Finding the Bandit in a Graph: Sequential Search-and-Stop

Pierre Perrault[1,2], Vianney Perchet[2,3], and Michal Valko[1]

[1]SequeL team, INRIA Lille - Nord Europe
[2]CMLA, ENS Paris-Saclay
[3]Criteo Research

Abstract

We consider the problem where an agent wants to find a hidden object that is randomly located in some vertex of a directed acyclic graph (DAG) according to a fixed but possibly unknown distribution. The agent can only examine vertices whose in-neighbors have already been examined. In scheduling theory, this problem is denoted by $1|prec|\sum w_j C_j$ (Graham et al., 1979). However, in this paper we address a *learning* setting where we allow the agent to stop before having found the object and restart searching on a new independent instance of the same problem. The goal is to maximize the total number of hidden objects found under a time constraint. The agent can thus skip an instance after realizing that it would spend too much time on it. Our contributions are both to the *search theory* and *multi-armed bandits*. If the distribution is known, we provide a quasi-optimal greedy strategy with the help of known computationally efficient algorithms for solving $1|prec|\sum w_j C_j$ under some assumption on the DAG. If the distribution is unknown, we show how to sequentially learn it and, at the same time, act near-optimally in order to collect as many hidden objects as possible. We provide an algorithm, prove theoretical guarantees, and empirically show that it outperforms the naïve baseline.

1 Introduction

We study the problem where an object, called hider, is randomly located in one vertex of a directed acyclic graph (DAG), and where an agent wants to find it by sequentially selecting vertices one by one, and examining them at a (possibly random) cost. The agent has a strong constraint: its search must respect *precedence constraints* imposed by the DAG, i.e., a vertex can be examined only if *all* its in-neighbors have already been examined. The goal of the agent is to minimize the expected total search cost incurred before finding the hider. This problem is a *single machine scheduling* problem (Lín, 2015), where a set of n jobs $[n] \triangleq \{1, \ldots, n\}$ have to be processed on a single machine that can process at most one job at a time. Once a job processing is started, it must continue without interruption until the processing is complete. Each job j has a cost $c_j > 0$, representing its processing time, and a weight $w_j \geq 0$, representing its importance; here, w_j is the probability that j contains the hider. The goal is to find a schedule that minimizes $\sum_{j=1}^{n} w_j C_j$, representing the expected cost suffered by the agent for finding the hider, where C_j is the completion time of job j.

The standard scheduling notation (Graham et al., 1979) denotes this problem as $1|prec|\sum w_j C_j$, and it was already shown to be NP-hard (Lawler, 1978; Lenstra and Rinnooy Kan, 1978). On the positive side, several polynomial-time α-approximations exist, depending on the assumption we take on the DAG (see e.g., the recent survey of Prot and Bellenguez-Morineau, 2017). For instance, the case of $\alpha = 2$ can be dealt without any additional assumption. On the other hand, there is an exact $\mathcal{O}(n \log n)$-time algorithm when the partially ordered set (poset) defined by the DAG is a series-parallel order (Lawler, 1978). More generally, when the poset has fractional dimension of at most f, there is a polynomial-time approximation with $\alpha = 2 - 2/f$ (Ambühl et al., 2011). In this work, we assume the DAG is such that an exact polynomial-time algorithm is available, for example, we can take two-dimensional poset (Ambühl and Mastrolilli, 2009).

The problem is also well known in *search theory* (Stone, 1976; Fokkink et al., 2016), one of the original disciplines of *operations research*. Here, the search space is a DAG. We thus fall within the network search setting (Kikuta and Ruckle, 1994; Gal, 2001; Evans and Bishop, 2013). When the DAG is an out-tree, the problem reduces to the *expanding search* problem introduced by Alpern and Lidbetter (2013).

The case of unknown distribution of the hider is usually studied within the field of *search games*, i.e., a zero-sum game where the agent picks the search, and plays against the hider, with search cost as payoff (Alpern and Gal, 2006; Alpern et al., 2013; Hohzaki, 2016). In our work, we deal with an unknown hider distribution by extending the stochastic setting to the sequential case, where at each round t, the agent faces a new, independent instance of the problem. The challenge is the need to learn the distribution through repeated interactions with the environment. Each instance, the agent has to perform a search based on the instances observed during the previous rounds. Furthermore, contrary to the typical search setting, the agent can additionally decide whether it wishes to abandon the search on the current instance and start a new one in the next round, even if the hider was not found. The goal of the agent is to collect as many hiders as possible, using a fixed budget B. This may be particularly useful, when the remaining vertices have large costs and it would not be cost-effective to examine them.

As a result, the hider may not be found in each round and the agent has to make a trade-off between exhaustive searches, which lead to a good estimation (*exploration*) and efficient searches, which leads to a good benefit/cost ratio (*exploitation*). The sequential *exploration-exploitation* trade-off is well studied in multi-armed bandits (MAB, Cesa-Bianchi and Lugosi, 2006) and has been applied to many fields including mechanism design (Mohri and Munoz, 2014), search advertising (Tran-Thanh et al., 2014) and personalized recommendation (Li et al., 2010). Our setting can thus be seen as an instance of stochastic combinatorial semi-bandits (Cesa-Bianchi and Lugosi, 2006, 2012; Gai et al., 2012; Gopalan et al., 2014; Kveton et al., 2015; Combes et al., 2015a; Wang and Chen, 2017; Valko, 2016). For this reason, we refer to a vertex $j \in [n]$ as an *arm*. We shall see, however, that this particular semi-bandit problem is challenging, first because the offline objective is not easy to optimize, and second, because it does not directly satisfy the standard key assumptions as *monotonicity* and *non-negativity* when using optimistic estimates. Moreover, due to the budget constraint, it is also an instance of *budgeted bandit*, also known as *bandits with knapsacks* (Badanidiyuru et al., 2013), in the case of single resource and infinite horizon. We thus evaluate the performance of a learning policy with the (classical) notion of *expected (cumulative budgeted) regret*. It measures the expected difference, in term of cumulative reward collected within the budget constraint B, between the learning policy and an *oracle* policy that knows a priori the exact parameters of the problem.

Motivations There are several motivations behind this setting. The *decision-theoretic troubleshooting* problem of giving a diagnosis for several devices having a malfunctioning component, and coming sequentially to the agent, is one example. Precedence constraints arise naturally in many *troubleshooting* applications, where there are restrictions imposed to the agent on the order of component tests, see e.g., Jensen et al., 2001. Moreover, allowing the agent to *stop* gives a new alternative to the so-called *service call* (Heckerman et al., 1995; Jensen et al., 2001) in order to deal with non-cost-effective vertices: Instead of giving a high cost to an extra action that will automatically find the fault in the device, we give it a zero cost, but do not reward such diagnostic failure. This way, we do not need to estimate any *call-service* cost, which could be useful, for example, when a new device is sent to the user if the diagnostic fails, with a cost that depends on a disutility for the user: loss of personal data, device reconfiguration, etc. Maximizing the number of hiders found is then analogous to maximizing the number of successful diagnoses.

Another motivation comes from online advertisement. There are several different actions that might generate a conversion from a user, such as sending one or several emails, displaying one or several ads on a website, buying keywords on search engines, etc. We assume that some precedence constraints are imposed between actions and that a conversion will occur if some sequence of actions is made, for instance, first, display an ad, then send the first email, and finally the second one. As a consequence, the conversion is "hidden", the precedence constraints restrict our access to it, and the agent aims at finding it. However, for some users, finding the correct sequence might be too expensive and it might be more interesting to abandon that specific user to focus on more promising ones.

Related settings Finally, there are several settings related to ours. One of them is stochastic probing (Gupta and Nagarajan, 2013), which differs in the fact that each arm can contain a hider, independently from each other. Another one is the framework of optimal discovery (Bubeck et al., 2013), widely studied in machine learning.

Our contributions One of our main contributions is a stationary *offline policy* (i.e., an algorithm that solves the problem when the distribution is known), for which we prove the approximation guarantees and adapt it in order to fit the online problem. In particular, we prove it is quasi-optimal, and use the exact algorithm for $1|prec|\sum w_j C_j$ to prove its computational efficiency. Next, we provide a solution when the distribution is unknown to the agent, based on combinatorial upper confidence bounds (`CUCB`) algorithm (Chen et al., 2016), and `UCB`-variance (`UCB-V`) of Audibert et al. (2009). Dealing with variance estimates allows us to sharp the bound on the expected cumulative regret, compared to the simple use of `CUCB`. We also propose a new method (that can be of independent interest) to avoid the usual $1/c_{\min}^2$ term in the expected regret bound (Tran-Thanh et al., 2012; Ding et al., 2013; Xia et al., 2016a,b; Watanabe et al., 2017), where $c_{\min}$ is the minimal expected cost paid over a single round.

2 Background

In this paper, we typeset vectors in bold and indicate components with indices, i.e., $\mathbf{a} = (a_1, \dots, a_n) \in \mathbb{R}^n$. Furthermore, we indicate randomness by underlining the relevant symbols (Hemelrijk, 1966).

We formalize in this section the setting we consider. We denote a finite DAG by $\mathcal{G} \triangleq ([n], \mathcal{E})$, where $[n]$ is its set of vertices, or arms, and $\mathcal{E}$ is its set of directed edges. For more generality, we assume arm costs are random and mutually independent. We denote $\underline{c}_j \in [0,1]$, with expectation $c_j \triangleq \mathbb{E}\left[\underline{c}_j\right] > 0$, the cost of arm j. We thus have $\mathbf{c} = \mathbb{E}\left[\underline{\mathbf{c}}\right] \in (0,1]^n$. We also assume that one specific vertex, called *hider*, is chosen at random, independently from $\underline{\mathbf{c}}$, accordingly to some fixed categorical (or multivariate Bernoulli) distribution parametrized by vector $\mathbf{w}$ satisfying[1] $\sum_{i=1}^n w_i = 1$ and $w_i \in [0,1]$. We remind that $\underline{\mathbf{w}} \sim \text{Bernoulli}(\mathbf{w})$ if, given $i \in [n]$ and with probability w_i, $\underline{w}_i = 1$ and $\underline{w}_j = 0$ for all $j \neq i$. Let $\mathcal{D}$ denote the joint distribution of $(\underline{\mathbf{c}}, \underline{\mathbf{w}})$.

For an (ordered) subset A of $[n]$, we denote by A^c, the complementary of A in $[n]$, and $|A|$ its cardinality. Moreover, if $\mathbf{x} \in \mathbb{R}^n$, we let $x_A \triangleq \sum_{i \in A} x_i$. Let $\mathcal{G}\langle A \rangle$ be the sub-DAG in $\mathcal{G}$ induced by $A \subset [n]$, i.e. the DAG with A as vertex set, and with (i,j) is an arc in $\mathcal{G}\langle A \rangle$ if and only if $(i,j) \in \mathcal{E}$. We call *support* of an ordered arm set $\mathbf{a} = (a_1, \dots, a_k)$ the corresponding non-ordered set. If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we write $\mathbf{x} \geq \mathbf{y}$ (resp., $\mathbf{x} \leq \mathbf{y}$) if $\mathbf{x} - \mathbf{y} \in \mathbb{R}_+^n$ (resp., $\mathbf{y} - \mathbf{x} \in \mathbb{R}_+^n$). We let $\mathbf{a}[j] \triangleq (a_1, a_2, \dots, a_j)$ for $j \leq |\mathbf{a}|$. In addition, we let $\mathbf{a}[\underline{\mathbf{w}}] \triangleq \mathbf{a}[j]$ if there is j such that $\underline{w}_{a_j} = 1$, and $\mathbf{a}[\underline{\mathbf{w}}] \triangleq \mathbf{a}$ otherwise. For two disjoint ordered arm sets $\mathbf{a}$ and $\mathbf{b}$, we let $\mathbf{ab} = (a_1, a_2, \dots, a_{|a|}, b_1, b_2, \dots, b_{|b|})$ be the concatenation of $\mathbf{a}$ and $\mathbf{b}$.

We assume that $\mathcal{G}$ allows a polynomial-time algorithm (w.r.t. n), denoted `Schedul`, for the problem of $1|prec|\sum w_j C_j$ with precedence constraints given by $\mathcal{G}$, i.e., for minimizing

$$d(\mathbf{s}) \triangleq \sum_{i=1}^{|\mathbf{s}|} w_{s_i} c_{\mathbf{s}[i]} = \sum_{i=1}^{|\mathbf{s}|} w_{s_i} \sum_{j=1}^{i} c_{s_j}$$

over linear extensions[2] $\mathbf{s} = (s_1, \dots, s_n)$ of the poset defined by $\mathcal{G}$ (that we call $\mathcal{G}$-linear extensions). $d(\mathbf{s})$ represents the expected cost $\mathbb{E}\left[\underline{c}_{\mathbf{s}[\underline{\mathbf{w}}]}\right]$ to pay for finding the hider, by searching arm s_1 first and paying $\underline{c}_{s_1}$, then s_2 by paying $\underline{c}_{s_2}$ if $\underline{w}_{s_1} = 0$, and so on until the hider is found (i.e., the last arm x searched is such that $\underline{w}_x = 1$).

We define a *search* in $\mathcal{G}$ as an ordering $\mathbf{s} = (s_1, \dots, s_k)$ of different arms such that for all $i \in [k]$, predecessors of s_i in $\mathcal{G}$ are included in $\{s_1, \dots, s_{i-1}\}$ (a search is a prefix of a $\mathcal{G}$-linear extension). We denote $\mathcal{S}_\mathcal{G}$ (or simply $\mathcal{S}$) the set of searches in $\mathcal{G}$. Support of a search is called *initial set*.

2.1 Protocol

The search problem we focus on is *sequential*. We consider an *agent*, also called a *learning algorithm* or a *policy* that knows $\mathcal{G}$ but that does not know $\mathcal{D}$. At each round t, an independent sample $(\underline{\mathbf{c}}^t, \underline{\mathbf{w}}^t)$ is drawn from $\mathcal{D}$. The aim of the agent is to search the hider (i.e., the arm x such that $\underline{w}_x^t = 1$) by constructing a *search* on $\mathcal{G}$. Since the hider may be located at some arm that doesn't belong to the search, it is not necessarily found over each round.

The search to be used by the agent can be computed based on all its previous observations, i.e., all the costs of explored vertices (and only those) and all the locations where the hider has been found or not. Obviously, the search cannot depend on non-observed quantities. For example, the agent may estimate $\mathbf{w}$ and $\mathbf{c}$ in order to choose the search accordingly. Each time an arm j is searched, the feedback $\underline{w}_j^t$ and $\underline{c}_j^t$ is given to the agent. Since several arms can be searched over one round, this problem falls into the family of *stochastic combinatorial semi-bandits*. The agent can keep searching until its budget, B, runs out. B is a positive number and does not need to be known to the agent in advance. The goal of the agent is to maximize the overall number of hiders found under the budget constraint.

The setting described above allows the agent to modify its behavior depending on the feedback it received during the current round. However, by independence assumption between random variables, the only feedback susceptible to modify the search the agent chose at the beginning of a round t is the observation of $\underline{w}_i^t = 1$ for some arm i. Even if nothing prevents the agent from continuing "searching" some arms after having seen such an event, it would not increase the number of hiders found (there is no more hider to find), while this would still decrease the remaining budget, and therefore it would have a pure exploratory purpose. Knowing this, an oracle policy that knows exactly $\mathcal{D}$ thus *selects* a search $\mathbf{s}$ at the beginning of round t, and then *performs* the search that follows $\mathbf{s}$ until either $\underline{w}_i^t = 1$ is observed or $\mathbf{s}$ is exhausted (i.e., no arms are left in $\mathbf{s}$). Therefore, the performed search is in fact $\mathbf{s}[\underline{\mathbf{w}}^t]$. We thus restrict ourselves to an agent that selects a search $\mathbf{s}$ at the beginning of each round t and then performs $\mathbf{s}[\underline{\mathbf{w}}^t]$ over this round. As a consequence, the selected search $\mathbf{s}$ is computed based on observations collected during previous rounds $t-1, t-2, \dots$, denoted $\mathcal{H}_t$, that we refer to as *history*.

