1} \mathbb{P}\left(\mathcal{X}_t(\tau') = 0\right)\right) \prod_{\tau'' \in [\tau-d_i, \tau-1]} \left(1 - \beta_{\tau''} \sum_{j \in \mathcal{C}} z_{i,j}(\tau'' - M_{\tau''})\right)\\
&\overset{(iv)}{\geq} \mathbb{P}\left(E\right) + (1 - \mathbb{P}\left(E\right))(1 - 1/d_i)^{(d_i - 1)} \overset{(v)}{\geq} \frac{1}{e}.
\end{aligned}$$

In the equality (i), we use if the $i$-th arm is unavailable for a contiguous stretch of length $d_i$ before $\tau$ (given by event $E$) then it will be available on $\tau$. Also, we break $E^c$ into mutually exclusive events. The inequality (ii) uses the events that the MC stays in state 0 from time $\tau'' = \tau'$ to $\tau$ to lower bound the probabilities. In inequality (iii) we further lower bound these probabilities by replacing $\tau'$ with $\tau - d_i$. For inequality (iv) we use $\beta_{\tau'} \sum_{j \in \mathcal{C}} z_{i,j}(\tau' - M_{\tau'}) \leq 1/d_i$ due to the LP constraint (**C1**), and the fact that $\beta_{\tau'} \leq 1$. Also, $\sum_{\tau'=\tau-d_i}^{\tau-1} \mathbb{P}\left(\mathcal{X}_t(\tau') = 0\right) \mathbb{P}\left(E^c\right)$. Finally, in (v) we minimize over $\mathbb{P}\left(E\right)$ and $d_i$ to obtain the bound.

Similar results hold for the MC $\mathcal{X}_{t'}$. Thus, we obtain that for any time $\tau \geq \nu_{\max}$, we have $\min\{\mathbb{P}\left(\mathcal{X}_t(\tau) = 0\right), \mathbb{P}\left(\mathcal{X}_{t'}(\tau) = 0\right)\} \geq 1/e$.

Therefore, at each time $\tau \geq \nu_{\max}$, we know that the two chains get coupled with probability at least $\frac{1}{e^2}$. Formally,

$$\begin{aligned}
\mathbb{P}\left(R_c \geq t' + 1\right) &\overset{(i)}{=} \mathbb{P}\left(\mathcal{X}_t(t') \neq \mathcal{X}_{t'}(t') \mid R_c \geq t'\right) \mathbb{P}\left(R_c \geq t'\right)\\
&\overset{(ii)}{=} (1 - \mathbb{P}\left(\mathcal{X}_t(t') = \mathcal{X}_{t'}(t') \mid R_c \geq t'\right)) \mathbb{P}\left(R_c \geq t'\right)\\
&\overset{(iii)}{\leq} (1 - \mathbb{P}\left(\mathcal{X}_t(t') = \mathcal{X}_{t'}(t') = 0 \mid R_c \geq t'\right)) \mathbb{P}\left(R_c \geq t'\right)\\
&\overset{(iv)}{\leq} (1 - \mathbb{P}\left(\mathcal{X}_t(t') = 0 \mid R_c \geq t'\right) \mathbb{P}\left(\mathcal{X}_{t'}(t') = 0 \mid R_c \geq t'\right)) \mathbb{P}\left(R_c \geq t'\right)\\
&\overset{(v)}{\leq} \left(1 - \frac{1}{e^2} \mathbb{I}\left(t' \geq \nu_{\max}\right)\right) \mathbb{P}\left(R_c \geq t'\right),
\end{aligned}$$

where (i) follows by definition of coupling, (iii) follows by the fact that $\{\mathcal{X}_t(t') = \mathcal{X}_{t'}(t') = 0\} \subseteq \{\mathcal{X}_t(t') = \mathcal{X}_{t'}(t')\}$ and (iv) follows by the fact that the two MCs evolve independently before round $R_c$. Finally, (v) follows by the fact that the probability of $\mathcal{X}_t$ (resp. $\mathcal{X}_{t'}$) being at state 0 is at least $\frac{1}{e}$, for any time $\tau \geq \nu_{\max}$ as shown above.

By repeating the arguments leading to (v) until we reach the event $\{R_c \geq \nu_{\max} - 1\}$ we have

$$\begin{aligned}
\mathbb{P}\left(R_c \geq t' + 1\right) &\leq \left(1 - \frac{1}{e^2}\mathbb{I}\left(t' \geq \nu_{\max}\right)\right)\mathbb{P}\left(R_c \geq t'\right) \\
&\leq \mathbb{P}\left(R_c \geq \nu_{\max} - 1\right)\prod_{\tau=\nu_{\max}}^{t'}\left(1 - \frac{1}{e^2}\right) \\
&\overset{(vi)}{\leq} \left(1 - \frac{1}{e^2}\right)^{t' - \nu_{\max} + 1} \\
&\overset{(vii)}{\leq} \left(1 - \frac{1}{e^2}\right)^{M_t - 2d_i},
\end{aligned}$$

where in (vi), we use the fact that $\mathbb{P}\left(R_c \geq 2d_i\right) \leq 1$. In (vii) we use the following derivations

$$\begin{aligned}
t' - \max\{\nu(t - M_t), \nu(t' - M_{t'})\} &\overset{(a)}{=} t' - \nu(t - M_t) \\
&\overset{(b)}{\geq} M_t + t' - t - d_i + 1. \\
&\overset{(c)}{\geq} M_t - 2d_i + 2.
\end{aligned}$$

The equality (a) in the above derivation holds since for $t \geq T_c$ and $t' \in [t - d_i + 1, t - 1]$, then by Fact 3 in Appendix F.1, it has to be that $t' - M_{t'} \leq t - M_t$ and, thus, $\nu(t' - M_{t'}) \leq \nu(t - M_t)$. Inequality (b) holds since $i$ becomes deterministically available in at most $d_i - 1$ time steps after $t - M_t$, i.e. $\nu(t - M_t) \leq t - M_t + d_i - 1$. The last inequality (c) holds as $t - t' \leq d_i - 1$.

Therefore, for concluding the proof of the lemma, we have:

$$\begin{aligned}
&\frac{\mathbb{P}\left(F^{\tilde{\pi}}_{i,t'} \mid H_{t-M_t}\right)}{\mathbb{P}\left(F^{\tilde{\pi}}_{i,t'} \mid H_{t'-M_{t'}}\right)} = \frac{q_{i,t'}(H_{t-M_t})}{q_{i,t'}(H_{t'-M_{t'}})} \leq 1 + \left|1 - \frac{q_{i,t'}(H_{t-M_t})}{q_{i,t'}(H_{t'-M_{t'}})}\right| \leq 1 + \left|1 - \frac{\mathbb{P}\left(\mathcal{X}_t(t') = 0\right)}{\mathbb{P}\left(\mathcal{X}_{t'}(t') = 0\right)}\right| \\
&\leq 1 + \frac{\mathbb{P}\left(R_c > t'\right)}{\mathbb{P}\left(\mathcal{X}_t(\tau) = 0\right)} \leq 1 + \frac{\left(1 - \frac{1}{e^2}\right)^{M_t - 2d_i}}{\frac{1}{e}} \leq 1 + e\left(\frac{e^2}{e^2 - 1}\right)^{2d_{\max}}\left(\frac{e^2}{e^2 - 1}\right)^{-M_t}.
\end{aligned}$$

The above results follow by use of triangle inequality and substituting the bounds derived so far. □

### F.4 Proof of Lemma 2

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

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

*Proof.* In the following proof, we start from the definition of $\alpha$-regret and we prove the regret upper bound of the statement, by applying a sequence of transformations: First, we incorporate the $\left(1 - \frac{d_{\max}-1}{d_{\max}-1+T}\right)$-multiplicative loss, due to the use of (**LP**) as an upper bound, into an $\mathcal{O}(d_{\max})$ additive term in the regret. Second, we upper bound the total regret due to the rounds such that $t - M_t \leq 0$, by another $\mathcal{O}(d_{\max})$ term in the regret. Then, focusing on each round such that $t \geq M_t$, we apply Lemma 1 in order to (approximately) express the regret of any such round by $\frac{d_i}{2d_i - 1}\left(z^*_{i,j} - z_{i,j}(t - M_t)\right)$, for any $i \in \mathcal{A}$ and $j \in \mathcal{C}$. We show that the total approximation loss for that case can be transformed into a constant additive loss in the regret. Finally, we notice that in the rounds, such that $t \geq M_t$, where $M_t$ is increased (by one unit as we show in Appendix F.1), the arm sampling is performed using the same extreme point solution as in the previous rounds. By observing that this can happen at most $\mathcal{O}(\log(T))$ times, we separate the rounds that use strictly updated UCB estimates, while we incorporate the rest as an $\mathcal{O}(\log(T)\Delta_{\max})$-additive loss in the regret bound.

In the following, we denote by $S^{\tilde{\pi}}_{i,t}$ the event that UCB-CBB samples arm $i \in \mathcal{A}$ at round $t$ and by $B^{\tilde{\pi}}_{i,t}$ the event that arm $i$ is not skipped at the round. Finally, we denote by $F^{\tilde{\pi}}_{i,t}$ the event that arm $i$ is available at round $t$.

**Incorporating Time-Dependent Approximation Loss.** The first step in proving the bound is to incorporate the $\left(1 - \frac{d_{\max}-1}{d_{\max}-1+T}\right)$-multiplicative loss, due to the use of (**LP**), into the regret. By definition of $\alpha$-regret, we have

$$\begin{aligned}\alpha\, \mathbf{Reg}^{\tilde{\pi}}_I(T) &= \alpha\, \mathbf{Rew}^*_I(T) - \mathbf{Rew}^{\tilde{\pi}}_I(T)\\ &= \frac{d_{\max}}{2d_{\max}-1}\left(1 - \frac{d_{\max}-1}{d_{\max}-1+T} + \frac{d_{\max}-1}{d_{\max}-1+T}\right)\mathbf{Rew}^*_I(T) - \mathbf{Rew}^{\tilde{\pi}}_I(T)\\ &\leq \frac{d_{\max}}{2d_{\max}-1}\left(1 - \frac{d_{\max}-1}{d_{\max}-1+T}\right)\mathbf{Rew}^*_I(T) - \mathbf{Rew}^{\tilde{\pi}}_I(T) + \frac{2}{3}\left(d_{\max}-1\right),\end{aligned}$$

where in the last inequality, we use the fact that

$$\frac{d_{\max}}{2d_{\max}-1}\frac{d_{\max}-1}{d_{\max}-1+T}\mathbf{Rew}^*_I(T) \leq \frac{d_{\max}}{2d_{\max}-1}\frac{d_{\max}-1}{T}\mathbf{Rew}^*_I(T) \leq \frac{d_{\max}}{2d_{\max}-1}\left(d_{\max}-1\right),$$

using that $\mathbf{Rew}^*_I(T) \leq T$ and the fact that for any possible $d_{\max}$, we have $\frac{d_{\max}}{2d_{\max}-1}\left(d_{\max}-1\right) \leq \frac{2}{3}\left(d_{\max}-1\right)$.

Now by applying the result of Theorem 1, we can further upper bound the $\alpha$-regret by using the fact that the algorithm FI-CBB produces, in expectation, a constant rate of regret over time. More specifically, by denoting $\mathbf{Rew}^{LP}_I$ the optimal solution to (**LP**), we have

$$\begin{aligned}\alpha\, \mathbf{Reg}^{\tilde{\pi}}_I(T) &\leq \frac{d_{\max}}{2d_{\max}-1}\left(1 - \frac{d_{\max}-1}{d_{\max}-1+T}\right)\mathbf{Rew}^*_I(T) - \mathbf{Rew}^{\tilde{\pi}}_I(T) + \frac{2}{3}\left(d_{\max}-1\right)\\ &\leq \frac{d_{\max}}{2d_{\max}-1}T \cdot \mathbf{Rew}^{LP}_I - \mathbf{Rew}^{\tilde{\pi}}_I(T) + \frac{2}{3}\left(d_{\max}-1\right) \qquad (18)\\ &\leq \sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mu_{i,j}\frac{d_i}{2d_i-1}z^*_{i,j} - \mathbf{Rew}^{\tilde{\pi}}_I(T) + \frac{2}{3}\left(d_{\max}-1\right), \qquad (19)\end{aligned}$$

where (18) follows by Lemma 5 and (19) by the fact that $\frac{d_{\max}}{2d_{\max}-1} \leq \frac{d_i}{2d_i-1}$ for any $i \in \mathcal{A}$.

**Simplifying the Expected Reward of ucb-cbb.** By the independence of the rewards $\{X_{i,j,t}\}_{\forall i,j,t}$, we have:

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} X_{i,j,t}\,\mathbb{I}\left(A^{\tilde{\pi}}_t = i, C_t = j\right)\right]\\ &= \sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[X_{i,j,t}\,\mathbb{I}\left(A^{\tilde{\pi}}_t = i, C_t = j\right)\right]\\ &= \sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\mathbb{E}\left[X_{i,j,t}\,\mathbb{I}\left(A^{\tilde{\pi}}_t = i, C_t = j\right)\middle| A^{\tilde{\pi}}_t, C_t\right]\right]\\ &= \sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\mathbb{E}\left[X_{i,j,t}\middle| A^{\tilde{\pi}}_t, C_t\right]\mathbb{I}\left(A^{\tilde{\pi}}_t = i, C_t = j\right)\right]\\ &= \sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\mu_{i,j}\,\mathbb{I}\left(A^{\tilde{\pi}}_t = i, C_t = j\right)\right]\\ &= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mu_{i,j}\,\mathbb{I}\left(A^{\tilde{\pi}}_t = i, C_t = j\right)\right].\end{aligned}$$

**Using Delayed Exploitation for Large Enough $t$.** The remainder of this proof is dedicated to bounding the difference between the expected reward collected by FI-CBB and UCB-CBB. More specifically, our goal is to directly associate the loss of any round $t$ with the suboptimality of the extreme point solution of $(\mathbf{LP})$ computed by UCB-CBB at the same round. More specifically, we are interested in upper bounding the term

$$\sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mu_{i,j}\frac{d_i}{2d_i-1}z^*_{i,j} - \mathbf{Rew}^{\tilde{\pi}}_I(T).$$

The first step is to lower bound the expected reward of UCB-CBB, namely,

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

Let $T_c$ be the minimum round such that $T_c \geq M_{T_c} + 1$. By the discussion in Appendix F.1, we know that $t \geq M_t + 1 \geq 2d_{\max}$ for any $t \geq T_c$.

