## A. Computational and Unconditional Hardness

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

*Proof.* We consider the simple instance of a single arm, denoted by $i$, which is always feasible to play (if available) and an infinite time horizon. The arm has a deterministic reward equal to 1, while its delay distribution can be described as:

$$D_i = \begin{cases} d, & \text{w.p. } p = \frac{1}{2} \\ 1, & \text{w.p. } 1-p, \end{cases}$$

where $d > 1$ is an integer. Note that in the above instance, we have $d_{\max} = d$.

In the above setting, assuming that the player does not have a priori knowledge of the delay realizations, we consider an online policy, $alg(q)$ defined as follows: Whenever arm $i$ is available, it is played with probability $q$. It is not hard to see that the *parameterized* policy $alg(q)$, captures any optimal policy of the player in the above asymptotic scenario. For any $q$, such a policy can be represented using a Markov Chain (MC) (c.f. (Puterman, 1994)) with $d$ states: $0, 1, \dots, d-1$. Each of these states indicates the number of subsequent rounds that remain until $i$ becomes available to play. At any round $t$, if the current state is 0 (i.e., the arm is available), the chain transitions from 0 to $d-1$ with probability $p \cdot q$. This corresponds to the case where the player chooses to play arm $i$ at time $t$ (an event that happens with probability $q$) and the delay realization at $t$ is $D_i = d$ (which happens independently with probability $p$). Alternatively, the chain transitions from state 0 to itself with probability $(1-q) + q(1-p) = 1 - p \cdot q$, which corresponds to the case where the player either does not play $i$ at time $t$, or plays $i$ at $t$ but the delay realization is 1. Finally, the MC transitions from any state $r \geq 1$ to state $r-1$ deterministically, as indicated in Figure 2.

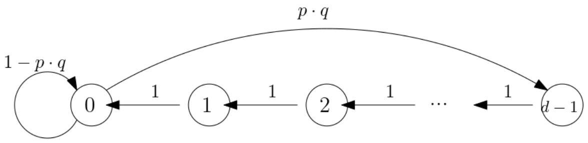

*Figure 2.* State transitions of policy $alg(q)$.

We assume that $alg(q)$ starts from state $r \in \{0, 1, ..., d-1\}$ with probability $\pi(r)$. The stationary distribution $\pi(r)$ corresponds to the solution of the following system equations: $\pi(1) = ... = \pi(d-1) = p \cdot q \cdot \pi(0)$ and $\sum_{r=0}^{d-1} \pi(r) = 1$. From this system it follows that: $\pi(0) = \frac{1}{1+p \cdot q \cdot (d-1)}$. Due to stationarity, for any time horizon $T > 0$, the expected average reward of $alg(q)$ is:

$$\text{Rew}_{avg}^{alg(q)} = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\left(\text{A}_t^{alg(q)} = i\right)\right] = q \cdot \pi(0) = \frac{q}{1 + p \cdot q \cdot (d-1)}.$$

Under the assumption that the delay realizations are not known a priori, as $T$ goes to infinity, an optimal online algorithm can be represented using the above MC (c.f. (Puterman, 1994)). To identify this policy we can maximize the expected average reward $\text{Rew}_{avg}^{alg(q)}$ w.r.t. $q \in [0, 1]$. The quantity $\text{Rew}_{avg}^{alg(q)}$ is maximized for $q = 1$. This verifies the intuition that, if the player is not a priori aware of the realizations of the delay, arm $i$ should be played every time it is available. For $q = 1$, the expected average reward collected by $alg(q)$ is: $\text{Rew}_{avg}^{alg(1)} = \frac{1}{1+p \cdot (d-1)}$.

We now turn our focus to algorithms that are a priori aware of all the delay realizations. Specifically, we consider the following (possibly sub-optimal) policy: At any time $t$, after observing the realization of $D_{i,t}$, play arm $i$ if and only if $D_{i,t} = 1$. The expected average reward of this policy within any time horizon $T$, with $T > 0$, is:

$$\text{Rew}_{avg}^{cl} = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\left(\text{A}_t^{cl} = i\right)\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\left(D_{i,t} = 1\right)\right] = 1 - p.$$

Thus, for the ratio of the expected average rewards of these two policies, we have that:

$$\frac{\text{Rew}_{avg}^{alg(1)}}{\text{Rew}_{avg}^{cl}} = \frac{1}{(1-p)(1+p\cdot(d-1))} = \frac{4}{d+1},$$

which concludes our proof. □

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

*Proof.* Suppose there is a polynomial $(\frac{k}{2})^{-\frac{1}{2}+\epsilon}$-approximation algorithm for the CBBSD problem, where $k$ is the number of arms. We consider an instance of the *Edge-Disjoint Path* (EDP) problem on a *directed* graph $\mathcal{G} = (V, E)$ with $|E| = k'$ and a set of $m$ pairs of vertices $\mathcal{T} = \{(s_i, t_i) | s_i, t_i \in V, i \in [m]\}$. By Theorem 2, it is NP-hard to distinguish whether all the $m$ pairs or at most $\frac{m}{k'^{1/2-\epsilon'}}$ pairs of the above instance can be connected by edge-disjoint paths, for any $\epsilon' > 0$. Our goal is to show that our $(\frac{k}{2})^{-\frac{1}{2}+\epsilon}$ approximation algorithm can be used to distinguish between these two cases.

First, we slightly transform the graph $\mathcal{G}$: We construct a directed graph $\mathcal{G}' = (V', E')$ such that $V' = V \cup \{s'_1, s'_2, \ldots, s'_m\}$ and $E' = E \cup \{(s'_1, s_1), (s'_2, s_2), \ldots, (s'_m, s_m)\}$, as indicated in Figure 3.

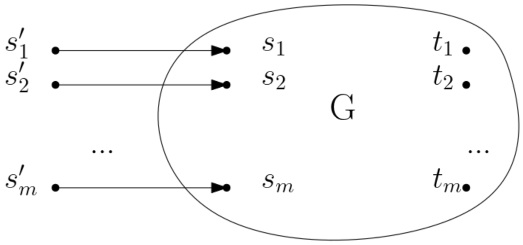

*Figure 3.* Construction of $G'$ from $G$.

We now consider an instance of the CBBSD problem where:

- Every directed edge of $\mathcal{G}'$ corresponds to an arm, that is, for the set of arms A we have that $\text{A} = E'$. The number of arms is denoted by $k$, thus $k = |A| = |E'| = k' + m$.
- The rewards of all arms are deterministically equal to 0, except for the arms that correspond to edges $(s'_i, s_i)$ for $i \in [m]$, which have a deterministic reward equal to 1. Moreover, all arms have deterministic delays equal to $m$.
- The feasible sets of arms correspond to edges that form a directed $(s'_i - t_i)$ path for a unique $i \in [m]$. That is, $\mathcal{I} = \{S \subseteq \text{A} \mid \exists i \in [m]$ unique s.t. $S$ forms a directed path from $s'_i$ to $t_i\}$.
- The time horizon is a multiple of $m$, i.e. $T = c \cdot m$ for some (polynomially bounded) $c \in \mathbb{N}_{\geq 1}$.

**Claim 1.** *We make the following observations on the above instance:*

1. *The family of feasible sets $\mathcal{I}$ does not satisfy the hereditary property.*
2. *Given a subset* F *of available arms, a maximum reward available feasible set in $\mathcal{I}(\text{F})$ can be computed in polynomial-time.*
3. *Any feasible arm-pulling schedule can be transformed in polynomial-time into a periodic schedule of period $m$ with at least the same reward.*

*Proof.* The first claim holds trivially, since a subset of a path from $s'_i$ to $t_i$ is not generally an $(s'_i - t_i)$ path.

For the second claim, we observe that the maximum possible reward collected by any algorithm during any round is $1$. This can be obtained by playing a path from $s'_i$ to $t_i$ for some $i \in [m]$, if such a path exists among the available edges. Given a graph $\mathcal{G}'$ and pairs of vertices $(s'_i, t_i)$ for $i \in [m]$, we can find a path, if one exists, in polynomial time (e.g. simply by performing a DFS for each one of the nodes $s'_i$).