Following Stone (1976), we refer to our problem as *sequential search-and-stop*. We now detail the overall objective for this problem: the agent wants to follow a policy π, that selects a search $\underline{\mathbf{s}}^t$ at round t (this choice can be random as it may depend on the past observations $\mathcal{H}_t$, as well as possible randomness from the algorithm),

[1] i.e., $\mathbf{w}$ belongs to the simplex of $\mathbb{R}^n$

[2] A linear extension of a poset is a total ordering consistent with the poset, i.e., if a is before b in the poset, then it is in a linear extension.

while maximizing the expected overall reward

$$F_B(\pi) \triangleq \mathbb{E}\left[\sum_{t=1}^{\tau_B-1} \underline{w}^t_{\mathbf{s}^t[\underline{\mathbf{w}}^t]}\right] = \mathbb{E}\left[\sum_{t=1}^{\tau_B-1} \sum_{i\in\mathbf{s}^t[\underline{\mathbf{w}}^t]} \underline{w}^t_i\right],$$

where τ_B is the random round at which the remaining budget becomes negative; i.e., if $\underline{B}^t \triangleq B - \sum_{u=1}^t \underline{c}^u_{\mathbf{s}^u[\underline{\mathbf{w}}^t]}$, $\underline{B}^{\tau_B-1} \geq 0$ and $\underline{B}^{\tau_B} < 0$. We evaluate the performance of a policy using the *expected (cumulative budgeted) regret* with respect to $F_B^\star$, the maximum value of F_B (among all possible oracle policies that know $\mathcal{D}$ and B), defined as

$$R_B(\pi) \triangleq F_B^\star - F_B(\pi).$$

Example 1. *One may wonder if there exist cases where it is interesting for the agent to stop the search earlier. Consider for instance the simplest non-trivial case with two arms and no precedence constraint. The costs are deterministically chosen to be ε and 1 and the location of the hider is chosen uniformly at random. An optimal search will always sample first the arm with $\varepsilon < 1$ cost. If it also samples the other, then the hider will be found with an expected cost of $\varepsilon + \frac{1}{2}$. However, if the agent always stops the search after the first arm, and reinitializes on a new instance by doing the same, the overall cost to find one hider is*

$$\sum_{t=1}^{\infty}\left(\frac{1}{2}\right)^t t\varepsilon = 2\varepsilon < \varepsilon + \frac{1}{2}, \qquad \textit{for } \varepsilon < \frac{1}{2}.$$

Therefore, stopping searches, even if the location of the hider is known, can be better than always trying to find it.

3 Offline oracle

In this section, we provide an algorithm for sequential search-and-stop when parameters $\mathbf{w}$ and $\mathbf{c}$ are given to the agent. We show that a simple stationary policy (i.e., the same search $\mathbf{s}^\star$ is selected at each round) can obtain almost the same expected overall reward as $F_B^\star$. We will denote by `Oracle`, an algorithm that takes $\mathbf{w}$, $\mathbf{c}$, and $\mathcal{G}$ as input and outputs $\mathbf{s}^\star$. This *offline* oracle will eventually be used by the agent for the *online* problem, i.e., when parameters are unknown. Indeed, at round t, the agent can approximate $\mathbf{s}^\star$ by the output $\mathbf{s}^t$ of `Oracle` $(\widetilde{\mathbf{w}}^t, \widetilde{\mathbf{c}}^t, \mathcal{G})$, where $\widetilde{\mathbf{w}}^t, \widetilde{\mathbf{c}}^t$ can be any guesses/estimates of the true parameters. Importantly, depending on the policy followed by the agent, $\widetilde{\mathbf{w}}^t$ may not stay in the simplex anymore. We will thus build `Oracle` such that an "acceptable" output is given for any input $(\widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}) \in \left(\mathbb{R}_+^n\right)^2$.

3.1 Objective design

A standard paradigm for designing a stationary approximation of the offline problem in budgeted multi-armed bandits is the following: $\mathbf{s}^\star$ has to minimize the ratio between the expected cost paid and the expected reward gain, over a single round, selecting $\mathbf{s}^\star$. We thus define, for $\mathbf{s} \in \mathcal{S}$,

$$J(\mathbf{s}) \triangleq \frac{d(\mathbf{s}) + (1 - w_{\mathbf{s}})c_{\mathbf{s}}}{w_{\mathbf{s}}} = \sum_{i=1}^{|\mathbf{s}|} \frac{c_{s_i}\left(1 - w_{\mathbf{s}[i-1]}\right)}{w_{\mathbf{s}}},$$

that is equal to $\mathbb{E}\left[\underline{c}_{\mathbf{s}[\underline{\mathbf{w}}]}\right]\mathbb{E}\left[\underline{w}_{\mathbf{s}[\underline{\mathbf{w}}]}\right]^{-1}$. Notice that we allow J to be equal to $+\infty$ (when $w_{\mathbf{s}} = 0$). We use the convention $J(\emptyset) = +\infty$, because there is no interest in choosing an empty search for a round. We define the optimal values of J on $\mathcal{S}$ as

$$J^\star \triangleq \min_{\mathbf{s}\in\mathcal{S}} J(\mathbf{s}), \quad \mathcal{S}^\star \triangleq \operatorname{argmin}_{\mathbf{s}\in\mathcal{S}} J(\mathbf{s}).$$

In Proposition 1, we provide guarantees on the stationary policy.

Proposition 1. *If $\pi^\star$ is the offline policy selecting $\mathbf{s}^\star \in \mathcal{S}^\star$ at each round t, then*

$$\frac{B-n}{J^\star} \leq F_B(\pi^\star) \leq F_B^\star \leq \frac{B+n}{J^\star}.$$

A proof is given in Appendix B and follows the one provided for Lemma 1 of Xia et al. (2016b). Intuitively, Proposition 1 states that the optimal overall expected reward that can be gained (i.e., the maximum expected number of hiders found) is approximately $\frac{B}{J^\star}$ (we assume that $B \gg n$). This is quite intuitive, since this quantity is actually the ratio between the overall budget and the minimum expected cost paid to find a *single* hider. Indeed, one can consider the related problem of minimizing the overall expected cost paid, over several rounds, to find a single hider. It can be expressed as an infinite-time horizon Markov decision process (MDP) with action space $\mathcal{S}$

and two states: whether the hider is found (which is the terminal state) or not. The goal is to choose a strategy $\mathbf{s}^1, \mathbf{s}^2, \ldots, \mathbf{s}^t, \ldots$, minimizing

$$\begin{aligned}\mathcal{J}\left(\mathbf{s}^1, \mathbf{s}^2, \ldots\right) &\triangleq \mathbb{E}\left[\sum_{t=1}^{\tau} c^t_{\mathbf{s}^t[\mathbf{w}^t]}\right] \\ &= \sum_{t=1}^{\infty}\left(w_{\mathbf{s}^t}\left(c_{\mathbf{s}^1} + \cdots + c_{\mathbf{s}^{t-1}}\right) + d\left(\mathbf{s}^t\right)\right)\prod_{u=1}^{t-1}\left(1 - w_{\mathbf{s}^u}\right),\end{aligned}$$

where the stopping time τ is the first round at which the hider is found. The Bellman equation is

$$\mathcal{J}\left(\mathbf{s}^1, \mathbf{s}^2, \mathbf{s}^3, \ldots\right) = d(\mathbf{s}^1) + (1 - w_{\mathbf{s}^1})\left(c_{\mathbf{s}^1} + \mathcal{J}(\mathbf{s}^2, \mathbf{s}^3, \ldots)\right),$$

from which we can deduce that there exists an optimal *stationary* strategy (Sutton and Barto, 1998) such that $\mathbf{s}^t = \mathbf{s}$ for all $t \in \mathbb{N}^*$. Therefore, we can minimize $\mathcal{J}\left(\mathbf{s}, \mathbf{s}, \ldots\right) = J(\mathbf{s})$ that gives the optimal value of $J^\star$.

As we already mentioned, `Oracle` aims at taking inputs $(\widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}) \in \left(\mathbb{R}^n_+\right)^2$. The first straightforward way to do is to consider

$$J\left(\mathbf{s}; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right) \triangleq \sum_{i=1}^{|\mathbf{s}|} \frac{\widetilde{c}_{s_i}\left(1 - \widetilde{w}_{\mathbf{s}[i-1]}\right)}{\widetilde{w}_{\mathbf{s}}}.$$

However, notice that with the definition above, $J\left(\cdot; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)$ could output negative values (if $\widetilde{w}_{[n]} > 1$), which is not desired, because the agent would then be enticed to search arms with a high cost. We thus need to design a non-negative extension of J to $(\widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}) \in \left(\mathbb{R}^n_+\right)^2$. One way is to replace $\left(1 - \widetilde{w}_{\mathbf{s}[i-1]}\right)$ by $\widetilde{w}_{(\mathbf{s}[i-1])^c}$, another is to consider $J\left(\mathbf{s}; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)^+$, where $x^+ \triangleq \max\{0, x\}$. There is a significant advantage of considering the second way, even if it is less natural than the first one, which is that[3] for $(\widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}) \in \left(\mathbb{R}^n_+\right)^2$,

$$J\left(\mathbf{s}; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)^+ \leq J\left(\mathbf{s}; \mathbf{w}, \mathbf{c}\right) = J(\mathbf{s}),$$

if $\widetilde{\mathbf{w}} \geq \mathbf{w}$ and $\widetilde{\mathbf{c}} \leq \mathbf{c}$. This property is known to be useful for analysis of many stochastic combinatorial semi-bandit algorithms (see e.g., Chen et al., 2016). Thus, we choose for `Oracle` the minimization of the surrogate $J\left(\cdot; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)^+$.

3.2 Algorithm and theoretical guarantees

We now provide `Oracle` in Algorithm 1 and claim that it minimizes $J\left(\cdot; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)^+$ over $\mathcal{S}$ in Theorem 1. Notice that `Oracle` needs to call the polynomial-time algorithm `Schedul` $(\widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}, \mathcal{G})$, that minimizes the objective function

$$d\left(\mathbf{s}; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right) \triangleq \sum_{i=1}^{|\mathbf{s}|} \widetilde{w}_{s_i}\widetilde{c}_{\mathbf{s}[i]}$$

over $\mathcal{G}$-linear extensions $\mathbf{s}$. Then, Algorithm 1 only computes the maximum value index of a list of size n that takes linear time. To give an intuition, $\mathbf{s}^\star$ follows the ordering given by `Schedul` $(\widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}, \mathcal{G})$, and stops at some point when it becomes more interesting to start a fresh new instance.

Algorithm 1 `Oracle`

Input: $\widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}$ and $\mathcal{G}$.
$\mathbf{s} \triangleq$ `Schedul` $(\widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}, \mathcal{G})$.
$i^\star \triangleq \operatorname{argmin}_{i \in [n]} J\left(\mathbf{s}[i]; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)^+$ (ties broken arbitrarily).
Output: the search $\mathbf{s}^\star \triangleq \mathbf{s}[i^\star]$.

Theorem 1. *For every* $(\widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}) \in \left(\mathbb{R}^n_+\right)^2$, *Algorithm 1 outputs a search minimizing* $J\left(\cdot; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)^+$ *over* $\mathcal{S}$.

We provide a proof of Theorem 1 in Appendix A. It mixes known concepts of scheduling theory, such as Sidney decompositions (Sidney, 1975), with new results about our objective function, such as the following property: If $\mathbf{xy}, \mathbf{xyz} \in \mathcal{S}$ with $\rho(\mathbf{z}) \geq \rho(\mathbf{y})$, then

$$J^+(\mathbf{xy}) \geq \min\left\{J^+(\mathbf{x}), J^+(\mathbf{xyz})\right\}.$$

[3]Notice this is not exactly a monotonicity property stated here, because we compare to a single point $(\mathbf{w}, \mathbf{c})$.

4 Online search-and-stop

We consider in this section the additional challenge where the distribution $\mathcal{D}$ is unknown and the agent must learn it, while minimizing $R_B(\pi)$ over sampling policies π, where B is a fixed budget. Recall that a policy π selects a search $\underline{\mathbf{s}}^t$ at the beginning of round t, using previous observations $\mathcal{H}_t$, and then performs the search $\underline{\mathbf{s}}^t[\underline{\mathbf{w}}^t]$ over the round. We will consider the problem as a variant of stochastic combinatorial bandits (Gai et al., 2012). The feedback received by an agent at round t is random, as usual in bandits, because it depends on $\underline{\mathbf{s}}^t$. However, in our case, it also depends on $\underline{\mathbf{w}}^t$, and thus it is not measurable w.r.t. $\mathcal{H}_t$. More precisely, $(\underline{w}_i^t, \underline{c}_i^t)$ is observed only for arms $i \in \underline{\mathbf{s}}^t[\underline{\mathbf{w}}^t]$. Notice that since $\underline{\mathbf{w}}^t$ is a one-hot vector, the agent can always deduce the value of $\underline{w}_i^t$ for all $i \in \underline{\mathbf{s}}^t$. As a consequence, we will maintain two types of counters for all arms $i \in [n]$ and all $t \geq 1$:

$$
\begin{aligned}
\underline{T}_{i,\mathbf{w}}^t &\triangleq \sum_{u=1}^{t} \mathbb{I}\{i \in \underline{\mathbf{s}}^u\}, \\
\underline{T}_{i,\mathbf{c}}^t &\triangleq \sum_{u=1}^{t} \mathbb{I}\{i \in \underline{\mathbf{s}}^u[\underline{\mathbf{w}}^u]\}.
\end{aligned} \tag{1}
$$

With the convention $1/0 = +\infty$, the corresponding empirical averages are then defined as

$$
\begin{aligned}
\overline{\underline{w}}_i^t &\triangleq \sum_{u=1}^{t} \frac{\mathbb{I}\{i \in \underline{\mathbf{s}}^u\}\underline{w}_i^u}{\underline{T}_{i,\mathbf{w}}^t}, \\
\overline{\underline{c}}_i^t &\triangleq \sum_{u=1}^{t} \frac{\mathbb{I}\{i \in \underline{\mathbf{s}}^u[\underline{\mathbf{w}}^u]\}\underline{c}_i^u}{\underline{T}_{i,\mathbf{c}}^t}.
\end{aligned} \tag{2}
$$

We propose an approach, similar to `UCB-V` of Audibert et al. (2009), based on `CUCB` by Chen et al. (2016), called `CUCBV`, that uses a variance estimation of $\mathbf{w}$ in addition to the empirical average. Notice that the variance of $\underline{w}_i$ for an arm i is $w_i(1 - w_i)$. In addition, since $\underline{w}_i$ is binary, the empirical variance of $\underline{w}_i$ after t rounds is $\overline{\underline{w}}_i^t(1 - \overline{\underline{w}}_i^t)$. For every round t and every edge $i \in [n]$, we define

$$
\dot{\underline{c}}_i^t \triangleq \left(\overline{\underline{c}}_i^{t-1} - \sqrt{\frac{0.5\zeta \log(t)}{\underline{T}_{i,\mathbf{c}}^{t-1}}}\right)^+, \quad \text{and}
$$

$$
\ddot{\underline{w}}_i^t \triangleq \min\left\{\overline{\underline{w}}_i^{t-1} + \sqrt{\frac{2\zeta \overline{\underline{w}}_i^{t-1}(1 - \overline{\underline{w}}_i^{t-1}) \log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}} + \frac{3\zeta \log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}, 1\right\},
$$

where we choose the exploration factor to be $\zeta \triangleq 1.2$ (notice that we could take any $\zeta > 1$ as shown by Audibert et al., 2009). We provide the policy $\pi_{\texttt{CUCBV}}$ that we consider in Algorithm 2.