We now fix any round $t \in [T]$ such that $t \geq T_c$. By using linearity of expectation, we can further simplify the expression of the expected reward of UCB-CBB, by conditioning on the history up to time $t - M_t$. For any fixed $i \in \mathcal{A}$ and $j \in \mathcal{C}$ we have:

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\mathbb{I}\left(A^{\tilde{\pi}}_t = i, C_t = j\right)\right]\\ &= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\mathbb{E}\left[\mathbb{I}\left(A^{\tilde{\pi}}_t = i, C_t = j\right)\middle| H_{t-M_t}\right]\right]\\ &= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\mathbb{E}\left[\mathbb{I}\left(S^{\tilde{\pi}}_{i,t}, B^{\tilde{\pi}}_{i,t}, F^{\tilde{\pi}}_{i,t}, C_t = j\right)\middle| H_{t-M_t}\right]\right]\\ &= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\mathbb{P}\left(S^{\tilde{\pi}}_{i,t}, B^{\tilde{\pi}}_{i,t}, F^{\tilde{\pi}}_{i,t}, C_t = j | H_{t-M_t}\right)\right]\\ &= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\mathbb{P}\left(S^{\tilde{\pi}}_{i,t}, C_t = j | H_{t-M_t}\right)\mathbb{P}\left(B^{\tilde{\pi}}_{i,t} | H_{t-M_t}\right)\mathbb{P}\left(F^{\tilde{\pi}}_{i,t} | H_{t-M_t}\right)\right] &(20)\\ &= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\mathbb{P}\left(S^{\tilde{\pi}}_{i,t} | H_{t-M_t}, C_t = j\right)\mathbb{P}\left(C_t = j | H_{t-M_t}\right)\beta_{i,t}\,\mathbb{P}\left(F^{\tilde{\pi}}_{i,t} | H_{t-M_t}\right)\right] &(21)\\ &= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\frac{z_{i,j}(t-M_t)}{f_j}f_j\beta_{i,t}\,\mathbb{P}\left(F^{\tilde{\pi}}_{i,t} | H_{t-M_t}\right)\right]\\ &= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\beta_{i,t}\,\mathbb{P}\left(F^{\tilde{\pi}}_{i,t} | H_{t-M_t}\right)\right],\end{aligned}$$

where in (20), we use the fact that the events $S^{\tilde{\pi}}_{i,t}, B^{\tilde{\pi}}_{i,t}$ and $F^{\tilde{\pi}}_{i,t}$ are independent conditioned on $H_{t-M_t}$. The reason is that the outcome of $S^{\tilde{\pi}}_{i,t}$ depends on the observed context and on the UCB indices computed before time $t-M_t$, while the outcome of the event $B^{\tilde{\pi}}_{i,t}$ has probability $\beta_{i,t}$, which is computable using only information from $H_{t-M_t}$. Finally, in (21), we use the fact that the observed context of round $t$ is independent of $H_{t-M_t}$, $B^{\tilde{\pi}}_{i,t}$ and $F^{\tilde{\pi}}_{i,t}$.

Clearly, by observing the history $H_{t-M_t}$, one can easily compute the first time arm $i \in \mathcal{A}$ becomes available after time $t-M_t$. If the arm is available at time $t-M_t$ and is not played, then we know that $\mathbb{P}\left(F^{\tilde{\pi}}_{i,t-M_t+1} | H_{t-M_t}\right) = 1$, while if the arm is blocked at time $t-M_t$, then it is played at some time $t' < t-M_t$ and, thus, $\mathbb{P}\left(F^{\tilde{\pi}}_{i,t'+d_i} | H_{t-M_t}\right) = 1$. The conditional probabilities of an arm being available, that is, $q_{i,t}(H_{t-M_t}) = \mathbb{P}\left(F^{\tilde{\pi}}_{i,t} \mid H_{t-M_t}\right)$ can be computed by Algorithm 3, as described in Appendix F.2. In short, given the fact that the algorithm uses at any round $t \geq T_c$ the extreme point computed in round $t - M_t$, for any $t' \in [t - M_t, t]$, the extreme points used are computable given $H_{t-M_t}$ and the algorithm can efficiently simulate any possible $\beta_{i,t'}$.

By the above analysis it follows that at any time $t \geq T_c$, we have:

$$\beta_{i,t} = \min\left\{1, \frac{d_i}{2d_i-1}\frac{1}{q_{i,t}(H_{t-M_t})}\right\} = \min\left\{1, \frac{d_i}{2d_i-1}\frac{1}{\mathbb{P}\left(F^{\tilde{\pi}}_{i,t} | H_{t-M_t}\right)}\right\}.$$

Similarly to the proof of Lemma 7, we distinguish between two cases on the value of $\beta_{i,t}$ conditioned on $H_{t-M_t}$:

**Case (a)** In the case where $1 > \frac{d_i}{2d_i-1}\frac{1}{\mathbb{P}\left(F^{\tilde{\pi}}_{i,t}|H_{t-M_t}\right)}$, we immediately get that:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\beta_{i,t}\,\mathbb{P}\left(F^{\tilde{\pi}}_{i,t}|H_{t-M_t}\right)\right]\\
&= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\frac{d_i}{2d_i-1}\frac{1}{\mathbb{P}\left(F^{\tilde{\pi}}_{i,t}|H_{t-M_t}\right)}\,\mathbb{P}\left(F^{\tilde{\pi}}_{i,t}|H_{t-M_t}\right)\right]\\
&= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\frac{d_i}{2d_i-1}z_{i,j}(t-M_t)\right].
\end{aligned}$$

**Case (b)** In the case where $1 \leq \frac{d_i}{2d_i-1}\frac{1}{\mathbb{P}\left(F^{\tilde{\pi}}_{i,t}|H_{t-M_t}\right)}$, we directly get that $\mathbb{P}\left(F^{\tilde{\pi}}_{i,t}|H_{t-M_t}\right) \leq \frac{d_i}{2d_i-1}$ and $\beta_{i,t}=1$. In order to get a lower bound on $\mathbb{P}\left(F^{\tilde{\pi}}_{i,t}|H_{t-M_t}\right)$, we attempt to upper bound $\mathbb{P}\left(\neg F^{\tilde{\pi}}_{i,t}|H_{t-M_t}\right)$ by union bound over the probability of each arm $i$ being played at some round $t' \in [t-d_i+1, t-1]$. More specifically:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\beta_{i,t}\,\mathbb{P}\left(F^{\tilde{\pi}}_{i,t}|H_{t-M_t}\right)\right]\\
&= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\,\mathbb{P}\left(F^{\tilde{\pi}}_{i,t}|H_{t-M_t}\right)\right]\\
&= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\left(1-\mathbb{P}\left(\neg F^{\tilde{\pi}}_{i,t}|H_{t-M_t}\right)\right)\right]\\
&\geq \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\left(1-\sum_{t'\in[t-d_i+1,t-1]}\mathbb{P}\left(A^{\tilde{\pi}}_{t'}=i|H_{t-M_t}\right)\right)\right]\\
&\geq \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\left(1-\sum_{t'\in[t-d_i+1,t-1]}\mathbb{P}\left(S^{\tilde{\pi}}_{i,t'},B^{\tilde{\pi}}_{i,t'},F^{\tilde{\pi}}_{i,t'}|H_{t-M_t}\right)\right)\right].
\end{aligned}$$

For each $t' \in [t-d_i+1, t-1]$, the events $S^{\tilde{\pi}}_{i,t'}$, $B^{\tilde{\pi}}_{i,t'}$ and $F^{\tilde{\pi}}_{i,t'}$ are independent conditioned on $H_{t-M_t}$, since the outcomes of $S^{\tilde{\pi}}_{i,t'}$ and $B^{\tilde{\pi}}_{i,t'}$ depend on the extreme points computed by UCB-CBB before time $t-M_t$. Moreover, since $M_t > d_i$, we have that $\mathbb{P}\left(S^{\tilde{\pi}}_{i,t'}|H_{t-M_t}\right) = \sum_{j'\in\mathcal{C}}\mathbb{P}\left(S^{\tilde{\pi}}_{i,t'}|C_{t'}=j',H_{t-M_t}\right)\mathbb{P}\left(C_{t'}|H_{t-M_t}\right) = \sum_{j'\in\mathcal{C}} f_{j'}\,\mathbb{P}\left(S^{\tilde{\pi}}_{i,t'}|C_{t'}=j',H_{t-M_t}\right)$, where the last equality follows by independence of $C_{t'}$ and $H_{t-M_t}$, for $M_t > d_i$. Finally, we have that $\mathbb{P}\left(S^{\tilde{\pi}}_{i,t'}|C_{t'}=j',H_{t-M_t}\right) = \frac{z_{i,j'}(t'-M_{t'})}{f_{j'}}$, since the probability of the event $S^{\tilde{\pi}}_{i,t'}$ depends on the extreme point computed at time $t'-M_{t'}$, and is computable conditioning on $H_{t-M_t}$ (see Fact 4 in Appendix F.1). By combining the aforementioned facts, we have:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\beta_{i,t}\,\mathbb{P}\left(F^{\tilde{\pi}}_{i,t}|H_{t-M_t}\right)\right]\\
&\geq \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\left(1-\sum_{t'\in[t-d_i+1,t-1]}\mathbb{P}\left(S^{\tilde{\pi}}_{i,t'},B^{\tilde{\pi}}_{i,t'},F^{\tilde{\pi}}_{i,t'}|H_{t-M_t}\right)\right)\right]\\
&= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\left(1-\sum_{t'\in[t-d_i+1,t-1]}\mathbb{P}\left(S^{\tilde{\pi}}_{i,t'}|H_{t-M_t}\right)\mathbb{P}\left(B^{\tilde{\pi}}_{i,t'}|H_{t-M_t}\right)\mathbb{P}\left(F^{\tilde{\pi}}_{i,t'}|H_{t-M_t}\right)\right)\right]\\
&= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\left(1-\sum_{t'\in[t-d_i+1,t-1]}\sum_{j'\in\mathcal{C}} f_{j'}\frac{z_{i,j'}(t'-M_{t'})}{f_{j'}}\beta_{i,t'}\,\mathbb{P}\left(F^{\tilde{\pi}}_{i,t'}|H_{t-M_t}\right)\right)\right]\\
&= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\left(1-\sum_{t'\in[t-d_i+1,t-1]}\sum_{j'\in\mathcal{C}} z_{i,j'}(t'-M_{t'})\beta_{i,t'}\,\mathbb{P}\left(F^{\tilde{\pi}}_{i,t'}|H_{t-M_t}\right)\right)\right].
\end{aligned}$$

By definition of $\beta_{i,t'}$, we have that $\beta_{i,t'} \leq \frac{d_i}{2d_i-1}\frac{1}{\mathbb{P}\left(F^{\tilde{\pi}}_{i,t'}|H_{t'-M_{t'}}\right)}$. Moreover, for any extreme point solution of (**LP**), by constraints (**C1**), we have that $\sum_{j'\in\mathcal{C}} z_{i,j'}(t'-M_{t'}) \leq \frac{1}{d_i}$. Therefore, the above relation becomes:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\beta_{i,t}\,\mathbb{P}\left(F^{\tilde{\pi}}_{i,t}|H_{t-M_t}\right)\right] \\
&\geq \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\left(1-\sum_{t'\in[t-d_i+1,t-1]}\sum_{j'\in\mathcal{C}} z_{i,j'}(t'-M_{t'})\beta_{i,t'}\,\mathbb{P}\left(F^{\tilde{\pi}}_{i,t'}|H_{t-M_t}\right)\right)\right] \\
&\geq \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\left(1-\frac{1}{d_i}\sum_{t'\in[t-d_i+1,t-1]}\frac{d_i}{2d_i-1}\frac{\mathbb{P}\left(F^{\tilde{\pi}}_{i,t'}|H_{t-M_t}\right)}{\mathbb{P}\left(F^{\tilde{\pi}}_{i,t'}|H_{t'-M_{t'}}\right)}\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\left(1-\frac{1}{2d_i-1}\sum_{t'\in[t-d_i+1,t-1]}\frac{\mathbb{P}\left(F^{\tilde{\pi}}_{i,t'}|H_{t-M_t}\right)}{\mathbb{P}\left(F^{\tilde{\pi}}_{i,t'}|H_{t'-M_{t'}}\right)}\right)\right].
\end{aligned}
$$

For any $t \geq T_C$ and $t' \in [t-d_i+1, t-1]$, by Lemma 1, we have:

$$
\frac{\mathbb{P}\left(F^{\tilde{\pi}}_{i,t'}|H_{t-M_t}\right)}{\mathbb{P}\left(F^{\tilde{\pi}}_{i,t'}|H_{t'-M_{t'}}\right)} \leq 1 + c_0\cdot c_1^{-M_t}. \tag{22}
$$