For the third claim, we consider a solution to the CBBSD instance of average reward equal to $\frac{m'}{m}$, for some $m'$. Then, there exists an interval of $m$ consecutive rounds of total reward at least $m'$. Given that $T$ is polynomial in $m$, we can observe all possible sub-intervals of $m$ consecutive rounds and choose the one of maximum total reward. Let $L \subseteq [T]$ with $|L| = m$ be the above sub-interval of total reward at least $m'$. It is not hard to see that by repeating the arm-pulling pattern in $L$ for $c = \frac{T}{m}$ times, consecutively, the resulting solution has average reward at least $\frac{m'}{m}$. It remains to verify that the resulting solution is feasible. First, we observe that each set of arms played at each round is feasible, since $L$ is part of a feasible solution. Moreover, given that the arms of each set selected at every time step are played exactly once in each period of $m$ rounds and the arm delays are all equal to $m$, the resulting schedule does not violate the blocking constraints. □

The above claim allows us to compare the reward of our approximation algorithm to the reward of the optimal algorithm within a period of duration $m$. Focusing on a period of $m$, we distinguish the following two cases:

**Case 1 (Yes instance):** Suppose that in the given EDP instance, all $m$ pairs can be routed by edge-disjoint paths. Then, one possible solution can be obtained by playing the sets of arms corresponding to each of the $m$ paths in a round-robin manner. It is easy to verify that the resulting periodic schedule is feasible and has an average reward of $1$. Let us denote by $\text{Rew}^*$ the average reward of an optimal algorithm, and by $\text{Rew}^{apx}$ the corresponding average reward collected by the $(\frac{k}{2})^{-\frac{1}{2}+\epsilon}$-approximation algorithm. By the above discussion, the average reward collected by an optimal algorithm must satisfy $\text{Rew}^* \geq 1$. Thus, the approximation algorithm collects an average reward such that:

$$\text{Rew}^{apx} \geq \frac{1}{(\frac{k}{2})^{\frac{1}{2}-\epsilon}} \text{Rew}^* \geq \frac{1}{(\frac{k}{2})^{\frac{1}{2}-\epsilon}}.$$

**Case 2 (No instance):** We now focus on the case where at most a $\frac{1}{k'^{1/2-\epsilon'}}$-fraction of the pairs in $\mathcal{T}$ can be connected by edge-disjoint paths. Let the number of edge-disjoint paths in $\mathcal{T}$ be $m'$, with $m' \leq \frac{m}{k'^{1/2-\epsilon'}}$. Then, no algorithm can collect an average reward greater that $\frac{m'}{m}$, because: (i) at every time step, any algorithm can collect a reward of at most 1, by playing a feasible path, and (ii) the arms that correspond to a feasible path become blocked for the next $m-1$ rounds after they are played. Thus, collecting an average reward greater than $\frac{m'}{m}$ (that is, collecting a reward greater than $m'$ within some period of duration $m$) suggests that there exist more than $m'$ edge-disjoint paths in $\mathcal{T}$. Based on the above observation, the reward collected by the approximation algorithm is:

$$\text{Rew}^{apx} \leq \text{Rew}^* \leq \frac{1}{(k')^{\frac{1}{2}-\epsilon'}} \leq \frac{1}{(\frac{k}{2})^{\frac{1}{2}-\epsilon'}},$$

where, for the last inequality, we use that $k' \geq \frac{k}{2}$, since $k = k' + m$ and $m \leq k'$. Finally, by choosing any $\epsilon' > 0$ such that $\epsilon' < \epsilon$, we are able to distinguish between the two cases, which concludes our proof. □

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

*Proof.* In the case where the feasible set of arms satisfies the hereditary property, we establish the hardness of the CBBSD problem through a reduction from Max-k-Cover. Consider any instance of Max-k-Cover, where $U = \{e_1, \dots, e_k\}$ is the ground set, $S_1, \dots, S_m \subseteq U$ is the given family of subsets and $l$ is the number of subsets we can collect in order to cover the maximum possible number of elements in $U$.

We consider the following instance of the CBBSD problem:

- The set of arms A, with $k = |A|$, contains one arm for each element in $U$.

- Every arm in A has a deterministic reward of 1 and a deterministic delay equal to $l$, i.e. the number of subsets we can collect in the Max-k-Cover instance.
- The feasible set is defined as $\mathcal{I} = \{S \subseteq \mathrm{A} \mid \exists i \in [m] \text{ s.t. } S \subseteq S_i\}$, that is, a subset of arms is feasible if it is contained in at least one set of the given family $S_1, \dots, S_m$.
- We define the time horizon to be $T = l \cdot \lceil g(k,m) \rceil$ with $g(k,m) \in poly(k,m)$, that is, an integer multiple of $l$ that is polynomial in $k$ and $m$.

Given the above construction, we make the following observations:

**Claim 2.** *For the above instance of CBBSD problem, we have:*

1. *The family of feasible sets $\mathcal{I}$ satisfies the hereditary property.*
2. *Given a subset* F *of available arms, computing the maximum reward feasible set in $\mathcal{I}$, that is contained in* F*, can be achieved in polynomial-time.*
3. *Any solution to the above instance can be transformed in polynomial-time into a periodic schedule of period* l *and with at least the same total reward.*

*Proof.* Showing the first claim is straightforward. By definition, for any set $S \in \mathcal{I}$ we have that $S \subseteq S_i$ for some $i \in [m]$. Therefore, any subset $S' \subset S$ is also contained in $S_i$ and, thus, $S' \in \mathcal{I}$.

For the second claim, given a set of available arms F, the maximizer of $\max_{i \in [m]}\{|F \cap S_i|\}$ can be computed in polynomial time by trying all sets $S_1, \dots, S_m$ and choose the one of maximum intersection with $F$.

Let us now focus on the third claim. Consider a solution to the CBBSD instance of average reward equal to $R$. Clearly, there exists an interval of $l$ consecutive rounds of total reward at least $l \cdot R$. Otherwise, the total reward over $T = l \cdot \lceil g(k,m) \rceil$ rounds should be strictly less than $l \cdot \lceil g(k,m) \rceil \cdot R = T \cdot R$ and, thus, the average reward cannot be equal to $R$. Since $T$ is polynomial in $l, m$ and $k$, we can observe all possible sub-intervals of $l$ consecutive rounds and choose the one of maximum total reward.

Let $L \subseteq [T]$ with $|L| = l$ be the above sub-interval of total reward at least $l \cdot R$ (and average reward $R$). It is not hard to see that by repeating the arm-pulling pattern in $L$ for $\frac{T}{l}$ times, consecutively, the resulting solution has average reward at least $R$. Thus, it remains to verify that the resulting solution is feasible. Clearly, given that $L$ is part of a feasible solution, the set of arms played at each round is in $\mathcal{I}$. Finally, given that each arm is played exactly once in each period of $l$ rounds and the arm delays are all deterministic and equal to $l$, the resulting solution satisfies the blocking constraints. □

We are now ready to prove our reduction. Let OPT be the optimal solution of Max-k-Cover and Rew$^*$ be the solution in the above instance of CBBSD.

Clearly, when OPT $\geq f$ for some integer $f$, then it has to be that Rew$^* \geq \frac{T}{l} f$. Indeed, let $\{T^*_1, \dots, T^*_l\} \subseteq \{S_1, \dots, S_m\}$ be the optimal solution of Max-k-Cover, such that $|\bigcup_{i \in [l]} T^*_l| \geq f$. Then, we can construct a feasible $l$-periodic solution to the corresponding instance of CBB as follows: In a single $l$-period, the algorithm consecutively plays the arms in $(T^*_1)$, $(T^*_2 \setminus T^*_1)$, $(T^*_3 \setminus T^*_2, T^*_1)$, $\dots$, $(T^*_l \setminus T^*_1, \dots T^*_{l-1})$. Clearly, the resulting solution is feasible and has total reward at least $\frac{T}{l} f$.

