Learning in Combinatorial Optimization: What and How to Explore

Sajad Modaresi
Duke University, sajad.modaresi@duke.edu

Denis Sauré
University of Chile, dsaure@dii.uchile.cl

Juan Pablo Vielma
MIT Sloan School of Management, jvielma@mit.edu

We study dynamic decision-making under uncertainty when, at each period, the decision maker faces a different instance of a combinatorial optimization problem. Instances differ in their objective coefficient vectors, which are unobserved prior to implementation. These vectors, however, are known to be random draws from an initially unknown distribution with known range. By implementing different solutions, the decision maker extracts information about the underlying distribution, but at the same time experiences the cost associated with said solutions. We show that resolving the implied *exploration vs. exploitation* trade-off efficiently is related to solving a *Lower Bound Problem* (LBP), which simultaneously answers the questions of *what* to explore and *how* to do so. For that, we establish a fundamental limit on the asymptotic performance of any admissible policy that is proportional to the optimal solution to LBP, and construct policies whose performance exhibits the same dependence on LBP. In addition, we construct more practical policies, amenable to implementation, which adaptively construct and solve a proxy for LBP at an exponentially decreasing frequency, thus enabling its implementation in real-time. We provide strong evidence of the practical tractability of said proxy and propose an oracle polynomial-time heuristic solution. We extensively test performance of our proposed policies and show that they significantly outperform relevant benchmark in the long-term and are competitive in the short-term.

Key words: Combinatorial Optimization, Multi-Armed Bandit, Mixed-Integer Programming.

1. Introduction

Motivation. Traditional solution approaches to many operational problems are based on combinatorial optimization problems and typically involve instantiating a deterministic mathematical program, whose solution is implemented repeatedly through time: nevertheless, in practice, instances are not usually known in advance. When possible, parameters characterizing said instances are estimated *off-line*, either by using historical data or from direct observation of the (idle) system. Unfortunately, off-line estimation is not always possible as, for example, historical data (if available) might only provide partial information pertaining previously implemented solutions. Consider, for instance, shortest path problems in network applications: repeated implementation of a given path might reveal cost information about arcs on such a path, but might provide no further information

about costs of other arcs in the graph. Similar settings arise, for example, in other network applications (e.g., tomography and connectivity) in which feedback about cost follows from instantiating and solving combinatorial problems such as spanning and Steiner trees.

Alternatively, parameter estimation might be conducted *on-line* using feedback associated with implemented solutions, and revisited as more information about the system's primitives becomes available. In doing so, one must consider the interplay between the performance of a solution and the feedback generated by its implementation: some parameters might only be reconstructed by implementing solutions that perform poorly (relative to the optimal solution). This is an instance of the *exploration vs. exploitation* trade-off that is at the center of many dynamic decision-making problems under uncertainty, and as such, it can be approached through the multi-armed bandit paradigm (Robbins 1952). However, there are salient features that distinguish our setting from the traditional bandit. In particular, the combinatorial structure induces correlation in the performance of different solutions, hence there might be multiple ways of estimating some parameters, each using feedback from a different set of solutions, and thus experiencing different performance. Also, because solutions are not upfront identical, the underlying combinatorial optimization problem might be invariant to changes in certain parameters, hence not all parameters might need to be estimated to solve said problem.

Unfortunately, the features above either prevent or discourage the use of known bandit algorithms. First, in the combinatorial setting, traditional algorithms might not be implementable as they would typically require solving an instance of the underlying combinatorial problem between decision epochs, for which, depending on the application, there might not be enough computational resources. Second, even with enough computational resources, such algorithms would typically call for implementing each feasible solution at least once, which in the settings of interest might take a prohibitively large number of periods and also result in poor performance.

Main Objectives and Assumptions. A thorough examination of the arguments behind results in the traditional bandit setting reveals that their basic principles are still applicable to the combinatorial setting. Thus, our objective can be seen as interpreting said principles and adapting them to the combinatorial setting with the goal of *developing efficient policies that are amenable to implementation*. In doing so, we also aim at understanding how the specifics of the underlying combinatorial problem affect achievable performance. For this, we consider settings in which an agent must implement solutions to a series of arriving instances of a given combinatorial problem (i.e., whose feasible solutions are structured subsets of a *ground* set), and there is initial uncertainty about said instances. In particular, we assume that instance uncertainty is restricted to cost-coefficients in the objective function. Hence, the feasible region is the same for each instance and known upfront by the agent. In this regard, we assume that cost-coefficients vary among

instances, but they are random draws from a common time-invariant distribution, which is initially unknown to the agent, except for its range. By implementing a solution, the agent receives *partial* feedback that depends on the solution implemented. Without loss of generality, we assume that the underlying combinatorial problem is that of cost minimization. Following the bulk of the bandit literature, we measure performance in terms of the cumulative *regret*, i.e., the cumulative cost incurred in excess of that of an oracle with prior knowledge of the cost distribution.

Main Contributions. From a methodological perspective, our contributions are as follows:

i) **We develop an asymptotically near-optimal policy:** We prove that no policy can achieve an asymptotic (on N which is the total number of instances) regret lower than $L \ln N$, where L is the optimal value of an instance-dependent optimization problem which we denote the *Lower Bound Problem* (LBP). To the best of our knowledge, this is the first lower bound for the stochastic combinatorial setting. Then, we show that an asymptotic performance of $L \ln N$ is not attainable in general, and develop a policy that admits an asymptotic (on N) regret upper bound of $(L + \gamma\ C)(\ln N^{1+\varepsilon})$, where L is the same constant as in the lower bound, C is the optimal value of another instance-dependent optimization problem, and γ and ε are arbitrary positive constants. Finally, we show that a performance of $L(\ln N^{1+\varepsilon})$ is attainable when the underlying combinatorial problem is a matroid.

ii) **We develop an implementable variant of the near-optimal policy:** The policies alluded above reconstruct LBP adaptively over time. However, this is often an exponentially-sized problem that is unlikely to be solvable in practice. For this reason, we simplify this optimization problem into a variant that distills two of its main goals: the determination of *what* should be explored and *how* to do so. The resulting variant is a combinatorial optimization problem which we denote the *Optimality Cover Problem* (OCP). While OCP is still an exponentially-sized problem, we provide strong evidence that it can be solved in practice. In particular, we show that OCP can be formulated as a Mixed-Integer Programming (MIP) problem that can be effectively tackled by state-of-the-art solvers. We also develop an oracle polynomial-time heuristic for OCP that uses a solution oracle for the underlying combinatorial optimization problem. Finally, we show that focusing exploration on the solution to OCP results in an asymptotic performance guarantee that is similar to that of the near-optimal policy.

iii) **We show that OCP leads to a policy with a finite-time performance that significantly outperforms other existing policies:** A key to the efficiency of the LBP- and OCP-based policies is that they do not necessarily explore every solution of the combinatorial problem (e.g., they do not necessarily explore all paths in a shortest path problem). However, they do explore all ground elements of the combinatorial optimization problems (e.g., all arcs in a shortest path problem) with an arbitrarily small but positive frequency. This exploration

is necessary to achieve convergence of certain estimates necessary for their asymptotic performance guarantees. However, such convergence is unnecessary (and usually unachievable) when evaluating finite-time performance. For this reason, we consider a simple policy that explores as dictated by OCP and rarely explores every ground element. Through extensive computational experiments we show that such policy significantly outperforms benchmark policies in both long- and short-term performance, even when OCP is solved with the oracle polynomial-time heuristic.

The optimal $\ln N$ scaling of the regret is well-known in the bandit literature (Lai and Robbins 1985) and can even be achieved in the combinatorial setting by traditional algorithms. The regret of such algorithms, however, is proportional to the number of solutions, which for combinatorial settings is typically exponential, which suggests that the dependence on N might not be the major driver of performance, especially in the finite time. Considering this, in this work we aim at studying the optimal scaling of the regret with respect to the combinatorial aspects of the setting. In doing so, our policies sacrifice the optimal dependence on N (by adding a sub-logarithmic term) to present performance bounds clearly, in terms of the underlying combinatorial aspects of the problem, thus facilitating their comparison to the fundamental performance limit. In this regard, through the LBP- and OCP-based policies and the associated lower bound we show that efficient exploration is achieved when exploration is focused on a "critical" subset of elements of the ground set. Our results speak of a fundamental principle in active learning, which is somewhat obscured in the traditional bandit setting: that of only exploring what is necessary to reconstruct the solution to the underlying problem, and doing so at the least possible cost.

The Remainder of the Paper. Next, we review related work. Section 3 formulates the problem, and reviews ideas from the classic bandit setting, interpreting them in our setting. In Section 4 we establish a fundamental limit on performance and propose a policy whose performance exhibits the *right* dependence with respect to the combinatorial structure of the setting. Section 5 presents a more practical policy, whose performance is similar to that of the policy in Section 4, but that is amenable to implementation. In Section 6 we discuss practical policy implementation, and Section 7 illustrates the results in the paper by means of numerical experiments. Finally, Section 8 presents extensions and concluding remarks. Proofs and supporting material are relegated to Appendices A and B.

2. Literature Review

Classical Bandit Settings. Introduced in Thompson (1933) and Robbins (1952), the multi-armed bandit setting is a classical framework for dynamic decision-making under uncertainty. In its basic formulation a gambler maximizes cumulative reward by pulling arms of a slot machine sequentially

over time when limited prior information on reward distributions is available. The gambler faces the classical exploration vs. exploitation trade-off: either pulling the arm thought to be the "best" at the risk of failing to actually identify such an arm, or trying other arms which allows identifying the best arm but hampers reward maximization.

The seminal work of Gittins (1979) shows that, for the case of independent and discounted arm rewards, and infinite horizon, the optimal policy is of the index type. Unfortunately, index-based policies are not always optimal (see Berry and Fristedt (1985), and Whittle (1982)) or cannot be computed in closed-form. In their seminal work, Lai and Robbins (1985) study asymptotically efficient policies for the undiscounted case. They establish a fundamental limit on achievable performance, which implies the (asymptotic) optimality of the order $\ln N$ dependence in the regret (see Kulkarni and Lugosi (1997) for a finite-sample minimax version of the result). Our proof of efficiency is based on the change of measure argument in this paper: see the discussion in Section 4.1. In the same setting, Auer et al. (2002) introduces the celebrated index-based UCB1 policy, which is both efficient and implementable. We revisit their results in the next section.

Envisioning each solution as an arm, our setting corresponds to a bandit with correlated rewards (and many arms): only a few papers address this case (see e.g., Ryzhov and Powell (2009) and Ryzhov et al. (2012)). Unlike in these papers, our focus is on asymptotic efficiency. Alternatively, envisioning each ground element as an arm, our setting can be seen as a bandit with multiple simultaneous pulls. Anantharam et al. (1987) extend the fundamental bound of Lai and Robbins (1985) to this setting and propose efficient allocations rules: see also Agrawal et al. (1990). Our setting imposes a special structure on the set of feasible simultaneous pulls, which prevents us from applying known results.

Bandit Problems with a Large Set of Arms. Bandit settings with a large number of arms have received significant attention in the last decade. In these settings, arms are typically endowed with some known structure that is exploited to improve upon the performance of traditional bandit algorithms.

A first strain of literature considers settings with a continuous set of arms, where exploring all arms is not feasible. Agrawal (1995) studies a multi-armed bandit in which arms represent points in the real line and their expected rewards are continuous functions of the arms. Mersereau et al. (2009) and Rusmevichientong and Tsitsiklis (2010) study bandits with possibly infinite arms when expected rewards are linear functions of an (unknown) scalar and a vector, respectively. Our paper also relates to the literature on linear bandit models (see e.g., Abernethy et al. (2008) and Dani et al. (2008)) as the model we study is a linear stochastic bandit with a finite (but combinatorial) number of arms. Our work differs from the bulk of this literature in the type of feedback obtained from implementing a solution (see Section 8 for a discussion about alternative feedback settings).

In a more general setting, Kleinberg et al. (2008) consider the case where arms form a metric space, and expected rewards satisfy a Lipschitz condition. See Bubeck et al. (2011) for a review of work in *continuum* bandits.

Bandit problems with some combinatorial structure have been studied in the context of assortment planning: in Rusmevichientong et al. (2010) and Sauré and Zeevi (2013) product assortments are implemented in sequence and (non-linear) rewards are driven by a choice model with initially unknown parameters. See Caro and Gallien (2007) for a similar formulation with linear rewards.

Gai et al. (2012) study combinatorial bandits when the underlying problem belongs to a restricted class, and extend the UCB1 policy to this setting. Their policy applies to the more general setting we study, and is used as a benchmark in our numerical experiments. They establish a performance guarantee that exhibits the right dependence on N, but that is expressed in terms of a polynomial of the size of the ground set A. We show that optimal performance dependence on the ground set is instead tied to the structure of the underlying combinatorial problem in a non-trivial manner.

Concurrent to our work, two papers examine the combinatorial setting: Chen et al. (2013) provide a tighter performance bound for the UCB1-type policy of Gai et al. (2012) applied to the general combinatorial case (still expressed as a polynomial of the size of the ground set); also, Liu et al. (2012) develop a version of our Cover-based policy (see Section 5) for network optimization problems (their ideas can be adapted to the general case as well) but under a different form of feedback. Their policy collects information through implementation of solutions in a *barycentric spanner* of the solution set, which in our feedback setting could be set as a solution cover: see further discussion in Section 8. Probable performance of their policy is essentially that of a static Cover-based policy, which is (asymptotically) always worse than or equal to that of its dynamic version, and might be arbitrarily worse than the OCP-based policy (see Proposition 1).

Drawing ideas from the literature of prediction with expert advice (see e.g., Cesa-Bianchi and Lugosi (2006)), Cesa-Bianchi and Lugosi (2012) study an adversarial combinatorial bandit where arms belong to a given finite set in $\mathbb{R}^d$ (see Auer et al. (2003) for a description of the adversarial bandit setting). Our focus instead is on *stochastic* (non-adversarial) settings. In this regard, our work leverages the additional structure imposed in the stochastic setting to developing efficient policies whose probable performance exhibits the "right" constant accompanying the $\ln N$ term.

Online Subset Selection. Broadly speaking, our work contributes to the literature of online learning with combinatorial number of alternatives. There are several studies that focus on similar online learning problems from the ranking and selection perspective. Ryzhov and Powell (2011) study information collection in settings where the decision maker selects individual arcs from a directed graph, and Ryzhov and Powell (2012) consider a more general setting where selection is made from a polyhedron. (See also Ryzhov et al. (2012).) The ideas in Harrison and Sunar (2013)

regarding selection of efficient learning mechanisms relate to the insight derived here. Also, see Jones et al. (1998), and the references within, for related work in the global optimization literature.

Finally, the concept of the optimality cover problem (OCP) is similar to (but not the same as) the idea behind the "optimistic constraint propagation" algorithm of Wen and Van Roy (2013). They propose to select a solution to implement based on the most optimistic feasible outcome subject to constrains. However, the model we study is different in that we consider a stochastic setting while Wen and Van Roy (2013) study a deterministic case.

3. Combinatorial Formulation vs. Traditional Bandits

3.1. Problem Formulation

Model Primitives and Basic Assumptions. We consider the problem of an agent who must implement solutions to a series of instances of a given combinatorial optimization problem. Without loss of generality, we assume that such a problem is that of cost minimization. Instances are presented sequentially through time, and we use n to index them according to their arrival times, so $n=1$ corresponds to the first instance, and $n=N$ to the last, where N denotes their (possibly unknown) total number. Each instance is uniquely characterized by a set of cost-coefficients, i.e., instance $n\in\mathbb{N}$ is associated with a cost-coefficient vector $B_n := (b_{a,n} : a\in A)\in\mathbb{R}^{|A|}$, a set of feasible solutions $\mathcal{S}$, and the full instance is defined as $f(B_n)$, where

$$f(B) : z^*(B) := \min\left\{\sum_{a\in S} b_a : S\in\mathcal{S}\right\} \quad B\in\mathbb{R}^{|A|}, \tag{1}$$

$\mathcal{S}$ is a family of subsets of elements of a given ground set A (e.g., arcs forming a path), S is the decision variable, and b_a is the *cost* associated with a ground element $a\in A$. We let $\mathcal{S}^*(B)$ be the set of optimal solutions to (1) and $z^*(B)$ be its optimal objective value (cost).