By using inequality (22), we get:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\beta_{i,t}\,\mathbb{P}\left(F^{\tilde{\pi}}_{i,t}|H_{t-M_t}\right)\right] \\
&\geq \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\left(1-\frac{1}{2d_i-1}\sum_{t'\in[t-d_i+1,t-1]}\left(1+c_0\cdot c_1^{-M_t}\right)\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\left(1-\frac{d_i-1}{2d_i-1}+\frac{d_i-1}{2d_i-1}c_0\cdot c_1^{-M_t}\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\left(\frac{d_i}{2d_i-1}+\frac{d_i-1}{2d_i-1}c_0\cdot c_1^{-M_t}\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[z_{i,j}(t-M_t)\left(\frac{d_i}{2d_i-1}+\frac{d_i-1}{2d_i-1}c_0\cdot c_1^{-M_t}\right)\right]
\end{aligned}
$$

By summing over all $t \in [T_c, T]$ and using the above analysis, we have:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=T_c}^{T}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} X_{i,j,t}\,\mathbb{I}\left(A^{\tilde{\pi}}_t=i, C_t=j\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=T_c}^{T}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\,\mathbb{I}\left(A^{\tilde{\pi}}_t=i, C_t=j\right)\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=T_c}^{T}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j} z_{i,j}(t-M_t)\beta_{i,t}\,\mathbb{P}\left(F^{\tilde{\pi}}_{i,t}|H_{t-M_t}\right)\right] \\
&\geq \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=T_c}^{T}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j} z_{i,j}(t-M_t)\left(\frac{d_i}{2d_i-1}-\frac{d_i-1}{2d_i-1}c_0\cdot c_1^{-M_t}\right)\right] \\
&\geq \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=T_c}^{T}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\frac{d_i}{2d_i-1} z_{i,j}(t-M_t)\right] - \sum_{t=[T]} c_0\cdot c_1^{-M_t},
\end{aligned} \tag{23}
$$

where in the last inequality we use the fact that

$$
\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=T_c}^{T}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\frac{d_i-1}{2d_i-1} z_{i,j}(t-M_t)c_0\cdot c_1^{-M_t}\right] \leq \sum_{t=T_c}^{T} c_0\cdot c_1^{-M_t}.
$$

Furthermore, by our choice of $M_t$, we have that

$$M_t = \lfloor 2\log_{c_1}(t) \rfloor + \lceil \log_{c_1}(c_0) \rceil + 1 \geq 2\log_{c_1}(t) + \log_{c_1}(c_0) = \log_{c_1}(c_0 \cdot t^2),$$

which implies that $\sum_{t=T_c}^{T} c_0 \cdot c_1^{-M_t} \leq \sum_{t\in[T]} c_0 \cdot c_1^{-\log_{c_1}(t^2\cdot c_0)} \leq \sum_{t=1}^{+\infty} \frac{1}{t^2} = \frac{\pi^2}{6}$. Therefore, inequality (23) becomes:

$$\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=T_c}^{T}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} X_{i,j,t}\,\mathbb{I}\left(A_t^{\tilde{\pi}}=i, C_t=j\right)\right] \geq \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=T_c}^{T}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\frac{d_i}{2d_i-1} z_{i,j}(t-M_t)\right] - \frac{\pi^2}{6}. \tag{24}$$

**Bounding Small t and Combining Everything.** By construction UCB-CBB, for the first rounds where $t \leq T_c - 1$, the algorithm selects arms and constructs non-skipping probabilities with respect to an initial extreme point solution $Z(0) = \{z_{i,j}(0)\}_{\forall i,j}$ to (**LP**). Since we cannot bound the expected reward of UCB-CBB for the these time steps, we accumulate this loss in the regret as follows:

$$\begin{aligned}
&\sum_{t=1}^{T_c-1}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\frac{d_i}{2d_i-1} z^*_{i,j} - \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=1}^{T_c-1}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\,\mathbb{I}\left(A_t^{\tilde{\pi}}=i, C_t=j\right)\right] \\
&\leq \sum_{t=1}^{T_c-1}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\frac{d_i}{2d_i-1} z^*_{i,j} \leq T_c - 1 
\end{aligned} \tag{25}$$

For the overall regret we have:

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\frac{d_i}{2d_i-1} z^*_{i,j} - \mathbf{Rew}_I^{\tilde{\pi}}(T) \\
&= \sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\frac{d_i}{2d_i-1} z^*_{i,j} - \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} X_{i,j,t}\,\mathbb{I}\left(A_t^{\tilde{\pi}}=i, C_t=j\right)\right] \\
&= \sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\frac{d_i}{2d_i-1} z^*_{i,j} - \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\,\mathbb{I}\left(A_t^{\tilde{\pi}}=i, C_t=j\right)\right] \\
&= \sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\frac{d_i}{2d_i-1} z^*_{i,j} - \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=1}^{T_c-1}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\,\mathbb{I}\left(A_t^{\tilde{\pi}}=i, C_t=j\right)\right] \\
&\qquad - \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=T_c}^{T}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\,\mathbb{I}\left(A_t^{\tilde{\pi}}=i, C_t=j\right)\right] \\
&\leq \sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\frac{d_i}{2d_i-1} z^*_{i,j} - \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=1}^{T_c-1}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\,\mathbb{I}\left(A_t^{\tilde{\pi}}=i, C_t=j\right)\right] \\
&\qquad - \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=T_c}^{T}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\frac{d_i}{2d_i-1} z_{i,j}(t-M_t)\right] + \frac{\pi^2}{6} && (26) \\
&\leq \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=T_c}^{T}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\frac{d_i}{2d_i-1}\left(z^*_{i,j} - z_{i,j}(t-M_t)\right)\right] + T_c - 1 + \frac{\pi^2}{6} && (27) \\
&= \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=T_c}^{T}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\frac{d_i}{2d_i-1}\left(z^*_{i,j} - z_{i,j}(t-M_t)\right)\right] + 3\cdot d_{\max} + 70 + \frac{\pi^2}{6} && (28) \\
&\leq \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=T_c}^{T}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j}\left(z^*_{i,j} - z_{i,j}(t-M_t)\right)\right] + 3\cdot d_{\max} + 70 + \frac{\pi^2}{6}, && (29)
\end{aligned}$$

where (26) follows by inequality (24) and (27) by inequality (25). Finally, equality (28) follows an upper bound on $T_c$ (given in Fact 2 of Appendix F.1) and inequality (29) by the fact that $\frac{d_i}{2d_i-1} \leq 1$ for any $i \in \mathcal{A}$.

**Synchronizing the Large Time Steps and Completing the Proof.** For completing the proof of the lemma, we focus on the quantity

$$\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=T_c}^{T}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mu_{i,j}\left(z^*_{i,j}-z_{i,j}(t-M_t)\right)\right].$$

Recall that for any $t \geq T_c$, the algorithm uses for arm sampling the extreme point solution $Z(t-M_t)$, computed using the indices $\{\bar{\mu}(t-M_t)\}_{i,j}$. As we show in Appendix F.1 (see Fact 1), for $t \geq T_c$, the value $M_t$ cannot be increased by more than one unit per round. Given any time interval $[t_1, t_2]$, with $t_1 \geq T_c$, we say that the UCB indices of the interval are *synchronized* (or, simply, we say that the interval is synchronized), if for any $t \in [t_1, t_2]$, there exists a integer constant $M'$, such that UCB-CBB at round $t$, uses information from time $t - M'$.

Let $t'$ be the first time that $M_t$ increases by one after time $T_c$. Clearly, the time interval $[T_c, t')$ is *synchronized* as the information used at each round from $T_c$ to $t'-1$ corresponds to times $T_c - M_{T_c}, T_c - M_{T_c} + 1, \dots, t' - 1 - M_{T_c}$. However, at time $t'$, given the fact that $M_{t'} = M_{T_c} + 1$, the index used corresponds, again, to time $t' - 1 - M_{T_c} = t' - M_{t'}$. Hopefully, by ignoring time $t'$, we can see that the index used at $t'+1$ corresponds to time $t' + 1 - M_{t'} = t' - M_{T_c}$, which remains synchronized with the interval before $t'$.

By repeating the above procedure, we ignore the non-synchronized rounds (that correspond to the unit increases of $M_t$) and we merge the remaining rounds into a single synchronized interval. Let $L$ be the number of non-synchronized time steps in $[T_c, T]$, which is formally defined as

$$L = |\{t \in [T_c + 1, T] \mid M_t = M_{t-1} + 1\}|.$$

By definition of $M_t$, the total number of non-synchronized time steps (as $t'$) can be upper bounded by $M_T$ , which, in turn, can be upper bounded by $2\log_{c_1}(T) + \log_{c_1}(c_0) + 2 \leq \frac{1}{3}\ln(T) + 9 + 2d_{\max}$.

Let $\Delta_{\max} = \sup_{Z\in\mathcal{Z}} \Delta_Z$, be the maximum suboptimiality gap over all the extreme points of $\mathcal{Z}$. The regret associated with each non-synchronized time step greater than $T_c$ can be upper bounded by $\Delta_{\max}$. By the above analysis, it follows directly that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=T_c}^{T}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mu_{i,j}\left(z^*_{i,j}-z_{i,j}(t-M_t)\right)\right]\\
&\leq \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=T_c}^{T-L}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mu_{i,j}\left(z^*_{i,j}+z_{i,j}(t-M_{T_c})\right)\right]+\left(\frac{1}{3}\ln(T)+9+2d_{\max}\right)\Delta_{\max}\\
&\leq \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=T_c-M_{T_c}}^{T-L-M_{T_c}}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mu_{i,j}\left(z^*_{i,j}+z_{i,j}(t)\right)\right]+\left(\frac{1}{3}\ln(T)+9+2d_{\max}\right)\Delta_{\max}
\end{aligned}$$

By combining the above inequality with (19) and (29), we can prove the following upper bound:

$$\begin{aligned}
\alpha\,\mathbf{Reg}^{\tilde{\pi}}_I(T) \leq &\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=T_c-M_{T_c}}^{T-L-M_{T_c}}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mu_{i,j}\left(z^*_{i,j}+z_{i,j}(t)\right)\right]+\frac{2}{3}(d_{\max}-1)\\
&+\left(\frac{1}{3}\ln(T)+9+2d_{\max}\right)\Delta_{\max}+3\cdot d_{\max}+70+\frac{\pi^2}{6}.
\end{aligned}$$

By noticing that $\Delta_{\max} \leq 1$, we can simplify the less important constants of the above bound as

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

Finally, we use that $T_c - M_{T_c} \geq 1$ and we let $M = L + M_c = \Theta(\log T + d_{\max})$, which leads to:

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

□

### F.5 Proof of Lemma 3

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

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

*Proof.* We fix an an arbitrary TP group $(i,j,l)$. Let $t_k$ be the time and $Z_k$ be the suboptimal extreme point used by UCB-CBB for sampling arms when the counter $N_{i,j,l}(t')$ is increased for the $k$-th time. Moreover, we denote by $Y_k = \mathbb{I}\left(A^{\tilde{\pi}}_{t_k} = i, C_{t_k} = j\right)$ the event that the TP group $(i,j,l)$ is triggered at time $t_k$, namely, arm $i$ is played under context $j$ and $2^{-l} \leq z^{Z_k}_{i,j} = z_{i,j}([t_k - M_{t_k}]^+) \leq 2^{-l+1}$, where $[t]^+ = \max\{t, 0\}$ for any integer $t$. We require concentration bounds for $\sum_{k=1}^{N_{i,j,l}(t)} Y_k$ conditioned on $N_{i,j,l}(t) = s$. The main roadblock in the analysis, comparing to [Wang and Chen, 2017], is that, due to the blocking constraints, the random variables $Y_k$ are not *mutually independent*. Indeed, if $|t_{k'} - t_k| < d_i$ then $Y_k$ and $Y_{k'}$ cannot be simultaneously equal to 1. In order to overcome the above issue, we opportunistically subsample the events $\{Y_k\}$ to ensure that the distance between two contiguous subsampled events, where $Y_k = 1$, is at least time $(d_i + 1)$ apart (inclusive of the first instance).

We first separate each *triggering* (i.e., arm pulling) event into two stages: *attempting to trigger* $Y'_k = \mathbb{I}\left(S^{\tilde{\pi}}_{i,t_k}, B^{\tilde{\pi}}_{i,t_k}, C_{t_k} = j\right)$, and *actual triggering* $Y_k = \mathbb{I}\left(S^{\tilde{\pi}}_{i,t_k}, B^{\tilde{\pi}}_{i,t_k}, C_{t_k} = j, F^{\tilde{\pi}}_{i,t_k}\right)$. Given this distinction, the second stage takes into account the blocking constraints, while the first stage takes into account the randomness introduced by nature and the random choices of UCB-CBB.

We partition the sequence $\{1, 2, \ldots, N_{i,j,l}(t)\}$, into $\lfloor N_{i,j,l}(t)/(2d_i - 1) \rfloor$ many windows of length $(2d_i - 1)$. The $\ell$-th window consists of the subsequence $\{\ell(2d_i - 1) + 1, \ldots, (\ell + 1)(2d_i - 1)\}$, starting from $\ell = 0$. Notice that for $s \geq 2 \cdot 2^l \geq 2d_i$, we have at least one such window, since $2^{-l} \leq \frac{1}{d_i}$.

We now define the indicator for the *triggering event* in window $\ell$, denoted by $\tilde{Y}_\ell$, and the *triggering time* (if the arm is triggered) in window $\ell$, denoted by $\tilde{t}_\ell$. In each window $\ell$, if there exists a $k$ in the last $d_i$ steps in the window (i.e. $k \in \{\ell(2d_i - 1) + d_i, \ldots, (\ell + 1)(2d_i - 1)\}$) such that the algorithm tries to trigger at time $t_k$ (i.e. $Y'_k = 1$), we set $\tilde{Y}_\ell = Y_k$ and time $\tilde{t}_\ell = t_k$. Otherwise, we set $\tilde{Y}_\ell = 0$ and $\tilde{t}_\ell = \ell(2d_i - 1) + d_i$. Thus, we have constructed an opportunistically subsampled sequence of tuples $(\tilde{Y}_\ell, \tilde{t}_\ell)$ for $0 \leq \ell \leq \lfloor N_{i,j,l}(t)/(2d_i - 1) \rfloor$, from the original subsequence $(Y_k, t_k)$. Clearly, $\sum_{\ell=0}^{\lfloor N_{i,j,l}(t)/(2d_i-1) \rfloor} \tilde{Y}_\ell$ constructs a lower bound for $T_{i,j}(t)$.