We now consider the case where Rew$^* \geq \frac{T}{l} f$ for some integer $f$. Then, for the optimal solution of Max-k-Cover, it has to be that OPT $\geq f$. This holds since, w.l.o.g., we can assume that the optimal solution of CBBSD has an $l$-periodic structure, as described in Claim 2. Therefore, since the total reward is $\frac{T}{l} f$ and there are exactly $\frac{T}{l}$ periods, the reward of each period must be at least $f$. By focusing on the first period, let $\{\mathrm{A}_1, \dots, \mathrm{A}_l\}$ be arms played at each round from 1 to $l$. Since $A_t \in \mathcal{I}$ for each $t \in [l]$, let $T_t$ be a set in $\{S_1, \dots, S_m\}$ such that $\mathrm{A}_t \subseteq T_t$. Clearly, the subfamily $\{T_1, \dots, T_l\}$ consists a feasible solution to the Max-k-Cover problem such that $|\bigcup_{t \in [l]} T_l| \geq f$.

By the above analysis we can conclude that, for $\epsilon > 0$, an efficient $\left(1 - \frac{1}{e} + \epsilon\right)$-approximation algorithm for CBBSD would imply a $\left(1 - \frac{1}{e} + \epsilon\right)$-approximation algorithm for Max-k-Cover. However, by Theorem 4, we know that this is impossible, unless P = NP. □

## B. Full-Information Setting

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

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

*Proof.* Recall that we denote by $\mathrm{A}_t^*$ and $\mathrm{B}_t^*$ the set of played and blocked arms, respectively, at any time $t$. The event that arm $i$ is played at round $t$ and the event that $i$ is blocked at $t$ are mutually exclusive, namely, $\mathbb{I}\left(i \in \mathrm{A}_t^*\right) + \mathbb{I}\left(i \in \mathrm{B}_t^*\right) \leq 1$. Further, an arm $i$ is blocked at round $t$ if and only if the arm has been played in some previous time step and is still unavailable at $t$. Therefore, at any round $t$, we get:

$$\mathbb{I}\left(i \in \mathrm{A}_t^*\right) + \sum_{t' < t} \mathbb{I}\left(i \in \mathrm{A}_{t'}^*\right) \mathbb{I}\left(D_{i,t'} > t - t'\right) \leq 1.$$

Note that the LHS of the inequality above cannot be greater than 1 because the events are mutually exclusive. Indeed, if an arm has been played at some round $t' < t$ and remains blocked until after round $t$, then this arm cannot have been played at any subsequent time point between $t'$ and $t$ (including $t$).

By taking expectation in the above expression, we have that:

$$\mathbb{P}\left(i \in \mathrm{A}_t^*\right) + \sum_{t' < t} \mathbb{E}\left[\mathbb{I}\left(i \in \mathrm{A}_{t'}^*\right) \mathbb{I}\left(D_{i,t'} > t - t'\right)\right] \leq 1, \quad \forall t \in [T] \tag{5}$$

For any two rounds $t$ and $t'$ such that $t' < t$, since we assume that the optimal algorithm is not aware of the delay realization of an arm before playing it, the events $\{i \in \mathrm{A}_{t'}^*\}$ and $\{D_{i,t'} > t - t'\}$ are conditionally independent given the history of arms played up to (and including) time $t' - 1$. Let us denote the history up to $t$ as $H_t^*$, then:

$$\begin{aligned}
\mathbb{E}\left[\mathbb{I}\left(i \in \mathrm{A}_{t'}^*\right) \mathbb{I}\left(D_{i,t'} > t - t'\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\left(i \in \mathrm{A}_{t'}^*\right) \mathbb{I}\left(D_{i,t'} > t - t'\right) \mid H_{t'-1}^*\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\left(i \in \mathrm{A}_{t'}^*\right) \mid H_{t'-1}^*\right] \mathbb{E}\left[\mathbb{I}\left(D_{i,t'} > t - t'\right) \mid H_{t'-1}^*\right]\right] \\
&= \mathbb{P}\left(i \in \mathrm{A}_{t'}^*\right) \mathbb{P}\left(D_{i,t'} > t - t'\right).
\end{aligned}$$

By summing (5) over all rounds $t \in [T]$ and since $\mathbb{P}\left(D_{i,t} > 0\right) = 1 \; \forall i \in \mathrm{A} \; \forall t \in [T]$, we get:

$$\sum_{t \in [T]} \sum_{t' \leq t} \mathbb{P}\left(i \in \mathrm{A}_{t'}^*\right) \mathbb{P}\left(D_{i,t'} > t - t'\right) = \sum_{t \in [T]} \mathbb{P}\left(i \in \mathrm{A}_t^*\right) + \sum_{t \in [T]} \sum_{t' < t} \mathbb{P}\left(i \in \mathrm{A}_{t'}^*\right) \mathbb{P}\left(D_{i,t'} > t - t'\right) \leq T. \tag{6}$$

Since the variables $D_{i,t}$ for $t \in [T]$ are i.i.d., we can consider any sample $D_i$ from the delay distribution of arm $i$ and have $\mathbb{P}\left(D_i > t - t'\right) = \mathbb{P}\left(D_{i,t} > t - t'\right)$. Moreover, since $D_i$ has support in $\{1, ..., d_i^{\max}\}$, we can write:

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{t'\leq t}\mathbb{P}\left(i\in A^*_{t'}\right)\mathbb{P}\left(D_i>t-t'\right)\\
&=\sum_{t\in[T]}\sum_{t'=\max\{1,t-d_i^{\max}\}}^{t}\mathbb{P}\left(i\in \mathrm{A}^*_{t'}\right)\mathbb{P}\left(D_i>t-t'\right)\\
&=\sum_{t'\in[T]}\sum_{t=t'}^{\min\{T,t'+d_i^{\max}-1\}}\mathbb{P}\left(i\in \mathrm{A}^*_{t'}\right)\mathbb{P}\left(D_i>t-t'\right)\\
&=\sum_{t'\in[T]}\mathbb{P}\left(i\in \mathrm{A}^*_{t'}\right)\sum_{t=t'}^{\min\{T,t'+d_i^{\max}-1\}}\mathbb{P}\left(D_i>t-t'\right)\\
&\geq\sum_{t'\in[T]}\mathbb{P}\left(i\in \mathrm{A}^*_{t'}\right)\sum_{t=t'}^{t'+d_i^{\max}-1}\mathbb{P}\left(D_i>t-t'\right)-\sum_{t'=T-d_i^{\max}+1}^{T}\sum_{t=t'}^{T}\mathbb{P}\left(D_i>t-t'\right)\\
&=\sum_{t'\in[T]}\mathbb{P}\left(i\in \mathrm{A}^*_{t'}\right)\mathbb{E}\left[D_i\right]-\sum_{t'=T-d_i^{\max}+1}^{T}\sum_{t=t'}^{T}\mathbb{P}\left(D_i>t-t'\right),
\end{aligned}$$

where the inequality is due to upper bounding the probability $\mathbb{P}\left(i\in A^*_{t'}\right)$ by 1. Further, in the last equality we use the fact that $D_i$ is a strictly positive random variable with bounded support in $\{1,\dots,d_i^{\max}\}$, thus:

$$\mathbb{E}\left[D_i\right]=\sum_{t=t'}^{t'+d_i^{\max}-1}\mathbb{P}\left(D_i>t-t'\right).$$

Therefore, inequality (6) becomes:

$$\sum_{t'\in[T]}\mathbb{P}\left(i\in \mathrm{A}^*_{t'}\right)\mathbb{E}\left[D_i\right]-\sum_{t'=T-d_i^{\max}+1}^{T}\sum_{t=t'}^{T}\mathbb{P}\left(D_i>t-t'\right)\leq T.$$

Finally, by rearranging terms in the above expression and dividing by $\mathbb{E}\left[D_i\right]\cdot T$, we conclude that:

$$\begin{aligned}
\frac{1}{T}\sum_{t'\in[T]}\mathbb{P}\left(i\in \mathrm{A}^*_{t'}\right)&\leq\frac{1}{\mathbb{E}\left[D_i\right]}+\frac{1}{T\,\mathbb{E}\left[D_i\right]}\sum_{t'=T-d_i^{\max}+1}^{T}\sum_{t=t'}^{T}\mathbb{P}\left(D_i>t-t'\right)\\
&\leq\frac{1}{\mathbb{E}\left[D_i\right]}+\frac{d_i^{\max}\,\mathbb{E}\left[D_i\right]}{T\,\mathbb{E}\left[D_i\right]}\\
&=\frac{1}{\mathbb{E}\left[D_i\right]}+\frac{d_i^{\max}}{T},
\end{aligned}$$

where in the last inequality we use the fact that the second sum $\sum_{t=t'}^{T}\mathbb{P}\left(D_i>t-t'\right)$ can be upper bounded by $\mathbb{E}\left[D_i\right]$, for all $t'$ specified in the first sum.