We assume that each element $b_{a,n}\in B_n$ is a random variable, independent and identically distributed across instances, and independent of other components in B_n. We let $F(\cdot)$ denote the common distribution of B_n for $n\in\mathbb{N}$, which we assume is initially *unknown*. We assume, however, that upper and lower bounds on its range are known upfront. That is, it is known that $l_a\le b_{a,n}\le u_a$ a.s. (with $l_a<u_a$), for all $a\in A$ and $n\in\mathbb{N}$. So as to ensure costs are indeed random, we further assume that

$$\mathbb{E}_F\{B_n\}\in\mathcal{B} := \prod_{a\in A}(l_a,u_a),$$

where $\mathbb{E}_F\{\cdot\}$ denotes expectation with respect to F. Furthermore, while our general approach and some of our results hold in more general settings, we assume for simplicity that the distributions of $b_{a,n}$ are absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}$. Let $\hat{b}_{a,n}$ denote a realization of the random variable $b_{a,n}$ and define $\hat{B}_n := \left(\hat{b}_{a,n} : a\in A\right)$.

We assume that, in addition to not knowing $F(\cdot)$, the agent does not observe $\hat{B}_n$ prior to implementing a solution. Instead, we assume that $\hat{B}_n$ is only revealed *partially* and *after* a solution is implemented. More specifically, we assume that if solution $S_n \in \mathcal{S}$ is implemented, only cost-coefficients associated with ground elements in S_n i.e., $\{\hat{b}_{a,n} : a \in S_n\}$ are observed by the agent and only after the corresponding cost is incurred.

REMARK 1. Note that the $b_{a,n}$'s are independent of S_n. While this accommodates several applications such as shortest path, Steiner tree, and knapsack problems, it may not accommodate applications such as assortment selection problem with discrete choice.

Finally, we assume that the agent is interested in minimizing the expected cumulative cost associated with implementing a sequence of solutions.

Full-Information Problem and Regret. Consider the case of a clairvoyant agent with prior knowledge about $F(\cdot)$. Such an agent, while still not capable of anticipating B_n, can solve for the solution that minimizes the expected cumulative cost: for instance $n \in \mathbb{N}$ (by the linearity of the objective function), it is optimal to implement $S_n \in \mathcal{S}^*(\mathbb{E}_F\{B_n\})$. That is, always implementing a solution to the problem where costs equal their expected values is the best among all non-anticipating (see below) solution sequences.

In practice, the agent does not know F upfront, hence no admissible policy incurs costs below those incurred by the clairvoyant agent, in expectation. Thus, we measure the performance of a policy in terms of its expected *regret*: let $\pi := (S_n)_{n=1}^{\infty}$ denote a non-anticipating policy, where $S_n : \Omega \to \mathcal{S}$ is a $\mathcal{F}_n$-measurable function that maps the available "history" at time n, $\mathcal{F}_n := \sigma(\{b_{a,m} : a \in S_m\,,\, m < n\})$, to a solution in $\mathcal{S}$. Given F and N, the expected regret of a policy π is

$$R^{\pi}(F,N) := \left(\sum_{n=1}^{N} \mathbb{E}_F\left\{\sum_{a\in S_n} b_{a,n}\right\}\right) - N\ z^*\left(\mathbb{E}_F\{B_n\}\right).$$

The regret represents the additional expected cumulative cost incurred by policy π relative to that incurred by a clairvoyant agent that knows F upfront (note that regret is always non-negative).

REMARK 2. Although the regret also depends on the combinatorial optimization problem through $\mathcal{S}$, we omit this dependence to simplify the notation.

Our exposition benefits from connecting the regret to the number of instances in which suboptimal solutions are implemented. To make this connection explicit, consider an alternative representation of the regret. For $S \in \mathcal{S}$ and $B \in \mathcal{B}$, let Δ_S^B denote the optimality gap associated with implementing solution S, when mean costs are given by B. That is,

$$\Delta_S^B := \sum_{a\in S} b_a - z^*(B).$$

(Note that the optimality gap associated with $S^* \in \mathcal{S}^*(B)$ is zero.) For $S \in \mathcal{S}$, let

$$T_n(S) := |\{m < n : S_m = S\}|$$

denote the number of times that the agent has implemented solution $S_m = S$ prior to instance n. Similarly, for $a \in A$, let

$$\tilde{T}_n(a) := |\{m < n : a \in S_m\}|$$

denote the number of times that the agent has selected element a prior to instance n (henceforth, we say ground element $a \in A$ is selected or tried at instance n if $a \in S_n$). Note that $\tilde{T}_n(a)$ and $T_n(S)$ are $\mathcal{F}_n$-adapted for all $a \in A$, $S \in \mathcal{S}$, and $n \in \mathbb{N}$. Using this notation we have that

$$R^\pi(F,N) = \sum_{S \in \mathcal{S}} \Delta_S^{\mathbb{E}_F\{B_n\}} \, \mathbb{E}_F\left\{T_{N+1}(S)\right\}. \tag{2}$$

3.2. Known Results for the Non-Combinatorial Bandit

Traditional multi-armed bandits correspond to settings where $\mathcal{S}$ is formed by ex-ante *identical* singleton subsets of A (i.e., $\mathcal{S} = \{\{a\} : a \in A\}$, l_a and u_a equal for all $a \in A$), thus the combinatorial structure is absent.

We restrict attention to *consistent* policies: a policy π is said to be *consistent* if $R^\pi(F,N) = o(N^\alpha)$ for all $\alpha > 0$, for every regular F (see below for a definition of regularity); this avoids considering policies that perform well in a particular setting at the expense of performing poorly in others. In this setting, the seminal work of Lai and Robbins (1985) (henceforth, LR) establishes an asymptotic lower bound on the regret attainable by any *consistent* policy when F is regular.

Definition 1 (Regularity). We say F is regular if $\mathbb{E}_F\{B_n\} \in \mathcal{B}$ and the density of $b_{a,n}$: (i) can be parametrized by its mean θ_a, and thus we denote it by $f_a(\cdot;\theta_a)$; (ii) $0 < I_a(\theta_a, \lambda_a) < \infty$ for all $l_a < \lambda_a < \theta_a < u_a$, $a \in A$, where $I_a(\theta_a, \lambda_a)$ denotes the Kullback-Leibler divergence (see e.g., Cover and Thomas (2006)) between $f_a(\cdot;\theta_a)$ and $f_a(\cdot;\lambda_a)$; and (iii) $I_a(\theta_a, \lambda_a)$ is continuous in $\lambda_a < \theta_a$ for all $\theta_a \in (l_a, u_a)$.

LR show that consistent policies must explore (pull) each element (arm) in A at least on order $\ln N$ times, hence, by (2), their regret must also be of at least order $\ln N$.

Theorem 1 **(Lai and Robbins 1985)**. *Suppose that F is regular and that $\mathcal{S} = \{\{a\} : a \in A\}$, then for any consistent policy π and for any $a \in A$ we have*

$$\liminf_{N\to\infty} \mathbb{P}_F\left\{\frac{\tilde{T}_N(a)}{\ln N} \geq K_a\right\} = 1, \tag{3}$$

where K_a is a positive finite constant depending on F. In addition, we have

$$\liminf_{N\to\infty} \frac{R^\pi(F,N)}{\ln N} \geq \sum_{a \in A} \Delta_{\{a\}}^{\mathbb{E}_F\{B_n\}} K_a. \tag{4}$$

In the above, K_a is the inverse of the Kullback-Leibler divergence between the original distribution F and a distribution F_a that makes a optimal (which always exists because arms are ex-ante identical). The argument behind the result above is the following: in order to distinguish F from distribution F_a, consistent policies cannot restrict the exploration of any given arm to a finite number of times (independent of N), and must explore all arms periodically. Thus, broadly speaking, balancing the exploration vs. exploitation trade-off in the traditional setting narrows down to answering *when* (or how frequently) to explore each element $a \in A$. (The answer to this question is given by LR's $\ln N/N$ exploration frequency).

Different policies have been shown to attain the logarithmic dependence on N in (4), and in general, there is a trade-off between computational complexity and larger leading constants (multiplying the $\ln N$ term). For instance, the index-based UCB1 algorithm introduced by Auer et al. (2002) is simple to compute and provides a finite-time theoretical performance guarantee.

THEOREM 2 **(Auer et al. 2002)**. *Suppose* $\mathbb{E}_F\{B_n\} \in (0,1)^{|A|}$. *Let* $\mathcal{S} = \{\{a\} : a \in A\}$ *and for each* $a \in A$, *let* $\tilde{K}_a := 8/(\Delta_{\{a\}}^{\mathbb{E}_F\{B_n\}})^2$. *Then the expected regret of policy UCB1 after* N *plays is such that*

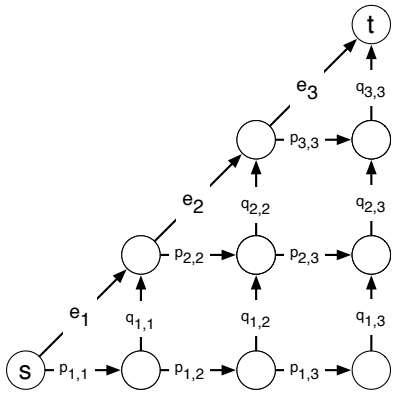
$$\frac{R^\pi(F,N)}{\ln N} \leq \sum_{a \in A} \Delta_{\{a\}}^{\mathbb{E}_F\{B_n\}} \tilde{K}_a + O(1/\ln N). \tag{5}$$

Furthermore, $K_a \leq \tilde{K}_a/16$.

3.3. Incorporating Combinatorial Aspects

A straightforward way to adapt algorithms designed for traditional multi-armed bandits to the combinatorial setting is to envision each solution in $\mathcal{S}$ as an arm. That is, the combinatorial setting can be seen as a traditional bandit with a combinatorial number of arms, and where arm rewards are correlated. Then, implementing off-the-shelf traditional index-based policies such as UCB1 for the combinatorial setting simply requires computing the index for each solution in $\mathcal{S}$ and implementing it. However, this approach has two important disadvantages in the combinatorial setting where $|\mathcal{S}|$ is normally exponential in $|A|$. The first disadvantage is that computing the exponential number of indices for each solution $\mathcal{S}$ is comparable to that of solving the underlying problem by enumeration which, in most cases of interest, is impractical. The second disadvantage is that because the traditional policies assume that all solutions are upfront identical, they have to periodically explore every solution in $\mathcal{S}$ with a frequency proportional to $\ln N/N$. However, because of the correlation between the solutions $\mathcal{S}$, this is no longer necessary in the combinatorial setting. We now illustrate this with two examples, where for simplicity of exposition, we ignore the exploration frequencies. That is, we assume that whatever elements in A selected for exploration, they are selected persistently over time (irrespective of how), so that their mean cost estimates are accurate.

Figure 1 Graph for Example 1.



EXAMPLE 1. Consider the digraph $G=(V,A)$ for $V=\{v_{i,j}: i,j\in\{1,\dots,k+1\}, i\le j\}$ and $A=\{e_i\}_{i=1}^{k}\cup\{p_{i,j}: i\le j\le k\}\cup\{q_{i,j}: i\le j\le k\}$ where $e_i=(v_{i,i},v_{i+1,i+1})$, $p_{i,j}=(v_{i,j},v_{i,j+1})$, and $q_{i,j}=(v_{i,j},v_{i+1,j})$. This digraph is depicted in Figure 1 for $k=3$. Let $\mathcal{S}$ be composed of all paths from node $s:=v_{1,1}$ to node $t:=v_{k+1,k+1}$.

Let $\epsilon<c\ll M$ and set $l_a=\epsilon$ and $u_a=\infty$ for every arc $a\in A$. Define F to be such that $\mathbb{E}_F\{b_{e_i,n}\}=c$, and $\mathbb{E}_F\left\{b_{p_{i,j},n}\right\}=\mathbb{E}_F\left\{b_{q_{i,j},n}\right\}=M$, for all $i\in\{1,\dots,k\}$ and $i\le j\le k$, $n\in\mathbb{N}$. The shortest (expected) path is $S^*(\mathbb{E}_F\{B_n\})=(e_1,e_2,\dots,e_k)$ with expected length (cost) $z^*(\mathbb{E}_F\{B_n\})=kc$, $|A|=k(k+2)$, and $|\mathcal{S}|$ corresponds to the number of $s-t$ paths, which is equal to $\frac{1}{k+2}\binom{2(k+1)}{(k+1)}\sim\frac{4^{k+1}}{(k+1)^{3/2}\sqrt{\pi}}$ (Stanley 1999).

A traditional bandit policy would need to explore all $\frac{1}{k+2}\binom{2(k+1)}{(k+1)}$ paths. However, the same exploration goal can be achieved while leveraging the combinatorial structure of the solution set to expedite estimation: a key observation is that one might conduct mean cost estimation for elements in the ground set, and then aggregate those to produce estimates for all solutions. A natural way of incorporating this observation is to explore a *minimal solution cover* $\mathcal{E}$ of A (i.e., $\mathcal{E}\subseteq\mathcal{S}$ such that each $a\in A$ belongs to at least one $S\in\mathcal{E}$ and $\mathcal{E}$ is minimal with respect to inclusion for this property). In Example 1 we can easily construct a solution cover $\mathcal{E}$ of size $k+1$, which is significantly smaller than $|\mathcal{S}|$.

An additional improvement follows from exploiting the ideas in the lower bound result in LR as well. To see this, note that, unlike in the traditional setting, solutions are not ex-ante identical in the combinatorial case. Thus, for some $a\in A$, there might not exist a distribution F_a such that $a\in S^*$ for some $S^*\in\mathcal{S}^*(\mathbb{E}_{F_a}\{B_n\})$. Moreover, even if such a distribution exists, one might be able to distinguish F from F_a without implementing solutions containing a. This opens up the possibility that information collection on some ground elements might be stopped after a finite number of instances, independent of N, without affecting asymptotic efficiency. This is illustrated in the following example.

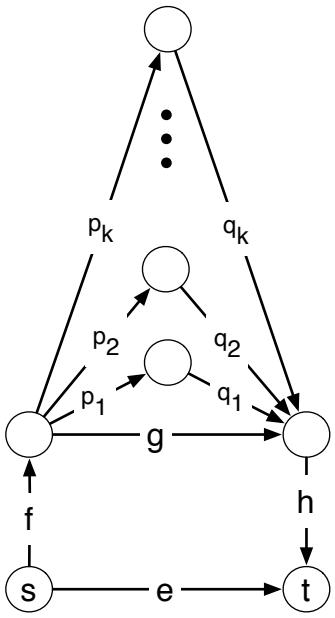


Figure 2 Graph for Example 2.

EXAMPLE 2. Let $G=(V,A)$ be the digraph depicted in Figure 2 and let $\mathcal{S}$ be composed of all paths from node s to node t. Set $l_a=0$ and $u_a=\infty$ for every arc $a\in A$, and F be such that $\mathbb{E}_F\{b_{e,n}\}=c$, $\mathbb{E}_F\{b_{g,n}\}=\epsilon$, $\mathbb{E}_F\{b_{f,n}\}=\mathbb{E}_F\{b_{h,n}\}=\frac{c+\epsilon}{2}$, $\mathbb{E}_F\{b_{p_i,n}\}=\mathbb{E}_F\{b_{q_i,n}\}=M$ for $n\in\mathbb{N}$ and for all $i\in\{1,\dots,k\}$ where $0<\epsilon\ll c\ll M$. The shortest (expected) path in this digraph is (e).

In Example 2, $|\mathcal{S}|=(k+2)$, and the only solution cover of A is $\mathcal{E}=\mathcal{S}$, which does not provide an advantage over traditional approaches. However, a cover is required only if we need to explore every element in A. Indeed, feedback obtained through exploration only needs to guarantee the optimality of path $\{e\}$ with respect to all *plausible* alternative distributions. However, because $f(\cdot)$ minimizes cost, it suffices to check only *one* possibility: that in which every unexplored element $a\in A$ has an expected cost equal to its lowest possible value l_a. In Example 2 we can check that every path other than $\{e\}$ uses arcs f and h and the sum of the expected costs of f and h is strictly larger than that of e. Together with the fact that the cost of every arc has a lower bound of zero, this implies that exploring f and h is sufficient to guarantee the optimality of $\{e\}$. We can explore f and h by implementing any path that contains them, but the cheapest way to do so is by implementing path $\{f,g,h\}$.