To avoid repetitive notations, let us denote by $\mathcal{H}_\ell = \{(\tilde{Y}_1, \tilde{t}_1), \ldots, (\tilde{Y}_{(\ell-1)}, \tilde{t}_{(\ell-1)})\}$ the subsequence from 0 upto (and excluding) the $\ell$-th entry in the sequence. We call the first event of observing at least one $Y'_k = 1$ in the $\ell$-th window as $\mathcal{E}_\ell$. As the sampling only happens at the later part of each window, the previous subsampling ensures that the random variables $\tilde{t}_\ell$ are at least $d_i$ time steps apart. We now claim that when $\tilde{Y}_\ell$ is set to $Y_k$ then irrespective of the past $(\tilde{Y}_{\ell'}, \tilde{t}_{\ell'})$, we have $\mathbb{P}[\tilde{Y}_\ell | \mathcal{E}_\ell, \mathcal{H}_\ell] \geq (1 - \frac{1}{d_i})^{(d_i - 1)} \geq \frac{1}{e}$. The above is true because, we know that when conditioned on history at least $d_i$ time steps apart we have $\mathbb{P}[F^{\tilde{\pi}}_{i,\tilde{t}_\ell} | H_{\tilde{t}_{(\ell-1)}}] \geq 1/e$. Phrased differently, if an arm is not deterministically blocked, then it is available with probability at least $1/e$.

Whenever the counter $N_{i,j,l}(t)$ is increased it is, by definition, due to an extreme point which plays the arm $(i,j)$ with probability at least $2^{-l}$, i.e. $\mathbb{P}\left(S^{\tilde{\pi}}_{i,t_k}, C_{t_k} = j\right) \geq 2^{-l}$. Moreover, $B^{\tilde{\pi}}_{i,t_k}$ is a Bernoulli r.v. with mean $\beta_{i,t_k} = \min\left(1, \frac{d_i}{2d_i - 1} \frac{1}{\mathbb{P}\left(F^{\tilde{\pi}}_{i,t_k} \mid H_{[t_k - M_{t_k}]^+}\right)}\right)$. Furthermore, it is not hard to see that $\beta_{i,t_k} \geq \frac{d_i}{2d_i - 1}$ and, thus, $B^{\tilde{\pi}}_{i,t_k}$ stochastically dominates an independent Bernoulli r.v. of mean $\frac{d_i}{2d_i - 1}$. Similarly, $\mathbb{P}\left(S^{\tilde{\pi}}_{i,t_k}\right)$ stochastically dominates an independent Bernoulli r.v. of mean $2^{-l}$. Therefore, the probability of event $\{Y_k = 1\}$ (trying to trigger arm $i$ at context $j$) is at least $\frac{d_i}{2d_i - 1} 2^{-l}$. We have:

$$\begin{aligned}
\mathbb{P}[\mathcal{E}_\ell|\mathcal{H}_\ell] &= 1 - \mathbb{P}[\mathcal{E}_\ell^c|\mathcal{H}_\ell] \\
&\geq 1 - (1 - \frac{d_i}{2d_i - 1}2^{-l})^{d_i+1} && (30)\\
&\geq \frac{d_i(d_i+1)}{2d_i-1}2^{-l} - \frac{d_i(d_i+1)}{2}(\frac{d_i}{2d_i-1})^2 2^{-2l} && (31)\\
&\geq \frac{d_i(d_i+1)}{2d_i-1}2^{-l} - \frac{d_i(d_i+1)}{2}\frac{d_i}{2d_i-1}2^{-l}\frac{1}{d_i} && (32)\\
&= \frac{d_i(d_i+1)}{2(2d_i-1)}2^{-l},
\end{aligned}$$

where (31) holds due to the Taylor expansion of $(1-x)^{d_i+1}$ around $x = 0$, and (32) follows by noticing that $2^{-l} \leq 1/d_i$ and $\frac{d_i}{2d_i-1} \leq 1$. Finally, (30) follows by the fact that for any extreme point, arm $i$ is played with probability at most $1/d_i$, given that $\sum_{j\in\mathcal{C}} z_{i,j} \leq 1/d_i$.

By combining the above inequalities, then for all $0 \leq \ell \leq \lfloor N_{i,j,l}(t)/(2d_i - 1)\rfloor$, we have

$$\begin{aligned}
\mathbb{E}[\tilde{Y}_\ell|\mathcal{H}_\ell] &\geq \mathbb{E}[\tilde{Y}_\ell|\mathcal{E}_\ell, \mathcal{H}_\ell]\mathbb{P}[\mathcal{E}_\ell|\mathcal{H}_\ell] \\
&\geq \frac{1}{e}\frac{d_i(d_i+1)}{2(2d_i-1)}2^{-l} \\
&\geq (2d_i - 1)2^{-l}\frac{1}{8e}.
\end{aligned}$$

The first inequality holds as $\tilde{Y}_\ell \geq 0$, and the second inequality is obtained by substituting the above appropriate lower bounds.

We next apply the multiplicative Chernoff bound for dependent random variables as stated in Theorem 5 to obtain the final concentration inequality. We use $\delta = 2/3$.

$$\begin{aligned}
\mathbb{P}\left(N_{i,j,l}(t) = s, T_{i,j}(t) \leq \frac{1}{3}\left\lfloor\frac{N_{i,j,l}(t)}{2d_i-1}\right\rfloor(2d_i-1)2^{-l}\frac{1}{8e}\right) &\leq \exp(-\frac{2}{9}\left\lfloor\frac{s}{2d_i-1}\right\rfloor(2d_i-1)2^{-l}\frac{1}{8e}) \\
&\leq \exp(-3\ln(t)) \\
&= \frac{1}{t^3},
\end{aligned}$$

where the second inequality holds for $s \geq 109 \cdot e \cdot 2^l \ln(t) \geq 108 \cdot e \cdot 2^l \ln(t) + 2d_i - 1$, where we use the fact that $2^l \geq d_i$. □

### F.6 Proof of Lemma 4

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

*Proof.* Recall that in any feasible extreme point solution of (**LP**), there exist $|\mathcal{A}||\mathcal{C}| = k\cdot m$ linearly independent inequalities that are tight (i.e., they are met with equality). By the structure of (**LP**), we know that at most $k$ of them can be from the set (**C1**) and at most $m$ can be from the set (**C2**). Therefore, the remaining tight inequalities should be nonnegativity constraints and, thus, they are of the form $z_{i,j} = 0$. This implies that at most $k + m$ variables can be nonzero and, therefore, that the support of any extreme point solution of (**LP**) has cardinality at most $k + m$. □

### F.7 Proof of Theorem 2 (Regret Upper Bound)

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

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

*Proof.* The proof of our regret bound follows closely the structure of [Wang and Chen, 2017]. In the following, we present a version of their proof simplified and adapted to our setting. We start from the upper bound on the regret given by Lemma 2. Then, we study this regret upper bound using techniques from [Wang and Chen, 2017] and making use of our Lemmas 3 and 4, in order to achieve tighter final regret bounds.

By Lemma 2, we have the following upper bound on the $\alpha$-regret

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

where $M = \Theta(\log(T) + d_{\max})$.

By using our definition of suboptimality gaps, we can express the first term of the above bound as

$$\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=1}^{T-M} \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{C}} \mu_{i,j}\left(z_{i,j}^{*} + z_{i,j}(t)\right)\right] = \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=1}^{T-M} \Delta_{Z(t)}\right],$$

where $\Delta_{Z(t)}$ is the suboptimality gap of the extreme point solution of $(\mathbf{LP})(t)$.

In the above summation, notice that for the computation of every $Z(t)$ for $t \in [T-M]$, the algorithm uses strictly updated UCB indices, since we have already excluded the rounds where indices are reused, due to the increases of $M_t$ (see Lemma 2).

We start by defining several important events that may occur during a run of our algorithm UCB-CBB. A reader familiar with the work of [Wang and Chen, 2017] should easily recognize their role. Recall that $T_{i,j}(t)$ denotes the number of times arm $i$ is played under context $j$ up to (and excluding) time $t$. Moreover, we denote by $N_{i,j,l}(t)$ the value of the counter that corresponds to the TP group $\mathcal{Z}_{i,j,l}$, at the beginning of round $t$.

**Definition 3** (Nice sampling). *We say that at the beginning of round $t$,* UCB-CBB *has a* nice sampling*, denoted by $\mathcal{N}_t^s$, if it is the case that:*

$$|\hat{\mu}_{i,j,T_{i,j}(t)} - \mu_{i,j}| \leq \sqrt{\frac{3\ln(t)}{2T_{i,j}(t)}}, \forall i \in \mathcal{A}, \forall j \in \mathcal{C}\,.$$

It is not hard to verify that, on any round $t$ such that $\mathcal{N}_t^s$ holds, we have:

$$\mu_{i,j} \leq \bar{\mu}_{i,j}(t) \leq \min\left\{1, \mu_{i,j} + 2\sqrt{\frac{3\ln(t)}{2T_{i,j}(t)}}\right\}, \forall i \in \mathcal{A}, \forall j \in \mathcal{C}\,.$$

The following lemma provides a lower bound to the probability that the UCB-CBB has a nice sampling at some time $t$.

**Lemma 8.** *The probability that* UCB-CBB *has a nice sampling at time $t$ is at least $\mathbb{P}(\mathcal{N}_t^s) \geq 1 - 2kmt^{-2}$.*

For the rest of this proof, we fix the constants $\mathrm{B}_1 = 109 \cdot e$ and $\mathrm{B}_2 = 24 \cdot e$. Moreover, for any real number $y$, we denote by $[y]^+ = \max\{y, 0\}$.

**Definition 4** (Nice triggering)**.** *We say that at the beginning of round $t$,* UCB-CBB *has a* nice triggering*, denoted by $\mathcal{N}_t^\tau$, if for any TP group $\mathcal{Z}_{i,j,l}$ associated with the pair $(i,j)$ and for any $1 \le l \le \left[\log_2(\frac{2(k+m)}{\Delta_{\min}^{i,j}})\right]^+$, given that $\sqrt{\frac{\mathrm{B}_1 \ln(t)}{N_{i,j,l}(t-1)2^{-l}}} \le 1$, it holds $T_{i,j}(t-1) \ge \frac{1}{\mathrm{B}_2} N_{i,j,l}(t-1)2^{-l}$.*

**Lemma 9.** *The probability that* UCB-CBB *does not have a nice triggering at time $t$ is at upper bounded by $\mathbb{P}(\neg\mathcal{N}_t^\tau) \le \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \left[\log_2(\frac{2(k+m)}{\Delta_{\min}^{i,j}})\right]^+ t^{-2}$.*

We consider the following functions:

$$\ell_{l,T}(\Delta) = \left\lfloor \frac{96 \cdot 2^{-l} \cdot \mathrm{B}_2(k+m)^2 \ln T}{\Delta^2} \right\rfloor$$

$$\kappa_{l,T}(\Delta, s) = \begin{cases} 4 \cdot 2^{-l}, & \text{if } s = 0, \\ 2\sqrt{\frac{4\,\mathrm{B}_1 \cdot 2^{-l} \ln(T)}{s}}, & \text{if } 1 \le s \le \ell_{l,T}(\Delta), \\ 0, & \text{if } s > \ell_{l,T}(\Delta). \end{cases}$$

For any extreme point $Z \in \mathcal{Z}$, we denote by $\tilde{Z} = \{(i,j) \in \mathcal{A} \times \mathcal{C}\,|z_{i,j}^Z > 0\}$ the set of arm context pairs in its support. Notice that by Lemma 4, for any extreme point $Z \in \mathcal{Z}$, we have $|\tilde{Z}| \le m + k$.

For any $Z \in \mathcal{Z}$, let $\Gamma_Z = \max_{(i,j)\in\tilde{Z}}\{\Delta_{\min}^{i,j}\}$ be the maximum $\Delta_{\min}^{i,j}$ over all pairs $(i,j) \in \tilde{Z}$. Our proof relies on the following technical lemma.