Algorithm 2 Combinatorial upper confidence bounds with variance estimates (`CUCBV`) for sequential search-and-stop

Input: $\mathcal{G}$.
Initialization: Set all counters $\underline{T}_{i,\mathbf{w}}^0$ and $\underline{T}_{i,\mathbf{c}}^0$ to 0 and empirical averages $\overline{\underline{w}}_i^0$ and $\overline{\underline{c}}_i^0$, for all $i \in [n]$.
for $t = 1..\infty$ **do**
  select $\underline{\mathbf{s}}^t$ given by `Oracle` $\left(\ddot{\underline{\mathbf{w}}}^t, \dot{\underline{\mathbf{c}}}^t, \mathcal{G}\right)$.
  perform $\underline{\mathbf{s}}^t[\underline{\mathbf{w}}^t]$.
  collect feedback and update counters and empirical average according to (1) and (2).
end for

4.1 Analysis

Notice that since an arm $i \in \underline{\mathbf{s}}^t$ is pulled (and thus $\underline{c}_i^t$ is revealed to the agent) with probability $1 - w^t_{\underline{\mathbf{s}}^t[i-1]}$ over round t, we fall into the setting of *probabilistically triggered arms* w.r.t. costs, described by Chen et al. (2016) and Wang and Chen (2017). Thus we could rely on these prior results. However, the main difficulty in our setting is that we also need to deal with probabilities $\underline{w}_i^t$, that the agent actually observes for every arm i in $\underline{\mathbf{s}}^t$, either because it actually pulls arm i, or because it deduces the value from other pulls of round t. In particular, if we follow the analysis of Chen et al. (2016) and Wang and Chen (2017), the double sum in the definition of J leads to expected regret bound that is quite large. Indeed, assuming that all costs are deterministically equal to 1, if we suffer an error of δ when approximating each w_i, then the global error can be as large as $\sum_{i=1}^{n}\sum_{j=1}^{i-1}\delta = \mathcal{O}(n^2\delta)$, contrary to just $\mathcal{O}(n\delta)$ for the approximation error w.r.t. costs, that is more common in classical combinatorial semi-bandits. Thus, here we rather combine this work with the variance estimates of w_i^t. Often, this does not provide a significant improvement over UCB in terms of expected cumulative regret (otherwise we could do the same for the costs), but since in our case, the variance is of order $1/n$, the gain is non-negligible.[4] We let $c_{\min} > 0$ be any deterministic lower bound on the set $\left\{c_{\underline{\mathcal{S}}^u[\underline{\mathbf{w}}^u]}, u \geq 1\right\}$. Furthermore, we let

$$T_B \triangleq \lceil 2B/c_{\min} \rceil$$

and for any search $\mathbf{s}$, we define the *gap* of $\mathbf{s}$ as

$$\begin{aligned}\Delta(\mathbf{s}) &\triangleq w_{\mathbf{s}} \left(\frac{J(\mathbf{s})}{J^\star} - 1 \right) \\ &= \frac{1}{J^\star} \sum_{i=1}^{|\mathbf{s}|} c_{s_i} \left(1 - \sum_{j=1}^{i-1} w_{s_j} \right) - \sum_{i=1}^{|\mathbf{s}|} w_{s_i} \geq 0,\end{aligned}$$

that represents the *local regret* of selecting a sub-optimal search $\mathbf{s}$ at some round. In addition, for each arm $i \in [n]$, we define

$$\Delta_{i,\min} \triangleq \inf_{\mathbf{s} \notin \mathcal{S}^\star:\ i \in \mathbf{s}} \Delta(\mathbf{s}) > 0.$$

We provide bounds for the expected cumulative regret of $\pi_{\texttt{CUCBV}}$ in Theorem 2. The first bound is $\mathcal{D}$-dependent, and is characterized by w_i and $\Delta_{i,\min}, i \in [n]$. Its main term scales logarithmically w.r.t. B. The second bound is true for any value of w_i and $\Delta_{i,\min}$. We only provide the leading term in B, neglecting the other ones and removing the universal multiplicative constants. Explicit statements can be found in Appendix C.

Theorem 2. *The expected cumulative regret of* CUCBV *is bounded as*

$$R_B(\pi_{\texttt{CUCBV}}) \lesssim n \left(1 + \frac{n}{J^\star}\right)^2 \sum_{i \in [n]} \frac{w_i}{\Delta_{i,\min}} \log T_B$$

$$\text{and} \quad R_B(\pi_{\texttt{CUCBV}}) \lesssim \sqrt{n} \left(1 + \frac{n}{J^\star}\right) \sqrt{T_B \log T_B}.$$

The proof is in Appendix C. Recall the main challenge comes from the estimation of $\mathbf{w}$ and not from $\mathbf{c}$. Our analysis uses *triggering probability groups* and the *reverse amortization trick* of Wang and Chen (2017) for dealing with costs. However, for hider probabilities, only the second trick is necessary.[5] We use it not only to deal with the slowly concentrating confidence term for the estimates of each arm i, but also to completely amortize the additional fast-rate confidence term due to variance estimation coming from the use of Bernstein's inequality. However, the analysis of Wang and Chen (2017) only considers the deterministic horizon. In our case, we need to deal with a *random-time* horizon. For that, notice that their regret upper bounds that hold in expectation are obtained by splitting the expectation into two parts. The first part is filtered with a high-probability event on which the regret grows as the logarithm of the random horizon and the second one filtered with a low-probability event, on which we bound the regret by a constant. Since the log function is concave, we can upper bound the expected regret by a term growing as the logarithm of the expectation of the random horizon, with Jensen's inequality. Finally, we upper bound the expectation of the random horizon to get the rate of $\log(T_B)$.

4.2 Tightness of our regret bounds

Since we succeeded to reduce the dependence on n in the expected regret with confidence bounds based on variance estimates, we can now ask whether this dependence in Theorem 2 is tight. We stress that our solution to sequential search-and-stop is *computationally efficient*. In particular, *both* the offline oracle optimization and the computation of the optimistic search $\mathbf{s}^t$ in the online part *are tractable*.

Whenever rewards are not arbitrary correlated (as is the case in our setting), we can potentially exploit these correlations in order to reduce the regret's dependence on n even further. This could be done by choosing a tighter confidence region such as a confidence ellipsoid (Degenne and Perchet, 2016), or a KL-confidence ball (Combes et al., 2015b), instead of coordinate-wise confidence intervals. Unfortunately, these do not lead to computationally efficient algorithms. Notice that given an *infinite* computational power, our dependence on n is not tight. In particular, there is an extra $\sqrt{n}$ factor in our gap-free bound (see Theorem 3). It is an open question whether a better *efficient* policy exists.

To show that we are only a $\sqrt{n}$ factor away, in the following theorem we provide a class of sequential search-and-stop problems (parametrized by n and B) on which the regret bound provided in Theorem 2 is tight up to a $\sqrt{n}$ factor (and a logarithmic one).

[4]The error δ is thus scaled by the standard deviation, of order $1/\sqrt{n}$, giving a global error of $\mathcal{O}(n^{1.5}\delta)$. We therefore recover the factor $n^{1.5}$ given in Theorem 2.

[5]When we select search $\mathbf{s}$, we all feedback $\underline{w}_i, i \in \mathbf{s}$ is received with probability 1, so *triggering probability groups* are not useful.

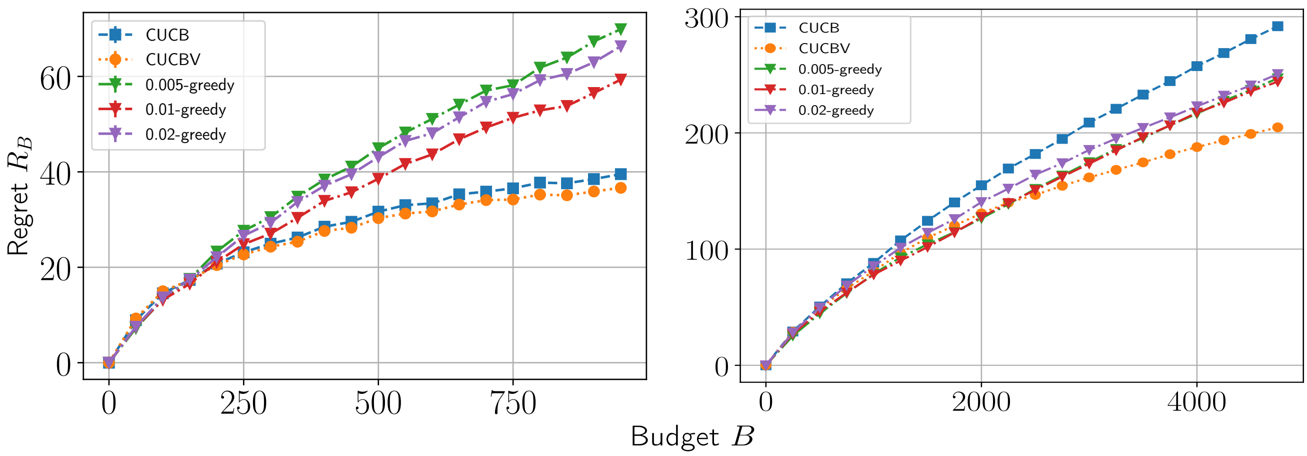

Figure 2: Regret for sequential search-and-stop. **Left:** $n = 10$. **Right:** $n = 20$.

Theorem 3. *For simplicity, let us assume that n is even and that B is a multiple of n. For any optimal online policy π, there is a sequential search-and-stop problem with n arms and budget B such that*

$$-4 + \frac{1}{28}\sqrt{\frac{B}{n}} \leq R_B(\pi) \lesssim \sqrt{B \log\left(\frac{B}{n}\right)}.$$

For the proof, we consider a DAG composed of two disjoint paths (Figure 1), with all costs deterministically set to 1 and with the hider located either at $a_{\frac{n}{2}}$ or $b_{\frac{n}{2}}$. This information is given to the agent. We then reduced this setting to a two-arm bandit over at least B/n rounds. The complete proof is in Appendix E.

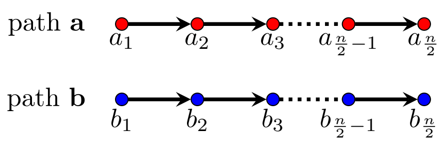

Figure 1: The DAG considered in Theorem 3

Notice that bounds provided in Theorem 3 *decrease* with n. This is because, in the sequential search-and-stop problem, the increasing dependence on n is counterbalanced by the fact that the number of rounds is of order B/n, and that $J^\star$ is of order n.

5 Experiments

In this section, we present experiments for sequential search-and-stop. We compare our `CUCBV` with two baselines. The first one is `CUCB` (Kveton et al., 2015), i.e., the selected search is given by `Oracle` $\left(\underline{\dot{\mathbf{w}}}^t, \underline{\dot{\mathbf{c}}}^t, \mathcal{G}\right)$, where

$$\forall i \in [n],\ \underline{\dot{w}}_i^t \triangleq \min\left\{\overline{w}_i^{t-1} + \sqrt{\frac{0.5\zeta\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}}, 1\right\}.$$

The second is known as ε-greedy, that acts greedily with probability $1-\varepsilon$ (exploitation), meaning that the agent selects the search given by `Oracle` $\left(\overline{\mathbf{w}}^t, \underline{\overline{\mathbf{c}}}^t, \mathcal{G}\right)$, and with probability ε (exploration), the agent selects a random search of size n (we uniformly select the next arm among available arms to continue the search until the hider is found). Notice that in the exploration step, we could potentially choose a *non-complete* search, by, for example, also taking a random stopping time. We chose not to do that since the ε-greedy would explore less this way in the exploration step. In experiments, we vary ε as 0.005, 0.01, and 0.02. We run simulations for all the algorithms without precedence constraints, i.e., when the DAG is an edgeless graph (or, equivalently, a dummy 'root' node is connected to everything). Notice that in this case, a search can be any ordered subset of arms (thus, the set of possible searches is of cardinality $n!$). This restriction does not remove complexity from the online problem, but rather from the offline one, so even in that case, the online problem is challenging. We take parameters $\mathbf{w}$ and $\mathbf{c}$ chosen uniformly at random in the simplex and $[0,1]^n$ conditionally that $\mathcal{S}^\star$ contains a search of size 10. We take $\zeta = 1.2$. We plot the expected cumulative regret curves for `CUCB`, `CUCBV`, and ε-greedy for different values of ε, all averaged over 100 simulations. In Figure 2, left, we take $n = 10$, and right, we take $n = 20$. We use a Bernoulli distribution for the cost $\underline{c}_i$ with means c_i. In the first case (Figure 2, left), we notice that both `CUCB` and `CUCBV` give similar results, both better than the ε-greedy approach and that this difference appears already for small budget value B. In the second case (Figure 2, right), since the number of arms considered doubles, we need to consider a larger budget to notice the difference between the algorithms. We note that since the optimal search is not complete, the difference between `CUCB` and `CUCBV` is significant, since the former explores too much, and is even worse than ε-greedy, while the latter makes use of the low variance of each $\underline{w}_i$ (of order $1/n$) to constrain its exploration in order to reach a better expected regret rate. In both cases, we found that ε-greedy with $\varepsilon = 0.01$ gives a lower expected regret compared to ε-greedy with twice the ε (0.02) and the half of it (0.005). Additional experiments are given in Appendix F.

6 Conclusion and future work

We presented sequential search-and-stop problem and provided a stationary offline solution. We gave theoretical guarantees on its optimality and proved that it is computationally efficient. We also considered the learning extension of the problem where the distribution of the hider and the cost are not known. We gave `CUCBV`, an upper-confidence bound approach, tailored to our case and provide expected regret guarantees with respect to the optimal policy. There are several possible extensions. We could consider several hiders rather than just one. Another would be to explore the Thomson sampling (Chapelle and Li, 2011; Agrawal and Goyal, 2012; Komiyama et al., 2015; Wang and Chen, 2018) in the learning case. However, the choice of the prior to use is not straightforward. For instance we can assign Beta prior for each arm, or consider a Dirichlet prior on the *whole* arm set. The Dirichlet seems appropriate because a draw $\widetilde{\mathbf{w}}$ from this prior is in the simplex. The main drawback however is the difficulty of *efficiently* updating such prior to get the posterior, because in the case when the hider is not found, the one-hot vector is not received entirely.

Acknowledgements V. Perchet has benefited from the support of the ANR (grant n.ANR-13-JS01-0004-01), of the FMJH Program Gaspard Monge in optimization and operations research (supported in part by EDF) and from the Labex LMH. The research presented was also supported by European CHIST-ERA project DELTA, French Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council, Inria and Otto-von-Guericke-Universität Magdeburg associated-team north-european project Allocate, and French National Research Agency projects ExTra-Learn (n.ANR-14-CE24-0010-01) and BoB (n.ANR-16-CE23-0003).

References

Agrawal, S. and Goyal, N. (2012). Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*.

Alpern, S., Fokkink, R., Gasieniec, L., Lindelauf, R., and Subrahmanian, V. S. (2013). *Search theory*. Springer.

Alpern, S. and Gal, S. (2006). *The theory of search games and rendezvous*, volume 55. Springer Science & Business Media.

Alpern, S. and Lidbetter, T. (2013). Mining Coal or Finding Terrorists: The Expanding Search Paradigm. *Operations Research*, 61(2):265–279.

Ambühl, C. and Mastrolilli, M. (2009). Single machine precedence constrained scheduling is a vertex cover problem. *Algorithmica (New York)*, 53(4):488–503.

Ambühl, C., Mastrolilli, M., Mutsanas, N., and Svensson, O. (2011). On the Approximability of Single-Machine Scheduling with Precedence Constraints. *Mathematics of Operations Research*, 36(4):653–669.

Audibert, J. Y., Munos, R., and Szepesvári, C. (2009). Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902.

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19(3):357–367.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. (2013). Bandits with knapsacks. In *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 207–216.

Bubeck, S., Ernst, D., and Garivier, A. (2013). Optimal discovery with probabilistic expert advice: finite time analysis and macroscopic optimality. *Journal of Machine Learning Research*, 14:601–623.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games.* Cambridge University Press.

Cesa-Bianchi, N. and Lugosi, G. (2012). Combinatorial bandits. In *Journal of Computer and System Sciences*, volume 78, pages 1404–1422.

Chapelle, O. and Li, L. (2011). An empirical evaluation of Thompson sampling. In *Neural Information Processing Systems*.

Chen, W., Wang, Y., and Yuan, Y. (2016). Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17.

Combes, R., Shahi, M. S. T. M., Proutiere, A., and Others (2015a). Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems*, pages 2116–2124.

Combes, R., Talebi, S., Proutière, A., and Lelarge, M. (2015b). Combinatorial Bandits Revisited. In *NIPS*, Montreal, Canada.

Degenne, R. and Perchet, V. (2016). Combinatorial semi-bandit with known covariance. *CoRR*, abs/1612.01859.

Ding, W., Qin, T., Zhang, X.-d., and Liu, T.-y. (2013). Multi-Armed Bandit with Budget Constraint and Variable Costs. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Evans, T. P. O. and Bishop, S. R. (2013). Static search games played over graphs and general metric spaces. *European Journal of Operational Research*, 231(3):667–689.