□

## C. Bandit Setting: Omitted Proofs

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

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

*Proof.* Let $H_t^{\tilde{\pi}}$ be the history of arm playing of our bandit algorithm up to (including) time $t$. The expected reward collected by our algorithm can be expressed as:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t\in[T]}\sum_{i\in\mathrm{A}} X_{i,t}\,\mathbb{I}\left(i\in\mathrm{A}_t^{\tilde{\pi}}\right)\right] &= \sum_{t\in[T]}\sum_{i\in\mathrm{A}}\mathbb{E}\left[X_{i,t}\,\mathbb{I}\left(i\in\mathrm{A}_t^{\tilde{\pi}}\right)\right]\\
&= \sum_{t\in[T]}\sum_{i\in\mathrm{A}}\mathbb{E}\left[\mathbb{E}\left[X_{i,t}\,\mathbb{I}\left(i\in\mathrm{A}_t^{\tilde{\pi}}\right)|H_{t-1}^{\tilde{\pi}}\right]\right]\\
&= \sum_{t\in[T]}\sum_{i\in\mathrm{A}}\mathbb{E}\left[\mathbb{E}\left[X_{i,t}|H_{t-1}^{\tilde{\pi}}\right]\mathbb{E}\left[\mathbb{I}\left(i\in\mathrm{A}_t^{\tilde{\pi}}\right)|H_{t-1}^{\tilde{\pi}}\right]\right]\\
&= \sum_{t\in[T]}\sum_{i\in\mathrm{A}}\mathbb{E}\left[\mu_i\,\mathbb{I}\left(i\in\mathrm{A}_t^{\tilde{\pi}}\right)\right]\\
&= \sum_{t\in[T]}\mathbb{E}\left[\mu(\mathrm{A}_t^{\tilde{\pi}})\right],
\end{aligned}$$

where we use the fact that, at each round $t\in[T]$, the reward $X_{i,t}$ and the choice of the set $\mathrm{A}_t^{\tilde{\pi}}$ are independent conditioned on the history $H_t^{\tilde{\pi}}$.

We recall the expected pulling rate of each arm $i\in\mathrm{A}$ by an optimal full information algorithm that satisfies Conditions 1 and 2, as defined in Section 4:

$$z_i=\mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]}\mathbb{I}\left(i\in\mathrm{A}_t^*\right)\right],$$

where $\mathrm{A}_t^*$ is the set of arms played by the optimal algorithm at time $t$.

A direct application of the definition of $\Gamma_t^{\tilde{\pi}}$, for every $t\in[T]$, gives:

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

By working along the lines of Theorem 6 and carrying the extra term of $-\Gamma_t^{\tilde{\pi}}$, we can see that for any round $t\in[T]$:

$$\begin{aligned}
\mathbb{E}\left[\mu(\mathrm{A}_t^{\tilde{\pi}})\right] &= \alpha\cdot\beta\cdot\mathbb{E}\left[\mu(\mathbf{OPT}_\mu(\mathrm{F}_t^{\tilde{\pi}}))\right]-\mathbb{E}\left[\Gamma_t^{\tilde{\pi}}\right]\\
&\geq \alpha\cdot\beta\cdot\mathbb{E}\left[\sum_{i\in\mathrm{F}_t^{\tilde{\pi}}}\mu_i z_i\right]-\mathbb{E}\left[\Gamma_t^{\tilde{\pi}}\right]\\
&= \alpha\cdot\beta\sum_{i\in\mathrm{A}}\mu_i z_i-\alpha\cdot\beta\,\mathbb{E}\left[\sum_{i\in\mathrm{B}_t^{\tilde{\pi}}}\mu_i z_i\right]-\mathbb{E}\left[\Gamma_t^{\tilde{\pi}}\right],
\end{aligned}$$

where the inequality above is due to Lemma 2 and expectations are taken over all sources of randomness (that is, reward and delay realizations).

By summing over $t\in T$ and using the fact that $\mathbb{E}\left[\sum_{t\in[T]}\sum_{i\in\mathrm{B}_t^{\tilde{\pi}}}\mu_i z_i\right]\leq\mathrm{Rew}^{\tilde{\pi}}(T)+\mathcal{O}(d_{\max}\cdot k)$ (which follows exactly as in Equation (4) in Theorem 6), we have:

$$\mathrm{Rew}^{\tilde{\pi}}(T)=\mathbb{E}\left[\sum_{t\in T}\mu(\mathrm{A}_t^{\tilde{\pi}})\right]\geq\alpha\beta\,\mathrm{Rew}^*(T)-\alpha\beta\,\mathrm{Rew}^{\tilde{\pi}}(T)-\mathcal{O}\left(d_{\max}\cdot k\right)-\mathbb{E}\left[\sum_{t\in T}\Gamma_t^{\tilde{\pi}}\right].$$

By rearranging terms in the expression above, we get that:

$$\mathrm{Rew}^{\tilde{\pi}}(T)\geq\frac{\alpha\cdot\beta}{1+\alpha\cdot\beta}\mathrm{Rew}_I^*(T)-\mathcal{O}\left(d_{\max}\cdot k\right)-\frac{1}{1+\alpha\cdot\beta}\mathbb{E}\left[\sum_{t\in[T]}\Gamma_t^{\tilde{\pi}}\right].$$

Thus, by definition of the $\rho$-regret, we can conclude that:

$$\rho\,\mathrm{Reg}^{\tilde{\pi}}(T) = \frac{\alpha\beta}{1+\alpha\beta}\,\mathrm{Rew}^{*}_{I}(T) - \mathrm{Rew}^{\tilde{\pi}}(T) \leq \frac{1}{1+\alpha\cdot\beta}\,\mathbb{E}\left[\sum_{t\in[T]} \Gamma^{\tilde{\pi}}_{t}\right] + \mathcal{O}(d_{\max}\cdot k).$$

□

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

*Proof.* This result is due to (Chen et al., 2016b), but we include its proof here for completeness.

For any $T_{i,t-1} > 0$ and $i \in \mathrm{A}$, using Hoeffding's inequality (see (Mitzenmacher & Upfal, 2005)) and the fact that $T_{i,t-1} \leq t-1$, we have that:

$$\mathbb{P}\left(|\hat{\mu}_{i,t-1} - \mu_i| \geq \sqrt{\frac{3\ln(t)}{2T_{i,t-1}}}\right) \leq \sum_{s=1}^{t-1} \mathbb{P}\left(|\hat{\mu}_{i,t-1} - \mu_i| \geq \sqrt{\frac{3\ln(t)}{2s}}\right) \leq 2te^{-3\ln t} \leq \frac{2}{t^2}.$$

The lemma follows by taking union bound over $i \in \mathrm{A}$ and using the above inequality:

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

□

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

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

*Proof.* The proof of this lemma follows closely the analysis of (Wang & Chen, 2017) for the case of a fixed availability set. However, in our case the availability set at every round is correlated with previous actions. To tackle this issue, we make the following relaxation: For each arm $i$, we consider the minimum possible suboptimality gap associated with arm $i$, $\Delta^{\min}_i$ (see Definition 7). This allows us to keep track of the number of rounds where our bandit algorithm can select a bad feasible solution containing arm $i$ (w.r.t. the availability set of the round), by examining whether $T_i \leq \ell_T(\Delta^{i}_{\min})$. This relaxation allows us to eventually drop the dependence on the availability set.