**Lemma 10.** (Suboptimality decomposition). *For any round $t \in [T]$, if $\{\Delta_{Z(t)} \ge \Gamma_{Z(t)}\}$ and $\mathcal{N}_t^s$, $\mathcal{N}_t^\tau$ hold, we have:*

$$\Delta_{Z(t)} \le \sum_{(i,j)\in\tilde{Z}(t)} \kappa_{l_{i,j},T}(\Delta_{\min}^{i,j}, N_{i,j,l_{i,j}}(t-1)),$$

*where $l_{i,j}$ the index of a TP group such that $Z(t) \in \mathcal{Z}_{i,j,l_{i,j}}$.*

We are now ready to prove the regret bound, with respect to $\{\Delta_{\min}^{i,j}\}_{\forall i,j}$ and $\Delta_{\max}$. For simplicity, we replace $T-M$ with $T$ in the regret upper bound of Lemma 2. Even though the rounds above $T-M$ might not correspond to UCB indices that were actually used in the run of UCB-CBB, we still get an upper bound to our regret by assuming a larger instance in the underlying combinatorial bandit problem. We have that:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mu_{i,j}\left(z^*_{i,j}-z_{i,j}(t)\right)\right]\\
&=\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\Delta_{Z(t)}\right]\\
&=\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\Delta_{Z(t)}\,\mathbb{I}\left(\Delta_{Z(t)}\geq\Gamma_{Z(t)}\right)\right]+\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\Delta_{Z(t)}\,\mathbb{I}\left(\Delta_{Z(t)}<\Gamma_{Z(t)}\right)\right]\\
&=\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\Delta_{Z(t)}\,\mathbb{I}\left(\Delta_{Z(t)}\geq\Gamma_{Z(t)}\right)\right]\\
&=\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\Delta_{Z(t)}\left(\mathbb{I}\left(\Delta_{Z(t)}\geq\Gamma_{Z(t)},\mathcal{N}^s_t\right)+\mathbb{I}\left(\Delta_{Z(t)}\geq\Gamma_{Z(t)},\neg\mathcal{N}^s_t\right)\right)\right]\\
&\leq\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\Delta_{Z(t)}\left(\mathbb{I}\left(\Delta_{Z(t)}\geq\Gamma_{Z(t)},\mathcal{N}^s_t\right)+\mathbb{I}\left(\neg\mathcal{N}^s_t\right)\right)\right]\\
&=\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\Delta_{Z(t)}\left(\mathbb{I}\left(\Delta_{Z(t)}\geq\Gamma_{Z(t)},\mathcal{N}^s_t,\mathcal{N}^\tau_t\right)+\mathbb{I}\left(\Delta_{Z(t)}\geq\Gamma_{Z(t)},\mathcal{N}^s_t,\neg\mathcal{N}^\tau_t\right)+\mathbb{I}\left(\neg\mathcal{N}^s_t\right)\right)\right]\\
&\leq\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\Delta_{Z(t)}\left(\mathbb{I}\left(\Delta_{Z(t)}\geq\Gamma_{Z(t)},\mathcal{N}^s_t,\mathcal{N}^\tau_t\right)+\mathbb{I}\left(\neg\mathcal{N}^\tau_t\right)+\mathbb{I}\left(\neg\mathcal{N}^s_t\right)\right)\right]\\
&\leq\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\Delta_{Z(t)}\,\mathbb{I}\left(\Delta_{Z(t)}\geq\Gamma_{Z(t)},\mathcal{N}^s_t,\mathcal{N}^\tau_t\right)\right]+\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\Delta_{Z(t)}\,\mathbb{I}\left(\neg\mathcal{N}^\tau_t\right)\right]+\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\Delta_{Z(t)}\,\mathbb{I}\left(\neg\mathcal{N}^s_t\right)\right],
\end{aligned}$$

where we use the fact that, if $\mathbb{I}\left(\Delta_{Z(t)}<\Gamma_{Z(t)}\right)$, it must be $\Delta_{Z(t)}=0$, since, otherwise, it should be that either $\tilde{Z}(t)=\emptyset$, or $\Delta_{Z(t)}<\Gamma_{Z(t)}=\max_{(i,j)\in\tilde{Z}(t)}\Delta^{i,j}_{\min}=\Delta^{i',j'}_{\min}$, for some $(i',j')\in\tilde{Z}(t)$. However, by the structure of (**LP**), we know that $\forall Z\in\mathcal{Z}$, $\tilde{Z}\neq\emptyset$, while the fact that $\Delta_{Z(t)}<\Delta^{i',j'}_{\min}$, for some $(i',j')\in\tilde{Z}(t)$, is a contradiction to the definition of $\Gamma_{Z(t)}$.

By Lemma 8, we have that: $\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\Delta_{Z(t)}\,\mathbb{I}\left(\neg\mathcal{N}^s_t\right)\right]\leq\Delta_{\max}\sum_{t\in[T]}\mathbb{P}\left(\neg\mathcal{N}^s_t\right)\leq\frac{\pi^2}{3}\cdot k\cdot m\cdot\Delta_{\max}$. Moreover, by Lemma 9, we have that $\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\Delta_{Z(t)}\,\mathbb{I}\left(\neg\mathcal{N}^\tau_t\right)\right]\leq\Delta_{\max}\sum_{t\in[T]}\mathbb{P}\left(\neg\mathcal{N}^\tau_t\right)\leq\frac{\pi^2}{6}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\log_2\left(\frac{2(k+m)}{\Delta^{i,j}_{\min}}\right)\Delta_{\max}$. Finally, in order to complete our bound, it suffices to upper bound

$$\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\Delta_{Z(t)}\,\mathbb{I}\left(\Delta_{Z(t)}\geq\Gamma_{Z(t)},\mathcal{N}^s_t,\mathcal{N}^\tau_t\right)\right].$$

For any arm-context pair such that $(i,j)\in Z(t)$ for some extreme point $Z(t)\in\mathcal{Z}$, we define $l^{(t)}_{i,j}$ such that $Z(t)\in\mathcal{Z}_{i,j,l^{(t)}_{i,j}}$. By Lemma 10, we have:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\Delta_{Z(t)}\,\mathbb{I}\left(\Delta_{Z(t)}\geq\Gamma_{Z(t)},\mathcal{N}^s_t,\mathcal{N}^\tau_t\right)\right]&\leq\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t\in[T]}\sum_{(i,j)\in\tilde{Z}(t)}\kappa_{l^{(t)}_{i,j},T}(\Delta^{i,j}_{\min},N_{i,j,l^{(t)}_{i,j}}(t-1))\right]\\
&=\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\sum_{l=1}^{+\infty}\sum_{s=0}^{N_{i,j,l}(T)-1}\kappa_{l,T}(\Delta^{i,j}_{\min},s)\right],
\end{aligned}$$

where the last equality follows by the fact that $N_{i,j,l_{i,j}}$ is increased if and only if $(i,j) \in \tilde{Z}(t)$. Now, for every arm $i \in \mathcal{A}$, context $j \in \mathcal{C}$ and $l \in \mathbb{N}_+$ and by definition of $\kappa_{l,T}(\Delta, s)$ we have:

$$\begin{aligned}
\sum_{s=0}^{N_{i,j,l}(T)-1} \kappa_{l,T}(\Delta^{i,j}_{\min}, s) &\leq \sum_{s=0}^{\ell_{l,T}(\Delta^{i,j}_{\min})} \kappa_{l,T}(\Delta^{i,j}_{\min}, s) && (33)\\
&= \kappa_{l,T}(\Delta^{i,j}_{\min}, 0) + \sum_{s=1}^{\ell_{l,T}(\Delta^{i,j}_{\min})} \kappa_{l,T}(\Delta^{i,j}_{\min}, s)\\
&= \kappa_{l,T}(\Delta^{i,j}_{\min}, 0) + \sum_{s=1}^{\ell_{l,T}(\Delta^{i,j}_{\min})} 2\sqrt{\frac{4\,\mathrm{B}_1 \ln(T) \cdot 2^{-l}}{s}}\\
&\leq \kappa_{l,T}(\Delta^{i,j}_{\min}, 0) + 4\sqrt{\mathrm{B}_1 \cdot 2^{-l} \cdot \ln(T)} \sum_{s=1}^{\ell_{l,T}(\Delta^{i,j}_{\min})} \sqrt{\frac{1}{s}}\\
&\leq \kappa_{l,T}(\Delta^{i,j}_{\min}, 0) + 8\sqrt{\mathrm{B}_1 \cdot 2^{-l} \cdot \ln(T)} \sqrt{\ell_{l,T}(\Delta^{i,j}_{\min})}, && (34)
\end{aligned}$$

where (33) follows by the fact that $\kappa_{l,T}(\Delta, s) = 0$, for $s \geq \ell_{l,T}(\Delta) + 1$, while (34), follows by the fact that for any integer $n \in \mathbb{N}_+$, we have: $\sum_{s=1}^{n} \sqrt{\frac{1}{s}} \leq \int_{s=0}^{n} \sqrt{\frac{1}{s}} ds = 2\sqrt{n}$. Using the definition of $\ell_{l,T}$, then (34) becomes:

$$\begin{aligned}
\sum_{s=0}^{N_{i,j,l}(T)-1} \kappa_{l,T}(\Delta^{i,j}_{\min}, s) &\leq \kappa_{l,T}(\Delta^{i,j}_{\min}, 0) + 8\sqrt{\mathrm{B}_1 \cdot 2^{-l} \cdot \ln(T)} \sqrt{\ell_{l,T}(\Delta^{i,j}_{\min})}\\
&\leq \kappa_{l,T}(\Delta^{i,j}_{\min}, 0) + 8\sqrt{\mathrm{B}_1 \cdot 2^{-l} \cdot \ln(T)} \sqrt{\frac{96 \cdot 2^{-l} \cdot \mathrm{B}_2 \cdot (k+m)^2 \ln(T)}{\left(\Delta^{i,j}_{\min}\right)^2}}\\
&= 4 \cdot 2^{-l} + 8\sqrt{96 \cdot \mathrm{B}_1 \cdot \mathrm{B}_2} \cdot 2^{-l} \cdot \frac{(k+m) \cdot \ln(T)}{\Delta^{i,j}_{\min}}.
\end{aligned}$$

By summing over all $l$, for each pair $(i,j)$, we have:

$$\begin{aligned}
\sum_{l=1}^{+\infty} \sum_{s=0}^{N_{i,j,l}(T)-1} \kappa_{l,T}(\Delta^{i,j}_{\min}, s) &\leq 4 \cdot \sum_{l=1}^{+\infty} 2^{-l} + 8\sqrt{96 \cdot \mathrm{B}_1 \cdot \mathrm{B}_2} \cdot \frac{(k+m) \cdot \ln(T)}{\Delta^{i,j}_{\min}} \sum_{l=1}^{+\infty} 2^{-l}\\
&\leq 4 + 8\sqrt{96 \cdot \mathrm{B}_1 \cdot \mathrm{B}_2} \cdot \frac{(k+m) \cdot \ln(T)}{\Delta^{i,j}_{\min}}.
\end{aligned}$$

By combining the aforementioned facts, we conclude that:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}} \left[ \sum_{t=1}^{T} \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{C}} \mu_{i,j} \left( z^*_{i,j} - z_{i,j}(t) \right) \right] &\leq 8\sqrt{96 \cdot \mathrm{B}_1 \cdot \mathrm{B}_2} \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{C}} \frac{(k+m) \ln(T)}{\Delta^{i,j}_{\min}} + 4 \cdot k \cdot m\\
&+ \frac{\pi^2}{6} \left( \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{C}} \log_2 \frac{2(k+m)}{\Delta^{i,j}_{\min}} + 2 \cdot k \cdot m \right) \Delta_{\max}
\end{aligned}$$

Finally, combining the above with the upper bound we get from Lemma 2, we get:

$$\begin{aligned}\alpha\,\mathbf{Reg}^{\tilde{\pi}}(T) &\leq \mathop{\mathbb{E}}_{\mathcal{R}_{N,\tilde{\pi}}}\left[\sum_{t=1}^{T-M}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mu_{i,j}\left(z^*_{i,j}-z_{i,j}(t)\right)\right] + + \frac{1}{3}\ln(T)\Delta_{\max} + 6d_{\max} + 71\\
&\leq 8\sqrt{96\cdot \mathrm{B}_1\cdot \mathrm{B}_2}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\frac{(k+m)\ln(T)}{\Delta^{i,j}_{\min}} + 4\cdot k\cdot m\\
&\qquad + \frac{\pi^2}{6}\left(\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\log_2\frac{2(k+m)}{\Delta^{i,j}_{\min}} + 2\cdot k\cdot m + \frac{2}{\pi^2}\ln(T)\right)\Delta_{\max} + 6d_{\max} + 71\\
&\leq 10898\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\frac{(k+m)\ln(T)}{\Delta^{i,j}_{\min}} + 4\cdot k\cdot m\\
&\qquad + \frac{\pi^2}{6}\left(\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\log_2\frac{2(k+m)}{\Delta^{i,j}_{\min}} + 2\cdot k\cdot m + \frac{2}{\pi^2}\ln(T)\right)\Delta_{\max} + 6d_{\max} + 71.\end{aligned}$$

The above regret bound completes our proof. □

### F.8 Proof of Lemma 8 [Nice Sampling]

**Lemma 8.** *The probability that* UCB-CBB *has a nice sampling at time $t$ is at least $\mathbb{P}(\mathcal{N}^s_t) \geq 1 - 2kmt^{-2}$.*

*Proof.* Let $\neg\mathcal{N}^s_t$ be the event that the algorithm does not have a nice sampling at some round $t\in[T]$. By union bound on the possible arm-context pairs, we have:

$$\begin{aligned}\mathbb{P}(\neg\mathcal{N}^s_t) &= \mathbb{P}\left(\exists i\in\mathcal{A}, j\in\mathcal{C},\ s.t.\ |\hat{\mu}_{i,j,T_{i,j}(t)} - \mu_{i,j}| > \sqrt{\frac{3\ln(t)}{2T_{i,j}(t)}}\right)\\
&\leq \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mathbb{P}\left(|\hat{\mu}_{i,j,T_{i,j}(t)} - \mu_{i,j}| > \sqrt{\frac{3\ln(t)}{2T_{i,j}(t)}}\right)\\
&\leq \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\sum_{s=1}^{t}\mathbb{P}\left(|\hat{\mu}_{i,j,s} - \mu_{i,j}| > \sqrt{\frac{3\ln(t)}{2s}}\right).\end{aligned}$$