Flajolet, A. and Jaillet, P. (2015). Logarithmic regret bounds for Bandits with Knapsacks.

Fokkink, R., Lidbetter, T., and Végh, L. A. (2016). On Submodular Search and Machine Scheduling.

Gai, Y., Krishnamachari, B., and Jain, R. (2012). Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *Transactions on Networking*, 20(5):1466–1478.

Gal, S. (2001). On the optimality of a simple strategy for searching graphs. *International Journal of Game Theory*, 29(4):533–542.

Gopalan, A., Mannor, S., and Mansour, Y. (2014). Thompson sampling for complex online problems. In *International Conference on Machine Learning*, pages 100–108.

Graham, R. L., Lawler, E. L., Lenstra, J. K., and Kan, A. H. (1979). Optimization and approximation in deterministic sequencing and scheduling: A survey. *Annals of Discrete Mathematics*, 5(C):287–326.

Gupta, A. and Nagarajan, V. (2013). A stochastic probing problem with applications. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7801 LNCS, pages 205–216.

Heckerman, D., Breese, J. S., and Rommelse, K. (1995). Decision-theoretic troubleshooting. *Communications of the ACM*, 38(3):49–57.

Hemelrijk, J. (1966). Underlining random variables. *Statistica Neerlandica*, 20(1):1–7.

Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30.

Hohzaki, R. (2016). Search games: Literature and survey. *Journal of the Operations Research Society of Japan*, 59(1):1–34.

Jensen, F. V., Kjaerulff, U., Kristiansen, B., Langseth, H., Skaanning, C., Vomlel, J., and Vomlelova, M. (2001). The SACSO methodology for troubleshooting complex systems. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AIEDAM*, 15(4):321–333.

Kikuta, K. and Ruckle, W. H. (1994). Initial point search on weighted trees. *Naval Research Logistics (NRL)*, 41(6):821–831.

Komiyama, J., Honda, J., and Nakagawa, H. (2015). Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-armed Bandit Problem with Multiple Plays.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. (2015). Tight regret bounds for stochastic combinatorial semi-bandits. In *International Conference on Artificial Intelligence and Statistics*.

Lawler, E. L. (1978). Sequencing jobs to minimize total weighted completion time subject to precedence constraints. *Annals of Discrete Mathematics*, 2(C):75–90.

Lenstra, J. K. and Rinnooy Kan, A. H. G. (1978). Complexity of Scheduling under Precedence Constraints. *Operations Research*, 26(1):22–35.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. *International World Wide Web Conference.*

Lín, V. (2015). Scheduling results applicable to decision-theoretic troubleshooting. *International Journal of Approximate Reasoning*, 56(PA):87–107.

Mohri, M. and Munoz, A. (2014). Optimal regret minimization in posted-price auctions with strategic buyers. In *Advances in Neural Information Processing Systems*, pages 1871–1879.

Prot, D. and Bellenguez-Morineau, O. (2017). A survey on how the structure of precedence constraints may change the complexity class of scheduling problems.

Sidney, J. B. (1975). Decomposition Algorithms for Single-Machine Sequencing with Precedence Relations and Deferral Costs. *Operations Research*, 23(2):283–298.

Smith, W. E. (1956). Various optimizers for single-stage production. *Naval Research Logistics (NRL)*, 3(1-2):59–66.

Stone, L. D. (1976). *Theory of optimal search*, volume 118. Elsevier.

Sutton, R. and Barto, A. (1998). *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA.

Tran-Thanh, L., Chapman, A., Rogers, A., and Jennings, N. R. (2012). Knapsack based optimal policies for budget-limited multi-armed bandits.

Tran-Thanh, L., Stein, S., Rogers, A., and Jennings, N. R. (2014). Efficient crowdsourcing of unknown experts using bounded multi-armed bandits. *Artificial Intelligence*, 214:89–111.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* Springer Series in Statistics. Springer New York, New York, NY.

Valko, M. (2016). *Bandits on graphs and structures.* habilitation, École normale supérieure de Cachan.

Wang, Q. and Chen, W. (2017). Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. In *Neural Information Processing Systems.*

Wang, S. and Chen, W. (2018). Thompson Sampling for Combinatorial Semi-Bandits.

Watanabe, R., Komiyama, J., Nakamura, A., and Kudo, M. (2017). Kl-ucb-based policy for budgeted multi-armed bandits with stochastic action costs. E100.A:2470–2486.

Xia, Y., Ding, W., Zhang, X.-D., Yu, N., and Qin, T. (2016a). Budgeted bandit problems with continuous random costs. In *Asian Conference on Machine Learning.*

Xia, Y., Qin, T., Ma, W., Yu, N., and Liu, T.-Y. (2016b). Budgeted multi-armed bandits with multiple plays. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2210–2216. AAAI Press.

A Proof of Theorem 1

Theorem 1. *For every $(\widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}) \in \left(\mathbb{R}_+^n\right)^2$, Algorithm 1 outputs a search minimizing $J\left(\cdot; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)^+$ over $\mathcal{S}$.*

Here, we might abbreviate $J\left(\cdot; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)^+$ into J^+, and $J\left(\cdot; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)$ into J, keeping in mind that our results will be valid for all $(\widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}) \in \left(\mathbb{R}_+^n\right)^2$. To prove Theorem 1 we first define the concept of *density*, well know in scheduling and search theory.

Definition 1 (Density)**.** *The density is the function defined on $A \in \mathcal{P}([n])$ by $\rho(A) \triangleq \widetilde{w}_A / \widetilde{c}_A$, and $\rho(\emptyset) = 0$.*

Density of $A \subset [n]$ can be understood as the quality/price ratio of that set of arms: the quality is the overall probability of finding the hider in it, while the price is the total cost to fully explore it. Without precedence constraint, the so-called Smith's rule of ratio (Smith, 1956) gives that $\mathbf{x}$ minimizes $d\left(\cdot; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)$ over linear orders (i.e., permutations of $[n]$) if and only if[6] $\rho(x_1) \geq \cdots \geq \rho(x_n)$. Sidney (1975) generalized this principle to any precedence constraint with the concept of Sidney decomposition. Recall that an initial set is the support of a search.

Definition 2 (Sidney decomposition)**.** *A Sidney decomposition $(X_1, X_2, \ldots, X_k)$ is an ordered partition of $[n]$ such that for all $i \in [k]$, X_i is an initial set of maximum density in $\mathcal{G}\langle X_i \sqcup \cdots \sqcup X_k \rangle$.*

Notice that the Sidney decomposition defines a more refined poset on $[n]$, with the extra constraint that an element of X_i must be processed before those of X_j for $i < j$. Any $\mathcal{G}$-linear extension that is also a linear extension of this poset is said to be *consistent* with the Sidney decomposition. The following theorem was proved by Sidney (1975):

Theorem 4 (Sidney, 1975)**.** *Every minimizer of $d\left(\cdot; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)$ over $\mathcal{G}$-linear extensions is consistent with some Sidney decomposition. Moreover, for every Sidney decomposition $(X_1, \ldots, X_k)$, there is a minimizer of $d\left(\cdot; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)$ over $\mathcal{G}$-linear extensions consistent with $(X_1, \ldots, X_k)$.*

Notice that Theorem 4 does not provide a full characterization of minimizers of $d\left(\cdot; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)$ over $\mathcal{G}$-linear extensions, but only a *necessary* condition. Nothing is stated about how to chose the ordering inside each X_i's, and this highly depends on the structure of $\mathcal{G}$ (Lawler, 1978; Ambühl and Mastrolilli, 2009; Ambühl et al., 2011). We are now ready to prove Theorem 1, thanks to Lemma 1, of which the proof is given in Appendix A.1.

Lemma 1. *For any Sidney decomposition $(X_1, \ldots, X_k)$, there exists $i \leq k$ and a search with support $X_1 \sqcup \cdots \sqcup X_i$ that minimizes J^+.*

Proof of Theorem 1. We know from first statement of Theorem 4 that $\mathbf{s} \triangleq \texttt{Schedul}\left(\widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}, \mathcal{G}\right)$ given in Algorithm 1 is consistent with some Sidney decomposition $(X_1, \ldots, X_k)$. Let $i \leq k$ and $\mathbf{x}$ minimizing J^+ of support $X_1 \sqcup \cdots \sqcup X_i$ given by Lemma 1. Let $\mathbf{s} = \mathbf{s}_1\mathbf{s}_2$ with $\mathbf{s}_1$ being the restriction of $\mathbf{s}$ to $X_1 \sqcup \cdots \sqcup X_i$ (and thus $\mathbf{s}_2$ is its restriction to $X_{i+1} \sqcup \cdots \sqcup X_k$). Let's prove that $\mathbf{s}_1$ is also a minimizer of J^+ by showing $J^+\left(\mathbf{s}_1\right) \leq J^+\left(\mathbf{x}\right)$, thereby concluding the proof. Since $0 \leq d\left(\mathbf{x}\mathbf{s}_2; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right) - d\left(\mathbf{s}_1\mathbf{s}_2; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right) = d\left(\mathbf{x}; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right) - d\left(\mathbf{s}_1; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)$, we have

$$\frac{d\left(\mathbf{s}_1; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right) + (1 - \widetilde{w}_{X_1 \sqcup \cdots \sqcup X_i})\widetilde{c}_{X_1 \sqcup \cdots \sqcup X_i}}{\widetilde{w}_{X_1 \sqcup \cdots \sqcup X_i}} \leq \frac{d\left(\mathbf{x}; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right) + (1 - \widetilde{w}_{X_1 \sqcup \cdots \sqcup X_i})\widetilde{c}_{X_1 \sqcup \cdots \sqcup X_i}}{\widetilde{w}_{X_1 \sqcup \cdots \sqcup X_i}},$$

i.e., $J\left(\mathbf{s}_1\right) \leq J\left(\mathbf{x}\right)$, and because $x \mapsto x^+$ is increasing on $\mathbb{R}$, we have $J^+\left(\mathbf{s}_1\right) \leq J^+\left(\mathbf{x}\right)$. □

The proof of Lemma 1 also uses Sidney's Theorem 4, but this time the second statement. However, although it provides a crucial analysis, with fixed support, concerning the order to choose for minimizing $d\left(\cdot; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)$ and therefore $J\left(\cdot; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)^+$, nothing is said about the support to choose. Thus, to prove Lemma 1, we also need the following Proposition 2, that gives the key support property satisfied by J^+.

Proposition 2 (Support property)**.** *If $\mathbf{xy}, \mathbf{xyz} \in \mathcal{S}$ with $\rho(\mathbf{z}) \geq \rho(\mathbf{y})$, then*

$$J^+(\mathbf{xy}) \geq \min\left\{J^+(\mathbf{x}), J^+(\mathbf{xyz})\right\}.$$

Proof. If $J(\mathbf{xyz}) < 0$, $J^+(\mathbf{xyz}) = 0 \leq J^+(\mathbf{xy})$. We thus suppose $J(\mathbf{xyz}) \geq 0$. Since $J(\mathbf{z}) \leq \frac{1}{\rho(\mathbf{z})}$,

$$0 \leq J(\mathbf{xyz}) = \frac{J(\mathbf{xy})\widetilde{w}_{\mathbf{xy}}}{\widetilde{w}_{\mathbf{xyz}}} + \frac{\widetilde{w}_{\mathbf{z}}J(\mathbf{z}) - \widetilde{w}_{\mathbf{xy}}\widetilde{c}_{\mathbf{z}}}{\widetilde{w}_{\mathbf{xyz}}} \leq \frac{J(\mathbf{xy})\widetilde{w}_{\mathbf{xy}}}{\widetilde{w}_{\mathbf{xyz}}} + \frac{\widetilde{w}_{\mathbf{z}}(1 - \widetilde{w}_{\mathbf{xy}})}{\rho(\mathbf{z})\widetilde{w}_{\mathbf{xyz}}}. \tag{3}$$

[6] One can see that $\sum_{\{i,j\} \in I(\sigma),\ i<j} \widetilde{c}_{s_i}\widetilde{c}_{s_j}\left(\rho(s_i) - \rho(s_j)\right)$ is the variation of d when swapping a linear order $\mathbf{s}$ by a permutation σ, where $I(\sigma)$ the set of inversions in σ.

Suppose that $1 - \widetilde{w}_{\mathbf{xy}} \geq 0$. If $J(\mathbf{x}) \geq J(\mathbf{xy})$, then

$$J(\mathbf{xy}) \geq \frac{1}{\widetilde{w}_{\mathbf{y}}} \left(J(\mathbf{xy})\widetilde{w}_{\mathbf{xy}} - J(\mathbf{x})\widetilde{w}_{\mathbf{x}}\right) = \sum_{i=1}^{|\mathbf{y}|} \frac{\widetilde{c}_{y_i}\left(1 - \widetilde{w}_{\mathbf{x}} - \widetilde{w}_{\mathbf{y}[i-1]}\right)}{\widetilde{w}_{\mathbf{y}}} \geq \frac{1 - \widetilde{w}_{\mathbf{xy}}}{\rho(\mathbf{y})}. \tag{4}$$

Thus, we have,

$$\begin{aligned}
J(\mathbf{xyz}) - J(\mathbf{xy}) \leq& \frac{J(\mathbf{xy})\widetilde{w}_{\mathbf{xy}}}{\widetilde{w}_{\mathbf{xyz}}} + \frac{\widetilde{w}_{\mathbf{z}}(1 - \widetilde{w}_{\mathbf{xy}})}{\rho(\mathbf{z})\widetilde{w}_{\mathbf{xyz}}} - J(\mathbf{xy}) && \text{using (3).} \\
=& \frac{-\widetilde{w}_{\mathbf{z}} J(\mathbf{xy})}{\widetilde{w}_{\mathbf{xyz}}} + \frac{\widetilde{w}_{\mathbf{z}}(1 - \widetilde{w}_{\mathbf{xy}})}{\rho(\mathbf{z})\widetilde{w}_{\mathbf{xyz}}} \\
\leq& \frac{\widetilde{w}_{\mathbf{z}}}{\widetilde{w}_{\mathbf{xyz}}} \left(\frac{-(1 - \widetilde{w}_{\mathbf{xy}})}{\rho(\mathbf{y})} + \frac{1 - \widetilde{w}_{\mathbf{xy}}}{\rho(\mathbf{z})}\right) \leq 0 && \text{using (4), and then } \rho(\mathbf{z}) \geq \rho(\mathbf{y}).
\end{aligned}$$

Finally, if $1 - \widetilde{w}_{\mathbf{xy}} \leq 0$, by (3), we have that $0 \leq J(\mathbf{xyz}) \leq \frac{J(\mathbf{xy})\widetilde{w}_{\mathbf{xy}}}{\widetilde{w}_{\mathbf{xyz}}} \leq J(\mathbf{xy})$. □

Example 2. *Now, as a preview, we can actually derive easily the proof of Lemma 1 when there is no precedence constraints, the idea in the general case being very similar. Let $(X_1, \ldots, X_k)$ be a Sidney decomposition. Then, if $x_{i,1}, \ldots, x_{i,j_i}$ are arms of X_i, we have*

$$\rho(x_{1,1}) = \cdots = \rho(x_{1,j_1}) \geq \cdots \geq \rho(x_{k,1}) = \cdots = \rho(x_{k,j_k}).$$

Let $\mathbf{s}^\star$ be a minimum-size minimizer of J^+ of support S. Assume S is not of the form given by Lemma 1, and let x be the first, for the order $(x_{1,1}, \ldots, x_{1,j_1}, \ldots, x_{k,1}, \ldots, x_{k,j_k})$, in some $X_i \backslash S$ while $S \cap (X_i \sqcup \cdots \sqcup X_k) \neq \emptyset$. By Proposition 2, we keep the optimality by either adding x to $\mathbf{s}^\star$ (which contradict the minimality of $|\mathbf{s}^\star|$), or by removing the suffix defined on $S \cap (X_i \sqcup \cdots \sqcup X_k)$, giving a support satisfying conclusion of Lemma 1.

A.1 Proof of Lemma 1

Lemma 1. *For any Sidney decomposition $(X_1, \ldots, X_k)$, there exists $i \leq k$ and a search with support $X_1 \sqcup \cdots \sqcup X_i$ that minimizes J^+.*

Before proving Lemma 1, we state some preliminaries about initial sets of the DAG $\mathcal{G}$.