Examples 1 and 2 show that in the combinatorial setting effective policies do not need to explore every solution in $\mathcal{S}$ or even every ground element in A. In particular, Example 2 shows that the question of *what* elements of A to explore (e.g., f and h) and *how* to explore these elements (e.g., through path $\{f,g,h\}$) become crucial to construct efficient policies in the combinatorial setting. However, we still need to answer the question of *when* to explore, or more precisely, what is the

relative exploration frequencies for the subset of A needed to be explored and the elements of $\mathcal{S}$ used to cover them. To achieve this we need to consider the analogs of K_a by extending the fundamental performance limit of LR from the traditional multi-armed bandits to the combinatorial setting.

4. Bounds on Achievable Performance

4.1. A Limit on Achievable Performance

Following the argument in the traditional bandit setting, consistent policies must explore those subsets of ground elements that have a chance to be part of an optimal solution. To formally define such subsets, for a cost vector $B \in \mathcal{B}$ we define $\mathcal{D}(B) := \{D \in \mathcal{D}'(B) : D' \notin \mathcal{D}'(B)\ \forall D' \subset D\}$, where

$$\mathcal{D}'(B) := \left\{ D \subseteq A : D \neq \emptyset, D \subseteq \bigcap_{S^* \in \mathcal{S}^*(B)} (A \setminus S^*)\ ,\ D \subseteq \bigcap_{S \in \mathcal{S}^*(B_D)} S \right\},\ \text{and}$$

$$B_D := (b_a : a \in A \setminus D) \cup (l_a : a \in D)\,.$$

By construction, $\mathcal{D}(\mathbb{E}_F\{B_n\})$ contains all subsets D of suboptimal ground elements such that they become part of every optimal solution if their costs are set to their lowest possible values, and that are minimal with respect to inclusion. That is, for any $D \in \mathcal{D}(\mathbb{E}_F\{B_n\})$ there exists an alternative distribution F_D under which all elements in D are part of any optimal solution. Because said elements are suboptimal, a consistent policy must distinguish F from F_D to attain asymptotic optimality. The following proposition, which we prove in Appendix A.1.1, shows that this can be accomplished by selecting *at least* one element in each set $D \in \mathcal{D}$ at a minimum frequency.

PROPOSITION 1. *For any consistent policy π, regular F, and $D \in \mathcal{D}(\mathbb{E}_F\{B_n\})$ we have that*

$$\lim_{N\to\infty} \mathbb{P}_F \left\{ \frac{\max\left\{\tilde{T}_{N+1}(a) : a \in D\right\}}{\ln N} \geq K_D\left(\mathbb{E}_F\{B_n\}\right) \right\} = 1, \tag{6}$$

where $K_D(\mathbb{E}_F\{B_n\})$ represents the inverse of the Kullback-Leibler divergence between F and F_D.

REMARK 3. The bound above is tighter when F_D is selected so as to maximize $K_D(\mathbb{E}_F\{B_n\})$. One can show that for $B \in \mathcal{B}$ (by the definition of regularity), the best bound for $D \in \mathcal{D}(B)$ is given by

$$K_D(B) = \inf \left\{ |D| \max_{a\in D} \left\{ I_a(b_a, \tilde{b}_a) \right\} : l_a \leq \tilde{b}_a \leq u_a,\ a \in D,\ \tilde{b}_a = b_a,\ a \in A \setminus D, D \in \bigcap_{S \in \mathcal{S}^*(\tilde{B})} S \right\}^{-1} \overset{(a)}{<} \infty, \tag{7}$$

where (a) follows from the fact that $K_D(B) \leq |D| \max_{a\in D} \left\{ I_a(b_a, \tilde{b}_a) \right\}$, where

$$\tilde{B} := (b_a,\ a \in A \setminus D) \cup \left(l_a + (b_a - l_a)\left(z^*(B) - z^*(B_D)\right) / \left(\sum_{a\in D} (b_a - l_a) \right) : a \in D \right).$$

Note that while Theorem 1 imposes lower bounds on the number of times that a solution (a singleton) is implemented, Proposition 1 imposes similar bounds, but on the number of times that certain subsets of A are selected. In other words, Proposition 1 characterizes *what* needs to be explored by a consistent policy. To transform this into a valid performance bound we additionally need to characterize *how* to explore these subsets in the most efficient way. In particular, in addition to selecting the set of ground elements that need to be explored, a consistent policy needs to determine a set of elements of $\mathcal{S}$ that contain or cover this set of ground elements. The following Lower Bound Problem (LPB) jointly determines the set of ground elements needed to be explored, a set of solutions that cover this set of ground elements, and their exploration frequencies. Furthermore, it does so in the most efficient way possible (i.e., by solving for the minimum-regret solution cover).

DEFINITION 2. For a cost vector $B \in \mathcal{B}$ define the lower bound problem as

$$LBP(B): z_L^*(B) := \min \sum_{S\in\mathcal{S}} \Delta_S^B \; y_S \tag{8a}$$

$$s.t. \qquad \max\{x_a : a \in D\} \geq K_D(B), \quad D \in \mathcal{D}(B) \tag{8b}$$

$$x_a \leq \sum_{S\in\mathcal{S}: a\in S} y_S, \quad a \in A \tag{8c}$$

$$x_a, y_S \in \mathbb{R}_+, \quad a \in A, S \in \mathcal{S}. \tag{8d}$$

Also, define $\Gamma_L(B)$ as the set of optimal solutions to $LBP(B)$.

For $B = \mathbb{E}_F\{B_n\}$, the set $\{a \in A : x_a > 0\}$ corresponds to the elements of the ground set that are explored to satisfy Proposition 1 and the actual values x_a represent the exploration frequencies $\tilde{T}_{N+1}(a)/\ln N$. Similarly, the set $\{S \in \mathcal{S} : y_S > 0\}$ corresponds to the solution cover (which we also call the *exploration set*) of the selected ground elements and y_S represent the exploration frequencies $T_{N+1}(S)/\ln N$. Indeed, constraints (8b) enforce exploration conditions (6) and constraints (8c) enforce the cover of the elements of A selected by (8b) .

For any consistent policy π, define $\zeta^\pi(F,N) := \sum_{S\in\mathcal{S}} \Delta_S^F \; T_{N+1}(S)$ to be the total additional cost (relative to an oracle agent) associated with that policy. Note that $\mathbb{E}_F\{\zeta^\pi(F,N)\} = R^\pi(F,N)$. The next proposition, which we prove in Appendix A.1.1, ties the asymptotic bounds in (6) to the solution to LBP and establishes an asymptotic bound on the regret of any consistent policy.

PROPOSITION 2. *For any consistent policy π and regular F we have that*

$$\lim_{N\to\infty} \mathbb{P}_F\Big(\zeta^\pi(F,N) \geq z_L^*\left(\mathbb{E}_F\{B_n\}\right)\ln N\Big) = 1.$$

Note that the result above establishes convergence in probability (hence it can be used to bound $\zeta^\pi(F,N)$, rather than just its expectation, which is the regret). The next result, whose proof we omit as it follows directly from Proposition 2 and Markov's inequality, shows that the regret of any consistent policy in the combinatorial setting is (at least) proportional to $z_L^*\left(\mathbb{E}_F\left\{B_n\right\}\right)\ln N$.

THEOREM 3. *The regret of any consistent policy π is such that for any regular F we have*

$$\liminf_{N\to\infty}\frac{R^\pi(F,N)}{\ln N}\geq z_L^*\left(\mathbb{E}_F\left\{B_n\right\}\right). \tag{9}$$

Going back to the discussion at the end of Section (3.3), we see that the fundamental limit on performance is deeply connected to both the combinatorial structure of problem $f(\cdot)$, as well as the range and mean of F. To see this, consider the setting in Example 2 with a slight modification: set now $l_f = l_h = c/2+\epsilon/4$. One can check that in this case $\mathcal{D}(\mathbb{E}\{B_n\})=\emptyset$ as any suboptimal solution has to use arcs f and h, whose cost lower bounds already ensure the optimality of solution $\{e\}$. Thus, in this case $z_L^*\left(\mathbb{E}_F\left\{B_n\right\}\right)=0$. This result suggests that, in this setting, a finite regret (independent of N) might be attainable. Indeed, this setting is such that active learning is not necessary, and information from implementation of an optimal solution to $f(\mathbb{E}_F\left\{B_n\right\})$ suffices to guarantee the optimality of said solution. The situation above is by no means particular to the shortest path problem. We formalize this in the next proposition, whose proof can be found in Appendix A.1.1.

PROPOSITION 3. *If $f(B)$ corresponds to a shortest path, minimum-cost spanning tree, minimum-cost perfect matching, generalized Steiner tree or knapsack problem, then there exists a family of instances where $z_L^*(B)=0$ while the minimum-sized cover of A is arbitrarily large.*

4.2. An Asymptotically Near-Optimal Policy

To match the lower bound of Theorem 3, given the construction of $LBP\left(\mathbb{E}_F\left\{B_n\right\}\right)$, it is natural to try allocating exploration efforts only to the solutions prescribed by a solution to $LBP\left(\mathbb{E}_F\left\{B_n\right\}\right)$ (i.e. those $S\in\mathcal{S}$ with $y_S>0$). Unfortunately, said solution is not readily available in practice, as it depends on the mean cost vector which is only partially estimated at any given time. Nonetheless, one might still attempt to focus exploration on the solution to $LBP\left(\bar{B}_n\right)$ for $\bar{b}_{a,n}:=\frac{1}{n-1}\sum_{l=1}^{n-1}\hat{b}_{a,l}$, hoping that said solution converges to that of $LBP\left(\mathbb{E}_F\left\{B_n\right\}\right)$. While this is indeed the case when $\bar{B}_n\to\mathbb{E}_F\left\{B_n\right\}$, collecting information only on solutions prescribed by the solution to $LBP\left(\bar{B}_n\right)$ does not suffice (in general) to accurately estimate the full mean cost vector, as the following example illustrates.

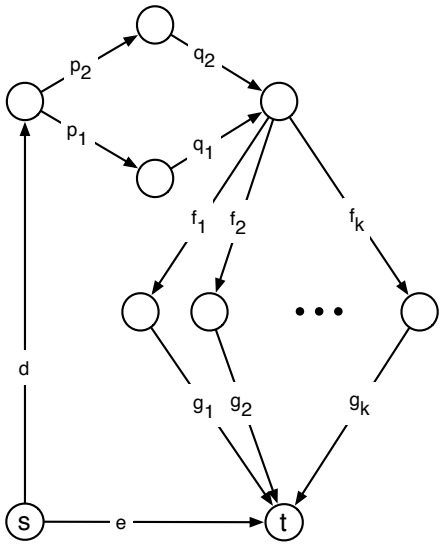


Figure 3 Graph for Example 3.

EXAMPLE 3. Let $G=(V,A)$ be the digraph depicted in Figure 3 and let $\mathcal{S}$ be composed of all paths from node s to node t. Set $l_a=0$ and $u_a=\infty$ for every arc $a\in A$, and F be such that for all $n\in\mathbb{N}$ we have $\mathbb{E}_F\{b_{e,n}\}=c$, $\mathbb{E}_F\{b_{d,n}\}=\mathbb{E}_F\{b_{p_1,n}\}=\mathbb{E}_F\{b_{q_1,n}\}=\epsilon/6$, $\mathbb{E}_F\{b_{p_2,n}\}=\mathbb{E}_F\{b_{q_2,n}\}=\frac{c-2\epsilon/3}{2}$, and $\mathbb{E}_F\{b_{f_i,n}\}=\mathbb{E}_F\{b_{g_i,n}\}=\frac{c+\epsilon/2}{2}$ for all $i\in\{1,\ldots,k\}$ where $0<\epsilon\ll c$. The shortest (expected) path in this digraph is (e).

For every $i\in\{1,\ldots,k\}$, define $S_i:=\{d,p_1,q_1,f_i,g_i\}$ and $\tilde{S}_i:=\{d,p_2,q_2,f_i,g_i\}$. In Example 3 we have that $\mathcal{D}=\{\{f_1\},\{f_2\},\ldots,\{f_k\},\{g_1\},\{g_2\},\ldots,\{g_k\}\}$. This, in turn, implies that the minimum-regret solution cover (i.e., exploration set) induced by the optimal solution to $LBP(\mathbb{E}_F\{B_n\})$ is $\{S_i\}_{i=1}^k$ with a regret of $k\epsilon$.

Suppose that we implement a policy that initially draws samples of the costs of p_1 and q_1 that are extremely high, so that the solution to $LBP\left(\bar{B}_n\right)$ consists of solutions $\{\tilde{S}_i\}_{i=1}^k$. There on, focusing exploration on the solution to $LBP\left(\bar{B}_n\right)$ might imply that no further samples of p_1 and q_1 are collected, thus $\bar{B}_n\to\tilde{B}$, with $\tilde{b}_a=\mathbb{E}_F\{b_{a,n}\}$ for all a in A except $a\in\{p_1,q_1\}$. One can see that in such a case, the exploration set (cover) that $LBP\left(\bar{B}_n\right)$ could converge to is $\{\tilde{S}_i\}_{i=1}^k$ with a regret of ck which is not an optimal solution to $LBP(\mathbb{E}_F\{B_n\})$.

Example 3 shows that convergence of $LBP\left(\bar{B}_n\right)$ to $LBP(\mathbb{E}_F\{B_n\})$ (and even $z_L^*(\bar{B}_n)$ to $z_L^*(\mathbb{E}_F\{B_n\})$) is not guaranteed if exploration is restricted to the solution to $LBP\left(\bar{B}_n\right)$. Thus, to assure convergence of $z_L^*(\bar{B}_n)$ to $z_L^*(\mathbb{E}_F\{B_n\})$ (so as to attain the asymptotic performance in the lower bound), one must collect samples on a subset of A that might contain more elements than those explored by the solution to $LBP(\mathbb{E}_F\{B_n\})$, and do so at a small but positive frequency.

While one might be able to formulate the problem of finding a subset of the ground set whose exploration incurs the least regret while guaranteeing the convergence of $LBP\left(\bar{B}_n\right)$, we instead choose to expand the exploration efforts to the whole ground set. By maintaining exploration frequencies on these additional elements small, the overall regret should still be driven by the cost of exploring the solution to $LBP\left(\bar{B}_n\right)$.

Following the discussion above, next we propose a policy that focuses exploration on the solution to $LBP\left(\bar{B}_n\right)$, but also at a lesser (tunable) degree on a solution cover of the ground set. Such an approach ensures the convergence of the solution to $LBP(\mathbb{E}_F\{B_n\})$ by guaranteeing that $\bar{B}_n \to \mathbb{E}_F\{B_n\}$ (see below for a more detailed discussion). To simplify the reconstruction of the aforementioned proxy (and the exposition), we make the following technical assumption, needed for proving a performance guarantee. We partially relax this assumption later in Section 5.

ASSUMPTION 1. *F is regular and $f_a(\cdot;\cdot)$ is known by the agent for all $a \in A$, and there exists a known finite constant K such that $K_D(\mathbb{E}_F\{B_n\}) \leq K$ for all $D \in \mathcal{D}(\mathbb{E}_F\{B_n\})$. In addition, there is no set $D \subseteq A$ such that $z^*(\mathbb{E}_F\{B_n\}) = z^*(\mathbb{E}_F\{B_n\}_D)$ with $\mathcal{S}^*(\mathbb{E}_F\{B_n\}) \neq \mathcal{S}^*(\mathbb{E}_F\{B_n\}_D)$.*

Knowing the parametric form of the cost density function for each arc reduces the burden of estimating $K_D(\mathbb{E}_F\{B_n\})$ as this can be performed by simply estimating $\mathbb{E}_F\{B_n\}$ (as is also the case for $\Delta_S^{\mathbb{E}_F\{B_n\}}$ and the set $\mathcal{D}(\mathbb{E}_F\{B_n\})$). A prior uniform bound on $K_D(\mathbb{E}_F\{B_n\})$ arises, for example, when a lower bound on the optimality gap of $f(\mathbb{E}_F\{B_n\})$ is known upfront; in practice this could be tied to a minimum tolerance error. The last part of Assumption 1 is necessary to correctly reconstruct the set of constraints (8b), and holds with probability one when, for example, mean costs are random draws from an absolutely continuous distribution: this suits most practical settings where mean costs are unknown and no particular structure for them is anticipated (note that any additional prior structural information on the mean cost vector might be taken advantage of).

Under Assumption 1, convergence of $z_L^*(\bar{B}_n)$ to $z_L^*(\mathbb{E}_F\{B_n\})$ is assured if $\bar{B}_n$ converges to $\mathbb{E}_F\{B_n\}$. As discussed in Example 1, this can be achieved by exploring a cover of A. We formalize the problem that finds the most efficient of these covers (i.e., a minimum-regret cover) in the following definition.