For any $s\in[t]$, $\hat{\mu}_{i,j,s}$ is the average of $s$ i.i.d. random variables, denoted by $X^{[1]}_{i,j},\dots,X^{[s]}_{i,j}$, drawn from the reward distribution of arm $i\in\mathcal{A}$, when it is played under context $j\in\mathcal{C}$. For any fixed $s\in[t]$ and for any pair $i\in\mathcal{A}, j\in\mathcal{C}$, we have:

$$\begin{aligned}\mathbb{P}\left(|\hat{\mu}_{i,j,s} - \mu_{i,j}| > \sqrt{\frac{3\ln(t)}{2s}}\right) &= \mathbb{P}\left(|\frac{\sum_{b\in[s]}X^{[b]}_{i,j}}{s} - \mu_{i,j}| > \sqrt{\frac{3\ln(t)}{2s}}\right)\\
&= \mathbb{P}\left(|\sum_{b\in[s]}X^{[b]}_{i,j} - \mu_{i,j}s| \geq \sqrt{\frac{3s\ln(t)}{2}}\right)\\
&\leq 2\exp\left(-2\frac{3s\ln(t)}{2s}\right) = t^{-3},\end{aligned}$$

where we use Hoeffding's inequality (see Appendix D) for upper bounding the last probability. By combining the above inequalities, we have:

$$\begin{aligned}\mathbb{P}(\neg\mathcal{N}^s_t) &\leq \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\sum_{s=1}^{t}\mathbb{P}\left(|\hat{\mu}_{i,j,s} - \mu_{i,j}| \geq \sqrt{\frac{3\ln t}{2s}}\right)\\
&\leq mkt^{-2}.\end{aligned}$$

□

### F.9 Proof of Lemma 9 [Nice Triggering]

**Lemma 9.** *The probability that* UCB-CBB *does not have a nice triggering at time $t$ is at upper bounded by $\mathbb{P}\left(\neg\mathcal{N}_t^\tau\right) \leq \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\left[\log_2(\frac{2(k+m)}{\Delta_{\text{min}}^{i,j}})\right]^+ t^{-2}$.*

*Proof.* Recall that $\mathrm{B}_1 = 109\cdot e$ and $\mathrm{B}_2 = 24\cdot e$ and consider the case where $t-1 \geq N_{i,j,l}(t-1) \geq \mathrm{B}_1 \cdot 2^l \ln(t)$. By union bound, we have:

$$\begin{aligned}
\mathbb{P}\left(\neg\mathcal{N}_t^\tau\right) &= \mathbb{P}\left(\exists i\in\mathcal{A}, \exists j\in\mathcal{C}, \exists l \in \left[1, \left[\log_2(\frac{2(k+m)}{\Delta_{\text{min}}^{i,j}})\right]^+\right], T_{i,j}(t-1) \leq \frac{1}{\mathrm{B}_2} N_{i,j,l}(t-1)2^{-l}\right) \\
&\leq \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \sum_{l=1}^{\left[\log_2(\frac{2(k+m)}{\Delta_{\text{min}}^{i,j}})\right]^+} \mathbb{P}\left(T_{i,j}(t-1) \leq \frac{1}{\mathrm{B}_2} N_{i,j,l}(t-1)2^{-l}\right) \\
&\leq \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \sum_{l=1}^{\left[\log_2(\frac{2(k+m)}{\Delta_{\text{min}}^{i,j}})\right]^+} \sum_{s=\lceil \mathrm{B}_1 \cdot 2^l \ln(t)\rceil}^{t} \mathbb{P}\left(N_{i,j,l}(t-1) = s, T_{i,j}(t-1) \leq \frac{1}{\mathrm{B}_2} N_{i,j,l}(t-1)2^{-l}\right). \qquad (35)
\end{aligned}$$

By Lemma 3, and since we consider only $N_{i,j,l}(t-1) \geq \mathrm{B}_1 \cdot 2^l \log(t) = 109\cdot e\cdot 2^l \log(t)$, inequality (35) can be further upper bounded by:

$$\begin{aligned}
\mathbb{P}\left(\neg\mathcal{N}_t^\tau\right) &\leq \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \sum_{l=1}^{\left[\log_2(\frac{2(k+m)}{\Delta_{\text{min}}^{i,j}})\right]^+} \sum_{s=\lceil \mathrm{B}_1 \cdot 2^l \ln(t)\rceil}^{t} \frac{1}{t_3} \\
&\leq \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \left[\log_2(\frac{2(k+m)}{\Delta_{\text{min}}^{i,j}})\right]^+ t^{-2}.
\end{aligned}$$

□

### F.10 Proof of Lemma 10 (Suboptimality Decomposition)

**Lemma 10.** (Suboptimality decomposition). *For any round $t\in[T]$, if $\{\Delta_{Z(t)} \geq \Gamma_{Z(t)}\}$ and $\mathcal{N}_t^s$, $\mathcal{N}_t^\tau$ hold, we have:*

$$\Delta_{Z(t)} \leq \sum_{(i,j)\in\tilde{Z}(t)} \kappa_{l_{i,j},T}(\Delta_{\text{min}}^{i,j}, N_{i,j,l_{i,j}}(t-1)),$$

*where $l_{i,j}$ the index of a TP group such that $Z(t) \in \mathcal{Z}_{i,j,l_{i,j}}$.*

*Proof.* Clearly, we are only interested in the rounds $t\in[T]$ such that $\Delta_{Z(t)} > 0$, since, otherwise, the inequality holds trivially. By optimality of (**LP**) at time $t$ (i.e. the solution of (**LP**) at time $t$ using the indices $\{\bar{\mu}_{i,j}(t)\}_{\forall i,j}$), we have that:

$$\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \bar{\mu}_{i,j}(t) z_{i,j}(t) \geq \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \bar{\mu}_{i,j}(t) z_{i,j}^*.$$

Moreover, by the nice sampling assumption on round $t$, we have:

$$\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \bar{\mu}_{i,j}(t) z_{i,j}^* \geq \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}} \mu_{i,j} z_{i,j}^*,$$

given that under $\mathcal{N}_t^s$, each index overestimates the actual mean value, namely, $\bar{\mu}_{i,j}(t) \geq \mu_{i,j}, \forall i\in\mathcal{A}, j\in\mathcal{C}$.

Finally, by definition of the suboptimality gap, we have that $\Delta_{Z(t)} = \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mu_{i,j}z^*_{i,j} - \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mu_{i,j}z_{i,j}(t)$. By combining the above facts, we get:

$$\begin{aligned}\Delta_{Z(t)} &= \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mu_{i,j}z^*_{i,j} - \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{C}}\mu_{i,j}z_{i,j}(t)\\ &\leq \sum_{(i,j)\in\tilde{Z}(t)}\left(\tilde{\mu}_{i,j}(t) - \mu_{i,j}\right)z_{i,j}(t).\end{aligned}$$

Now, by assumption that $\Delta_{Z(t)} \geq \Gamma_{Z(t)} = \max_{(i,j)\in\tilde{Z}}\{\Delta^{i,j}_{\min}\}$ and using the above inequality, we have:

$$\begin{aligned}\Delta_{Z(t)} &\leq -\Gamma_{Z(t)} + 2\sum_{(i,j)\in\tilde{Z}(t)}\left(\bar{\mu}_{i,j}(t) - \mu_{i,j}\right)z_{i,j}(t)\\ &= 2\sum_{(i,j)\in\tilde{Z}(t)}\left(\left(\bar{\mu}_{i,j}(t) - \mu_{i,j}\right)z_{i,j}(t) - \frac{\Gamma_{Z(t)}}{2|\tilde{Z}(t)|}\right)\\ &\leq 2\sum_{(i,j)\in\tilde{Z}(t)}\left(\left(\bar{\mu}_{i,j}(t) - \mu_{i,j}\right)z_{i,j}(t) - \frac{\Delta^{i,j}_{\min}}{2(k+m)}\right),\end{aligned}$$

where in the last inequality, we use the fact that, by Lemma 4, we have $|\tilde{Z}(t)| \leq k+m$, and that for any pair $(i,j)\in\tilde{Z}(t)$, we have $\Gamma_{Z(t)} \geq \Delta^{i,j}_{\min}$.

For any $(i,j)\in\tilde{Z}(t)$, let $l_{i,j}$ be the index such that $Z(t)\in\mathcal{Z}_{i,j,l_{i,j}}$. For each $(i,j)\in\tilde{Z}(t)$, we are trying to upper bound $2\left(\left(\bar{\mu}_{i,j}(t)-\mu_{i,j}\right)z_{i,j}(t) - \frac{\Delta^{i,j}_{\min}}{2(k+m)}\right)$, by distinguishing between two cases on the value of $l_{i,j}$.

**Case (a):** $1 \leq l_{i,j} \leq \lceil \log\frac{2(k+m)}{\Delta^{i,j}_{\min}}\rceil_+$. By $\mathcal{N}^s_t$, we have that $\bar{\mu}_{i,j}(t) - \mu_{i,j} \leq 2\sqrt{\frac{3\ln(t)}{2T_{i,j}(t)}}$, while by definition of TB groups, we have $z_{i,j}(t) \leq 2^{-l_{i,j}+1}$. We further distinguish between sub-cases.

**Sub-case (i):** $N_{i,j,l_{i,j}}(t-1) = 0$. In that case, we have that $\kappa_{l_{i,j},T}(\Delta^{i,j}_{\min}, 0) = 4\cdot 2^{-l_{i,j}}$ and, thus:

$$\begin{aligned}2\left(\left(\bar{\mu}_{i,j}(t)-\mu_{i,j}\right)z_{i,j}(t) - \frac{\Delta^{i,j}_{\min}}{2(k+m)}\right) &\leq 2\left(\bar{\mu}_{i,j}(t)-\mu_{i,j}\right)z_{i,j}(t)\\ &\leq 2\cdot 2^{-l_{i,j}+1}\\ &= 4\cdot 2^{-l_{i,j}}\\ &= \kappa_{l_{i,j},T}(\Delta^{i,j}_{\min}, 0).\end{aligned}$$

**Sub-case (ii):** $\sqrt{\frac{\mathrm{B}_1\ln(t)}{N_{i,j,l_{i,j}}(t-1)2^{-l_{i,j}}}} \geq 1$. Then we have that:

$$\begin{aligned}2\left(\left(\bar{\mu}_{i,j}(t)-\mu_{i,j}\right)z_{i,j}(t) - \frac{\Delta^{i,j}_{\min}}{2(k+m)}\right) &\leq 2\left(\bar{\mu}_{i,j}(t)-\mu_{i,j}\right)z_{i,j}(t)\\ &\leq 2\cdot 2^{-l_{i,j}+1}\\ &\leq 2\cdot 2^{-l_{i,j}+1}\sqrt{\frac{\mathrm{B}_1\ln(t)}{N_{i,j,l_{i,j}}(t-1)2^{-l_{i,j}}}}\\ &\leq 2\cdot\sqrt{\frac{4\,\mathrm{B}_1\cdot 2^{-l_{i,j}}\ln(t)}{N_{i,j,l_{i,j}}(t-1)}}\\ &= \kappa_{l_{i,j},T}(\Delta^{i,j}_{\min}, N_{i,j,l_{i,j}}(t-1)).\end{aligned}$$

**Sub-case (iii):** $\sqrt{\frac{\mathrm{B}_1 \ln(t)}{N_{i,j,l_{i,j}}(t-1) 2^{-l_{i,j}}}} \leq 1$. Then by $\mathcal{N}_t^{\tau}$ and $\mathcal{N}_t^{s}$, we have:

$$\bar{\mu}_{i,j}(t) - \mu_{i,j} \leq 2\sqrt{\frac{3\ln(t)}{2T_{i,j}(t-1)}} \leq 2\sqrt{\frac{3\,\mathrm{B}_2 \ln(t)}{2N_{i,j,l_{i,j}}(t-1) \cdot 2^{-l_{i,j}}}}.$$

Therefore, we have that:

$$\left(\bar{\mu}_{i,j}(t) - \mu_{i,j}\right) z_{i,j}(t) \leq \min\left\{\sqrt{\frac{24\,\mathrm{B}_2 \ln(t) \cdot 2^{-l_{i,j}}}{N_{i,j,l_{i,j}}(t-1)}}, 2 \cdot 2^{-l_{i,j}}\right\}.$$

Now, in the case where $N_{i,j,l_{i,j}}(t-1) \geq \ell_{l_{i,j},T}(\Delta_{\min}^{i,j}) + 1$, we have: $\sqrt{\frac{24\,\mathrm{B}_2 \ln(t) \cdot 2^{-l_{i,j}}}{N_{i,j,l_{i,j}}(t-1)}} \leq \sqrt{\frac{24\,\mathrm{B}_2 \ln(t) \cdot 2^{-l_{i,j}} (\Delta_{\min}^{i,j})^2}{96 \cdot \mathrm{B}_2\, 2^{-l_{i,j}} (k+m)^2 \ln(T)}} \leq \sqrt{\frac{(\Delta_{\min}^{i,j})^2}{4(k+m)^2}} = \frac{\Delta_{\min}^{i,j}}{2(k+m)}$, and, thus, $2\left(\left(\bar{\mu}_{i,j}(t) - \mu_{i,j}\right) z_{i,j}(t) - \frac{\Delta_{\min}^{i,j}}{2(k+m)}\right) \leq 2\left(\frac{\Delta_{\min}^{i,j}}{2(k+m)} - \frac{\Delta_{\min}^{i,j}}{2(k+m)}\right) \leq 0 = \kappa_{l_{i,j},T}(\Delta_{\min}^{i,j}, N_{i,j,l_{i,j}}(t-1))$. In the case where $N_{i,j,l_{i,j}}(t-1) \leq \ell_{l_{i,j},T}(\Delta_{\min}^{i,j})$, we simply use $2\left(\left(\bar{\mu}_{i,j}(t) - \mu_{i,j}\right) z_{i,j}(t) - \frac{\Delta_{\min}^{i,j}}{2(k+m)}\right) \leq 2\left(\left(\bar{\mu}_{i,j}(t) - \mu_{i,j}\right) z_{i,j}(t) - \frac{\Delta_{\min}^{i,j}}{2(k+m)}\right) \leq \sqrt{\frac{96\,\mathrm{B}_2 \ln(t) \cdot 2^{-l_{i,j}}}{N_{i,j,l_{i,j}}(t-1)}} \leq 2\sqrt{\frac{4\,\mathrm{B}_1 \ln(t) \cdot 2^{-l_{i,j}}}{N_{i,j,l_{i,j}}(t-1)}}$.