Observe that, by definition of $\mathcal{N}_t$, at any time $t$, if $\mathcal{N}_t$ holds, then $\forall i \in \mathrm{A}$ the UCB indices at round $t$ can be bounded as follows:

$$|\,\bar{\mu}_{i,t-1} - \mu_i| \leq \min\left\{2\sqrt{\frac{3\ln t}{2T_{i,t-1}}}, 1\right\}, \tag{7}$$

and:

$$\bar{\mu}_{i,t-1} \geq \mu_i. \tag{8}$$

Then, if $\mathcal{Q}_t$ holds, we have that:

$$\bar{\mu}(\mathrm{A}^{\tilde{\pi}}_t) \geq \alpha\cdot\bar{\mu}(\mathrm{OPT}_{\bar{\mu}}(\mathrm{F}^{\tilde{\pi}}_t)) \geq \alpha\cdot\bar{\mu}(\mathrm{OPT}_{\mu}(\mathrm{F}^{\tilde{\pi}}_t)),$$

where for the last inequality we use the definition of $\mathrm{OPT}_{\bar{\mu}}$. By applying Equation (8), we get that:

$$\bar{\mu}(\mathrm{A}_t^{\tilde{\pi}}) \geq \alpha \cdot \bar{\mu}(\mathrm{OPT}_{\mu}(\mathrm{F}_t^{\tilde{\pi}})) \geq \alpha \cdot \mu(\mathrm{OPT}_{\mu}(\mathrm{F}_t^{\tilde{\pi}})) = \mu(\mathrm{A}_t^{\tilde{\pi}}) + \Delta(\mathrm{A}_t^{\tilde{\pi}}, \mathrm{F}_t^{\tilde{\pi}}),$$

where for the last equality we use the definition of $\Delta(\mathrm{A}_t^{\tilde{\pi}}, \mathrm{F}_t^{\tilde{\pi}})$.

By reordering terms, we get that:

$$\Delta(\mathrm{A}_t^{\tilde{\pi}}, \mathrm{F}_t^{\tilde{\pi}}) \leq \bar{\mu}(\mathrm{A}_t^{\tilde{\pi}}) - \mu(\mathrm{A}_t^{\tilde{\pi}}) \leq \sum_{i \in \mathrm{A}_t^{\tilde{\pi}}} |\, \bar{\mu}_i - \mu_i|. \tag{9}$$

The event $\mathrm{A}_t^{\tilde{\pi}} \in \mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}})$ dictates that a bad feasible solution has been played at round $t$. Then, using the definition of $\Delta^i_{\text{min}}$, for the suboptimality gap of the set selected at round $t$ we have that:

$$\Delta(\mathrm{A}_t^{\tilde{\pi}}, \mathrm{F}_t^{\tilde{\pi}}) \geq \max_{i \in \mathrm{A}_t^{\tilde{\pi}}} \Delta^i_{\text{min}}.$$

If we apply this inequality to Equation (9), we get:

$$\sum_{i \in \mathrm{A}_t^{\tilde{\pi}}} |\, \bar{\mu}_i - \mu_i| - \max_{i \in \mathrm{A}_t^{\tilde{\pi}}} \Delta^i_{\text{min}} \geq 0. \tag{10}$$

Now, using Equations (9) and (10) we can bound the suboptimality gap of the bad set selected at round $t$ as follows:

$$\Delta(\mathrm{A}_t^{\tilde{\pi}}, \mathrm{F}_t^{\tilde{\pi}}) \leq \sum_{i \in \mathrm{A}_t^{\tilde{\pi}}} |\, \bar{\mu}_i - \mu_i| \leq 2 \sum_{i \in \mathrm{A}_t^{\tilde{\pi}}} |\, \bar{\mu}_i - \mu_i| - \max_{i \in \mathrm{A}_t^{\tilde{\pi}}} \Delta^i_{\text{min}} \leq 2 \sum_{i \in \mathrm{A}_t^{\tilde{\pi}}} (|\, \bar{\mu}_i - \mu_i| - \frac{\Delta^i_{\text{min}}}{2|\, \mathrm{A}_t^{\tilde{\pi}}\,|}) \leq 2 \sum_{i \in \mathrm{A}_t^{\tilde{\pi}}} \Big(|\, \bar{\mu}_i - \mu_i| - \frac{\Delta^i_{\text{min}}}{2r}\Big),$$

where in the last inequality we use that $|\, \mathrm{A}_t^{\tilde{\pi}}\,| \leq r$, by definition of $r = \max_{S \in \mathcal{I}}\{|S|\}$. Using Equation (7), the expression above becomes:

$$\Delta(\mathrm{A}_t^{\tilde{\pi}}, \mathrm{F}_t^{\tilde{\pi}}) \leq 2 \sum_{i \in \mathrm{A}_t^{\tilde{\pi}}} \Big( \min\{2\sqrt{\frac{3 \ln t}{2T_{i,t-1}}}, 1\} - \frac{\Delta^i_{\text{min}}}{2r} \Big) = \sum_{i \in \mathrm{A}_t^{\tilde{\pi}}} \Big( \min\{\sqrt{\frac{24 \ln t}{T_{i,t-1}}}, 2\} - \frac{\Delta^i_{\text{min}}}{r} \Big).$$

Now, observe that if $T_{i,t-1} \leq \ell_T(\Delta^i_{\text{min}})$, then $\min\{\sqrt{\frac{24 \ln t}{T_{i,t-1}}}, 2\} - \frac{\Delta^i_{\text{min}}}{r} \leq \min\{\sqrt{\frac{24 \ln t}{T_{i,t-1}}}, 2\} = \kappa_T(\Delta^i_{\text{min}}, T_{i,t-1})$. On the other hand, if $T_{i,t-1} > \ell_T(\Delta^i_{\text{min}})$, then $\min\{\sqrt{\frac{24 \ln t}{T_{i,t-1}}}, 2\} \leq \sqrt{\frac{24 \ln t}{T_{i,t-1}}} \leq \sqrt{\frac{24 \ln t}{24 \ln T r^2 / (\Delta^i_{\text{min}})^2}} \leq \frac{\Delta^i_{\text{min}}}{r}$, thus $\min\{\sqrt{\frac{24 \ln t}{T_{i,t-1}}}, 2\} - \frac{\Delta^i_{\text{min}}}{r} \leq 0 = \kappa_T(\Delta^i_{\text{min}}, T_{i,t-1})$. Therefore, we conclude that:

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

□

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

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

*and the following* distribution-independent *bound:*

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

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

*Proof.* We first focus on the distribution-dependent bound of the regret. According to Lemma 3, to bound the $\rho$-approximate regret of our bandit algorithm, it suffices to bound the loss accumulated due to the instantaneous regret over time. This loss can be rewritten as:

$$\begin{aligned}\mathbb{E}\left[\sum_{t\in[T]}\Gamma_t^{\tilde{\pi}}\right] &= \mathbb{E}\left[\sum_{t\in[T]}\Big(\alpha\cdot\beta\cdot\mu(\mathrm{OPT}_\mu(\mathrm{F}_t^{\tilde{\pi}}))-\mu(\mathrm{A}_t^{\tilde{\pi}})\Big)\right]\\ &= \mathbb{E}\left[\sum_{t\in[T]}\Big(\alpha\cdot\mu(\mathrm{OPT}_\mu(\mathrm{F}_t^{\tilde{\pi}}))-\mu(\mathrm{A}_t^{\tilde{\pi}})+\alpha\cdot(\beta-1)\cdot\mu(\mathrm{OPT}_\mu(\mathrm{F}_t^{\tilde{\pi}}))\Big)\right].\end{aligned}$$