DEFINITION 3. For a cost vector $B \in \mathcal{B}$ define the cover problem as

$$Cover(B): z_C^*(B) := \min \sum_{S\in\mathcal{S}} \Delta_S^B \; y_S \tag{10a}$$

$$s.t. \qquad 1 \leq \sum_{S\in\mathcal{S}:a\in S} y_S, \quad a \in A \tag{10b}$$

$$y_S \in \{0,1\}, S \in \mathcal{S}. \tag{10c}$$

Also, define $\Gamma_C(B)$ as the set of optimal solutions to $Cover(B)$.

The proposed policy, which we refer to as the *LBP* policy, formulates and solves $LBP\left(\bar{B}_n\right)$ and $Cover\left(\bar{B}_n\right)$, and focuses exploration efforts (at different degrees) on the solutions to said problems. To enforce the logarithmic exploration frequency found in Theorem 3, the policy uses an idea known as the *doubling trick* (Cesa-Bianchi and Lugosi 2006, Chapter 2.3). This approach also allows us to minimize the number of times that the underlying combinatorial problem and auxiliary exploration problems $LBP\left(\bar{B}_n\right)$ and $Cover\left(\bar{B}_n\right)$ need to be solved. The doubling trick divides the horizon into cycles of growing length so that cycle i starts at time n_i where $(n_i)_{i\in\mathbb{N}}$ is a strictly increasing sequence of positive integers such that $n_1 = 1$ and $n_{i+2} - n_{i+1} \geq n_{i+1} - n_i$ for all $i \in \mathbb{N}$. Within each cycle we first solve $f\left(\bar{B}_n\right)$, $LBP\left(\bar{B}_n\right)$ and $Cover\left(\bar{B}_n\right)$, and then ensure that the appropriate exploration frequencies are achieved (in expectation). The frequency of exploration can then be controlled by varying the increment in length of the cycles (e.g., to achieve exploration frequencies proportional to $\ln N/N$ we can use cycles of exponentially increasing lengths). Let $\Phi := \{n_i : i \in \mathbb{N}\}$. The *LBP* policy is described by Algorithm 1. There, we define

$$p_S := \begin{cases} y_S/(n_{i+1} - n_i) & \text{if} \quad \sum_{S'\in\mathcal{S}} y_{S'} \leq (n_{i+1} - n_i) \\ y_S/\sum_{S'\in\mathcal{S}} y_{S'} & \text{otherwise} \end{cases}$$

for $S \in \mathcal{S} \setminus \mathcal{S}^*(\bar{B}_n)$. We also define $p_{S^*} := (1 - \sum_{S\in\mathcal{S}\setminus\mathcal{S}^*(\bar{B}_n)} p_S)/|\mathcal{S}^*(\bar{B}_n)|$ for $S^* \in \mathcal{S}^*(\bar{B}_n)$.

Algorithm 1 LBP policy $\pi^*(\gamma)$

Set $i = 0$, and take $b_{a,1}$ randomly from $[l_a, u_a], \forall a \in A$
for $n = 1$ to N **do**
  if $n \in \Phi$ **then**
    Set $i = i + 1$
    Set $S^* \in \mathcal{S}^*\left(\bar{B}_n\right)$ [Update exploitation set]
    Set $\mathcal{E} \in \Gamma_C\left(\bar{B}_n\right)$ [Update Cover-exploration set]
    Set $(x, y) \in \Gamma_L\left(\bar{B}_n\right)$ [Update LBP-exploration set]
  end if
  if $\tilde{T}_n(a) < \gamma\ i$ for some $a \in A$ **then**
    Set $S_n = S$ with $S \in \mathcal{E}$ such that $a \in S$ [Cover-based exploration]
  else
    Implement $S_n = S$ with probability p_S, $S \in \mathcal{S}$ [LBP-based exploration/Exploitation]
  end if
end for

The LBP policy admits the following performance guarantee, which we prove in Appendix A.1.2.

THEOREM 4. *Consider $\gamma \in (0,1)$ and $\varepsilon > 0$ arbitrary. Suppose that F is regular and Assumption 1 holds, and let $\pi^*(\gamma)$ denote the LBP policy when we choose $n_i := \max\{\lfloor e^{i^{1/(1+\varepsilon)}} \rfloor, n_{i-1}+1\}$ for all $i \geq 2$, then*

$$\lim_{N\to\infty} \frac{R^{\pi^*(\gamma)}(F,N)}{(\ln N)^{1+\varepsilon}} \leq z_L^*\left(\mathbb{E}_F\left\{B_n\right\}\right) + \gamma z_C^*\left(\mathbb{E}_F\left\{B_n\right\}\right). \tag{11}$$

4.3. Performance Gap Analysis

Optimal Scaling with Respect to N. While it is possible to achieve the optimal $\ln N$ dependence in the bound above (through the selection of a different set Φ and the introduction of additional tunable parameters), this comes at the price of additional constants in front of the second term in the right-hand side above. We introduce an additional sub-logarithmic term to the optimal scaling, so as to avoid introducing terms that emanate in part from the proof techniques, and so as to have a bound that reflects a fundamental insight about the result: asymptotic regret arises from suboptimal exploration which in our policy is distributed between the solution to LBP problem and, at a lower frequency, the solution to *Cover* problem.

Improved Upper Bounds. By setting γ arbitrarily close to zero, one can set the leading constant in the right-hand side of (11) arbitrarily close to that in Theorem 3 up to sub-logarithmic terms. However, it is not possible to set $\gamma = 0$ in general, as illustrated in Example 3, as this would not guarantee convergence on the solution to LBP.

It is possible, however, to reduce the gap between the leading constants in Theorems 3 and 4. For that, instead of complementing exploration on the solution to LBP with the solution to *Cover*, one can find a minimum-regret solution set that fulfills condition (6) and is robust to changes in the mean cost of unexplored ground elements. That is, one can design a policy whose regret admits a bound of the form

$$\lim_{N\to\infty} \frac{R^{\pi^*(\gamma)}(F,N)}{(\ln N)^{1+\varepsilon}} \leq z_R^*\left(\gamma, \mathbb{E}_F\left\{B_n\right\}\right),$$

for $\gamma > 0$, where $z_R^*(\gamma, B)$ is the optimal solution to a "robust" variation of $LBP(B)$ (we include such a formulation in Appendix A.4), and such that

$$z_L^*(B) \leq z_R^*(\gamma, B) \leq z_L^*(B) + \gamma z_C^*(B).$$

While we do not prove such bounds here (this requires more convoluted, lengthier arguments), the insight derived from it remains the same: regret emanates from suboptimal exploration. There are settings, however, where samples from exploring the solution to LBP suffices to guarantee the optimality of said solution. For example, one can show that if $f(\cdot)$ is a *weighted basis* or *independent set matroid* minimization problem, feedback from the solution to LBP suffices to guarantee its optimality. In such cases one has that $z_L^*(B) = z_R^*(\gamma, B)$ for all $\gamma > 0$ and $B \in \mathcal{B}$, thus a variation

of the LBP policy focusing exploration exclusively on a ϱ-optimal solution to the robust LBP problem (and not a cover) admits a performance bound of

$$\lim_{N\to\infty} \frac{R^{\pi^*(\gamma,\varrho)}(F,N)}{(\ln N)^{1+\varepsilon}} \leq (1+\varrho) z_L^*\left(\mathbb{E}_F\left\{B_n\right\}\right),$$

for $\varrho > 0$ and $\varepsilon > 0$ arbitrary (see Remark 5 for the role of ϱ in said algorithm).

Improved Lower Bounds. As shown above, in general it is not possible to improve the leading constant in (11) as finding and validating an optimal solution to $LBP(\mathbb{E}_F\{B_n\})$ might require knowledge of the mean costs of ground elements that are not explored by said solution. Hence, to find an optimal solution of $LBP(\mathbb{E}_F\{B_n\})$ we may need complementary exploration through a cover or a robust version of $LBP(\mathbb{E}_F\{B_n\})$. In contrast, our theoretical lower bound assumes advance knowledge of these unexplored costs, which allows it to bypass this complementary exploration. This difference is precisely the source of the gap between the leading constants in (11) and (9). It may be possible to derive an improved lower bound by not assuming such an advance knowledge. Unfortunately, it is not clear how to derive such a bound using the techniques in this paper or previous work on bandits.

5. A Practical Policy

A significant obstacle for the implementation of the LBP policy is the ability to solve the auxiliary formulation LBP repeatedly over time. Indeed, while LBP is a continuous optimization problem it has an exponential number of constraints (8b) that do not have a clear separation procedure. In addition, the maximum in constraint (8b) is known to be notoriously difficult to handle (Toriello and Vielma 2012). For this reason, we instead concentrate on developing practical policies inspired by the exploration principles behind Theorems 3 and 4.

To make the LBP policy in Algorithm 1 practical we need a version of $LBP\left(\bar{B}_n\right)$ that can be solved effectively with modern optimization techniques. To achieve this, we distill the core combinatorial aspects of $LBP\left(\bar{B}_n\right)$ by concentrating on *what* and *how* to explore, while somewhat ignoring the question of when to explore (e.g., the precise exploration frequencies).

To answer the question of what to explore, we note from Proposition 1 that consistent policies must try at least one element in each $D \in \mathcal{D}(\mathbb{E}_F\{B_n\})$ at a minimum frequency so as to distinguish F from an alternative distribution F_D that makes the set D part of any optimal solution. To this end, note that mean cost estimates for these frequently explored elements should converge to their true mean values, and that ought to suffice to guarantee the optimality of the elements in $\mathcal{S}^*(\mathbb{E}_F\{B_n\})$.

We consider an alternative mechanism to guarantee the optimality of $\mathcal{S}^*(\mathbb{E}_F\{B_n\})$. Such a mechanism directly identifies elements that need to be explored frequently. Suppose that exploration is

focused on a set $C \subseteq A$: elements outside that set would not be permanently sampled, so there will be no guarantees on where their true mean costs lie. Thus, frequent exploration can be restricted to a set C as long as this guarantees the optimality of $\mathcal{S}^*(\mathbb{E}_F\{B_n\})$ *independent* of where true mean costs of elements outside C are. Taking a pessimistic approach to those possibilities, we conclude that a set C must be such that

$$z^*\left(\mathbb{E}_F\{B_n\}\right) \leq z^*\left(\mathbb{E}_F\{B_n\}_{A\setminus C}\right).$$

One can check that $D \cap C \neq \emptyset$ for all $D \in \mathcal{D}(\mathbb{E}_F\{B_n\})$ for such a set C. This, in turn, implies that setting $x_a = K$ for all $a \in C$, with K as in Assumption 1 should lead to a feasible solution to LBP. This motivates the following definition.

Definition 4 (**Critical Set**). A subset $C \subseteq A$ is a *sufficient ground exploration set* (or simply *sufficient set*) for a cost vector $B \in \mathcal{B}$ if and only if

$$z^*(B) \leq z^*(B_{A\setminus C}). \tag{12}$$

A subset $C \subseteq A$ is a *critical set* if and only if it is a sufficient set that is minimal with respect to inclusion.

We may use condition (12) to simplify $LBP\left(\bar{B}_n\right)$ by just enforcing the exploration of a critical set (what to explore). Once the critical set is identified we can explore it efficiently (in terms of cost) by implementing a minimum-regret cover (exploration set) of it (how to explore). Both the selection of the critical set and its minimum-regret cover can be achieved simultaneously through the following combinatorial optimization problem.

Definition 5. For a given cost vector $B \in \mathcal{B}$, we let the *Optimality Cover Problem* (henceforth, OCP) be the optimization problem given by

$$OCP(B): z^*_{OCP}(B) := \min \sum_{S\in\mathcal{S}} \Delta^B_S \, y_S \tag{13a}$$

$$s.t. \qquad x_a \leq \sum_{S\in\mathcal{S}: a\in S} y_S, \quad a \in A \tag{13b}$$

$$\sum_{a\in S} (l_a(1-x_a) + b_a x_a) \geq z^*(B), \quad S \in \mathcal{S} \tag{13c}$$

$$x_a,\, y_S \in \{0,1\}, \quad a \in A, S \in \mathcal{S}, \tag{13d}$$

Also, define $\Gamma_{OCP}(B)$ as the set of optimal solutions to $OCP(B)$.

By construction, a feasible solution (x, y) to OCP corresponds to incidence vectors of a critical set $C \subseteq A$ and a solution cover $\mathcal{G}$ of such a set. That is, $(x, y) := (x^C, y^{\mathcal{G}})$ where $x^C_a = 1$ if $a \in C$

and zero otherwise, and $y^{\mathcal{G}}_S = 1$ if $S \in \mathcal{G}$ and zero otherwise. In what follows we refer to a solution (x, y) to OCP and the induced pair of sets $(C, \mathcal{G})$ interchangeably.

Constraints (13c) guarantee the optimality of $\mathcal{S}^*(B)$ even if costs of elements outside C are set to their lowest possible values (i.e., $b_a = l_a$ for all $a \notin C$), and constraints (13b) guarantee that $\mathcal{G}$ covers C (i.e., $a \in S$ for some $S \in \mathcal{G}$, for all $a \in C$). Finally, (13a) ensures that the regret associated with implementing the solutions in $\mathcal{G}$ is minimized. Note that when solving (13), one can impose $y_S = 1$ for all $S \in \mathcal{S}^*(B)$ without affecting the objective function, thus one can restrict attention to solutions that cover optimal elements of A.

REMARK 4. There is a clear connection between LBP and OCP. This is formalized in the next Lemma, whose proof can be found in Appendix A.2.

LEMMA 1. *Consider a cost vector $B \in \mathcal{B}$. An optimal solution to the linear relaxation of $OCP(B)$ is also optimal to formulation $LBP(B)$ when one replaces $K_D(B)$ by $K > 0$ for all $D \in \mathcal{D}(B)$.*

Proof of Lemma 1 shows that a feasible solution to (8) can be mapped into a feasible solution to the linear relaxation of OCP (via proper augmentation), and vice versa. The above elucidates that OCP is a version of LBP that imposes equal exploration frequencies across all solutions. Besides this, the formulations are essentially equivalent up to a minor difference: optimal solutions to OCP *must* cover all optimal ground elements; this, however, can be done without affecting performance in both formulations and hence it is inconsequential.

We obtain a practical policy based on OCP by simply exploring the elements of $\Gamma_{OCP}(\bar{B}_n)$ with frequency $(\ln N)^{1+\varepsilon}/N$ (more precisely, by implementing $\mathcal{G}$ for some $(C, \mathcal{G}) \in \Gamma_{OCP}\left(\bar{B}_n\right)$). Enforcing this exploration frequency and updating $\Gamma_{OCP}(\bar{B}_n)$ can again be achieved through the doubling trick. However, to obtain an asymptotic performance bound on this policy we need additional modifications similar to those for the LBP policy in Algorithm 1. The resulting policy, which we refer to as the *hybrid* policy (this name choice will become apparent shortly), is depicted in Algorithm 2.

To prove an asymptotic performance bound on the policy described by Algorithm 2 we need a relaxed version of Assumption 1.

ASSUMPTION 2. *There is no set $D \subseteq A$ such that $z^*(\mathbb{E}_F\{B_n\}) = z^*(\mathbb{E}_F\{B_n\}_D)$ with $\mathcal{S}^*(\mathbb{E}_F\{B_n\}) \neq \mathcal{S}^*(\mathbb{E}_F\{B_n\}_D)$.*

Note that Assumption 2 ensures that Constraint (13c) is not active for any $S \notin \mathcal{S}^*(\mathbb{E}_F\{B_n\})$ and any vectors x and y satisfying (13b) and (13d).

As discussed in Section 4.2, the assumption holds when, for example, mean costs are randomly drawn from an absolutely continuous distribution. This suits most practical settings where mean costs are unknown and no particular structure for them is anticipated.