Proposition 3. *A is an initial set in $\mathcal{G}$ if and only if for all $a \in A$, the predecessors of a in $\mathcal{G}$ are also in A.*

Proof. The direct sense is clear. Suppose now that for all $a \in A$, the predecessors of a in $\mathcal{G}$ are also in A. Consider $\mathbf{a} = (a_1, \ldots, a_{|A|})$ a linear extension of $\mathcal{G}\langle A \rangle$. Then it is a search, and predecessors of any a_i in $\mathcal{G}$ are in $\{a_1, \ldots, a_{i-1}\} \cup A^c$, thus in $\{a_1, \ldots, a_{i-1}\}$ by assumption. Therefore, $\mathbf{a}$ is a search in $\mathcal{G}$ and A is an initial set. □

Let us recall that $\mathcal{L} \subset \mathcal{P}([n])$ is a lattice if $X, Y \in \mathcal{L} \Rightarrow (X \cap Y \in \mathcal{L}$ and $X \cup Y \in \mathcal{L})$.

Proposition 4. *The set of initial sets in $\mathcal{G}$ is a lattice.*

Proof. Let X and Y be two initial sets in $\mathcal{G}$. If $x \in X \cup Y$ (respectively $x \in X \cap Y$), then predecessors of x are included in predecessors of X or (respectively and) the predecessors of Y, i.e., in X or (respectively and) Y, so in $X \cup Y$ (respectively $X \cap Y$). □

Even if we do not use the following proposition,[7] we provide it nonetheless, since it illustrates how to handle density ρ.

Proposition 5. *The set of initial sets of maximum density in $\mathcal{G}$ is a lattice.*

Proof. We use the fact that for $a, b \geq 0$ and $a', b' > 0$, $\frac{a+b}{a'+b'} \leq \max\left\{\frac{a}{a'}, \frac{b}{b'}\right\}$, with equality if and only if $\frac{a}{a'} = \frac{b}{b'}$. Indeed, if X and Y are two initial sets of maximum density in $\mathcal{G}$, then

$$\frac{\widetilde{w}_X}{\widetilde{c}_X} = \frac{\widetilde{w}_X + \widetilde{w}_Y}{\widetilde{c}_X + \widetilde{c}_Y} = \frac{\widetilde{w}_{X \cap Y} + \widetilde{w}_{X \cup Y}}{\widetilde{c}_{X \cap Y} + \widetilde{c}_{X \cup Y}} \leq \max\left\{\frac{\widetilde{w}_{X \cup Y}}{\widetilde{c}_{X \cup Y}}, \frac{\widetilde{w}_{X \cap Y}}{\widetilde{c}_{X \cap Y}}\right\}.$$

$X \cap Y$ and $X \cup Y$ are initial sets. Therefore, by maximality, $\max\left\{\frac{\widetilde{w}_{X \cup Y}}{\widetilde{c}_{X \cup Y}}, \frac{\widetilde{w}_{X \cap Y}}{\widetilde{c}_{X \cap Y}}\right\} \leq \frac{\widetilde{w}_X}{\widetilde{c}_X}$. Since the equality holds, it needs to be the case that $\frac{\widetilde{w}_{X \cup Y}}{\widetilde{c}_{X \cup Y}} = \frac{\widetilde{w}_{X \cap Y}}{\widetilde{c}_{X \cap Y}} = \frac{\widetilde{w}_X}{\widetilde{c}_X}$. □

[7] Theorem 4 does need this proposition.

Proof of Lemma 1. Let j be the largest integer such that there is a search minimizing J^+ of the form $\mathbf{x}_1 \cdots \mathbf{x}_j \mathbf{a}$ with $\mathbf{x}_i$ of support X_i for all $i \in [j]$, and $\mathbf{a}$ of support A. Let $\mathbf{s}$ be such search, with $|\mathbf{s}|$ being the smallest possible. By Theorem 4, we know there exists a minimizer of the form $\mathbf{x}_{j+1}\mathbf{y}$ of $d\left(\mathbf{x}_1 \cdots \mathbf{x}_j\ \cdot\ ; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)$ over $\mathcal{G}\langle X_{j+1} \sqcup A\rangle$-linear extensions, with $\mathbf{x}_{j+1}$ of support X_{j+1}. $X_{j+1} \cap A$ is an initial set of $\mathcal{G}\langle X_{j+1} \sqcup \cdots \sqcup X_k\rangle$, therefore

$$\rho(X_{j+1} \cap A) \leq \rho(X_{j+1}) = \rho\left((X_{j+1} \cap A) \sqcup (X_{j+1} \backslash A)\right) \leq \rho(X_{j+1} \backslash A),$$

and thus $\rho(A) \leq \rho(X_{j+1}) \leq \rho(X_{j+1} \backslash A)$. If we let $\mathbf{b}$ be a search of $\mathcal{G}\langle (X_{j+1} \backslash A) \sqcup X_{j+2} \sqcup \cdots \sqcup X_k\rangle$ with support $X_{j+1} \backslash A$, then by Proposition 2, associated with $d\left(\mathbf{x}_1 \cdots \mathbf{x}_j \mathbf{x}_{j+1}\mathbf{y}; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right) \leq d\left(\mathbf{x}_1 \cdots \mathbf{x}_j \mathbf{ab}; \widetilde{\mathbf{w}}, \widetilde{\mathbf{c}}\right)$, we have that

$$J^+(\mathbf{s}) \geq \min\left\{J^+(\mathbf{x}_1 \cdots \mathbf{x}_j), J^+(\mathbf{x}_1 \cdots \mathbf{x}_j \mathbf{ab})\right\} \geq \min\left\{J^+(\mathbf{x}_1 \cdots \mathbf{x}_j), J^+(\mathbf{x}_1 \cdots \mathbf{x}_j \mathbf{x}_{j+1}\mathbf{y})\right\},$$

contradicting either the definition of j or the minimality of $|\mathbf{s}|$. $\square$

B Proof of Proposition 1

Proposition 1. *If $\pi^\star$ is the offline policy selecting $\mathbf{s}^\star \in \mathcal{S}^\star$ at each round t, then*

$$\frac{B-n}{J^\star} \leq F_B(\pi^\star) \leq F_B^\star \leq \frac{B+n}{J^\star}\cdot$$

Proof. If we let $B^0 = B$, then for any offline policy π, if we denote by $\mathbf{s}^t$ the search selected by π at round t (we saw that an optimal policy *selects* at the begining of a round a search and then *performs* it), and if we let $\underline{B}^t = B - \sum_{u=1}^t \underline{c}^u_{\mathbf{s}^u[\underline{\mathbf{w}}^t]}$ be the remaining budget at time t,

$$\begin{aligned}
F_B(\pi) = \sum_{t=1}^\infty \mathbb{E}\left[\sum_{i\in\mathbf{s}^t} \mathbb{I}\{\underline{B}^t \geq 0,\ \underline{w}_i^t = 1\}\right] &\leq \sum_{t=1}^\infty \mathbb{E}\left[\sum_{i\in\mathbf{s}^t} \mathbb{I}\{\underline{B}^{t-1} \geq 0,\ \underline{w}_i^t = 1\}\right] &(5)\\
&= \sum_{t=1}^\infty \mathbb{E}\left[\sum_{i\in\mathbf{s}^t} \mathbb{I}\{\underline{B}^{t-1} \geq 0\}\underline{w}_i\right] &(6)\\
&= \sum_{t=1}^\infty \mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1} \geq 0\}\underline{w}_{\mathbf{s}^t}\right]\\
&= \sum_{t=1}^\infty \mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1} \geq 0\}\frac{d(\mathbf{s}^t) + (1-\underline{w}_{\mathbf{s}^t})\underline{c}_{\mathbf{s}^t}}{J(\mathbf{s}^t)}\right]\\
&\leq \sum_{t=1}^\infty \mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1} \geq 0\}\frac{d(\mathbf{s}^t) + (1-\underline{w}_{\mathbf{s}^t})\underline{c}_{\mathbf{s}^t}}{J^\star}\right]\\
&= \frac{1}{J^\star}\mathbb{E}\left[\sum_{t=1}^{\underline{\tau}_B}(d(\mathbf{s}^t) + (1-\underline{w}_{\mathbf{s}^t})\underline{c}_{\mathbf{s}^t})\right] &(7)\\
&= \frac{1}{J^\star}\mathbb{E}\left[\sum_{t=1}^{\underline{\tau}_B}\underline{c}_{\mathbf{s}^t[\underline{\mathbf{w}}^t]}\right]\\
&\leq \frac{1}{J^\star}\mathbb{E}\left[\sum_{t=1}^{\underline{\tau}_B-1}\underline{c}^t_{\mathbf{s}^t[\underline{\mathbf{w}}^t]} + \underline{c}^{\underline{\tau}_B}_{\mathbf{s}^{\underline{\tau}_B}[\underline{\mathbf{w}}^t]}\right]\\
&\leq \frac{B+n}{J^\star}, &(8)
\end{aligned}$$

where (5) uses $\underline{B}^t \geq 0 \Rightarrow \underline{B}^{t-1} \geq 0$, (6) is obtained by conditioning on previously sampled arms, (7) uses the random round $\underline{\tau}_B$ such that $\underline{B}^{\underline{\tau}_B-1} \geq 0$ and $\underline{B}^{\underline{\tau}_B} < 0$, and (8) uses the definition of $\underline{B}^{\underline{\tau}_B-1}$ and $\underline{c}_i^t \leq 1$. Now, for the lower bound, we have that

$$\begin{aligned}
F_B(\pi^\star) &\geq \sum_{t=1}^\infty \mathbb{E}\left[\sum_{i\in\mathbf{s}^\star} \mathbb{I}\{\underline{B}^{t-1} \geq n,\ \underline{w}_i^t = 1\}\right] &(9)\\
&= \sum_{t=1}^\infty \mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1} \geq n\}\underline{w}_{\mathbf{s}^\star}\right] &(10)\\
&= \frac{1}{J^\star}\mathbb{E}\left[\sum_{t=1}^{\underline{\tau}}\underline{c}^t_{\mathbf{s}^\star[\underline{\mathbf{w}}^t]}\right] &(11)
\end{aligned}$$

$$\geq \frac{B-n}{J^\star}, \tag{12}$$

where (9) uses $\underline{B}^{t-1} \geq n \Rightarrow \underline{B}^t \geq 0$, (10) uses the same derivation as previously, (11) uses $\underline{\tau}$, the random round such that $\underline{B}^{\underline{\tau}-1} \geq n$ and $\underline{B}^{\underline{\tau}} < n$, and (12) is by definition of $\underline{B}^{\underline{\tau}}$. $\square$

C Proof of Theorem 2

Theorem 2. *The expected cumulative regret of* `CUCBV` *is bounded as*

$$R_B(\pi_{\texttt{CUCBV}}) \lesssim n\left(1+\frac{n}{J^\star}\right)^2 \sum_{i\in[n]} \frac{w_i}{\Delta_{i,\min}} \log T_B$$
$$\text{and} \quad R_B(\pi_{\texttt{CUCBV}}) \lesssim \sqrt{n}\left(1+\frac{n}{J^\star}\right)\sqrt{T_B \log T_B}.$$

We let $\beta(t) \triangleq \inf_{1<\alpha\leq 3} \min\left\{\frac{\log(t)}{\log(\alpha)}, t\right\} t^{-\frac{\zeta}{\alpha}}$. In the proof of Theorem 2, we make several uses of the following concentration inequalities that use the same peeling argument for their proof as Theorem 1 of Audibert et al. (2009) applied to original *anytime* inequalities.

Fact 1 (Theorem 1 of Audibert et al., 2009)**.** *Let $(\underline{x}^t)$ be iid centered random variables with common support $[0,1]$ with common support $[0,1]$, $\overline{\underline{x}}^t \triangleq \frac{1}{t}(\underline{x}^1+\cdots+\underline{x}^t)$ and let $\underline{v}^t \triangleq \frac{1}{t}\sum_{u=1}^t \left(\overline{\underline{x}}^t - \underline{x}^u\right)^2$, then*

$$\mathbb{P}\left[\exists u \leq t,\ \overline{\underline{x}}^u > \sqrt{\frac{2\underline{v}^u\zeta\log(t)}{u}} + \frac{3\zeta\log(t)}{u}\right] \leq 2\beta(t).$$

Fact 2 (Hoeffding, 1963; Azuma, 1967)**.** *Let $(\underline{x}^t)$ be a martingale difference sequence with common support $[0,1]$, and let $\overline{\underline{x}}^t \triangleq \frac{1}{t}(\underline{x}^1+\cdots+\underline{x}^t)$, then*

$$\mathbb{P}\left[\exists u \leq t,\ \overline{\underline{x}}^u > \sqrt{\frac{\zeta\log(t)}{2u}}\right] \leq \beta(t).$$

Fact 3 (Bernstein inequality)**.** *Let $(\underline{x}^t)$ be a martingale difference sequence with common support $[0,1]$, $\sigma^2 \triangleq \mathbb{V}(\underline{x}^t)$, and let $\overline{\underline{x}}^t \triangleq \frac{1}{t}(\underline{x}^1+\cdots+\underline{x}^t)$, then*

$$\mathbb{P}\left[\exists u \leq t,\ \overline{\underline{x}}^u > \sqrt{\frac{2\sigma^2\zeta\log(t)}{u}} + \frac{\zeta\log(t)}{3u}\right] \leq \beta(t).$$

Before we dive into the proof of Theorem 2, we first state a lemma that gives a high-probability control on the error that is made when estimating w_i.

Lemma 2. $\mathbb{P}\left[\ddot{\underline{w}}_i^{t-1} - w_i > \sqrt{\frac{8\zeta w_i(1-w_i)\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}} + \frac{13.3\zeta\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}\right] \leq 2\beta(t).$

Proof. Let

$$\underline{r} \triangleq \frac{8\zeta\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}} + 2\sqrt{\left(\frac{\sqrt{7}\zeta\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}\right)^2 + \frac{2\zeta w_i(1-w_i)\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}},$$
$$\underline{\delta} \triangleq \sqrt{\frac{8\zeta w_i(1-w_i)\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}} + \frac{13.3\zeta\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}, \quad \text{and}$$
$$\varepsilon(u) \triangleq \sqrt{\frac{8w_i(1-w_i)\zeta\log(t)}{u}} + \frac{2\zeta\log(t)}{3u}.$$

We have that

$$\mathbb{P}\left[\ddot{\underline{w}}_i^{t-1} - w_i > \underline{\delta}\right]$$
$$= \mathbb{P}\left[\min\left\{\overline{\underline{w}}_i^{t-1} + \sqrt{\frac{2\zeta\overline{\underline{w}}_i^{t-1}(1-\overline{\underline{w}}_i^{t-1})\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}} + \frac{3\zeta\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}, 1\right\} - w_i > \underline{\delta}\right]$$

$$\leq \mathbb{P}\left[\underline{\overline{w}}_i^{t-1} + \sqrt{\frac{2\zeta \underline{\overline{w}}_i^{t-1}(1-\underline{\overline{w}}_i^{t-1})\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}} + \frac{3\zeta\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}} - w_i > \underline{\delta}\right]$$
$$\leq \mathbb{P}\left[\underline{\overline{w}}_i^{t-1} + \sqrt{\frac{2\zeta\left(w_i(1-w_i)+\underline{\delta}/2\right)\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}} + \frac{3\zeta\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}} - w_i > \underline{\delta}\right]$$
$$+ \mathbb{P}\left[\underline{\overline{w}}_i^{t-1}(1-\underline{\overline{w}}_i^{t-1}) > w_i(1-w_i) + \underline{\delta}/2\right].$$

Since $\underline{\overline{w}}_i^{t-1}(1-\underline{\overline{w}}_i^{t-1}) = \underline{\overline{w}}_i^{t-1} - 2w_i\underline{\overline{w}}_i^{t-1} + w_i^2 - \left(w_i - \underline{\overline{w}}_i^{t-1}\right)^2$,

$$\mathbb{P}\left[\underline{\overline{w}}_i^{t-1}(1-\underline{\overline{w}}_i^{t-1}) \geq w_i(1-w_i)+\underline{\delta}/2\right] \leq \mathbb{P}\left[\underline{\overline{w}}_i^{t-1} - 2w_i\underline{\overline{w}}_i^{t-1} + w_i^2 \geq w_i(1-w_i)+\underline{\delta}/2\right].$$