Focusing on the term $\alpha\cdot\mu(\mathrm{OPT}_\mu(\mathrm{F}_t^{\tilde{\pi}}))-\mu(\mathrm{A}_t^{\tilde{\pi}})$ of the loss, observe that this quantity is positive only when the algorithm plays a bad feasible solution w.r.t. the availability set $\mathrm{F}_t^{\tilde{\pi}}$. Thus, we can write:

$$\begin{aligned}\mathbb{E}\left[\sum_{t\in[T]}\Gamma_t^{\tilde{\pi}}\right] &= \mathbb{E}\left[\sum_{t\in[T]}\Big(\alpha\cdot\mu(\mathrm{OPT}_\mu(\mathrm{F}_t^{\tilde{\pi}}))-\mu(\mathrm{A}_t^{\tilde{\pi}})+\alpha\cdot(\beta-1)\cdot\mu(\mathrm{OPT}_\mu(\mathrm{F}_t^{\tilde{\pi}}))\Big)\right]\\ &\leq \mathbb{E}\left[\sum_{t\in[T]}\Big(\mathbb{I}\left(\mathrm{A}_t^{\tilde{\pi}}\in\mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}})\right)(\alpha\cdot\mu(\mathrm{OPT}_\mu(\mathrm{F}_t^{\tilde{\pi}})-\mu(\mathrm{A}_t^{\tilde{\pi}})))+\alpha\cdot(\beta-1)\cdot\mu(\mathrm{OPT}_\mu(\mathrm{F}_t^{\tilde{\pi}}))\Big)\right]\\ &= \mathbb{E}\left[\sum_{t\in[T]}\Big(\mathbb{I}\left(\mathrm{A}_t^{\tilde{\pi}}\in\mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}})\right)\Delta(\mathrm{A}_t^{\tilde{\pi}},\mathrm{F}_t^{\tilde{\pi}})+\alpha\cdot(\beta-1)\cdot\mu(\mathrm{OPT}_\mu(\mathrm{F}_t^{\tilde{\pi}}))\Big)\right]. \qquad (11)\end{aligned}$$

Now, we can use the idea of (Wang & Chen, 2017) to treat the parts of the regret that result from oracle fails or non-representative sampling separately. However, in our case, we also have to address the fact that our availability set changes over time, depending on past actions. We decompose the indicator in Equation (11) based on whether the events $\mathcal{Q}_t$ and $\mathcal{N}_t$ hold, as follows:

$$\begin{aligned}(11) \leq \mathbb{E}&\left[\sum_{t\in[T]}\mathbb{I}\left(\mathcal{Q}_t,\mathcal{N}_t,\mathrm{A}_t^{\tilde{\pi}}\in\mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}})\right)\cdot\Delta(\mathrm{A}_t^{\tilde{\pi}},\mathrm{F}_t^{\tilde{\pi}})\right] &(12)\\ +\mathbb{E}&\left[\sum_{t\in[T]}\mathbb{I}\left(\neg\mathcal{Q}_t,\mathrm{A}_t^{\tilde{\pi}}\in\mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}})\right)\cdot\Delta(\mathrm{A}_t^{\tilde{\pi}},\mathrm{F}_t^{\tilde{\pi}})+\alpha\cdot(\beta-1)\cdot\mu(\mathrm{OPT}_\mu(\mathrm{F}_t^{\tilde{\pi}}))\right] &(13)\\ +\mathbb{E}&\left[\sum_{t\in[T]}\mathbb{I}\left(\neg\mathcal{N}_t,\mathrm{A}_t^{\tilde{\pi}}\in\mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}})\right)\cdot\Delta(\mathrm{A}_t^{\tilde{\pi}},\mathrm{F}_t^{\tilde{\pi}})\right]. &(14)\end{aligned}$$

We bound each one of the above terms, separately. For (13), to address the issue that the availability set changes at each round, we consider the suboptimality gap of the worst feasible solution contained in the availability set *F*, $\Delta_{max}(F)$, namely:

$$\Delta_{\max}(F)=\max_{i\in\mathrm{A},\ S\in S_{i,B}(F)}\Delta(S,F).$$

We use the above definition to upper bound the suboptimality gap, $\Delta(\mathrm{A}_t^{\tilde{\pi}}, \mathrm{F}_t^{\tilde{\pi}})$, appearing in Equation (13). We have that:

$$\begin{aligned}
(13) &\leq \mathbb{E}\left[\sum_{t\in[T]} \Big(\mathbb{I}\left(\neg\mathcal{Q}_t\right)\cdot\Delta_{\max}(\mathrm{F}_t^{\tilde{\pi}}) + \alpha\cdot(\beta-1)\cdot\mu(\mathrm{OPT}_\mu(\mathrm{F}_t^{\tilde{\pi}}))\Big)\right] \\
&= \mathbb{E}\left[\sum_{t\in[T]}\sum_{F\subseteq \mathrm{A}} \mathbb{E}\left[\mathbb{I}\left(\neg\mathcal{Q}_t\right)\cdot\Delta_{\max}(F) + \alpha\cdot(\beta-1)\cdot\mu(\mathrm{OPT}_\mu(F)) \mid \mathrm{F}_t^{\tilde{\pi}} = F\right]\mathbb{I}\left(\mathrm{F}_t^{\tilde{\pi}} = F\right)\right] \\
&= \mathbb{E}\left[\sum_{t\in[T]}\sum_{F\subseteq \mathrm{A}} \Big(\mathbb{P}\left(\neg\mathcal{Q}_t \mid \mathrm{F}_t^{\tilde{\pi}} = F\right)\cdot\Delta_{\max}(F) + \alpha\cdot(\beta-1)\cdot\mu(\mathrm{OPT}_\mu(F))\Big)\mathbb{I}\left(\mathrm{F}_t^{\tilde{\pi}} = F\right)\right] \\
&\leq \mathbb{E}\left[\sum_{t\in[T]}\sum_{F\subseteq \mathrm{A}} \Big((1-\beta)\cdot\Delta_{\max}(F) + \alpha\cdot(\beta-1)\cdot\mu(\mathrm{OPT}_\mu(F))\Big)\mathbb{I}\left(\mathrm{F}_t^{\tilde{\pi}} = F\right)\right] \\
&\leq 0, \qquad (15)
\end{aligned}$$

where to obtain the first equality we use the law of total expectation. The second inequality holds since the success probability of the oracle is at least $\beta$ independently of the availability set, thus, $\mathbb{P}\left(\neg\mathcal{Q}_t \mid \mathrm{F}_t^{\tilde{\pi}} = F\right) \leq 1-\beta$. Finally, the last inequality follows by the fact that, for the suboptimality gap of any bad feasible set $S$ (w.r.t. an availability set $F$), we have $\Delta(S,F) \leq \alpha\cdot\mu(\mathrm{OPT}_\mu(F))$, thus, $\Delta_{\max}(F) \leq \alpha\cdot\mu(\mathrm{OPT}_\mu(F))$.