Algorithm 2 hybrid policy $\pi_h(\gamma, \varrho)$

Set $i = 0$, $C = A$, $\mathcal{E}$ a minimal cover of A, $\mathcal{G} = \mathcal{E}$, and take $b_{a,1}$ randomly from $[l_a, u_a], \forall a \in A$
for $n = 1$ to N **do**
  if $n \in \Phi$ **then**
    Set $i = i + 1$
    Set $S^* \in \mathcal{S}^*\left(\bar{B}_n\right)$ [Update exploitation set]
    Set $\mathcal{E} \in \Gamma_C\left(\bar{B}_n\right)$ [Update Cover-exploration set]
    if $(C, \mathcal{G})$ is not a ϱ-optimal solution to OCP $\left(\bar{B}_n\right)$ **then**
      Set $(C, \mathcal{G}) \in \Gamma_{OCP}\left(\bar{B}_n\right)$ [Update OCP-exploration set]
    end if
  end if
  if $\tilde{T}_n(a) < \gamma\, i$ for some $a \in A$ **then**
    Set $S_n = S$ with $S \in \mathcal{E}$ such that $a \in S$ [Cover-based exploration]
  else if $\gamma < 1$ and $\tilde{T}_n(a) < i$ for some $a \in C$ **then**
    Set $S_n = S$ with $S \in \mathcal{G}$ such that $a \in S$ [OCP-based exploration]
  else
    Implement $S_n = S^*$ [Exploitation]
  end if
end for

REMARK 5. Note that parameter ϱ in Algorithm 2 allows the policy to converge to an optimal solution to $OCP(\mathbb{E}_F\{B_n\})$ (because there might exist multiple optimal solutions, the "Update OCP-exploration set" step ensures that the policy settles on one of them).

Under Assumption 2 we obtain the following performance bound, which we prove in Appendix A.2.

THEOREM 5. *Let $\pi_h(\gamma, \varrho)$ denote the hybrid policy and suppose that Assumption 2 holds. If we choose $n_i := \max\{\lfloor e^{i^{1/(1+\varepsilon)}} \rfloor, n_{i-1} + 1\}$ with $\varepsilon > 0$ arbitrary, for all $i \geq 2$, and we select ϱ to be smaller than the minimum optimality gap for OCP $(\mathbb{E}_F\{B_n\})$, then for $\gamma \in (0, 1)$*

$$\lim_{N\to\infty} \frac{R^{\pi_h(\gamma,\varrho)}(F, N)}{(\ln N)^{1+\varepsilon}} \leq z^*_{OCP}\left(\mathbb{E}_F\{B_n\}\right) + \gamma z^*_C\left(\mathbb{E}_F\{B_n\}\right).$$

Similar to the LBP policy in Algorithm 1, the exploration of the cover is needed to ensure convergence of $z^*_{OCP}\left(\bar{B}_n\right)$ to $z^*_{OCP}\left(\mathbb{E}_F\{B_n\}\right)$. Hence, one could improve the performance of the policy by focusing exploration on a robust version of OCP. Because our emphasis is on practical

policies, we instead concentrate on the evaluation of the practical implementation of the hybrid policy and its finite-time performance comparison to the benchmark.

As discussed in the previous section, setting $\gamma > 0$ is necessary to achieve convergence of certain estimates, necessary for ensuring asymptotic efficiency. However, such convergence is unnecessary (and usually unachievable) when evaluating finite-time performance. For this reason, we consider a simpler, less conservative policy by setting $\gamma = 0$, that explores as dictated by OCP and rarely explores every ground element. We refer to this policy as the *OCP-based* policy.

Similarly, we also consider the policy arising from setting $\gamma = 1$, which gives rise to a policy we refer to as the *Cover-based* policy: one can show that such a limiting policy admits a performance bound of

$$\lim_{N\to\infty} \frac{R^{\pi_h(\gamma,\varrho)}(F,N)}{(\ln N)^{1+\varepsilon}} \leq z_C^*\left(\mathbb{E}_F\left\{B_n\right\}\right).$$

We evaluate the practical impact of parameter γ in the numerical experiments in Section 7 by considering both the *Cover-based* and *OCP-based* policies. As will be illustrated, our experiments show that the OCP-based policy significantly outperforms benchmark policies in both long- and short-term performance, even when OCP is solved with an oracle polynomial-time heuristic. We discuss the details of the heuristic and numerical experiments in Sections 6 and 7, respectively.

6. Computational Aspects for Practical Policy Implementation

In this section, we address computational aspects for the practical implementation of all variants of the hybrid policy $\pi_h(\gamma,\varrho)$. We provide strong evidence that, for a large class of combinatorial problems, our policies scale reasonably well. For this, we focus our attention on the practical solvability of OCP and $Cover$ problems, which our policies solve repeatedly during the horizon, for many cost vector inputs B. Note that $f(B)$, $OCP(B)$ and $Cover(B)$ have generic combinatorial structures and hence could be extremely hard to solve. Thus, practical tractability of said problems is essential for implementation.

We begin by delineating a time-asynchronous version of policy $\pi_h(\gamma,\varrho)$, which is implementable in real-time and highlights the importance of solving OCP and $Cover$ effectively. Then, we focus our attention on settings where $f(B)$ is theoretically tractable, i.e., it is solvable in polynomial time. This class includes problems such as shortest path, network flow, matching, and spanning tree problems (Schrijver 2003). For these problems we develop polynomial-sized mixed-integer programming (MIP) formulations of OCP and $Cover$, which can be effectively tackled by state-of-the-art solvers.

We also present an oracle polynomial-time heuristic for OCP and $Cover$. This heuristic requires a polynomial number of calls to an oracle for solving $f(B)$. It therefore runs in polynomial time when $f(B)$ is polynomialy solvable. Furthermore, it provides a practical solution method for OCP

and $Cover$ when $f(B)$ is not expected to be solvable in polynomial time, but is frequently tractable in practice (e.g., medium-size instances of NP-complete problems such as the traveling salesman (Applegate et al. 2011), Steiner tree (Magnanti and Wolsey 1995, Koch and Martin 1998, Carvajal et al. 2013), and set cover problems (Etcheberry 1977, Hoffman and Padberg 1993, Balas and Carrera 1996)).

6.1. A Time-Constrained Asynchronous Policy

Depending on the application, real-time implementation might require choosing a solution $S_n \in \mathcal{S}$ prior to the *exogenous* arrival of instance B_n. However, the solution times for OCP, $Cover$, or even $f(B)$, could be longer than the time available to the executing policy. For example, most index-based policies must solve an instance of $f(B)$ between successive arrivals, which might not be possible in practice. Fortunately, a key feature of our proposed policies is that the frequency at which OCP, $Cover$ and $f(B)$ need to be solved decreases exponentially. Indeed, such problems are solved at the beginning of each cycle and the length of cycle i is $\Theta\left(\exp\left(i^{1/(1+\varepsilon)}\right)\right)$. Hence, as cycles elapse, there will be eventually enough time to solve these problems.

Nonetheless, as described in Algorithm 2, the policy cannot proceed until the corresponding problems are solved. However, one can easily modify the policy so that it begins solving $f(B)$, OCP and/or $Cover$ at the beginning of a cycle, but continues to implement solutions while these problems are being solved (such solutions might be computed either upfront or in previous cycles). Solution to these problems update incumbent solutions as they become available, which for long cycles would be at the beginning of the next one. Algorithm 4, which can be found in Appendix A.3 presents one such possible modification.

6.2. MIP Formulations of OCP and Cover for Polynomially-Solvable Problems

In this section we assume that $f(B)$ is polynomially solvable. However, this does not imply that neither OCP nor $Cover$ are tractable or practically solvable, as they might contain an exponential (in $|A|$) number of variables and constraints. The following theorem, whose proof can be found in Appendice A.3.2, ensures that both OCP and $Cover$ remain in NP, the class of non-deterministic polynomially-solvable problems (see e.g., Cook et al. (1998)).

THEOREM 6. *If $f(B)$ is in P, then $OCP(B)$ and $Cover(B)$ are in NP.*

Regarding the precise theoretical complexity of OCP and $Cover$, the next result, whose proof is relegated to Appendix A.3.3, establishes that at least for a particular class of problems in P there is no jump in theoretical complexity between f and $OCP/Cover$.

THEOREM 7. *$OCP(B)$ and $Cover(B)$ are in P for weighted basis or independent set matroid minimization problems.*

While it is possible to establish a non-trivial jump in theoretical complexity for problems within P, we deem the study of the theoretical complexity of $OCP/Cover$ for different problems outside the scope of the paper. Instead, here we focus on their practical solvability. For this, we first establish the existence of polynomial-sized MIP formulations when $f(B)$ admits a linear programming (LP) formulation. Then, we address the case when $f(B)$ admits a polynomial-sized extended LP formulation, and finally, the case when $f(B)$ does not admit such an extended formulation.

Problems with LP Formulations. We present a polynomial-sized formulation of OCP and $Cover$ when $f(B)$ admits a polynomial-sized LP formulation. For that, let I be an arbitrary finite set and $x \in \{0,1\}^{|I|}$; we let the support of x be $\text{supp}(x) := \{i \in I : x_i = 1\}$.

PROPOSITION 4. *Let y^S be the incidence vector of $S \in \mathcal{S}$, $M \in \mathbb{R}^{m \times |A|}$, and $d \in \mathbb{R}^m$ be such that $\{y^S\}_{S\in\mathcal{S}} = \left\{y \in \{0,1\}^{|A|} : My \le d\right\}$ and $\text{conv}\left(\{y^S\}_{S\in\mathcal{S}}\right) = \left\{y \in [0,1]^{|A|} : My \le d\right\}$. Then a MIP formulation of $OCP(B)$ is given by*

$$\min \quad \sum_{i=1}^{|A|} \left(\sum_{a\in A} b_a y_a^i - z^*(B) \right) \tag{14a}$$

$$s.t. \quad x_a \le \sum_{i=1}^{|A|} y_a^i, \qquad a \in A \tag{14b}$$

$$My^i \le d, \qquad i \in \{1,\dots,|A|\} \tag{14c}$$

$$M^T w \le \text{diag}(l)(\mathbf{1} - x) + \text{diag}(b)x \tag{14d}$$

$$d^T w \ge z^*(B) \tag{14e}$$

$$x_a,\, y_a^i \in \{0,1\},\, w \in \mathbb{R}^m, \qquad a \in A, i \in \{1,\dots,|A|\}, \tag{14f}$$

where for $v \in \mathbb{R}^r$, $\text{diag}(v)$ is the $r \times r$ diagonal matrix with v as its diagonal, and $\mathbf{1}$ is a vector of ones. A formulation for $Cover(B)$ is obtained by setting $x_a = 1$ for all $a \in A$ and removing (14d)–(14e).

In the above, x represents the incidence vector of a critical set. Such a condition is imposed via LP duality, using constraints (14d) and (14e), and eliminates the necessity of introducing constraint (13c) for each solution in $\mathcal{S}$. Similarly, each y^i represents the incidence vector of a solution $S \in \mathcal{S}$ for OCP. A formal proof of the validity of this formulation is included in Appendix A.3.4.

Formulation (14) has $O(|A|^2)$ variables and $O(m|A|)$ constraints. If m is polynomial in the size of the input of $f(B)$, then we should be able to solve (14) directly with a state-of-the-art integer programming (IP) solver. If m is exponential, but the constraints in the LP formulation can be

separated effectively, we should still be able to effectively deal with (14c) within a branch-and-cut algorithm. However, in such a case one would have an exponential number of w variables, which would force us to use a more intricate, and potentially less effective, branch-and-cut-and-price procedure. Nonetheless, when $f(B)$ does not admit a polynomial-sized LP formulation, one can still provide formulations with a polynomial number of variables, many of them also having a polynomial number of constraints. We discuss such cases next.

Problems with Polynomial-Sized Extended Formulations. The first way to construct polynomial-sized IP formulations of OCP and $Cover$ is to exploit the fact that many polynomially-solvable problems with LP formulations with an exponential number of constraints also have polynomial-sized *extended* LP formulations (i.e., formulations that use a polynomial number of auxiliary variables). A standard example of this class of problems is the spanning tree problem, where m in the LP formulation required by Proposition 4 is exponential in the number of nodes of the underlying graph. However, in the case of spanning trees, we can additionally use a known polynomial-sized extended formulation of the form $P := \left\{ y \in [0,1]^{|A|} : \exists\, z \in \mathbb{R}^p, \quad Cy + Dz \leq d \right\}$ where $C \in \mathbb{R}^{m' \times |A|}$, $D \in \mathbb{R}^{m' \times p}$ and $d \in \mathbb{R}^{m'}$, with both m' and p being only cubic on the number of nodes (and hence polynomial in $|A|$) (Martin 1991, e.g.). This formulation satisfies $\{y^S\}_{S \in \mathcal{S}} = P \cap \{0,1\}^{|A|}$ and $\operatorname{conv}\left(\{y^S\}_{S \in \mathcal{S}}\right) = P$. Then, a MIP formulation with a polynomial number of variables and constraints of OCP (and hence $Cover$) for the spanning tree problem is obtained by replacing (14c) with $Cy^i + Dz^i \leq d$, replacing (14d) with $C^T w \leq \operatorname{diag}(l)(\mathbf{1} - x) + \operatorname{diag}(b)x$ and $D^T w \leq 0$, and adding the polynomial number of variables z^i for $i \in \{1, \ldots, |A|\}$. Similar techniques can be used to construct polynomial-sized formulations for other problems with polynomial-sized extended LP formulations.

Problems without Polynomial-Sized Extended Formulations. It has recently been shown that there is no polynomial-sized extended LP formulations for the non-bipartite perfect matching problem (Rothvoß 2013*a*). Hence, we cannot use the techniques in the previous paragraph to construct polynomial-sized IP formulations of OCP and $Cover$ for matching. Fortunately, a simple linear programming observation and a result by Ventura and Eisenbrand (2003) allow constructing a version of (14) with a polynomial number of variables. The observation is that a solution y^* is optimal for $\max\{b^T y : My \leq d\}$ if and only if it is optimal for $\max\{b^T y : M_i^T y \leq d_i \quad \forall i \in I(y^*)\}$ where $I(y^*) := \{i \in \{1, \ldots, m\} : M_i^T y^* = d_i\}$ is the set of active constraints at y^*, and M_i is the i-th row of M. The number of active constraints can still be exponential for matching. However, for each perfect matching y^*, Ventura and Eisenbrand (2003) give explicit $C \in \mathbb{R}^{m' \times |A|}$, $D \in \mathbb{R}^{m' \times p}$ and $d \in \mathbb{R}^{m'}$, such that m' and p are polynomial in $|A|$ and $\left\{ y \in [0,1]^{|A|} : \exists\, z \in \mathbb{R}^p, \quad Cy + Dz \leq d \right\} = \left\{ y \in \mathbb{R}^{|A|} : M_i^T y \leq d_i \quad \forall i \in I(y^*) \right\}$. Using these matrices and vectors we can then do a replacement of (14d) analog to that for spanning trees to obtain a version of (14) with a polynomial number

of variables. We would still have an exponential number of constraints in (14c), but these can be separated in polynomial time for matching, so OCP and $Cover$ for matching could be effectively solved by branch-and-cut.

Perfect matching is the only explicit polynomially-solvable combinatorial optimization problem that is known not to admit a polynomial-sized extended LP formulation. However, Rothvoß (2013*b*) shows that there must exist a family of matroid problems without a polynomial-sized extended LP formulation. Fortunately, Theorem 7 shows that $OCP/Cover$ for matroids can be solved in polynomial time. We are not aware of any other polynomially-solvable combinatorial optimization problem which require non-trivial results to formulate OCP or $Cover$ with a polynomial number of variables.

Remark 6. Further improvements and extensions to (14) can be achieved. We give two such examples in Appendices A.3.5 and A.3.6. The first example shows how (14) for OCP and $Cover$ can be extended to the case when $f(B)$ is not in P, but admits a compact IP formulation. The second example gives a linear-sized formulation of OCP and $Cover$ for shortest path problems.

6.3. Oracle Polynomial-Time Heuristic

To further illustrate the potential practicality of policies based on OCP we develop a heuristic, that only requires a polynomial number of queries to an oracle for $f(B)$ (plus a polynomial number of additional operations). Furthermore, the version of the heuristic for OCP always returns a solution that is equal and possibly arbitrarily better than a minimal cover of A.

We begin by describing the heuristic for OCP in Algorithm 3. This heuristic first sets all costs to their lowest possible values, and successively solves instances of $f(B)$, each time incorporating the incumbent solution into the cover $\mathcal{E}$, adding its ground elements to C, and updating the cost vector accordingly. The procedure stops when the feedback from C *suffices* to guarantee the optimality of the best solution (i.e., when $z^*(\tilde{B}) \geq z^*(B)$). To achieve *efficiency* of such a feedback, the heuristic then prunes elements in C that are not required to guarantee sufficiency of the feedback.