For the first term, notice that

$$\sqrt{\frac{2\zeta\left(w_i(1-w_i)+\underline{\delta}/2\right)\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}} + \frac{3\zeta\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}} \leq \underline{\delta}/2.$$

Indeed, this holds because $\underline{\delta}$ is greater than $\underline{r}$, the greatest root of second-degree polynomial

$$X^2/4 - \frac{4\zeta\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}X + \left(\frac{3\zeta\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}\right)^2 - \frac{2\zeta w_i(1-w_i)\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}.$$

Hence, $\mathbb{P}\left[\underline{\ddot{w}}_i^{t-1} - w_i > \underline{\delta}\right]$ is bounded by

$$\mathbb{P}\left[\underline{\overline{w}}_i^{t-1} - w_i > \underline{\delta}/2\right] + \mathbb{P}\left[\underline{\overline{w}}_i^{t-1} - 2w_i\underline{\overline{w}}_i^{t-1} + w_i^2 > w_i(1-w_i)+\underline{\delta}/2\right]$$
$$\leq \mathbb{P}\left[\underline{\overline{w}}_i^{t-1} - w_i > \frac{\varepsilon\left(\underline{T}_{i,\mathbf{w}}^{t-1}\right)}{2}\right] + \mathbb{P}\left[\underline{\overline{w}}_i^{t-1} - 2w_i\underline{\overline{w}}_i^{t-1} + w_i^2 > w_i(1-w_i) + \frac{\varepsilon\left(\underline{T}_{i,\mathbf{w}}^{t-1}\right)}{2}\right] \quad \text{since } \varepsilon\left(\underline{T}_{i,\mathbf{w}}^{t-1}\right) \leq \underline{\delta}$$
$$\leq \mathbb{P}\left[\exists u \leq t,\ \frac{1}{u}\sum_{v=1}^{u}\underline{w}_i^v - w_i > \frac{\varepsilon(u)}{2}\right] + \mathbb{P}\left[\exists u \leq t,\ \frac{1}{u}\sum_{v=1}^{u}\left(\underline{w}_i^v - w_i\right)^2 - w_i(1-w_i) > \frac{\varepsilon(u)}{2}\right]$$
$$\leq 2\beta(t),$$

where the last inequality uses Bernstein's inequality (Fact 3) twice, noticing that

$$\mathbb{V}\left(\frac{1}{u}\sum_{v=1}^{u}\left(\underline{w}_i^v - w_i\right)^2\right) \leq w_i(1-w_i).$$

□

Proof of Theorem 2. We start with showing a lower bound on the expected reward of any policy π,

$$F_B(\pi) \geq \sum_{t\geq1}\mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1} \geq n\}w_{\underline{\mathbf{s}}^t}\right] \tag{13}$$
$$= \sum_{t\geq1}\mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1}\geq n,\ \underline{\mathbf{s}}^t \in \mathcal{S}^\star\}w_{\underline{\mathbf{s}}^t}\right] + \sum_{t\geq1}\mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1}\geq n,\ \underline{\mathbf{s}}^t \notin \mathcal{S}^\star\}w_{\underline{\mathbf{s}}^t}\right]$$
$$= \frac{1}{J^\star}\sum_{t\geq1}\mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1}\geq n,\ \underline{\mathbf{s}}^t \in \mathcal{S}^\star\}\left(d(\underline{\mathbf{s}}^t) + (1-w_{\underline{\mathbf{s}}^t})c_{\underline{\mathbf{s}}^t}\right)\right]$$
$$+ \frac{1}{J^\star}\sum_{t\geq1}\mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1}\geq n,\ \underline{\mathbf{s}}^t \notin \mathcal{S}^\star\}\left(d(\underline{\mathbf{s}}^t) + (1-w_{\underline{\mathbf{s}}^t})c_{\underline{\mathbf{s}}^t}\right)\right]$$
$$- \sum_{t\geq1}\mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1}\geq n,\ \underline{\mathbf{s}}^t \notin \mathcal{S}^\star\}\Delta\left(\underline{\mathbf{s}}^t\right)\right]$$
$$= \frac{1}{J^\star}\sum_{t\geq1}\mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1}\geq n\}\left(d(\underline{\mathbf{s}}^t) + (1-w_{\underline{\mathbf{s}}^t})c_{\underline{\mathbf{s}}^t}\right)\right]$$
$$- \sum_{t\geq1}\mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1}\geq n,\ \underline{\mathbf{s}}^t \notin \mathcal{S}^\star\}\Delta\left(\underline{\mathbf{s}}^t\right)\right]$$
$$\geq \frac{B-n}{J^\star} - \sum_{t\geq1}\mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1}\geq n,\ \underline{\mathbf{s}}^t \notin \mathcal{S}^\star\}\Delta\left(\underline{\mathbf{s}}^t\right)\right], \tag{14}$$

with (13) obtained as (9) and (10), and (14) as (12). Therefore, since $F_B^\star \le (B+n)/J^\star$ by Proposition 1, we have that

$$R_B(\pi) - \frac{2n}{J^\star} \le \sum_{t\ge1} \mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1} \ge n,\ \underline{\mathbf{s}}^t \notin \mathcal{S}^\star\}\Delta\left(\underline{\mathbf{s}}^t\right)\right] \le \sum_{t\ge1} \mathbb{E}\left[\mathbb{I}\{\underline{B}^t \ge 0\}\Delta\left(\underline{\mathbf{s}}^t\right)\right] = \mathbb{E}\left[\sum_{t=1}^{\underline{\tau}_B - 1} \Delta\left(\underline{\mathbf{s}}^t\right)\right].$$

At this part of the analysis, the know techniques (Xia et al., 2016a,b) are based on the following variant of Hoeffding's inequality, in order to bound $\underline{\tau}_B$ by T_B with high probability.

Fact 4 (Hoeffding, 1963; Flajolet and Jaillet, 2015)**.** *Let $\underline{x}^1, \dots, \underline{x}^t$ be the random variables with common support $[0,1]$ and such that there exists $a \in \mathbb{R}$ with $\forall u \in [t]$, $\mathbb{E}\left[\underline{x}^u | \underline{x}^1, \dots, \underline{x}^{u-1}\right] \ge a$. Let $\overline{\underline{x}}^t \triangleq \frac{1}{t}(\underline{x}^1 + \dots + \underline{x}^t)$, then*

$$\forall \varepsilon \ge 0\ ,\ \mathbb{P}\left[\overline{\underline{x}}^t - a \le -\varepsilon\right] \le e^{-2\varepsilon^2 t}.$$

Indeed, this would decompose the expected regret into a term $\mathbb{E}\left[\sum_{t=1}^{T_B} \Delta\left(\underline{\mathbf{s}}^t\right)\right]$, and another of order $e^{-c_{\min}B}/c_{\min}^2$. Although the second term decreases exponentially fast to 0 when $B \to \infty$, the dependence on $1/c_{\min}^2$ is undesirable and artificial. Therefore, we bound $\mathbb{E}\left[\underline{\tau}_B\right]$ directly instead.

$$\begin{aligned}
\mathbb{E}\left[\underline{\tau}_B\right] = 1 + \mathbb{E}\left[\sum_{t\ge1} \mathbb{I}\{\underline{B}^t \ge 0\}\right] &= 1 + \sum_{t\ge1} \mathbb{P}\left[B - tc_{\min} + tc_{\min} \ge \sum_{u=1}^{t} \underline{c}_{\underline{s}^u[\underline{\mathbf{w}}^u]}\right] \\
&\le T_B + 1 + \sum_{t \ge T_B+1} \exp\left(\frac{-2(B - tc_{\min})^2}{t}\right) &(15)\\
&\le T_B + 1 + \sum_{t \ge T_B+1} \exp\left(\frac{-c_{\min}^2 t}{2}\right) &(16)\\
&\le T_B + 1 + \frac{2}{c_{\min}^2} \exp\left(\frac{c_{\min}^2}{2} - \frac{c_{\min}^2(T_B+1)}{2}\right) &(17)\\
&\le T_B + 1 + \frac{2}{c_{\min}^2} \exp\left(-c_{\min}B\right), &(18)
\end{aligned}$$

where (15) makes use of Fact 4, (16) is obtained because $2(B - tc_{\min})^2 \ge c_{\min}^2 t^2/2$ for $t \ge 2B/c_{\min}$ and we get (17) since $1/(1 - e^{-c_{\min}^2/2}) \le 2e^{c_{\min}^2/2}/c_{\min}^2$.

Finally, we use Jensen's inequality in the expected regret bound with the random time horizon $\underline{\tau}_B - 1$. More precisely, we bound $\mathbb{E}\left[\sum_{t=1}^{\underline{\tau}_B-1} \Delta\left(\underline{\mathbf{s}}^t\right)\right]$ that has a factor $\mathbb{E}\left[\log\left(\underline{\tau}_B\right)\right]$, but since log is a concave function, we have $\mathbb{E}\left[\log\left(\underline{\tau}_B\right)\right] \le \log\left(\mathbb{E}\left[\underline{\tau}_B\right]\right)$, and this last term is of order $\log(T_B)$.

C.1 Bound on $\Delta\left(\underline{\mathbf{s}}^t\right)$

Since $\underline{\mathbf{s}}^t$ minimizes $J\left(\cdot; \ddot{\underline{\mathbf{w}}}^t, \dot{\underline{\mathbf{c}}}^t\right)^+$, then $J\left(\underline{\mathbf{s}}^t; \ddot{\underline{\mathbf{w}}}^t, \dot{\underline{\mathbf{c}}}^t\right)^+ \le J\left(\mathbf{s}^\star; \ddot{\underline{\mathbf{w}}}^t, \dot{\underline{\mathbf{c}}}^t\right)^+$. Therefore,

$$\begin{aligned}
\Delta\left(\underline{\mathbf{s}}^t\right) &= \frac{1}{J^\star}\left(\sum_{i=1}^{|\underline{\mathbf{s}}^t|} c_{\underline{s}_i^t}\left(1 - w_{\underline{\mathbf{s}}^t[i-1]}\right) - J^\star w_{\underline{\mathbf{s}}^t}\right) \\
&= \frac{1}{J^\star}\left(\sum_{i=1}^{|\underline{\mathbf{s}}^t|} c_{\underline{s}_i^t}\left(1 - w_{\underline{\mathbf{s}}^t[i-1]}\right) - J^\star \ddot{\underline{w}}_{\underline{\mathbf{s}}^t}^t\right) + \ddot{\underline{w}}_{\underline{\mathbf{s}}^t}^t - w_{\underline{\mathbf{s}}^t} \\
&= \frac{1}{J^\star}\left(\sum_{i=1}^{|\underline{\mathbf{s}}^t|} c_{\underline{s}_i^t}\left(1 - w_{\underline{\mathbf{s}}^t[i-1]}\right) - J\left(\mathbf{s}^\star; \ddot{\underline{\mathbf{w}}}^t, \dot{\underline{\mathbf{c}}}^t\right)^+ \ddot{\underline{w}}_{\underline{\mathbf{s}}^t}^t\right) + \frac{J\left(\mathbf{s}^\star; \ddot{\underline{\mathbf{w}}}^t, \dot{\underline{\mathbf{c}}}^t\right)^+ - J^\star}{J^\star} \ddot{\underline{w}}_{\underline{\mathbf{s}}^t}^t + \ddot{\underline{w}}_{\underline{\mathbf{s}}^t}^t - w_{\underline{\mathbf{s}}^t} \\
&\le \frac{1}{J^\star}\left(\sum_{i=1}^{|\underline{\mathbf{s}}^t|} c_{\underline{s}_i^t}\left(1 - w_{\underline{\mathbf{s}}^t[i-1]}\right) - J\left(\underline{\mathbf{s}}^t; \ddot{\underline{\mathbf{w}}}^t, \dot{\underline{\mathbf{c}}}^t\right)^+ \ddot{\underline{w}}_{\underline{\mathbf{s}}^t}^t\right) + \frac{J\left(\mathbf{s}^\star; \ddot{\underline{\mathbf{w}}}^t, \dot{\underline{\mathbf{c}}}^t\right)^+ - J^\star}{J^\star} \ddot{\underline{w}}_{\underline{\mathbf{s}}^t}^t + \ddot{\underline{w}}_{\underline{\mathbf{s}}^t}^t - w_{\underline{\mathbf{s}}^t} \\
&\le \frac{1}{J^\star}\sum_{i=1}^{|\underline{\mathbf{s}}^t|}\left(c_{\underline{s}_i^t}\left(1 - w_{\underline{\mathbf{s}}^t[i-1]}\right) - \dot{\underline{c}}_{\underline{s}_i^t}^t\left(1 - \ddot{\underline{w}}_{\underline{\mathbf{s}}^t[i-1]}^t\right)\right) + \frac{J\left(\mathbf{s}^\star; \ddot{\underline{\mathbf{w}}}^t, \dot{\underline{\mathbf{c}}}^t\right)^+ - J^\star}{J^\star} \ddot{\underline{w}}_{\underline{\mathbf{s}}^t}^t + \ddot{\underline{w}}_{\underline{\mathbf{s}}^t}^t - w_{\underline{\mathbf{s}}^t} \\
&= \frac{1}{J^\star}\left(\sum_{i=1}^{|\underline{\mathbf{s}}^t|}\left(c_{\underline{s}_i^t} - \dot{\underline{c}}_{\underline{s}_i^t}^t\right)\left(1 - w_{\underline{\mathbf{s}}^t[i-1]}\right) + \sum_{i=1}^{|\underline{\mathbf{s}}^t|} \dot{\underline{c}}_{\underline{s}_i^t}^t\left(\ddot{\underline{w}}_{\underline{\mathbf{s}}^t[i-1]}^t - w_{\underline{\mathbf{s}}^t[i-1]}\right)\right)
\end{aligned}$$

$$+\frac{J\left(\mathbf{s}^\star;\ddot{\underline{\mathbf{w}}}^t,\dot{\underline{\mathbf{c}}}^t\right)^+ - J^\star}{J^\star}\ddot{\underline{w}}^t_{\underline{\mathbf{s}}^t}+\ddot{\underline{w}}^t_{\underline{\mathbf{s}}^t}-w_{\underline{\mathbf{s}}^t}$$

$$\leq \frac{1}{J^\star}\left(\sum_{i=1}^{|\underline{\mathbf{s}}^t|}\left(c_{\underline{s}^t_i}-\dot{\underline{c}}^t_{\underline{s}^t_i}\right)\left(1-w_{\underline{\mathbf{s}}^t[i-1]}\right)+(n+J^\star)\left(\ddot{\underline{w}}^t_{\underline{\mathbf{s}}^t}-w_{\underline{\mathbf{s}}^t}\right)\right)+\frac{J\left(\mathbf{s}^\star;\ddot{\underline{\mathbf{w}}}^t,\dot{\underline{\mathbf{c}}}^t\right)^+ - J^\star}{J^\star}\ddot{\underline{w}}^t_{\underline{\mathbf{s}}^t}$$

$$=\Delta_{\mathbf{c}}\left(\underline{\mathbf{s}}^t\right)+\Delta_{\mathbf{w}}\left(\underline{\mathbf{s}}^t\right)+\frac{J\left(\mathbf{s}^\star;\ddot{\underline{\mathbf{w}}}^t,\dot{\underline{\mathbf{c}}}^t\right)^+ - J^\star}{J^\star}\ddot{\underline{w}}^t_{\underline{\mathbf{s}}^t},$$

where

$$\Delta_{\mathbf{c}}\left(\underline{\mathbf{s}}^t\right)\triangleq\frac{1}{J^\star}\sum_{i=1}^{|\underline{\mathbf{s}}^t|}\left(c_{\underline{s}^t_i}-\dot{\underline{c}}^t_{\underline{s}^t_i}\right)\left(1-w_{\underline{\mathbf{s}}^t[i-1]}\right)$$

and

$$\Delta_{\mathbf{w}}\left(\underline{\mathbf{s}}^t\right)\triangleq\frac{n+J^\star}{J^\star}\left(\ddot{\underline{w}}^t_{\underline{\mathbf{s}}^t}-w_{\underline{\mathbf{s}}^t}\right).$$