For the term (14), using the definition of $\Delta_{\max}$ and Lemma 4, we have that:

$$(14) \leq \mathbb{E}\left[\sum_{t\in[T]}\mathbb{I}\left(\neg\mathcal{N}_t\right)\cdot\Delta_{\max}\right] = \sum_{t\in[T]}\mathbb{P}\left(\neg\mathcal{N}_t\right)\cdot\Delta_{\max} \leq \sum_{t\in[T]}\frac{2k}{t^2}\Delta_{\max} \leq \frac{\pi^2 k}{3}\Delta_{\max}\,. \qquad (16)$$

The term in (12) can be bounded by utilizing Lemma 5:

$$\begin{aligned}
(12) &= \mathbb{E}\left[\sum_{t\in[T]}\mathbb{I}\left(\mathcal{Q}_t, \mathcal{N}_t, \mathrm{A}_t^{\tilde{\pi}} \in \mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}})\right)\cdot\Delta(\mathrm{A}_t^{\tilde{\pi}}, \mathrm{F}_t^{\tilde{\pi}})\right] \\
&\leq \mathbb{E}\left[\sum_{t\in[T]}\sum_{i\in\mathrm{A}_t^{\tilde{\pi}}}\mathbb{I}\left(\mathcal{Q}_t, \mathcal{N}_t, \mathrm{A}_t^{\tilde{\pi}} \in \mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}})\right)\kappa_T(\Delta^i_{\mathrm{min}}, T_{i,t-1})\right] \\
&\leq \mathbb{E}\left[\sum_{i\in\mathrm{A}}\sum_{s=0}^{T_{i,T}}\kappa_T(\Delta^i_{\mathrm{min}}, s)\right], \qquad (17)
\end{aligned}$$

where the last inequality holds because a bad set containing arm $i$ can be played at most $T_{i,T}$ times. Then, we substitute $\kappa_T(\Delta^i_{\mathrm{min}}, s)$ and use an integral to upper bound the inner sum, following the idea of (Wang & Chen, 2017):

$$\begin{aligned}
(17) &\leq \mathbb{E}\left[\sum_{i\in\mathrm{A}}\sum_{s=0}^{\ell_T(\Delta^i_{\mathrm{min}})}\kappa_T(\Delta^i_{\mathrm{min}}, s)\right] = 2k + \sum_{i\in\mathrm{A}}\sum_{s=1}^{\ell_T(\Delta^i_{\mathrm{min}})}\sqrt{\frac{24\ln(T)}{s}} \leq 2k + \sum_{i\in\mathrm{A}}\int_{s=0}^{\ell_T(\Delta^i_{\mathrm{min}})}\sqrt{\frac{24\ln(T)}{s}}\,ds \\
&\leq 2k + \sum_{i\in\mathrm{A}}2\sqrt{24\ln(T)\ell_T(\Delta^i_{\mathrm{min}})} \leq 2k + \sum_{i\in\mathrm{A}}2\sqrt{24\ln(T)\frac{24r^2\ln(T)}{(\Delta^i_{\mathrm{min}})^2}} \leq 2k + \sum_{i\in\mathrm{A}}48\frac{r}{\Delta^i_{\mathrm{min}}}\ln(T)\,. \qquad (18)
\end{aligned}$$

We conclude our proof for the distribution-dependent bound by combining the above results with Lemma 3.

We now focus on the distribution-independent bound of the regret. The following part is similar to the proof for the distribution-dependent bound. We rely on Lemma 3 to bound the $\rho$-approximate regret by bounding the expected loss due to the instantaneous regret over time. By Equation (11) we have that:

$$\mathbb{E}\left[\sum_{t\in[T]}\Gamma_t^{\tilde{\pi}}\right]\leq\mathbb{E}\left[\sum_{t\in[T]}\Big(\mathbb{I}\left(\mathrm{A}_t^{\tilde{\pi}}\in\mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}})\right)\Delta(\mathrm{A}_t^{\tilde{\pi}},\mathrm{F}_t^{\tilde{\pi}})+\alpha\cdot(\beta-1)\cdot\mu(\mathrm{OPT}_\mu(\mathrm{F}_t^{\tilde{\pi}}))\Big)\right].$$

As before, we distinguish the following cases based on whether the events $\mathcal{Q}_t$ and $\mathcal{N}_t$ hold:

$$\mathbb{E}\left[\sum_{t\in[T]}\Gamma_t^{\tilde{\pi}}\right]\leq\mathbb{E}\left[\sum_{t\in[T]}\mathbb{I}\left(\mathcal{Q}_t,\mathcal{N}_t,\mathrm{A}_t^{\tilde{\pi}}\in\mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}})\right)\cdot\Delta(\mathrm{A}_t^{\tilde{\pi}},\mathrm{F}_t^{\tilde{\pi}})\right] \tag{19}$$

$$+\mathbb{E}\left[\sum_{t\in[T]}\Big(\mathbb{I}\left(\neg\mathcal{Q}_t,\mathrm{A}_t^{\tilde{\pi}}\in\mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}})\right)\cdot\Delta(\mathrm{A}_t^{\tilde{\pi}},\mathrm{F}_t^{\tilde{\pi}})+\alpha\cdot(\beta-1)\cdot\mu(\mathrm{OPT}_\mu(\mathrm{F}_t^{\tilde{\pi}}))\Big)\right] \tag{20}$$

$$+\mathbb{E}\left[\sum_{t\in[T]}\mathbb{I}\left(\neg\mathcal{N}_t,\mathrm{A}_t^{\tilde{\pi}}\in\mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}})\right)\cdot\Delta(\mathrm{A}_t^{\tilde{\pi}},\mathrm{F}_t^{\tilde{\pi}})\right]. \tag{21}$$

The terms (20) and (21) can be bounded using Equations (15) and (16), respectively. As in (Wang & Chen, 2017), we define $M=\sqrt{(48kr\ln T)/T}$. We use this definition to rewrite the term (19) as:

$$(19)=\mathbb{E}\left[\sum_{t\in[T]}\mathbb{I}\left(\mathcal{Q}_t,\mathcal{N}_t,\mathrm{A}_t^{\tilde{\pi}}\in\mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}}),\Delta(\mathrm{A}_t^{\tilde{\pi}},\mathrm{F}_t^{\tilde{\pi}})\geq M\right)\cdot\Delta(\mathrm{A}_t^{\tilde{\pi}},\mathrm{F}_t^{\tilde{\pi}})\right] \tag{22}$$

$$+\mathbb{E}\left[\sum_{t\in[T]}\mathbb{I}\left(\mathcal{Q}_t,\mathcal{N}_t,\mathrm{A}_t^{\tilde{\pi}}\in\mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}}),\Delta(\mathrm{A}_t^{\tilde{\pi}},\mathrm{F}_t^{\tilde{\pi}})< M\right)\cdot\Delta(\mathrm{A}_t^{\tilde{\pi}},\mathrm{F}_t^{\tilde{\pi}})\right]. \tag{23}$$

The second term of the above equation can be bounded using the fact that $\Delta(\mathrm{A}_t^{\tilde{\pi}},\mathrm{F}_t^{\tilde{\pi}})<M$, as follows:

$$(23)\leq\mathbb{E}\left[\sum_{t\in[T]}\mathbb{I}\left(\mathrm{A}_t^{\tilde{\pi}}\in\mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}}),\Delta(\mathrm{A}_t^{\tilde{\pi}},\mathrm{F}_t^{\tilde{\pi}})< M\right)\cdot\Delta(\mathrm{A}_t^{\tilde{\pi}},\mathrm{F}_t^{\tilde{\pi}})\right]\leq T\cdot M=T\sqrt{(48\cdot r\cdot k\ln T)/T}=\sqrt{48\cdot r\cdot k\cdot T\ln T}. \tag{24}$$