Note that in each iteration of the first loop, Algorithm 3 calls an oracle for $f(B)$ and adds at least one ground element to C. Similarly, in the second loop, the heuristic calls such an oracle once for every element in C. Hence, the procedure calls such an oracle at most $2|A|$ times. Thus, the heuristic makes a linear number of calls to the oracle for $f(B)$. In particular, if $f(B)$ is in P, then the heuristic runs in *polynomial time*.

The performance of the heuristic ultimately depends on the specifics of a setting. For instance, in the setting of Example 1, the heuristic returns, in the worst case, a solution with $|\mathcal{G}| = k$, which is of the order of a cover of A. In the setting of Example 2 on the other hand, the heuristic returns a

Algorithm 3 Oracle Polynomial-Time Heuristic

Set $\tilde{B} := \left(\tilde{b}_a : a \in A\right) = (l_a : a \in A)$, $\mathcal{G} = \emptyset$, $C = \emptyset$.
while $z^*\left(\tilde{B}\right) < z^*(B)$ **do**
  Select $S \in \mathcal{S}^*\left(\tilde{B}\right)$ and set $\tilde{b}_a = b_a$ for all $a \in S$
  $\mathcal{G} \leftarrow \mathcal{G} \cup \{S\}$ and $C \leftarrow C \cup S$
end while
for $a \in C$ **do**
  if $z^*\left(\tilde{B}_{\{a\}}\right) \geq z^*(B)$ **then**
    $C \leftarrow C \setminus \{a\}$ and $\tilde{b}_a \leftarrow l_a$
  end if
end for

solution with $|\mathcal{G}| = 2$ (in such a setting a cover of A is of order k). It is not hard to identify settings where the heuristic performs arbitrarily better than any cover of A. However, in the numerical experiments we will see that the heuristic can return a solution that is much closer to a cover of A than an optimal solution to OCP.

We finally note that the heuristic in Algorithm 3 can be modified as follows for solving the *Cover* problem: the first loop should be implemented while $A \not\subseteq C$ and the second loop is no longer needed. The resulting set $\mathcal{G}$ provides a cover of A.

7. Numerical Experiments

In this section we study if the principles behind the asymptotically efficient policies from Sections 4 and 5 yield policies with good finite-time performance. We divide the numerical experiments in two classes: long-term and short-term experiments. For each setting, we first describe the benchmark policies and then present numerical results for settings of the shortest path, Steiner tree and knapsack problems. In both classes of experiments we consider four simple policies based on the principles from Sections 4 and 5, which we now describe in detail.

7.1. Four Simple Policies

We consider four policies each of which is a version of the policy $\pi_h(\gamma, \varrho = 0)$ described in Algorithm 2 with $n_i := \max\left\{\lfloor e^{i/H} \rfloor, n_{i-1} + 1\right\}$, for all $i \geq 2$ for $H > 0$. Such choice of $\Phi := \{n_i : i \in \mathbb{N}\}$ enforces exploration frequencies proportional to $\ln N/N$ instead of the $(\ln N)^{1+\varepsilon}/N$ enforced by the policy in Theorem 5. Such policies still satisfy an asymptotic performance guarantee, but the constants (multiplying the $\ln N$ term) are significantly more complicated than those in Theorem 5.

Hence, we omit them as we are interested in finite-time performance in this section. We report results for $H = 5$: preliminary tests using $H = \{5, 10, 20\}$ always resulted in logarithmic regrets.

We consider two *Cover-based* policies ($\gamma = 1$), and two *OCP-based* policies ($\gamma = 0$): See the end of Section 5. On one hand, for the first Cover-based policy, we compute a minimum-size cover of A (which does not depend on cost vector B) at time $t = 0$ which we keep for all periods; we refer to this policy as the *static cover* policy. The second Cover-based policy does use $Cover\left(\bar{B}_n\right)$ which needs to be solved at the beginning of each cycle: we refer to this second policy as the *dynamic cover* policy.

On the other hand, the first OCP-based policy solves $OCP\left(\bar{B}_n\right)$ to optimality at the beginning of each cycle. We simply refer to this policy as *OCP*. The second OCP-based policy solves $OCP\left(\bar{B}_n\right)$ heuristically using Algorithm 3 and we simply refer to it as *heuristic*.

All four policies start with an initialization phase in which each solution in a common minimum-size cover of A is implemented. When choosing a solution from the exploration set to implement, in case of a tie, the OCP-based policies select the solution that contains the most number of critical elements. In case of a second tie, they select a solution with the smallest average cost.

7.2. Long-Term Experiments

7.2.1. Benchmark Policies and Implementation Details

Benchmark Policies. Our benchmark policies are versions of UCB1, adapted to improve performance in the combinatorial setting. Recall that UCB1 implements solution S_n for instance n, where

$$S_n \in \arg\min \left\{ \bar{b}_{S,n} - \sqrt{2\ln(n-1)/T_n(S)} \right\},$$

and $\bar{b}_{S,n}$ denotes an estimate of the expected cost of solution $S \in \mathcal{S}$ at period n, computed at the solution level (this is the average cost incurred in previous implementations of solution S). We improve performance of UCB1 by: (i) conducting parameter estimation at the ground element level; (ii) adjusting confidence interval length to better reflect the amount of information used in estimating parameters; (iii) adjusting said length so that confidence bounds remain within the bounds implied by the range of F; and (iv) reducing the solution set so that it only includes solutions that are minimal with respect to inclusion. The resulting policy, which we denote UCB1+, implements solution S_n for instance n, where

$$S_n \in \arg\min_{S \in \mathcal{S}} \left\{ \max \left\{ \sum_{a \in S} \bar{b}_{a,n} - \sqrt{2\ln(n-1)/(\min_{a \in S} \left\{ \tilde{T}_n(a) \right\})}, \sum_{a \in S} l_a \right\} \right\}.$$

In a similar setting, Gai et al. (2012) propose another adaptation of UCB1: a modified version of such a policy implements

$$S_n \in \arg\min_{S \in \mathcal{S}} \left\{ \sum_{a \in S} \max \left\{ \bar{b}_{a,n} - \sqrt{(\mathcal{K}+1)\ln(n-1)/\tilde{T}_n(a)}, l_a \right\} \right\}$$

for instance n, for some positive finite constant $\mathcal{K}$. We denote this policy as Extended UCB1+.

Implementation Details. We report results when the marginals of F are exponential (we normalize the mean costs of the ground elements so that the maximum solution cost is at most one): we tested many cost distributions and observed consistent performance. We implemented UCB1+ and Extended UCB1+ with and without truncating indices at the implied lower bounds. Here, we present the point-wise minimum regret among both versions of each policy. Finally, we set $\mathcal{K} = 1$ in Extended UCB1+, as this selection outperformed the recommendation in Gai et al. (2012), and also is the natural choice for extending the UCB1 policy. UCB1+ and Extended UCB1+ start with an initialization phase that implements every solution in the same minimum-size cover of A used in the initialization phase of the policies described in Section 7.1.

The figures in this section report average performance for $N = 2000$ over 100 replications, and dotted lines represent 95% confidence intervals.

All policies were implemented in MATLAB R2011b. Shortest path problems were solved using Dijkstra's algorithm except when implementing UCB1+ (note that because of the index computation, $f(\cdot)$ must be solved by enumeration). For Steiner tree and knapsack problems, we solved standard IP formulations using GUROBI 5.0 Optimizer. The OCP policy solves formulation (13) of OCP using GUROBI 5.0 Optimizer. All experiments ran on a machine with an Intel(R) Xeon(R) 2.80GHz CPU and 16GB of memory. The average running time for a single replication ranged from less than 5 seconds for static cover policy to around 1.5 minutes for the OCP policy. (Note, however, that while the running times of simple and OCP policies grow (roughly) logarithmically with the horizon, those of UCB1+ and Extended UCB1+ grow linearly.)

7.2.2. Settings and Results

The settings are comprised of the shortest path problems in Example 1 for $k = 3$ (as shown in Figure 1), and Examples 2 and 3 for $k = 20$, followed by randomly generated instances (structures and costs) of shortest path, Steiner tree and knapsack problems. We observed consistent performance of our policies across these settings: here we only show a representative from each class. The random settings are complementary to Examples 1 and 2 in that the optimal critical subsets are large and hence the OCP and heuristic policies do not have an immediate advantage.

Examples 1, 2 and 3. Figure 4 depict the average performance of six different policies on Examples 1 (left), 2 (center) and 3 (right), respectively.

On Example 1, the OCP and heuristic policies perform significantly better than the other policies. The static cover policy provides a slightly better performance than its dynamic counterpart (the minimum-size cover has 4 elements). The situation is essentially the same on Example 2, only

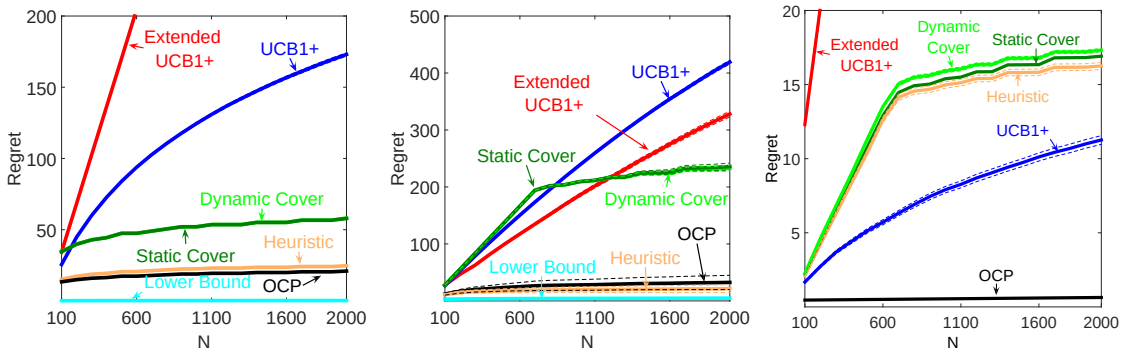


Figure 4 Average performance of different policies on Examples 1 (left), 2 (center) and 3 (right).

that this time Extended UCB1+ initially outperforms the static and dynamic cover policies (in this setting, the minimum-size cover is equal to $\mathcal{S}$, which is of size 22). In contrast, the solution to $OCP(\mathbb{E}_F\{B_n\})$ is only of size 2, which helps it achieve the best performance. (Note that for this setting, the heuristic tends to find the actual optimal solution to $OCP(\mathbb{E}_F\{B_n\})$ even with unreliable estimates.) On Example 3, the heuristic solution to OCP coincides with the minimum-regret cover of $\mathcal{S}$, thus the performance of heuristic coincides with those of the static and dynamic cover policies, which in turn are outperformed by UCB1+ (note that this latter policy rarely uses the arcs p_2 and q_2, since the costs of p_1 and q_1 are close to 0).

As we discussed before, the lower bound in Theorem 3 is asymptotic, so it is not clear whether the lower bound is meaningful in the finite time. However, we plot the lower bound for the three shortest path examples in Figure 4. As can be noted from the graph, in Examples 1 and 2, the lower bound is in fact meaningful and the regret of the OCP and heuristic policies is much closer to the lower bound than the other benchmark policies. In Example 3, however, the lower bound is not meaningful, that is the lower bound is larger than the regret of all policies as it only provides an asymptotic lower bound on regret.

In terms of efficient information collection, one can divide the set of ground elements (arcs) into three classes: those that are part of the optimal solution (Optimal arcs), those that are covered by at least one optimal solution to $OCP(\mathbb{E}_F\{B_n\})$ (exploration arcs), and the rest (*uninformative* arcs). Table 1 shows the average number of times that each type of arc (shown in columns called "Opt.", "Exp.", and "Uninf.", respectively) is tested up to horizon $N = 2000$ by each policy. Note that the OCP and heuristic policies spend significantly less time exploring uninformative arcs. Table 1 also shows the average length of implemented solutions (i.e., the average number of arcs in the implemented solutions) for different policies (the column called "Length").

Figure 5 depicts box plots of the 100 different cumulative regrets at the final time period $N = 2000$ (i.e., sample path final regrets) for OCP, UCB1+ and Extended UCB1+. We observe that the OCP

	Example 1				Example 2				Example 3			
	Opt.	Exp.	Uninf.	Length	Opt.	Exp.	Unin.	Length	Opt.	Exp.	Unin.	Length
OCP	1958.93	470.67	2.25	3.06	1858.25	548.12	4.55	1.19	140.03	214.50	1.00	4.72
Heuristic	1951.62	472.18	3.38	3.07	1918.43	524.20	3.32	1.11	106.83	215.94	35.71	4.79
Dyn. Cover	1885.88	482.03	38.00	3.17	1159.20	734.66	37.15	2.21	119.47	214.68	38.09	4.76
Stat. Cover	1886.52	481.91	37.81	3.17	1128.79	749.88	37.15	2.24	142.95	212.59	37.19	4.71
UCB1+	1660.75	533.35	42.12	3.51	474.31	929.80	66.61	3.19	92.45	217.75	24.61	4.82
Ext. UCB1+	791.31	684.36	364.72	4.81	870.88	795.78	53.76	2.67	14.87	219.02	151.79	4.97

Table 1 Average number of trials of different arcs up to horizon $N = 2000$, and also average solution size over different policies on Examples 1, 2 and 3.

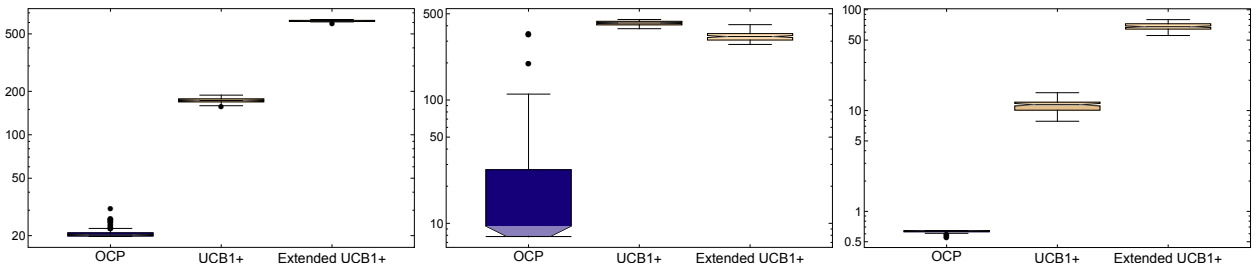


Figure 5 Box plots of performance for OCP and benchmark policies on Examples 1 (left), 2 (center) and 3 (right).

policy significantly outperforms UCB1+ and Extended UCB1+ not only on average, but also for (almost) all replications.

Shortest Path Problem. We consider a shortest path problem on a randomly generated layered graph (Ryzhov and Powell 2011). The graph consists of a source node, a destination node, and 5 layers in between, each containing 4 nodes. In each layer, every node (but those in the last layer) is connected to 3 randomly chosen nodes in the next layer. The source node is connected to every node in the first layer and every node in the last layer is connected to the destination node. Mean arc costs are selected randomly from the set $\{0.1, 0.2, \ldots, 1\}$ and then normalized. The representative graph is such that $|A| = 56$, $|\mathcal{S}| = 324$, and while the minimum-size cover of A is of size 13, the solution to $OCP(\mathbb{E}_F\{B_n\})$ is of size 16 even though the implied critical subset has 40 arcs. The left panel in Figure 6 depicts the average performance of different policies on this setting. We see that the OCP and heuristic policies outperform the benchmark. (Note, however, that UCB1+ outperforms the static and dynamic cover policies, in the short term.)

Knapsack Problem. Here the set A represents items that might go into the knapsack to maximize total utility. The solution set $\mathcal{S}$ consists of the subsets of items whose total weights do not exceed the knapsack weight limit. Weight and utility of items, as well as the weight limit, are selected randomly. The representative setting is such that $|A| = 20$, $|\mathcal{S}| = 24680$, the minimum-size cover is of size 4, and the solution to $OCP(\mathbb{E}_F\{B_n\})$ is of size 8 with an implied critical subset of

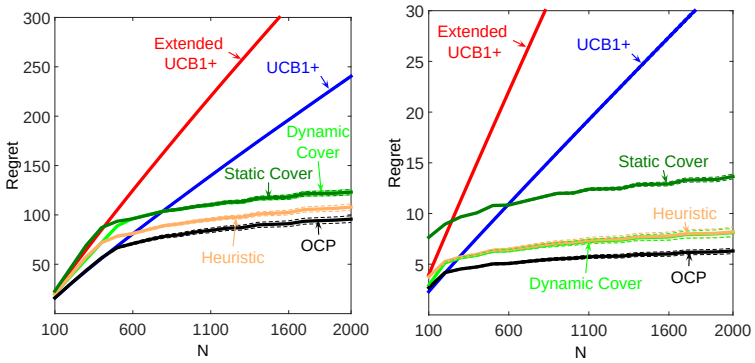


Figure 6 Average performance of different policies on the representative from the shortest path (left) and knapsack (right) settings.

size 17. The right panel in Figure 6 depicts the average performance of different policies on the representative for the knapsack setting. We see that the OCP policy outperforms the benchmark, with the heuristic and dynamic cover policies being close seconds.