**Case (b):** $l_{i,j} \geq \lceil \log \frac{2(k+m)}{\Delta_{\min}^{i,j}} \rceil_+ + 1$. Using the fact that $\bar{\mu}_{i,j}(t) - \mu_{i,j} \leq 1$ and the definition of TB groups, we have:

$$\left(\bar{\mu}_{i,j}(t) - \mu_{i,j}\right) z_{i,j}(t) \leq z_{i,j}(t) \leq 2^{-l_{i,j}+1} \leq 2^{-\lceil \log \frac{2(k+m)}{\Delta_{\min}^{i,j}} \rceil_+} \leq 2^{-\log \frac{2(k+m)}{\Delta_{\min}^{i,j}}} \leq \frac{\Delta_{\min}^{i,j}}{2(k+m)}.$$

By using the non-negativity of $\kappa_{l_{i,j},T}(\Delta_{\min}^{i,j}, N_{i,j,l_{i,j}}(t-1))$, the above implies that:

$$\frac{\Delta_{\min}^{i,j}}{2(k+m)} - \frac{\Delta_{\min}^{i,j}}{2(k+m)} \leq 0 \leq \frac{1}{2}\kappa_{l_{i,j},T}(\Delta_{\min}^{i,j}, N_{i,j,l_{i,j}}(t-1)),$$

and, thus,

$$2\left(\left(\bar{\mu}_{i,j}(t) - \mu_{i,j}\right) z_{i,j}(t) - \frac{\Delta_{\min}^{i,j}}{2(k+m)}\right) \leq \kappa_{l_{i,j},T}(\Delta_{\min}^{i,j}, N_{i,j,l_{i,j}}(t-1)).$$

□

# G HARDNESS RESULTS: OMITTED PROOFS

## G.1 Proof of Theorem 3

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

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

*Proof.* We now prove an upper bound on the (asymptotic) competitive ratio of the full-information case of our problem. It suffices to provide an instance $I$, such that the ratio between the expected reward collected by an (asymptotically) optimal online policy, denoted by $\lim_{T\to+\infty}\mathbf{Rew}_I^{\mathrm{opt}}(T)$, and by an optimal clairvoyant policy, denoted by $\lim_{T\to+\infty}\mathbf{Rew}_I^{*}(T)$, is upper bounded by $\frac{d_{\max}}{2d_{\max}-1}$. Recall, that a clairvoyant policy has a priori knowledge of all context realizations, $\{C_t\}_{\forall t\in[T]}$.

Consider the following instance $I$. Let $\mathcal{A}$ be a set of $k$ arms and let arm $i^m$ such that $i^m=\arg\max_{i'\in\mathcal{A}}d_{i'}$, namely, an arm of maximum possible delay. Let $\mathcal{C}=\{1,2\}$ be a set of two contexts, such that $f_1=\epsilon$ and $f_2=1-\epsilon$, for some small $\epsilon\in(0,1)$. We assume that the rewards $\{X_{i,j,t}\}_{\forall i\in\mathcal{A},j\in\mathcal{C},t\in[T]}$ are constants, while the rewards of all arms except for $i^m$, i.e., $\mathcal{A}\setminus\{i^m\}$, are identically equal to zero for any possible context. The above implies that, without loss of generality, neither the optimal clairvoyant policy, nor the optimal online policy ever play these arms and, thus, we can assume that only arm $i^m$ is played. For arm $i^m$, we have that $X_{i^m,1,t}=\mu_{i^m,1}=\frac{R}{\epsilon}$ for some fixed $R>0$ and $X_{i^m,2,t}=\mu_{i^m,2}=1$, for all $t\in[T]$. We note that the reward $\frac{R}{\epsilon}$ may be greater than 1, which can be fixed by dividing all the rewards by $\left(1+\frac{R}{\epsilon}\right)$, in order to keep them within range $[0,1]$.

In this proof, we compute the average reward collected by an optimal (non-clairvoyant) online and we lower bound the average reward collected by an optimal clairvoyant policy for instance $I$. Given that $i^m$ is the only arm played in both cases, we focus only on this arm and we simplify the notation by referring to it as $i$.

**Online Policy.** We focus on arm $i$ and consider the behavior of a specific online policy, denoted by $\mathrm{alg}(q_1,q_2)$. This online policy starts at time $t=r$ for $r\in\{0,\dots,d_i-1\}$ with probability $\pi(r)$ (to be specified later). At each time, if the arm $i$ is available and the context is 1 (resp., 2) it plays the arm $i$ with probability $q_1$ (resp., $q_2$).

The behavior of this online policy $\mathrm{alg}(q_1,q_2)$ can be analyzed using a Markov chain. Specifically, the Markov chain has $d_i$ states, $0,1,\dots,d_i-1$, where each state $r$ indicates the fact that arm $i$ is blocked (i.e., not available) for the next $r$ rounds. Let $q_1,q_2\in[0,1]$ (determined by the policy $\mathrm{alg}(q_1,q_2)$) denote the probabilities that arm $i$ is played, if available, given that the context is 1 and 2, respectively. At each time $t$, the Markov chain moves from state 0 to state $(d_i-1)$ with probability $(q_1f_1+q_2f_2)$ and gains the expected reward $(\mu_{i,1}q_1f_1+\mu_{i,2}q_2f_2)$. Otherwise, the Markov chain remains in state 0 with probability $(1-q_1f_1-q_2f_2)$. Given that at some time $t$, the state is $r$, for $r\geq1$, the Markov chain deterministically moves to the state $(r-1)$ (collecting zero reward). Let $\pi(r)$ be the stationary probability of state $s_r$ in the above Markov chain, which is parameterized by $q_1$ and $q_2$. This is the same $\pi(r)$ that is used in the definition of the policy $\mathrm{alg}(q_1,q_2)$.

We can compute the probability $\pi(0)$ by solving the system: $\sum_{r\in\{0,\dots,d_i-1\}}\pi(r)=1$ and that $\pi(1)=\pi(2)=\cdots=\pi(d_i-1)=\pi(0)\left(q_1f_1+q_2f_2\right)$. Recall that the expected reward $(\mu_{i,1}q_1f_1+\mu_{i,2}q_2f_2)$ is collected only when the Markov chain is at state 0 (and moves to state $d_i-1$). Finally, we let the Markov chain start from stationary state, i.e. at time $t=0$ the Markov chain is in state $r$ w.p. $\pi(r)$. Due to stationarity, the expected average reward for the above online policy $\mathrm{alg}(q_1,q_2)$, for any time horizon $T$, denoted by $\mathbf{Rew}_I^{\mathrm{alg}(q_1,q_2)}(T)$, can be expressed as:

$$\mathbf{Rew}_I^{\mathrm{alg}(q_1,q_2)}=\mathop{\mathbb{E}}_{\mathcal{R}_{N,\mathrm{alg}(q_1,q_2)}}\left[\frac{1}{T}\sum_{t\in[T]}\sum_{j\in\mathcal{C}}\mathbb{I}\left(A_t^{\pi(q_1,q_2)}=i,C_t=j\right)\right]=\frac{Rq_1+(1-\epsilon)q_2}{1+(d_i-1)(\epsilon q_1+(1-\epsilon)q_2)}$$

We have already argued that for the setting under consideration, there exists an optimal online policy which only plays arm $i$. Further from the theory of Markov decision processes (MDP), as the time horizon $T$ tends to infinity, there exists an optimal online policy (which only plays arm $i$) that is represented by the above stationary Markov chain (c.f., [Puterman, 2014]). In particular, this optimal online policy can be designed by maximizing the time-average expected reward over the probabilities $q_1$ and $q_2$. Therefore, computing the optimal time-average expected reward in our setting can be formulated as the following optimization program:

$$\textbf{maximize: } f(q_1, q_2) = \frac{Rq_1 + (1-\epsilon)q_2}{1 + (d_i - 1)(\epsilon q_1 + (1-\epsilon)q_2)} \textbf{ s.t. } q_1, q_2 \in [0, 1]. \tag{36}$$

The following lemma specifies the solution of the above optimization problem for a specific range of $(R, \epsilon)$.

**Lemma 11.** *For $R > \epsilon + \frac{1}{d_i - 1}$, the optimal solution to the mathematical program* (36) *is attained by setting* $(q_1, q_2) = (1, 0)$ *and its value is equal to* $\frac{R}{1+(d_i-1)\epsilon}$.

**Lower Bound on the Optimal Clairvoyant Policy.** We now need to compute the expected average reward of an optimal clairvoyant policy on our instance, namely, a policy that has a priori knowledge of the context realizations of all rounds. However, given that providing a characterization of the optimal solution for any possible context realization is a difficult task, we instead attempt to lower bound the optimal expected reward. For this reason, we study a simpler and (possibly) suboptimal clairvoyant policy. This policy is based on partitioning the time horizon into blocks of size $B$, and, then, treating each block, separately, using a simple strategy.

We define the block size to be $B = kd_i$, where $k \in \mathbb{N}_+$ is a positive natural number such that $k \geq 2$ and $d_i$ is the delay of arm $i$. We further assume without loss of generality that the time horizon $T$, which we later extend to infinity, is a multiple of the block size $B$. Our algorithm works separately, in each of the $\frac{T}{B}$ blocks, according to the following simple rule:
**Case (a)**: If context 1 appears at exactly one time $t'$ within the first $B - d_i$ rounds of the block, then the algorithm plays the arm on time $t'$ and nothing else.
**Case (b)**: If context 1 does not appear at all within the $B$ rounds of the block, the algorithm plays arm $i$ exactly $k - 1$ times (every $d_i$ times), starting from the first round of the and excluding the last $d_i$ rounds of the block.
**Case (c)**: In any other case, the algorithm takes no action during the $B$ rounds of the block.

It is important to notice that, in all the aforementioned cases, no action is taken within the last $d_i$ rounds of each block. This allows us to study the expected reward of each block independently.

The average reward collected by the above policy is at least,

$$\frac{1}{B}\left(\sum_{t=1}^{B+1-d_i} \frac{R}{\epsilon}\epsilon(1-\epsilon)^{(B-1)} + \frac{(B-d_i)}{d_i}(1-\epsilon)^B\right) = R(1 - \frac{d_i}{B})(1-\epsilon)^{(B-1)} + (\frac{1}{d_i} - \frac{1}{B})(1-\epsilon)^B.$$

Therefore, for $R > \epsilon + \frac{1}{d_i-1}$ and using Lemma 11, the ratio of reward of the online policy over the designed clairvoyant policy is upper bounded by:

$$\frac{\frac{R}{1+(d_i-1)\epsilon}}{R(1-\frac{d_i}{B})(1-\epsilon)^{(B-1)} + (\frac{1}{d_i} - \frac{1}{B})(1-\epsilon)^B}.$$

We now consider a series of instances, where $B = d_i\lceil\frac{1}{\sqrt{\epsilon}}\rceil$, $R = 2\epsilon + \frac{1}{d_i-1}$ and $\epsilon$ approaches 0. The limiting competitive ratio of an optimal online algorithm becomes:

$$\lim_{\epsilon\to 0} \frac{\frac{R}{1+(d_i-1)\epsilon}}{R(1-\frac{d_i}{B})(1-\epsilon)^{(B-1)} + (\frac{1}{d_i} - \frac{1}{B})(1-\epsilon)^B} = \frac{\frac{1}{d_i-1}}{\frac{1}{d_i-1} + \frac{1}{d_i}} = \frac{d_i}{2d_i - 1},$$

where we use the fact that $\lim_{\epsilon\to 0}(1-\epsilon)^{\frac{1}{\sqrt{\epsilon}}} = 1$.

Given that $d_i = d_{\max}$ for instance $I$, we can conclude that the optimal asymptotic competitive ratio of the full-information case of our problem can be upper bounded by $\frac{d_{\max}}{2d_{\max}-1}$. □

## G.2 Proof of Lemma 11

**Lemma 11.** *For $R > \epsilon + \frac{1}{d_i - 1}$, the optimal solution to the mathematical program* (36) *is attained by setting* $(q_1, q_2) = (1, 0)$ *and its value is equal to* $\frac{R}{1+(d_i-1)\epsilon}$.

*Proof.* Let $f(q_1, q_2) = \frac{Rq_1 + (1-\epsilon)q_2}{1+(d_i-1)(\epsilon q_1 + (1-\epsilon)q_2)}$. Taking the partial derivative of $f(q_1, q_2)$ with respect to $q_1$, we get:

$$\begin{aligned}\frac{\partial f(q_1, q_2)}{\partial q_1} &= \frac{R\,(1+(d_i-1)(q_1\epsilon + q_2(1-\epsilon))) - \epsilon(d_i-1)(Rq_1 + q_2(1-\epsilon))}{(1+(d_i-1)(q_1\epsilon + q_2(1-\epsilon)))^2} \\ &= \frac{R + (d_i-1)q_2(1-\epsilon)(R-\epsilon)}{(1+(d_i-1)(q_1\epsilon + q_2(1-\epsilon)))^2}.\end{aligned}$$

Therefore, for $R > \epsilon$ we have $\frac{\partial f(q_1,q_2)}{\partial q_1} > 0$ for all $q_1, q_2 \in [0,1]$ and, thus, the optimal solution in this case is attained at $q_1^* = 1$. We now take the derivative of $f(q_1, q_2)$ with respect to $q_2$:

$$\begin{aligned}\frac{\partial f(q_1, q_2)}{\partial q_2} &= \frac{(1-\epsilon)(1+(d_i-1)(q_1\epsilon + q_2(1-\epsilon))) - (d_i-1)(1-\epsilon)(Rq_1 + q_2(1-\epsilon))}{(1+(d_i-1)(q_1\epsilon + q_2(1-\epsilon)))^2} \\ &= \frac{(1-\epsilon)(1-(d_i-1)q_1(R-\epsilon))}{(1+(d_i-1)(q_1\epsilon + q_2(1-\epsilon)))^2}.\end{aligned}$$

Therefore, for $R > \epsilon$ at $q_1^* = 1$ we have $\frac{\partial f(q_1,q_2)}{\partial q_2} > 0$, when $(d_i - 1)(R - \epsilon) < 1$, and $\frac{\partial f(q_1,q_2)}{\partial q_2} \leq 0$. Therefore, for $R < \epsilon + \frac{1}{d_i-1}$ we have the optimal at $q_2^* = 1$ and the optimal value is $\frac{R+1-\epsilon}{d_i}$. For $R > \epsilon + \frac{1}{d_i-1}$ we have the optimal at $q_2^* = 0$ and the optimal value is $\frac{R}{1+(d_i-1)\epsilon}$. □

# H SIMULATIONS

We simulate the UCB-CBB algorithm for 60 sample paths and $10k$ iterations on different instances, and report the mean, 25% and 75% trajectories of cumulative $\alpha$-regret. The $\alpha$-regret is defined empirically, using the solution of the LP as an upper bound on the optimal average reward.