The first term can be bounded using Lemma 6, following the same procedure as in the distribution-dependent part:

$$\begin{aligned}(22)&=\mathbb{E}\left[\sum_{t\in[T]}\mathbb{I}\left(\mathcal{Q}_t,\mathcal{N}_t,\mathrm{A}_t^{\tilde{\pi}}\in\mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}}),\Delta(\mathrm{A}_t^{\tilde{\pi}},\mathrm{F}_t^{\tilde{\pi}})\geq M\right)\cdot\Delta(\mathrm{A}_t^{\tilde{\pi}},\mathrm{F}_t^{\tilde{\pi}})\right]\\&\leq\mathbb{E}\left[\sum_{t\in[T]}\sum_{i\in\mathrm{A}_t^{\tilde{\pi}}}\mathbb{I}\left(\mathcal{Q}_t,\mathcal{N}_t,\mathrm{A}_t^{\tilde{\pi}}\in\mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}}),\Delta(\mathrm{A}_t^{\tilde{\pi}},\mathrm{F}_t^{\tilde{\pi}})\geq M\right)\kappa_T(M,T_{i,t-1})\right]\\&\leq\mathbb{E}\left[\sum_{i\in\mathrm{A}}\sum_{s=0}^{T_{i,T}}\kappa_T(M,s)\right]\\&\leq 2k+\sum_{i\in\mathrm{A}}48\frac{r}{M}\ln T\\&=2k+\sqrt{48\cdot k\cdot r\cdot T\ln T},\end{aligned} \tag{25}$$

where the last inequality is derived similarly to (18). We conclude our proof by combining the above Equations (15), (16), (24) and (25) with Lemma 3:

$$\begin{aligned}
\rho \operatorname{Reg}^{\tilde{\pi}}(T) &\leq \frac{1}{1+\alpha\cdot\beta}\,\mathbb{E}\left[\sum_{t\in[T]}\Gamma_t^{\tilde{\pi}}\right] + \mathcal{O}(d_{\max}\cdot k)\\
&\leq \frac{1}{1+\alpha\cdot\beta}\Big(\frac{\pi^2 k}{3}\Delta_{\max} + \sqrt{48\cdot k\cdot r\cdot T\ln T} + 2k + \sqrt{48\cdot k\cdot r\cdot T\ln T}\Big) + \mathcal{O}(d_{\max}\cdot k)\\
&\leq \frac{14\sqrt{k\cdot r\cdot T\ln T}}{1+\alpha\cdot\beta} + \frac{k}{1+\alpha\cdot\beta}(2+\frac{\pi^2}{3}\Delta_{\max}) + \mathcal{O}(d_{\max}\cdot k).
\end{aligned}$$

□

**Lemma 6.** *For any $t\in[T]$, if the event $\{\mathcal{Q}_t, \mathcal{N}_t, \mathrm{A}_t^{\tilde{\pi}} \in \mathcal{S}_B(\mathrm{F}_t^{\tilde{\pi}}), \Delta(\mathrm{A}_t^{\tilde{\pi}}, \mathrm{F}_t^{\tilde{\pi}}) \geq M\}$ holds, then $\Delta(\mathrm{A}_t^{\tilde{\pi}}, \mathrm{F}_t^{\tilde{\pi}}) \leq \sum_{i\in \mathrm{A}_t^{\tilde{\pi}}}\kappa_T(M, T_{i,t-1})$.*

*Proof.* This proof resembles the proof of Lemma 5: From Equation (9) and the fact that $\Delta(\mathrm{A}_t^{\tilde{\pi}}, \mathrm{F}_t^{\tilde{\pi}}) \geq M$ we obtain: $\Delta(\mathrm{A}_t^{\tilde{\pi}}, \mathrm{F}_t^{\tilde{\pi}}) \leq 2\sum_{i\in\mathrm{A}_t^{\tilde{\pi}}}|\bar{\mu}_i - \mu_i| - M$. This can be rewritten as:

$$\Delta(\mathrm{A}_t^{\tilde{\pi}}, \mathrm{F}_t^{\tilde{\pi}}) \leq 2\sum_{i\in\mathrm{A}_t^{\tilde{\pi}}}|\bar{\mu}_i - \mu_i| - M \leq 2\sum_{i\in\mathrm{A}_t^{\tilde{\pi}}}(|\bar{\mu}_i - \mu_i| - \frac{M}{2|\mathrm{A}_t^{\tilde{\pi}}|}) \leq 2\sum_{i\in\mathrm{A}_t^{\tilde{\pi}}}\Big(|\bar{\mu}_i - \mu_i| - \frac{M}{2r}\Big),$$

where in the last inequality we use that $|\mathrm{A}_t^{\tilde{\pi}}| \leq r$, by definition of $r$. Then, using Equation (7), the expression above becomes:

$$\Delta(\mathrm{A}_t^{\tilde{\pi}}, \mathrm{F}_t^{\tilde{\pi}}) \leq 2\sum_{i\in\mathrm{A}_t^{\tilde{\pi}}}\Big(\min\{2\sqrt{\frac{3\ln t}{2T_{i,t-1}}}, 1\} - \frac{M}{2r}\Big) = \sum_{i\in\mathrm{A}_t^{\tilde{\pi}}}\Big(\min\{\sqrt{\frac{24\ln t}{T_{i,t-1}}}, 2\} - \frac{M}{r}\Big).$$

Finally, similarly to Lemma 5, we note that when $T_{i,t-1} \leq \ell_T(M)$, then $\min\{\sqrt{\frac{24\ln t}{T_{i,t-1}}}, 2\} - \frac{M}{r} \leq \kappa_T(M, T_{i,t-1})$. However, when $T_{i,t-1} > \ell_T(M)$ then $\min\{\sqrt{\frac{24\ln t}{T_{i,t-1}}}, 2\} \leq \sqrt{\frac{24\ln t}{T_{i,t-1}}} \leq \frac{M}{r}$, thus $\min\{\sqrt{\frac{24\ln t}{T_{i,t-1}}}, 2\} - \frac{M}{r} \leq 0 = \kappa_T(M, T_{i,t-1})$, which completes the proof.

□

## D. Experimental Setting

Our results are evaluated on two experimental settings, where the feasible sets are defined by matching and knapsack constraints. In both settings we use $k = 50$ arms and time horizon $T = 3000$. The results are averaged over 50 trials, which include the randomness in the selection of the constraints of feasible sets, as well as the realizations of the rewards and delays.

In the matching setting, the arms correspond to edges of a graph $G(V, E)$ with $|V| = 16$ nodes. The edges are selected uniformly at random among all possible edges, such that the resulting graph is connected. Each edge $e$ is associated with a mean reward $\mu_e$ selected uniformly in the interval $[1, 10]$ (all mean rewards are normalized afterwards) and a random delay $d_e$ selected uniformly from $\{1, ..., 5\}$. The reward realizations of any arm $e$ are independent Bernoulli random variables Bernoulli$(\mu_e)$, and the delay realizations are independent binomial random variables $B(n = 2d_e, p = 1/2)$. Algorithm 1 has access to an exact oracle for the underlying matching problem over the availability set at every round. The exact oracle deterministically returns the maximum-reward matching over the available set of arms, given any reward vector for the arms. In order to compute the regret of Algorithm 1 we compute an upper bound of the expected reward collected by the full-information optimal algorithm. This upper bound corresponds to an algorithm collecting at every round a reward equal to the solution of the following LP:

$$\begin{aligned} \text{maximize} \quad & \sum_{e=1}^{k} \mu_e x_e \\ \text{subject to} \quad & \sum_{x_e \text{ s.t. } e=(i,j)\in E} x_e \ \leq \ 1, \qquad \forall i \in V \\ & 0 \ \leq x_e \ \leq \frac{1}{d_e}, \qquad \forall e \in E. \end{aligned}$$

In the knapsack setting, each arm corresponds to one of $50$ items. Each item $i$ is associated with a mean reward $\mu_i$ selected uniformly in the interval $[1, 10]$ (again, all mean rewards are normalized afterwards) and a random delay $d_i$ selected uniformly from $\{1, \dots, 20\}$. Moreover, the capacity is $C = 20$ and each arm is associated with a random integer weight, selected uniformly in $\{1, \dots, \frac{20 \cdot C}{k}\}$. Again, the reward realizations of any arm $i$ are independent Bernoulli random variables Bernoulli$(\mu_i)$, and the delay realizations are independent binomial random variables $B(n = 2d_i, p = 1/2)$. Algorithm 1 has access to an exact oracle for the underlying discrete knapsack problem, which deterministically returns the maximum-reward subset of available arms, such that the capacity constraint is not violated, given any reward vector for the arms. The regret of algorithm 1 is computed against an upper bound of the expected reward collected by the full information optimal algorithm. This upper bound corresponds to the algorithm collecting at every round a reward equal to the solution of the fractional knapsack variant, where the weight of each arm is set to $\frac{1}{d_i}$.