Minimum Steiner Tree Problem. We consider a generalized version of the Steiner tree problem (Williamson and Shmoys 2011), where for a given undirected graph with non-negative edge costs and a set of pairs of vertices, the objective is to find a minimum-cost subset of edges (tree) such that every given pair is connected in the set of selected edges. The graphs as well as the pairs of vertices are generated randomly, as well as the mean cost values. The representative setting is such that $|A| = 18$, $|\mathcal{S}| = 10651$, the minimum-size cover is of size 2. The left panel in Figure 7 depicts average performance when all cost lower bounds are set to zero. In this representative setting, we have that the solution to $OCP(\mathbb{E}_F\{B_n\})$ is of size 7 with an implied critical subset of size 17. In this case, all arcs (but those trivially suboptimal) are critical, thus the OCP policy is essentially equivalent to the dynamic cover policy, which is corroborated by our results. The right panel in Figure 7 depicts average performance when lower bounds are chosen randomly. The representative setting is such that the solution to $OCP(\mathbb{E}_F\{B_n\})$ is of size 5 with an implied critical subset of size 12. Note that the OCP policy outperforms the benchmark as it successfully limits exploration to a critical set. Also note that the non-concave behavior of the regret curve of UCB1+ arises only in the transient as a by-product of truncation, and it disappears at around $n = 1200$.

Sample Path Regret Comparison. We check whether the findings in Figure 5 extend outside the instances of the Examples in the paper. For that we compared the sample path final regrets (i.e., at time period $N = 2000$) of OCP policy with those of UCB1+ and Extended UCB1+ policies. Out of 700 sample paths in the numerical experiments in Section 7.2.2, OCP policy outperforms the UCB1+ and Extended UCB1+ policies in 700 (i.e., 100% of sample paths) and 697 (i.e., 99.6% of sample paths), respectively.

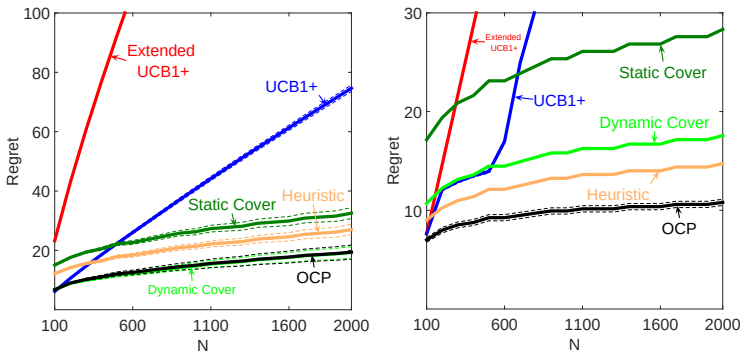


Figure 7 Average performance of different policies on the representative from the Steiner tree setting with zero (left) and positive (right) lower bounds.

7.3. Experiment with Size of the Ground Set

Next, we evaluate how performance of various policies vary with size of the ground set. For this we select both benchmark policies UCB1+ and Extended UCB1+, and OCP which is the version of our proposed policies that has the best performance and highest computational cost.

We experiment with a layered graph (see the previous section for a description) with $\mathfrak{L}$ layers, 2 nodes in each layer, complete connections between layers, and a direct arc from the source s to sink t. We experiment with $\mathfrak{L} = 2, 4, 6, 8, 10$ which results in $|A| = 9, 17, 25, 33, 41$ and $|\mathcal{S}| = 5, 17, 65, 257, 1025$, respectively.

We add a direct $s-t$ arc (path) to the original description of the layered graph so as to isolate the effect of size of the ground set on the performance of different policies. To this end, we let the expected cost of the $s-t$ arc (path) be 0.1, while all other arcs have an expected cost of $0.2/(\mathfrak{L}+1)$ where $\mathfrak{L}$ is the number of layers. Therefore, the $s-t$ path is the expected shortest path while all other paths (each of which has $\mathfrak{L}+1$ arcs) have an expected cost of 0.2, regardless of the size of the ground set. Thus, increasing the size of the ground set does not affect the cost (regret) of different paths in different instances. We run the experiments for $N = 20000$ and 40 replications.

For the OCP policy, we solve OCP using the linear-sized formulation (A-25). For all choices of $\mathfrak{L}$ we obtain a behavior similar to the graph on the left panel of Figure 9. That is, the cumulative regret of all three policies grow similar to a function $\mathfrak{K}\ln(n)$ for some policy-dependent constant $\mathfrak{K}$. (Note that the OCP policy significantly outperforms the benchmark policies regardless of the size of the ground set.) We consider two estimates of $\mathfrak{K}$: (i) K_{Final}, which we find by dividing the average final regret, which we denote by $\hat{R}(20000)$, by $\ln(20000)$, that is, $K_{Final} := \hat{R}(20000)/\ln(20000)$; (*ii*) K_{LS}, which is found by fitting the function $K_{LS}\ln(n)$ to the sample of average regrets for $n = 100, 200, \ldots, 20000$ and by minimizing the sum of squared errors, that is

$$K_{LS} := \min_{k} \left\{ \sum_{n=1}^{200} \left(k\ln(100n) - \hat{R}(100n) \right)^2 \right\}.$$

We present the value of both constants for the three policies and varying $|A|$ in Figure 8. We also present the average performance and computation time of different policies for the instance with $\mathfrak{L} = 10$ ($|A| = 41$ and $|S| = 1025$) as a representative setting in Figure 9 as we observed similar behavior in other instances. As can be seen in the left panel of Figure 9 (and also from

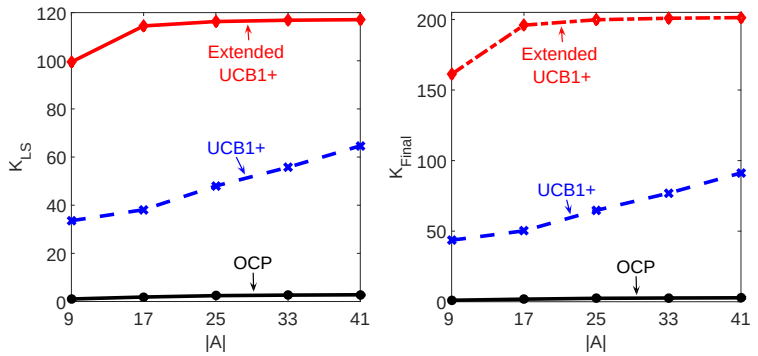


Figure 8 Constants K_{LS} (left) K_{Final} (right) when increasing the size of the ground set.

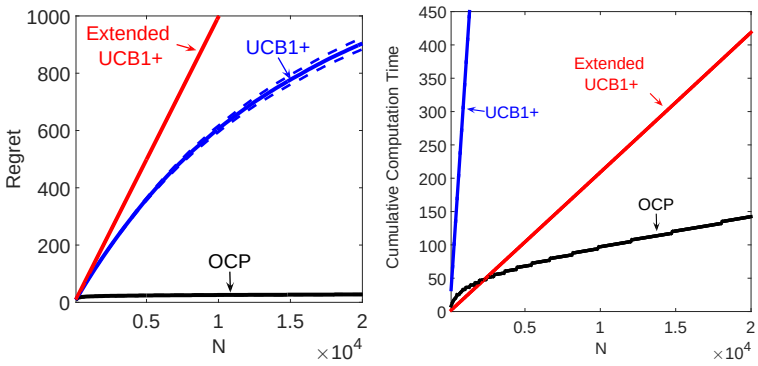


Figure 9 Average performance (left) and computation time (right) as a function of N for the instance with $\mathfrak{L} = 10$, $|A| = 41$, and $|S| = 1025$.

Figure 8), the OCP policy significantly outperforms both benchmark policies regardless of the size of the ground set. Moreover, the constants K_{LS} and K_{Final} are significantly smaller for the OCP policy than those for the benchmark policies. In addition, such constants grow with a much smaller rate for the OCP policy than the benchmarks. Moreover, as illustrated by the right panel of Figure 9, the computation time of the OCP policy grows logarithmically with N. Furthermore, there is a significant variation if we consider computation times. This is shown in Table 2, which presents the average running time for a complete replication for each policy. This time includes all calculations required by the policy (e.g., for OCP policy it includes the solution time of all

	$\|A\|$				
	9	17	25	33	41
OCP	75.54	79.43	81.18	92.60	142.38
UCB1+	65.47	127.38	376.56	1483.71	6686.70
Extended UCB1+	103.59	190.64	267.22	342.93	418.83

Table 2 Average total computation time (in seconds) for each replication of $N = 20000$.

instances of $OCP\left(\bar{B}_n\right)$ and $f\left(\bar{B}_n\right)$ as dictated by the corresponding version of Algorithm 2). We can see that the OCP policy runs faster than both benchmark policies for (almost) all instances (we note that although for much larger instances, one expects the Extended UCB1+ to run faster than the OCP policy, the Extended UCB1+ performs very poorly regardless of the size of the instance). Moreover, UCB1+, which is the more "competitive" benchmark policy in terms of performance, is significantly slower than the OCP policy. These observations further pronounce the practical advantage of the OCP policy both in terms of performance (i.e., regret) and computation time.

7.4. Short-Term Experiments

7.4.1. Benchmark Policies and Implementation Details

Benchmark Policies. Our benchmark policies are adaptations of the Knowledge-Gradient (KG) policy in Ryzhov et al. (2012) and the Gittins index approximation in Lai (1987) to our setting. Both policies require prior knowledge of the time horizon N, and because of this, several runs of the benchmark policies are necessary to construct their cumulative regret curves.

The KG policy requires a prior distribution for the cost and hyper-parameters. In our implementation, we use the Exponential-Gamma conjugate prior for each ground element. That is, the algorithm assumes that $b_{a,n}$ follows an exponential distribution with rate μ_a, but this rate itself is random, and initially distributed according to a Gamma distribution with parameters $\alpha_{a,0}$ and $\beta_{a,0}$. At time n, the posterior distribution of μ_a is a Gamma with parameters

$$\alpha_{a,n} = \alpha_{a,0} + \tilde{T}_n(a), \quad \beta_{a,n} = \beta_{a,0} + \sum_{m<n: a\in S_m} b_{a,m}, \quad a \in A.$$

Thus at time n, the KG algorithm implements solution S_n^{KG}, where

$$S_n^{KG} \in \arg\min_{S\in\mathcal{S}} \left\{ \sum_{a\in S} \frac{\beta_{a,n}}{\alpha_{a,n}-1} - (N-n)\ \mathbb{E}_S^n \left\{ \min_{S'\in\mathcal{S}} \left\{ \sum_{a\in S'} \frac{\beta_{a,n}}{\alpha_{a,n}-1} \right\} - \min_{S'\in\mathcal{S}} \left\{ \sum_{a\in S'} \frac{\beta_{a,n+1}}{\alpha_{a,n+1}-1} \right\} \right\} \right\},$$

where the expectation is taken with respect to $\{b_{a,n} : a \in S\}$. The expectation above corresponds to the knowledge gradient term $v_S^{KG,n}$ in the notation of Ryzhov et al. (2012). Unlike in that paper, there is no closed-form expression for $v_S^{KG,n}$ in our setting. Our plain vanilla implementation of the KG algorithm computes such a term via Monte Carlo simulation, and performs the outer

minimization via enumeration. The complexity of the implementation limited the size of the settings we tested.

The second benchmark is an approximation based on the Gittins index rule which in the finite-horizon undiscounted settings takes the form of an *average productivity* index (see Niño-Mora (2011)), and although it is not optimal in general, it is still applied heuristically. Our implementation assigns an index to each ground element, and computes the index of a solution as the sum of the indices of the ground elements it includes. The policy requires a parametric representation of the uncertainty. To mimic a setting where the functional form of reward distributions is unknown, we consider the approximation in Lai (1987) based on normally distributed rewards and use Normal/Normal-Gamma conjugate priors (this is motivated by a central limit argument): in our approximation, the index of a ground element $a \in A$ at the arrival of instance n is given by

$$g^a_{n,N}(\mu_{a,n}, \lambda_{a,n}, \alpha_{a,n}, \beta_{a,n}) = \left(\mu_{a,n} - \sqrt{\frac{\beta_{a,n}}{(\alpha_{a,n}-1)\lambda_{a,n}}}\; h\left(\frac{\lambda_{a,n} - \lambda_{a,0}}{N-n+1+\lambda_{a,n}-\lambda_{a,0}} \right) \right)^+,$$

where $\mu_{a,n}$ and $\lambda_{a,n}$ are the mean and variance of the normal posterior, respectively, $\alpha_{a,n}$ and $\beta_{a,n}$ are the hyper-parameters of the Gamma posterior, respectively, and $h(\cdot)$ approximates the boundary of an underlying optimal stopping problem. The policy implements solution S^{Gitt}_n, where

$$S^{Gitt}_n \in \arg\min_{S \in \mathcal{S}} \left\{ \sum_{a \in S} g^a_{n,N}(\mu_{a,n}, \lambda_{a,n}, \alpha_{a,n}, \beta_{a,n}) \right\}.$$

Implementation Details. The implementation details are as in the long-term experiments. The average running time for a single replication ranged from around one second for the OCP policy to around 2 seconds for Gittins to less than 10 minutes for KG. We exclude the results for the benchmark policies of the long-term experiments, because they were consistently outperformed by the OCP policy.

7.4.2. Settings and Results

We consider randomly generated (structure and costs) settings of shortest path, Steiner tree and knapsack problems. We observed consistent performance of the policies across settings, and show only a representative setting for each class of problems. There, the total number of periods is selected so as to visualize the value at which the OCP policy begins outperforming the benchmark. In all settings, the benchmark policies initially provide a better performance compared to the OCP policy, but the latter policy eventually surpasses the benchmarks for moderate values of N. The same holds true for the case of the heuristic policy.

Shortest Path Problem. The left panel at Figure 10 depicts the average performances for a shortest path problem in a layered graph with 5 layers, each with 4 nodes, and 2 connections

between each inner layer. The representative setting is such that $|A| = 40$, $|\mathcal{S}| = 64$, the minimum-size cover is of size 9, and the solution to $OCP(\mathbb{E}_F\{B_n\})$ is of size 10 with an implied critical subset of size 23.

Minimum Steiner Tree Problem. The central panel at Figure 10 depicts the average performances on a representative from the Steiner tree setting. The representative setting is such that $|A| = 9$, $|\mathcal{S}| = 50$, the minimum-size cover is of size 2, and the solution to $OCP(\mathbb{E}_F\{B_n\})$ is of size 4 with an implied critical subset of size 8.

Knapsack Problem. Figure 10 depicts the average performances on a representative from the knapsack setting. (Here we report on the average behavior over 500 replications so that the confidence intervals do not cross.) The representative setting is such that $|A| = 11$, $|\mathcal{S}| = 50$, the minimum-size cover is of size 7, and the solution to $OCP(\mathbb{E}_F\{B_n\})$ is of size 2 with an implied critical subset of size 5.

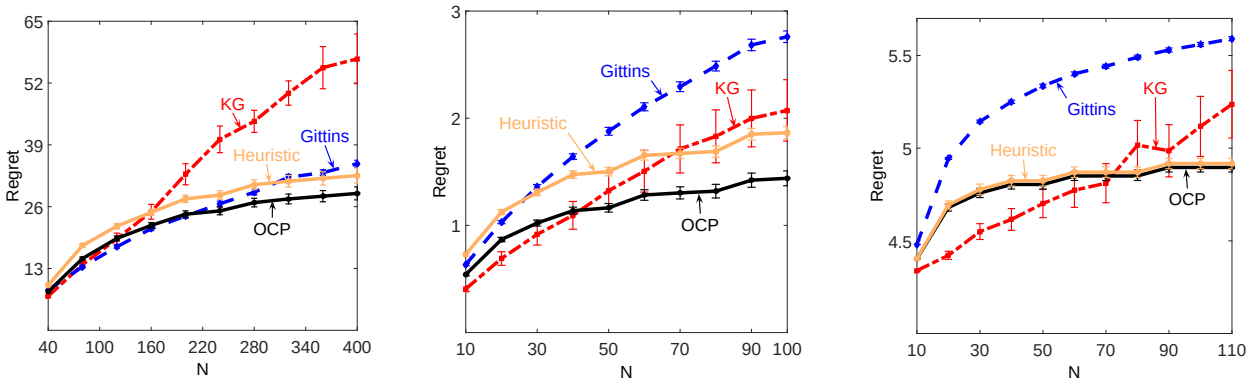


Figure 10 Average performance of different policies on the representative from the shortest path (left), Steiner tree (center) and knapsack (right) settings.