In addition, we report three other quantities:

(i) The empirical probability that the LP solution causes round skipping. Recall at any time $t$ and having observed context $j_t \in \mathcal{C}$, UCB-CBB samples arms using the extreme point $\{z_{j_t i}([t-M_t]^+)\})i \in \mathcal{A}$, and may return no arm if $\frac{1}{f_{j_t}} \sum_{i\in\mathcal{A}} z_{j_t i}([t-M_t]^+) < 1$. We denote this time-series by *lp skip* in the figures.

(ii) The empirical probability that the adaptive skipping technique actually skips a round to ensure future availability, even after an arm is sampled using the extreme point. We denote this time-series by *skip* in the figures.

(iii) The empirical blocking probability, namely, the time-average number of attempts to play an arm that fail due to blocking. We denote this quantity by *block* in the figures.

**A greedy heuristic:** We compare our algorithm with a UCB-based greedy algorithm, called UCB-GREEDY, that plays, among the available arms of each round, the arm of highest UCB index, given the observed context $j_t \in \mathcal{C}$, namely, $i_t^g = \arg\max_{i\in\mathcal{F}_t} \bar{\mu}_{i,j_t}(t)$, where $j_t$ is the context and $\mathcal{F}_t$ is the set of available arms at time $t$. Clearly, this heuristic does not use delayed exploitation, unlike UCB-CBB, no adaptive LP rounding is involved. For this algorithm *lp skip* and *skip* both equal 0 by construction, whereas *blocking* may occur in the case where no arm is available.

In the rest of this section, we provide simulations for empirically verifying that, even for small instances of our problem, a naive greedy approach like UCB-GREEDY fails to provide sublinear $\alpha$-regret. A larger instance of 100 arm-context pairs can be found in Section 6.

## H.1 Integral instances

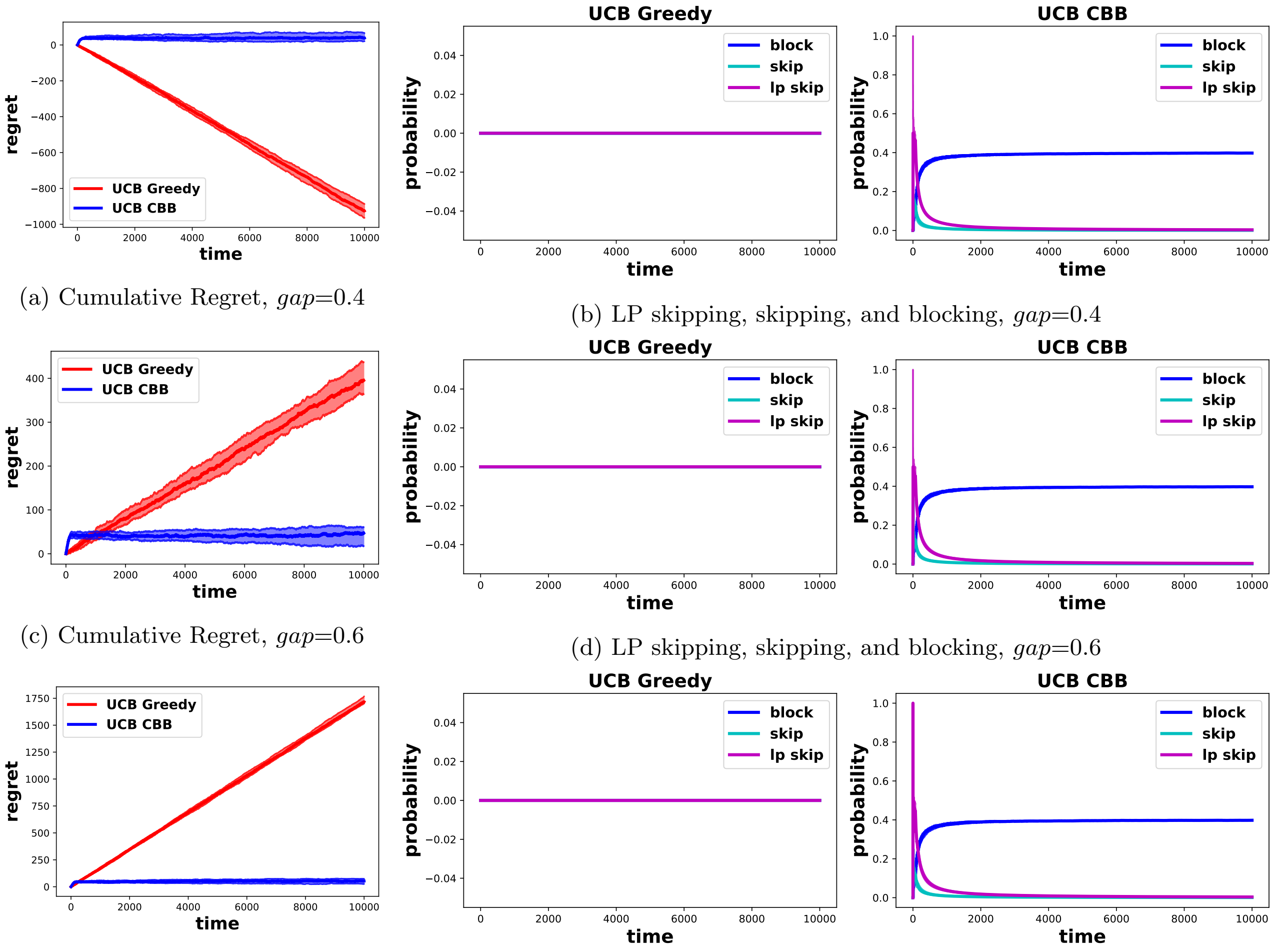
In this class, we provide simulations for demonstrating a pathological scenario for UCB-GREEDY, verifying the fact that selective round skipping is necessary for achieving optimal competitive guarantees. We consider 3 arms each of delay $d = 3$, and 3 contexts that appear with equal probability $f = 1/3$, while the rewards are generated by Bernoulli distributions (see Figure 2). Each arm is associated with a unique context of high mean reward, while for the rest of the arm-context pairs the mean reward is relatively smaller. Specifically, for arm $i \in [3]$ and context $i$, we consider mean reward equal to $\mu_{i,i} = 0.9$, whereas, for the remaining arm-context pairs have mean $\mu_{i,j} = 0.9 - gap$ for $i \neq j$, with $gap = 0.4$ for Figure 2a, 2b, $gap = 0.6$ for Figure 2c, 2d, and $gap = 0.8$ for Figure 2e, 2f.

In these cases, the (**LP**) admits a solution whose support yields a matching between arms and contexts, where arm $i$ is matched to context $i$ for $i \in [3]$. As a result, the marginal probabilities used by UCB-CBB for sampling arms are integral (i.e., equal to 1 for the arm that is matched with the current context and 0, otherwise). We see the UCB-CBB algorithm has a 0.6-Regret that grows logarithimically for all instances, whereas for the UCB-GREEDY algorithm, the 0.6-Regret is positive linear for $gap = 0.8$, and 0.6; but is negative linear for $gap = 0.4$. As it appears, UCB-GREEDY beats the UCB-CBB algorithm in the cumulative regret for $gap = 0.4$, as the effect of choosing the optimal matching in UCB-CBB is countered by the effect of adaptively skipping at a rate $\frac{2}{5}$. On the other end, for $gap = 0.8, 0.6$ the UCB-CBB algorithm performs better in the cumulative regret as the effect of choosing the optimal matching outweighs the effect of adaptive skipping. We note that this instance is dense, as $\sum_i \frac{1}{d_i} = 1$. Therefore, it is natural that UCB-GREEDY performs better when facing instances of smaller gaps among the rewards. In all the cases, the UCB-GREEDY incurs no blocking, whereas the UCB-CBB algorithm converges to an empirical blocking rate of $\frac{2}{5}$.

## H.2 Non-Integral Instances

In this class, we consider two different instances of 3 arms and 3 contexts each.

In the first instance, for each context $i \in [3]$ arm $i$ has mean reward 0.9, whereas all other arm-context pairs



(a) Cumulative Regret, *gap*=0.4

(b) LP skipping, skipping, and blocking, *gap*=0.4

(c) Cumulative Regret, *gap*=0.6

(d) LP skipping, skipping, and blocking, *gap*=0.6

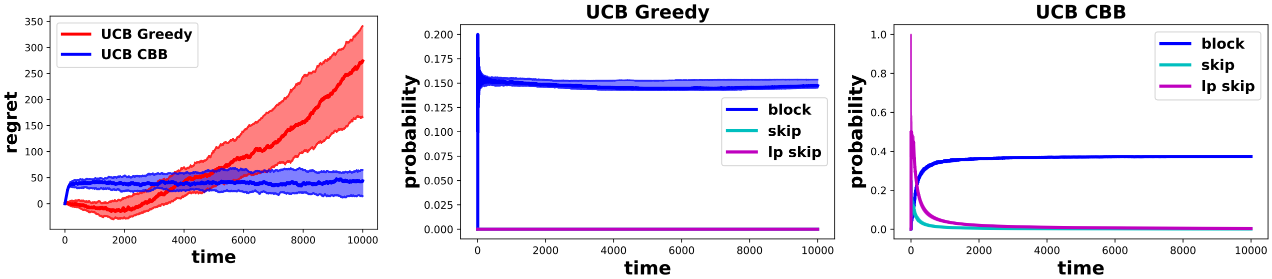
(e) Cumulative Regret, *gap*=0.8

(f) LP skipping, skipping, and blocking, *gap*=0.8

Figure 2: Integral instance on 3 arms, each of delay 3, and 3 equiprobable contexts for varying gap in the mean rewards.

have mean 0.3. The contexts are equi-probable, whereas the arms have delays 2, 3 and 6. For this instance, Figure 3 shows that the $\alpha$-regret is logarithmic for the UCB-CBB algorithm and positive linear for UCB-GREEDY. We also observe the convergence in blocking probability for both algorithms in the same figure. We note that the UCB-GREEDY also incurs blocking for this dense ($\sum_i \frac{1}{d_i} = 1$) instance.

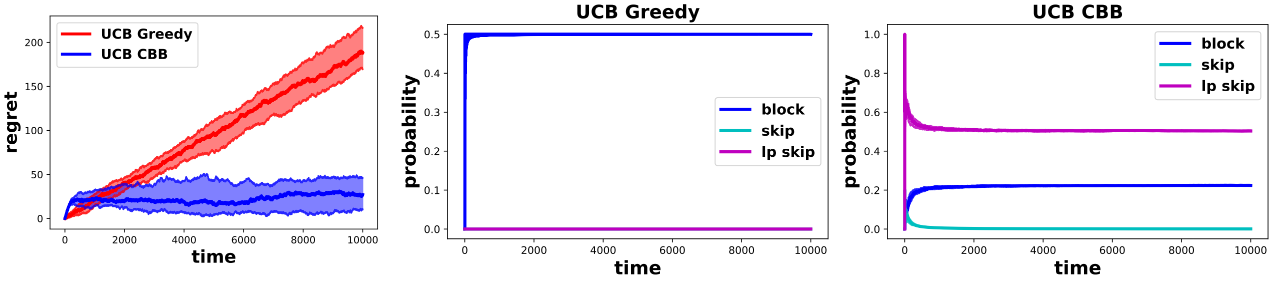
In the second instance with 3 arms and 3 contexts, for each context $i \in [3]$ arm $i$ has mean u.a.r. $[0.5, 0.9]$, whereas all other arm-context pairs have mean u.a.r. $[0, 0.3]$. In that case, the context probabilities are again chosen randomly on a simplex. All the arms have delay equal to 6. We note that this instance is non-dense, i.e. $\sum_i \frac{1}{d_i} = 1/2 < 1$. For this setting, Figure 4 shows a logarithmic $\alpha$-regret for the UCB-CBB algorithm and a linear regret for UCB-GREEDY. Both algorithms converge to non-zero probability of blocking. We note that UCB-GREEDY incurs 0.5 blocking for this non-dense instance, as compared to 0.22 blocking in UCB-CBB. This happens as UCB-CBB conserves arm $i$ for context $i$, which can be seen through high *lp block* for UCB-CBB and low regret. On the other hand, UCB-GREEDY myopically plays the best available arm at each time slot, incurring high blocking rate and high regret.



(a) Cumulative Regret

(b) LP skipping, skipping, and blocking

Figure 3: Non-integral instances with 3 arms and 3 contexts. The delays of the arms are either 2, 3, and 6. The contexts are equiprobable. The best arm per context has arm-mean 0.9, whereas all other arm-context pairs have means 0.3.



(a) Cumulative Regret

(b) LP skipping, skipping, and blocking

Figure 4: Non-integral instances with 3 arms and 3 contexts. All the arms have delay 6. The context probabilities are selected randomly. The best arm per context has mean in $[0.5, 0.9]$, whereas all other arm-context pairs have means in $[0, 0.3]$.