We now define events

$$\mathcal{A}^t\triangleq\left\{\forall i\in\underline{\mathbf{s}}^t,\ \frac{13.3\zeta\log(t)}{\underline{T}^{t-1}_{i,\mathbf{w}}}\leq\frac{J^\star\Delta\left(\underline{\mathbf{s}}^t\right)}{2n(n+J^\star)}\right\},$$
$$\mathcal{M}^t\triangleq\left\{\ddot{\underline{\mathbf{w}}}^t\geq\mathbf{w},\dot{\underline{\mathbf{c}}}^t\leq\mathbf{c}\right\},$$
$$\mathcal{N}^t_{\mathbf{c}}\triangleq\left\{\forall i\in\underline{\mathbf{s}}^t,\ \ c_i-\dot{\underline{c}}^t_i\leq\sqrt{\frac{2\zeta\log(t)}{\underline{T}^{t-1}_{i,\mathbf{c}}}}\right\},$$

and

$$\mathcal{N}^t_{\mathbf{w}}\triangleq\left\{\forall i\in\underline{\mathbf{s}}^t,\ \ddot{\underline{w}}^{t-1}_i-w_i\leq\sqrt{\frac{8\zeta w_i(1-w_i)\log(t)}{\underline{T}^{t-1}_{i,\mathbf{w}}}}+\frac{13.3\zeta\log(t)}{\underline{T}^{t-1}_{i,\mathbf{w}}}\right\}.$$

Using Lemma 5 of Chen et al. (2016) , with Jensen's inequality, we have that

$$\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_B-1}\Delta\left(\underline{\mathbf{s}}^t\right)\mathbb{I}\{\neg\mathcal{A}^t\}\right]\leq\sum_{i\in[n]}\mathbb{E}\left[\frac{32\log(\mathcal{T}_B+n)n(n+J^\star)\left(1+\log\left(\frac{n}{\Delta_{i,\min}}\right)\right)}{J^\star}\right]$$
$$\leq\sum_{i\in[n]}\frac{32\log(\mathbb{E}\left[\mathcal{T}_B\right]+n)n(n+J^\star)\left(1+\log\left(\frac{n}{\Delta_{i,\min}}\right)\right)}{J^\star},$$

Moreover, $\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_B-1}\Delta\left(\underline{\mathbf{s}}^t\right)\right]$ is bounded by $\Delta\mathbb{E}\left[\mathcal{T}_B\right]$ for $\underline{\mathbf{s}}^t$ with gap $\Delta\left(\underline{\mathbf{s}}^t\right)$ smaller than some Δ and bounded using previous inaquality for other $\underline{\mathbf{s}}^t$ (with $\frac{1}{\Delta_{i,\min}}\leq\frac{1}{\Delta}$). Maximizing over Δ then gives that $\Delta\mathbb{E}\left[\mathcal{T}_B\right]=a\log(\mathbb{E}\left[\mathcal{T}_B\right]+n)\left(1+\log(\frac{n}{\Delta})\right)$ with $a=32n^2\left(\frac{n+J^\star}{J^\star}\right)$. Thus $\Delta\mathbb{E}\left[\mathcal{T}_B\right]\geq a\log\left(\mathbb{E}\left[\mathcal{T}_B\right]+n\right)$ and $\frac{1}{\Delta}\leq\frac{\mathbb{E}[\mathcal{T}_B]}{a\log(\mathbb{E}[\mathcal{T}_B]+n)}$. We finally get

$$\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_B-1}\Delta\left(\underline{\mathbf{s}}^t\right)\mathbb{I}\{\neg\mathcal{A}^t\}\right]\leq 64n^2\left(\frac{n+J^\star}{J^\star}\right)\log(\mathbb{E}\left[\mathcal{T}_B\right]+n)\left(1+\log\left(\frac{\mathbb{E}\left[\mathcal{T}_B\right]}{32n\left(\frac{n+J^\star}{J^\star}\right)\log\left(\mathbb{E}\left[\mathcal{T}_B\right]+n\right)}\right)\right).$$

By Hoeffding's inequality (Fact 2), and Theorem 1 of Audibert et al. (2009, Fact 1), we have that $\mathcal{M}^t$ holds with probability at least $1-3n\beta(t)$. $\mathcal{N}^t_{\mathbf{c}}$ holds with probability at least $1-n\beta(t)$ by Hoeffding's inequality (Fact 2), and $\mathcal{N}^t_{\mathbf{w}}$ holds with probability at least $1-2n\beta(t)$ by Lemma 2. Thus,

$$\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_B-1}\Delta\left(\underline{\mathbf{s}}^t\right)\left(\mathbb{I}\{\neg\mathcal{M}^t\}+\mathbb{I}\{\neg\mathcal{N}^t_{\mathbf{c}}\}+\mathbb{I}\{\neg\mathcal{N}^t_{\mathbf{w}}\}\right)\right]\leq\frac{n}{J^\star}\sum_{t>0}\frac{\log t}{\log(1.1)}6nt^{-\frac{1.2}{1.1}}$$
$$\leq 7613\frac{n^2}{J^\star}.$$

Suppose now $\mathcal{M}^t$, $\mathcal{N}^t_{\mathbf{w}}$, $\mathcal{N}^t_{\mathbf{c}}$, and $\mathcal{A}^t$ hold. Moreover, we suppose that $\{\Delta\left(\underline{\mathbf{s}}^t\right)>0\}$ holds (if it doesn't, then the local regret is nul). We have that $J\left(\mathbf{s}^\star;\ddot{\underline{\mathbf{w}}}^t,\dot{\underline{\mathbf{c}}}^t\right)^+\leq J^\star$ thanks to $\mathcal{M}^t$. Thus,

$$\Delta\left(\underline{\mathbf{s}}^t\right)\leq\Delta_{\mathbf{w}}\left(\underline{\mathbf{s}}^t\right)+\Delta_{\mathbf{c}}\left(\underline{\mathbf{s}}^t\right)$$

$$
\begin{aligned}
&\leq -\Delta\left(\underline{\mathbf{s}}^t\right) + 2\Delta_{\mathbf{w}}\left(\underline{\mathbf{s}}^t\right) + 2\Delta_{\mathbf{c}}\left(\underline{\mathbf{s}}^t\right) \\
&= \frac{2}{J^\star}\sum_{i\in\underline{\mathbf{s}}^t}(n+J^\star)\cdot\left(\ddot{\underline{w}}_i^t - w_i - \frac{J^\star\Delta\left(\underline{\mathbf{s}}^t\right)}{2\left|\underline{\mathbf{s}}^t\right|(n+J^\star)}\right) + 2\Delta_{\mathbf{c}}\left(\underline{\mathbf{s}}^t\right) \\
&\leq \frac{2}{J^\star}\sum_{i\in\underline{\mathbf{s}}^t}(n+J^\star)\cdot\min\left\{\sqrt{\frac{8\zeta w_i(1-w_i)\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}} + \frac{13.3\zeta\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}} - \frac{J^\star\Delta\left(\underline{\mathbf{s}}^t\right)}{2n(n+J^\star)}, 1\right\} + 2\Delta_{\mathbf{c}}\left(\underline{\mathbf{s}}^t\right) \\
&\leq \frac{2}{J^\star}\sum_{i\in\underline{\mathbf{s}}^t}(n+J^\star)\cdot\min\left\{\sqrt{\frac{8\zeta w_i(1-w_i)\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}}, 1\right\} \\
&\quad + \frac{2}{J^\star}\sum_{i\in[|\underline{\mathbf{s}}^t|]}\min\left\{\sqrt{\frac{2\zeta\log(t)}{\underline{T}_{\underline{\mathbf{s}}_i^t,\mathbf{c}}^{t-1}}}, 1\right\}\left(1-w_{\underline{\mathbf{s}}^t[i-1]}\right).
\end{aligned} \tag{19}
$$

C.2 Use of Wang and Chen (2017) results

From this point, since $\left(1-w_{\underline{\mathbf{s}}^t[i-1]}\right)$ is the probability of getting cost feedback from arm i, the analysis given by Theorem 1 of Wang and Chen (2017) takes care of the second term, while the analysis of their Theorem 4 takes care of the first. We restate their results in Theorem 6 and 5, respectively. We want to use these results with $\mathcal{B}^t$ being the intersection of events assumed to hold, and with $M_i = \Delta_{i,\min}$. On the one hand, we apply second result of each theorem, for the first with

$$
\lambda = \frac{2(n+J^\star)}{J^\star} \text{ and } \Lambda_i^2 = \frac{8\zeta w_i(1-w_i)}{1.5}
$$

and for the second with

$$
\lambda = \frac{2}{J^\star} \text{ and } \Lambda_i^2 = \frac{2\zeta}{1.5}.
$$

We thus get

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_B-1}\Delta\left(s^t\right)\right] \leq &\frac{1}{J^\star}\left(71\sqrt{n}(J^\star+n)+31n\right)\sqrt{\mathbb{E}\left[\mathcal{T}_B\right]\log(\mathbb{E}\left[\mathcal{T}_B\right])} \\
&+ 64n^2\left(\frac{n+J^\star}{J^\star}\right)\log(\mathbb{E}\left[\mathcal{T}_B\right])\left(1+\log\left(\frac{\mathbb{E}\left[\mathcal{T}_B\right]}{32\left(\frac{n+J^\star}{J^\star}\right)\log(\mathbb{E}\left[\mathcal{T}_B\right])}\right)\right) \\
&+ \frac{\pi^2 n^2}{3J^\star}\left[\log_2\left(\frac{\mathbb{E}\left[\mathcal{T}_B\right]}{18\log\left(\mathbb{E}\left[\mathcal{T}_B\right]\right)}\right)\right]^+ + \frac{8n\left(1+n+J^\star\right)}{J^\star} \\
&+ 7613\frac{n^2}{J^\star}.
\end{aligned}
$$

On the other hand, we can multiply (19) by 4, to get that $2\Delta\left(\underline{\mathbf{s}}^t\right)$ is bounded by $A+B$, where

$$
A = \frac{8(n+J^\star)}{J^\star}\sum_{i\in\underline{\mathbf{s}}^t}\min\left\{\sqrt{\frac{8\zeta w_i(1-w_i)\log(t)}{\underline{T}_{i,\mathbf{w}}^{t-1}}}, 1\right\} - \sup_{i\in\underline{\mathbf{s}}^t}\Delta_{i,\min}
$$

and

$$
B = \frac{8}{J^\star}\sum_{i\in[|\underline{\mathbf{s}}^t|]}\min\left\{\sqrt{\frac{2\zeta\log(t)}{\underline{T}_{\underline{\mathbf{s}}_i^t,\mathbf{c}}^{t-1}}}, 1\right\}\left(1-w_{\underline{\mathbf{s}}^t[i-1]}\right) - \sup_{i\in\underline{\mathbf{s}}^t}\Delta_{i,\min}.
$$

We then apply first result of each theorem, for A with

$$
\lambda = \frac{4(n+J^\star)}{J^\star} \text{ and } \Lambda_i = \frac{8\zeta w_i(1-w_i)}{1.5}
$$

and for B with

$$
\lambda = \frac{4}{J^\star} \text{ and } \Lambda_i = \frac{2\zeta}{1.5}.
$$

We thus finally get

$$
\mathbb{E}\left[\sum_{t=1}^{\mathbb{E}[\mathcal{T}_B]}\Delta\left(s^t\right)\right] \leq \frac{1}{J^{\star 2}}\sum_{i\in[n]}\left(\frac{2048\zeta w_i(1-w_i)n(J^\star+n)^2}{\Delta_{i,\min}} + \frac{6144\zeta n}{\Delta_{i,\min}}\right)\log(\mathbb{E}\left[\mathcal{T}_B\right])
$$

$$+\frac{1}{J^\star}\sum_{i\in[n]} 32\log(\mathbb{E}[\underline{\tau}_B])n(n+J^\star)\left(1+\log\left(\frac{n}{\Delta_{i,\min}}\right)\right)$$
$$+\frac{\pi^2 n}{3J^\star}\sum_{i\in[n]}\left\lceil\log_2\left(\frac{8n}{J^\star\Delta_{i,\min}}\right)\right\rceil^+ + \frac{8n+4n(n+J^\star)}{J^\star}$$
$$+7613\frac{n^2}{J^\star}.$$

Final results hold with the bound obtained on $\mathbb{E}[\underline{\tau}_B]$ in (18). $\square$

D Wang and Chen (2017) results

We built on the results of Wang and Chen (2017) for combinatorial multi-armed bandits with probabilistically triggered arms (CMAB-T). In particular, Wang and Chen (2017) give expected regret bounds under specific assumptions that our setting satisfies. In CMAB-T, at each round t, the agent selects some action $\underline{\mathbf{s}}^t$ and a random subset of *arms* is triggered. The corresponding feedback is given to the agent which then goes to the next round. We denote by $\sigma(\underline{\mathbf{s}}^t)$, the set of arms that have a positive probability of being triggered if $\underline{\mathbf{s}}^t$ is selected, and $\underline{\sigma}(\underline{\mathbf{s}}^t) \subset \sigma(\underline{\mathbf{s}}^t)$ the random subset of arms i that are actually triggered and for which we maintain a counter $\underline{T}_i^t$. We restate two results of Wang and Chen (2017) that hold under following assumptions. Notice that we generalize their results to a random horizon and then use Jensen's inequality. For a round $t \geq 1$, we let $\mathcal{B}^t$ be any event. We let $\mathbf{M} \in (0,\infty)^n$ and for an action $\mathbf{s}$, $M^{\mathbf{s}} = \sup_{i\in\sigma(\mathbf{s})} M_i$. We let $\underline{\tau}$ a (possibly random) round, and $\mathbf{\Lambda} \in \mathbb{R}_+^n$. $H^{\mathbf{s}^t}$ and λ are non-negative numbers, $H^{\mathbf{s}^t}$ (deterministically) depends on $\mathbf{s}^t$. We write $\underline{\mathbf{s}}^t$ be the action chosen at round t.

Theorem 5. *Suppose that* $\forall \mathbf{s}, \forall i \in \sigma(\mathbf{s})$, $\mathbb{P}[i \in \underline{\sigma}(\mathbf{s})] = 1$. *If for all* t, *under event* $\mathcal{B}^t$,

$$H^{\underline{\mathbf{s}}^t} \leq \sum_{i\in\sigma(\underline{\mathbf{s}}^t)} \lambda \min\left\{2\Lambda_i\sqrt{\frac{1.5\log(t)}{\underline{T}_i^{t-1}}}, 1\right\},$$

then

$$\mathbb{E}\left[\sum_{t=1}^{\underline{\tau}-1}\left(2H^{\underline{\mathbf{s}}^t} - M^{\underline{\mathbf{s}}^t}\right)\mathbb{I}\{\mathcal{B}^t\}\right] \leq \sum_{i\in[n]}\mathbb{E}\left[\frac{48n\Lambda_i^2\lambda^2\log(\underline{\tau})}{M_i}\right] + 2\lambda n \leq \sum_{i\in[n]}\frac{48n\Lambda_i^2\lambda^2\log(\mathbb{E}[\underline{\tau}])}{M_i} + 2\lambda n \quad \textit{and}$$

$$\mathbb{E}\left[\sum_{t=1}^{\underline{\tau}} H^{\underline{\mathbf{s}}^t}\mathbb{I}\{\mathcal{B}^t\}\right] \leq 14\lambda\|\mathbf{\Lambda}\|_2\sqrt{n\mathbb{E}[\underline{\tau}]\log(\mathbb{E}[\underline{\tau}])} + 2\lambda n.$$

Theorem 6. *If for all* t, *under event* $\mathcal{B}^t$, $H^{\underline{\mathbf{s}}^t} \leq \sum_{i\in\sigma(\underline{\mathbf{s}}^t)}\mathbb{P}[i\in\sigma(\underline{\mathbf{s}}^t)]\,\lambda\min\left\{2\Lambda_i\sqrt{\frac{1.5\log(t)}{\underline{T}_i^{t-1}}}, 1\right\}$, *then*

$$\mathbb{E}\left[\sum_{t=1}^{\underline{\tau}-1}\left(2H^{\underline{\mathbf{s}}^t} - M^{\underline{\mathbf{s}}^t}\right)\mathbb{I}\{\mathcal{B}^t\}\right] \leq \sum_{i\in[n]}\frac{576n\Lambda_i^2\lambda^2\log(\mathbb{E}[\underline{\tau}])}{M_i} + \sum_{i\in[n]}\left\lceil\log_2\left(\frac{2\lambda n}{M_i}\right)\right\rceil^+ \times \frac{\pi^2\lambda n}{6} + 4n\lambda \quad \textit{and}$$

$$\mathbb{E}\left[\sum_{t=1}^{\underline{\tau}-1} H^{\underline{\mathbf{s}}^t}\mathbb{I}\{\mathcal{B}^t\}\right] \leq 12\lambda\|\mathbf{\Lambda}\|_2\sqrt{n\mathbb{E}[\underline{\tau}]\log(\mathbb{E}[\underline{\tau}])} + n\left\lceil\log_2\left(\frac{\mathbb{E}[\underline{\tau}]}{18\log(\mathbb{E}[\underline{\tau}])}\right)\right\rceil^+ \times \frac{\pi^2\lambda n}{6} + 2n\lambda.$$

E Proof of Theorem 3

Theorem 3. *For simplicity, let us assume that* n *is even and that* B *is a multiple of* n. *For any optimal online policy* π, *there is a sequential search-and-stop problem with* n *arms and budget* B *such that*

$$-4 + \frac{1}{28}\sqrt{\frac{B}{n}} \leq R_B(\pi) \lesssim \sqrt{B\log\left(\frac{B}{n}\right)}.$$

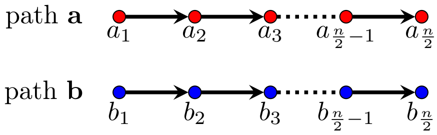

Figure 1: The DAG considered in Theorem 3

Proof. Let $0 < \varepsilon < 1/4$. We consider a DAG composed of two disjoint paths (Figure 1), both with $n/2$ nodes. We denote the two paths by **a** and **b**. We deterministically set all the costs to 1, $w_i = 0$ for $i \notin \{a_{\frac{n}{2}}, b_{\frac{n}{2}}\}$. All this information is given to the agent. Notice that this does not make the problem harder.

Now consider two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ defined by

$$\mathcal{D}_1: \quad w_{a_{\frac{n}{2}}} \triangleq \frac{1}{2} + \varepsilon, \quad w_{b_{\frac{n}{2}}} \triangleq \frac{1}{2} - \varepsilon \quad \text{and} \quad \mathcal{D}_2: \quad w_{a_{\frac{n}{2}}} \triangleq \frac{1}{2} - \varepsilon, \quad w_{b_{\frac{n}{2}}} \triangleq \frac{1}{2} + \varepsilon.$$

Notice that an optimal online policy does not modify its behavior during a round t, since after having seen $\underline{w}_i^t = 1$, continuing searching would only give information about cost distribution which is known by the problem definition, and no additional information about the rewards. Therefore, there is *an* optimal policy that selects some search $\mathbf{s}$ and perform $\mathbf{s}[\underline{\mathbf{w}}^t]$ over round t. Observe that $\mathbf{s}^\star = \mathbf{ab}$ for $\mathcal{D}_1$ and $\mathbf{s}^\star = \mathbf{ba}$ for $\mathcal{D}_2$. We have $J^\star = \frac{3}{4}n - \varepsilon n \geq \frac{1}{2}n$ for both $\mathcal{D}_1$ and $\mathcal{D}_2$.

We now show that we can restrict ourselves to policies that take searches in $\{\mathbf{ab}, \mathbf{ba}\}$.

- First, an optimal online policy does not select a search that would not include at least one of the leaves $\{a_{\frac{n}{2}}, b_{\frac{n}{2}}\}$ for a round. Therefore, it has a full information on $\underline{w}$. Indeed, such a search is noninformative and does not bring any reward while having a cost.

- Second, for a policy π that does not select a search in $\{\mathbf{ab}, \mathbf{ba}\}$ for some round t, we construct π' that acts like π except for this round t where it selects **ab** if π would see the leaf $a_{\frac{n}{2}}$ first, and **ba** otherwise, i.e., if π would first see the leaf $b_{\frac{n}{2}}$. Now compare both policies on the same realization of $\underline{\mathbf{w}}^1, \underline{\mathbf{w}}^2, \ldots$. We claim that the global reward of π' is never smaller than that of π. By symmetry, assume that π sees $a_{\frac{n}{2}}$ first within round t and thus π' selects **ab**.

  – If $\underline{w}_{a_{\frac{n}{2}}}^t = 1$ or $\left(\underline{w}_{b_{\frac{n}{2}}}^t = 1 \text{ and } \pi \text{ visits } b_{\frac{n}{2}} \text{ within round } t\right)$, both policies obtain the same reward of 1 within round t, but π' pays less than π.

  – If $\underline{w}_{b_{\frac{n}{2}}}^t = 1$ and π does not visit $b_{\frac{n}{2}}$ within round t, π gains 0 and pays at least $n/2$, whereas π' gains 1 and pays n within round t. Thus, the budget of π compared to π' is augmented by at most $n/2$, with which it can increase its reward by at most 1.

  The overall reward of π' remains higher than that of π for both cases.

A direct consequence of the restriction to $\{\mathbf{ab}, \mathbf{ba}\}$ is that $c_{\min} = n/2$, giving the upper bound in Theorem 3 by invoking the result of Theorem 2.

Now for a policy π using searches from $\{\mathbf{ab}, \mathbf{ba}\}$, we have

$$\begin{aligned}
F_B(\pi) &= \sum_{t=1}^{\infty} \mathbb{E}\left[\sum_{i \in \mathbf{s}^t} \mathbb{I}\{\underline{B}^t \geq 0,\ \underline{w}_i^t = 1\}\right] \leq \sum_{t=1}^{\infty} \mathbb{E}\left[\sum_{i \in \mathbf{s}^t} \mathbb{I}\{\underline{B}^{t-1} \geq 0,\ \underline{w}_i^t = 1\}\right] \\
&= \sum_{t \geq 1} \mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1} \geq 0,\ \underline{\mathbf{s}}^t \in \mathcal{S}^\star\} w_{\underline{\mathbf{s}}^t}\right] + \sum_{t \geq 1} \mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1} \geq 0,\ \underline{\mathbf{s}}^t \notin \mathcal{S}^\star\} w_{\underline{\mathbf{s}}^t}\right] \\
&= \frac{1}{J^\star} \sum_{t \geq 1} \mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1} \geq 0,\ \underline{\mathbf{s}}^t \in \mathcal{S}^\star\} \left(d(\underline{\mathbf{s}}^t) + (1 - w_{\underline{\mathbf{s}}^t}) c_{\underline{\mathbf{s}}^t}\right)\right] \\
&\quad + \frac{1}{J^\star} \sum_{t \geq 1} \mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1} \geq 0,\ \underline{\mathbf{s}}^t \notin \mathcal{S}^\star\} \left(d(\underline{\mathbf{s}}^t) + (1 - w_{\underline{\mathbf{s}}^t}) c_{\underline{\mathbf{s}}^t}\right)\right] - \sum_{t \geq 1} \mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1} \geq 0,\ \underline{\mathbf{s}}^t \notin \mathcal{S}^\star\} \Delta\left(\underline{\mathbf{s}}^t\right)\right] \\
&= \frac{1}{J^\star} \sum_{t \geq 1} \mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1} \geq 0\} \left(d(\underline{\mathbf{s}}^t) + (1 - w_{\underline{\mathbf{s}}^t}) c_{\underline{\mathbf{s}}^t}\right)\right] - \sum_{t \geq 1} \mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1} \geq 0,\ \underline{\mathbf{s}}^t \notin \mathcal{S}^\star\} \Delta\left(\underline{\mathbf{s}}^t\right)\right] \\
&\leq \frac{B + n}{J^\star} - \sum_{t \geq 1} \mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1} \geq 0,\ \underline{\mathbf{s}}^t \notin \mathcal{S}^\star\} \Delta\left(\underline{\mathbf{s}}^t\right)\right].
\end{aligned}$$

As a result we get

$$\begin{aligned}R_B(\pi) = F_B^\star - F_B(\pi) &\geq \frac{B-n}{J^\star} - \frac{B+n}{J^\star} + \sum_{t\geq 1}\mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1}\geq 0,\ \underline{\mathbf{s}}^t\notin\mathcal{S}^\star\}\Delta\left(\underline{\mathbf{s}}^t\right)\right]\\ &= -\frac{2n}{J^\star} + \sum_{t\geq 1}\mathbb{E}\left[\mathbb{I}\{\underline{B}^{t-1}\geq 0,\ \underline{\mathbf{s}}^t\notin\mathcal{S}^\star\}\Delta\left(\underline{\mathbf{s}}^t\right)\right]\cdot\end{aligned}$$

Since we restrict π to take a search in $\{\mathbf{ab},\mathbf{ba}\}$, we have a single gap (the same for $\mathcal{D}_1$ and $\mathcal{D}_2$)

$$\Delta = \frac{\frac{n}{2}\left(\frac{1}{2}-\varepsilon\right)+n\left(\frac{1}{2}+\varepsilon\right)}{\frac{n}{2}\left(\frac{1}{2}+\varepsilon\right)+n\left(\frac{1}{2}-\varepsilon\right)} - 1 = \frac{1.5+\varepsilon}{1.5-\varepsilon} - 1 = \frac{2\varepsilon}{1.5-\varepsilon} \geq \frac{4\varepsilon}{3}\cdot \tag{20}$$

Furthermore we can bound the number of rounds from below by B/n. To proceed we use high-probability Pinsker inequality (Tsybakov, 2009, Lemma 2.6).

Fact 5 (high-probability Pinsker inequality)**.** *Let P and Q be probability measures on the same measurable space, and let $\mathcal{A}$ be an event. Then*

$$P(\mathcal{A}) + Q(\neg\mathcal{A}) \geq \frac{1}{2}\exp\left(-KL\left(P\|Q\right)\right),$$

where KL is the Kullback-Leibler divergence.

We let $R_{1,B}(\pi)$ be the regret of π for distribution $\mathcal{D}_1$ and similarly, $R_{2,B}(\pi)$ for $\mathcal{D}_2$. If $\mathbb{P}_1$ and $\mathbb{P}_2$ denote the probability when random variable are samples from $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively, we have

$$\begin{aligned}\max\left\{R_{1,B}(\pi), R_{2,B}(\pi)\right\} &\geq \frac{R_{1,B}(\pi)+R_{2,B}(\pi)}{2}\\ &\geq -\frac{2n}{J^\star} + \frac{\Delta}{2}\sum_{t=1}^{B/n}\left(\mathbb{P}_1\left[\underline{B}^{t-1}\geq 0, \mathbf{s}^t = \mathbf{ba}\right] + \mathbb{P}_2\left[\underline{B}^{t-1}\geq 0, \mathbf{s}^t = \mathbf{ab}\right]\right)\\ &\geq -\frac{2n}{J^\star} + \frac{\varepsilon}{3}\sum_{t=1}^{B/n}\exp\left(-\mathrm{KL}\left(\mathcal{D}_1^{\otimes t}\|\mathcal{D}_2^{\otimes t}\right)\right) \qquad (21)\\ &= -\frac{2n}{J^\star} + \frac{\varepsilon}{3}\sum_{t=1}^{B/n}\exp\left(-t\mathrm{KL}\left(\mathcal{D}_1\|\mathcal{D}_2\right)\right),\end{aligned}$$

where (21) is due to Fact 5 and (20). Then,

$$\begin{aligned}\mathrm{KL}\left(\mathcal{D}_1\|\mathcal{D}_2\right) &= \left(\frac{1}{2}+\varepsilon\right)\log\left(\frac{\frac{1}{2}+\varepsilon}{\frac{1}{2}-\varepsilon}\right) + \left(\frac{1}{2}-\varepsilon\right)\log\left(\frac{\frac{1}{2}-\varepsilon}{\frac{1}{2}+\varepsilon}\right)\\ &\leq 2\varepsilon\left(\frac{\frac{1}{2}+\varepsilon}{\frac{1}{2}-\varepsilon}\right) - 2\varepsilon\left(\frac{\frac{1}{2}-\varepsilon}{\frac{1}{2}+\varepsilon}\right) = \frac{4\varepsilon^2}{\frac{1}{4}-\varepsilon^2} \leq \frac{64}{3}\varepsilon^2 \qquad \text{(because } \log(x)\leq x-1\text{)}.\end{aligned}$$

Thus, with $J^\star \geq \frac{n}{2}$, we have

$$\begin{aligned}\max\left\{R_{1,B}(\pi), R_{2,B}(\pi)\right\} \geq -4 + \frac{\varepsilon}{3}\sum_{t=1}^{B/n}\exp\left(-\frac{64}{3}t\varepsilon^2\right) &\geq -4 + \frac{\varepsilon\left(1-\exp\left(-\frac{64}{3}\frac{B\varepsilon^2}{n}\right)\right)}{3\left(\exp\left(\frac{64}{3}\varepsilon^2\right)-1\right)}\\ &\geq -4 + \frac{1-\exp\left(-\frac{64}{3}\frac{B\varepsilon^2}{n}\right)}{64\varepsilon}\\ &\geq -4 + \min\left\{\frac{1}{128\varepsilon}, \frac{\varepsilon B}{6n}\right\}.\end{aligned}$$

□

Taking $\varepsilon = \sqrt{(6n)/(128B)}$, the lower bound becomes

$$\max\left\{R_{1,B}(\pi), R_{2,B}(\pi)\right\} \geq -4 + \sqrt{\frac{B}{768n}} \geq -4 + \frac{1}{28}\sqrt{\frac{B}{n}}.$$

F Additional experiments

We show here further experiments that underline a strong asymptotic difference between `CUCBV` and `CUCB`. The setting is the same as in the experiments in Section 5, but this time we fix $n = 10$ and vary the size of an optimal search. In particular, we let the size of the optimal search be 1 in Figure 3, left, and 10 in Figure 3, right. Notice, that in this experiment, when there is an optimal search of size 1, both algorithms suffer a quite large expected regret compared to the case where there is an optimal search of size 10. This suggests that the problem is simpler in the first case. On the other hand, in both cases, `CUCB` and `CUCBV` are quite comparable for a low budget, but a significant difference arises for higher budget values.

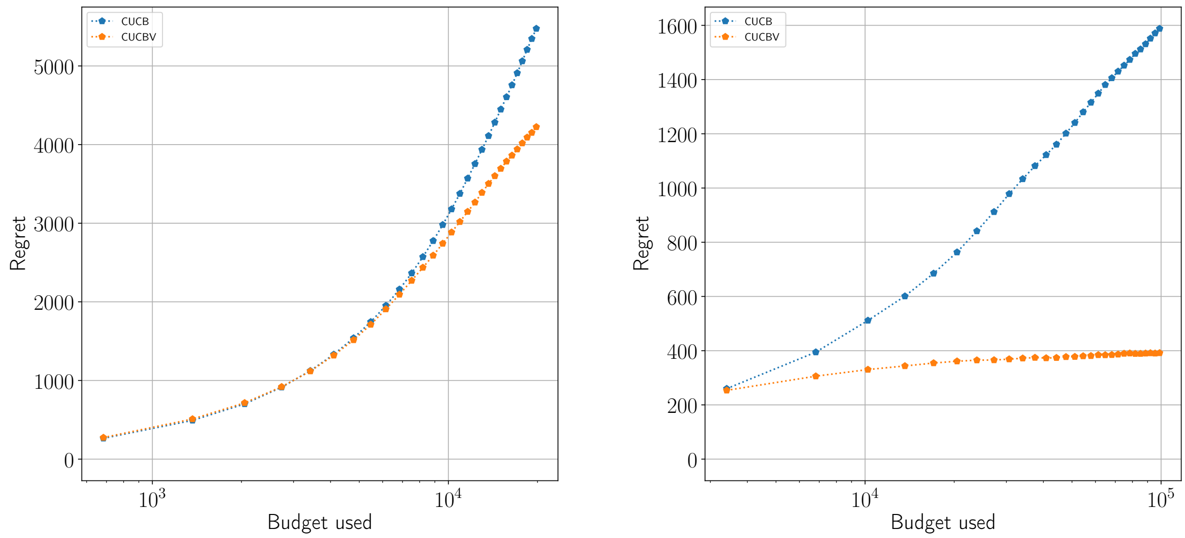

Figure 3: Regret for sequential search-and-stop. **Left:** $|\mathbf{s}^\star| = 1$. **Right:** $|\mathbf{s}^\star| = 10$.