8. Final Remarks and Extensions

In this paper we have studied a class of sequential decision-making problems where the underlying single-period decision problem is a combinatorial optimization problem, and there is initial uncertainty about its objective coefficients. By framing the problem as a *combinatorial* multi-armed bandit, we have adapted key ideas behind results in the classical bandit setting to develop a theoretical policy that is asymptotically near-optimal. We also developed simpler and more practical variants of this policy. In doing so, we have shown that in addition to answering the question of *when* (with what frequency) to explore, which is key in the traditional setting, in the combinatorial setting one must also answer the questions of *what* and *how* to explore. We answer such questions

by explicitly solving for the cheapest optimality guarantee for the optimal solution to the underlying combinatorial problem (i.e., by solving OCP). We have shown evidence that the proposed policies are scalable and implementable in practice, and our numerical experiments show that they perform significantly well compared to relevant benchmark, both in the long- and short-term.

Finally, we note that the flexibility of the OCP-based policies allows them to be easily extended or combined with other techniques that consider similar what-and-how-to-explore questions. For instance, the OCP-based policy can be easily combined with the *barycentric spanner* of Awerbuch and Kleinberg (2004) to extend our results from element-level observations to set- or solution-level observations as follows. For a particular application it might be the case that the decision maker only has access, for example, to the *total* cost incurred by implementing solution S_n. We begin by showing how a cover-based policy can be adapted to this last setting. For a set of ground elements $S \subseteq A$, let $I_S \in \{0,1\}^{|A|}$ denote the incidence vector of the ground set (so that $S = \operatorname{supp}(I_S)$). We say a solution set $\mathcal{E}$ *recovers* a set $E \subseteq A$ if for each $a \in E$, there exists a vector $\gamma(a) := (\gamma_S(a), S \in \mathcal{E})$ such that

$$\sum_{S \in \mathcal{E}} \gamma_S(a) I_S = I_{\{a\}}. \tag{15}$$

Without loss of generality, one can assume that each ground element is recovered by at least one solution set. Let $\mathcal{E}$ be a solution set that recovers A, and let $\gamma := (\gamma(a), a \in A)$ be such that $\sum_{S \in \mathcal{E}} \gamma_S(a) I_S = I_{\{a\}}$, for all $a \in A$. One can implement a cover-based policy with $\mathcal{E}$ playing the role of a cover while replacing the estimate $\bar{b}_{a,n} := \frac{1}{n-1} \sum_{l=1}^{n-1} \hat{b}_{a,l}$ with

$$\bar{b}_{a,n} := \sum_{S \in \mathcal{E}} \frac{\gamma_S(a)}{T_n(S)} \sum_{l < n : S_l = S} \sum_{a \in S} \hat{b}_{a,l}, \quad a \in A. \tag{16}$$

The estimate above reconstructs the expected cost of each solution in $\mathcal{E}$ and uses (15) to translate such estimates to the ground-element level. Implementing this modification requires precomputing a solution set $\mathcal{E}$ recovering A. Such a set can be selected so that $|\mathcal{E}| \leq |A|$, and computed by solving $O(|A|)$ instances of $f(\cdot)$ (see e.g., the algorithm in Awerbuch and Kleinberg (2004)).

The idea above can also be used to extend the OCP-based policy to this new setting. For that we could consider the estimates in (16) and $(C, \mathcal{E})$ to be solution to an alternative version of OCP where in addition to (13b)-(13d), one imposes that $\mathcal{E}$ recovers C, that is

$$OCP'(B): \min \sum_{S \in \mathcal{S}} \Delta_S^B \; y_S \tag{17a}$$

$$s.t. \sum_{S \in \mathcal{S}} \gamma_S(a) I_S = x_a I_{\{a\}}, \quad a \in A \tag{17b}$$

$$\gamma_S(a) \leq Q \; y_S, \quad S \in \mathcal{S}, a \in A \tag{17c}$$

$$-\gamma_S(a) \leq Q \; y_S, \quad S \in \mathcal{S}, a \in A \tag{17d}$$

$$\sum_{a\in S}(l_a(1-x_a)+b_a x_a)\geq z^*(B),\quad S\in\mathcal{S} \tag{17e}$$

$$x_a,\, y_S\in\{0,1\},\, \gamma_S(a)\in\mathbb{R},\quad a\in A, S\in\mathcal{S}, \tag{17f}$$

where Q is an instance-dependent constant, whose size is polynomial in the size of the instance. The additional constraints(17b)-(17d) in OCP' ensure that the solution set $\mathcal{E}$ recovers the critical subset C. Like OCP, the formulation above can be specialized to accommodate the combinatorial structure of $f(\cdot)$ (as shown in Section 6.2). The performance guarantee in Theorem 5 would remain valid with the constants associated with OCP'. We anticipate that the challenge of solving OCP' effectively is comparable to that of solving OCP.

9. Acknowledgments

We thank Costis Maglaras, the associate editor, and the three anonymous referees for their thoughtful and constructive comments, which helped us improve the quality of our work in various fronts. This research is supported in part by the National Science Foundation [Grant CMMI-1233441], the Chilean Millennium Institute of Complex Engineering Systems (ICM: P05-004-F FIN. ICM-FIC) and the Business Intelligence Research Center (CEINE) at the University of Chile.

References

Abernethy, J., Hazan, E. and Rakhlin, A. (2008), Competing in the dark: An efficient algorithm for bandit linear optimization., *in* 'COLT', pp. 263–274.

Agrawal, R. (1995), 'The continuum-armed bandit problem', *SIAM J. Control Optim.* **33**(6), 1926–1951.

Agrawal, R., Hegde, M. and Teneketzis, D. (1990), 'Multi-armed bandit problems with multiple plays and switching cost', *Stochastics: An International Journal of Probability and Stochastic Processes* **29**(4), 437–459.

Anantharam, V., Varaiya, P. and Walrand, J. (1987), 'Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part I: IID rewards', *Automatic Control, IEEE Transactions on* **32**(11), 968–976.

Applegate, D., Bixby, R., Chvátal, V. and Cook, W. (2011), *The Traveling Salesman Problem: A Computational Study*, Princeton Series in Applied Mathematics, Princeton University Press.

Auer, P., Cesa-Bianchi, N. and Fischer, P. (2002), 'Finite-time Analysis of the Multiarmed Bandit Problem', *Machine Learning* **47**(2-3), 235–256.

Auer, P., Cesa-bianchi, N., Freund, Y. and Schapire, R. E. (2003), 'The non-stochastic multi-armed bandit problem', *SIAM Journal on Computing* **32**, 48–77.

Awerbuch, B. and Kleinberg, R. D. (2004), Adaptive routing with end-to-end feedback: distributed learning and geometric approaches, *in* 'Proceedings of the thirty-sixth annual ACM symposium on Theory of computing', STOC '04, ACM, New York, NY, USA, pp. 45–53.

Balas, E. and Carrera, M. C. (1996), 'A dynamic subgradient-based branch-and-bound procedure for set covering', *Operations Research* **44**, 875–890.

Berry, D. and Fristedt, B. (1985), *Bandit Problems*, Chapman and Hall, London, UK.

Bubeck, S., Munos, R., Stoltz, G. and Szepesvári, C. (2011), 'X-armed bandits', *Journal of Machine Learning Research* **12**, 1655–1695.

Caro, F. and Gallien, J. (2007), 'Dynamic assortment with demand learning for seasonal consumer goods', *Management Science* **53**, 276–292.

Carvajal, R., Constantino, M., Goycoolea, M., Vielma, J. P. and Weintraub, A. (2013), 'Imposing connectivity constraints in forest planning models', *Operations Research* **61**(4), 824–836.

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

Cesa-Bianchi, N. and Lugosi, G. (2012), 'Combinatorial bandits', *Journal of Computer and System Sciences* .

Chen, W., Wang, Y. and Yuan, Y. (2013), Combinatorial multi-armed bandit: General framework, results and applications, *in* 'Proceedings of the 30th International Conference on Machine Learning (ICML-13)', pp. 151–159.

Cook, W. J., Cunningham, W. H., Pulleyblank, W. R. and Schrijver, A. (1998), *Combinatorial optimization*, John Wiley & Sons, Inc., New York, NY, USA.

Cover, T. and Thomas, J. (2006), *Elements of Information theory*, John Wiley & Sons, Inc., Hoboken, NJ.

Dani, V., Hayes, T. P. and Kakade, S. M. (2008), Stochastic linear optimization under bandit feedback., *in* 'COLT', pp. 355–366.

Etcheberry, J. (1977), 'The set-covering problem: A new implicit enumeration algorithm', *Operations research* **25**, 760–772.

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

Gittins, J. (1979), 'Bandit processes and dynamic allocation rules', *Journal of the Royal Statistical Society* **41**, 148–177.

Harrison, J. and Sunar, N. (2013), Investment timing with incomplete information and multiple means of learning. Working paper, Stanford University.

Hoffman, K. L. and Padberg, M. (1993), 'Solving airline crew scheduling problems by branch-and-cut', *Management Science* **39**, 657–682.

Jones, D., Schonlau, M. and Welch, W. (1998), 'Efficient global optimization of expensive black-box functions', *Journal of Global Optimization* **13**, 455–492.

Kleinberg, R., Slivkins, A. and Upfal, E. (2008), 'Multi-armed bandits in metric spaces', *CoRR* **abs/0809.4882**.

Koch, T. and Martin, A. (1998), 'Solving steiner tree problems in graphs to optimality', *Networks* **32**(3), 207–232.

Kulkarni, S. and Lugosi, G. (1997), Minimax lower bounds for the two-armed bandit problem, *in* 'Decision and Control, 1997., Proceedings of the 36th IEEE Conference on', Vol. 3, IEEE, pp. 2293–2297.

Lai, T. L. (1987), 'Adaptive treatment allocation and the multi-armed bandit problem', *The Annals of Statistics* pp. 1091–1114.

Lai, T. L. and Robbins, H. (1985), 'Asymptotically efficient adaptive allocation rules', *Advances in Applied Mathematics* **6**(1), 4–22.

Liu, K., Vakili, S. and Zhao, Q. (2012), Stochastic online learning for network optimization under random unknown weights. Working paper.

Magnanti, T. L. and Wolsey, L. A. (1995), *Optimal trees*, Vol. 7 of *Handbooks in Operational Research and Management Science*, North-Holland, Amsterdam, pp. 503–615.

Martin, R. K. (1991), 'Using separation algorithms to generate mixed integer model reformulations', *Operations Research Letters* **10**, 119–128.

Mersereau, A., Rusmevichientong, P. and Tsitsiklis, J. (2009), 'A structured multiarmed bandit problem and the greedy policy', *IEEE Transactions on Automatic Control* **54**(12), 2787–2802.

Niño-Mora, J. (2011), 'Computing a classic index for finite-horizon bandits', *INFORMS Journal on Computing* **23**(2), 254–267.

Robbins, H. (1952), 'Some aspects of the sequential design of experiments', *Bulletin of the American Mathematical Society* **58**, 527–535.

Rothvoß, T. (2013*a*), 'The matching polytope has exponential extension complexity', *arXiv preprint arXiv:1311.2369* .

Rothvoß, T. (2013*b*), 'Some 0/1 polytopes need exponential size extended formulations', *Mathematical Programming* **142**, 255–268.

Rusmevichientong, P., Shen, Z. and Shmoys, D. (2010), 'Dynamic assortment optimization with a multinomial logit choice model and capacity constraint', *Operations Research* **58**(6), 1666–1680.

Rusmevichientong, P. and Tsitsiklis, J. (2010), 'Linearly parameterized bandits', *Mathematics of Operations Research* **35**(2), 395–411.

Ryzhov, I. O. and Powell, W. B. (2009), The knowledge gradient algorithm for online subset selection, *in* 'Proceedings of the 2009 IEEE International Symposium on Adaptive Dynamic Programming and Reinforcement Learning', pp. 137–144.

Ryzhov, I. O. and Powell, W. B. (2011), 'Information collection on a graph', *Operations Research* **59**(1), 188–201.

Ryzhov, I. O. and Powell, W. B. (2012), 'Information collection for linear programs with uncertain objective coefficients', *SIAM Journal on Optimization* **22**(4), 1344–1368.

Ryzhov, I. O., Powell, W. B. and Frazier, P. I. (2012), 'The knowledge gradient algorithm for a general class of online learning problems', *Operations Research* **60**(1), 180–195.

Sauré, D. and Zeevi, A. (2013), 'Optimal dynamic assortment planning with demand learning', *Manufacturing & Service Operations Management* **15**(3), 387–404.

Schrijver, A. (2003), *Combinatorial Optimization - Polyhedra and Efficiency*, Springer.

Stanley, R. (1999), *Enumerative combinatorics, Volume 2*, Cambridge studies in advanced mathematics, Cambridge University Press.

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

Toriello, A. and Vielma, J. P. (2012), 'Fitting piecewise linear continuous functions', *European Journal of Operational Research* **219**, 86 – 95.

Ventura, P. and Eisenbrand, F. (2003), 'A compact linear program for testing optimality of perfect matchings', *Operations Research Letters* **31**(6), 429–434.

Wen, Z. and Van Roy, B. (2013), Efficient exploration and value function generalization in deterministic systems, *in* 'Advances in Neural Information Processing Systems', pp. 3021–3029.

Whittle, P. (1982), *Optimization over time: Vol I*, John Wiley and Sons Ltd.

Williamson, D. P. and Shmoys, D. B. (2011), *The Design of Approximation Algorithms*, Cambridge University Press.

Appendix A: Omitted Proofs and Complementary Material

A.1. Ommitted Proofs from Section 4

A.1.1. A Limit on Achievable Performance

PROPOSITION 1. *For any consistent policy π, regular F, and $D \in \mathcal{D}(\mathbb{E}_F\{B_n\})$ we have that*

$$\lim_{N\to\infty} \mathbb{P}_F\left\{\frac{\max\left\{\tilde{T}_{N+1}(a) : a \in D\right\}}{\ln N} \geq K_D\left(\mathbb{E}_F\{B_n\}\right)\right\} = 1, \tag{6}$$

where $K_D(\mathbb{E}_F\{B_n\})$ represents the inverse of the Kullback-Leibler divergence between F and F_D.

We begin with some preliminaries.

Preliminaries. Define $\Theta_a := (l_a, u_a)$ and let θ_a denote the "true" parameter for $f_a, a \in A$ (i.e., $\theta_a = \mathbb{E}\{b_{a,n}\}$). For $\lambda_a \in \Theta_a$, we have that

$$I_a(\theta_a, \lambda_a) = \int_{-\infty}^{\infty} [\ln(f_a(x_a;\theta_a)/f_a(x_a;\lambda_a))] f_a(x_a;\theta_a)\, dx_a.$$

Finally, define $\theta := (\theta_a : a \in A)$ and let $\mathbb{E}_\lambda$ and P_λ denote the expectation and probability induced when each f_a receives the parameter $\lambda_a \in \Theta_a$, and define $\mathcal{S}^*_\lambda := \mathcal{S}^*(\mathbb{E}_\lambda\{B_n\})$ for $\lambda := \{\lambda_a : a \in A\}$.

Proof of the result. Consider $D \in \mathcal{D}(\theta)$ as defined in Section 4.1, and take $\lambda \in \mathcal{B}$ so that $\lambda_a = \theta_a$ for $a \notin D$, and that $D \subseteq S^*$ for all $S^* \in \mathcal{S}^*_\lambda$. By the consistency of π, one has that

$$\mathbb{E}_\lambda\left\{N - \sum_{S^*\in\mathcal{S}^*_\lambda} T_{N+1}(S)\right\} = o(N^\alpha),$$

for any $\alpha > 0$. By construction, each optimal solution under λ tries each $a \in D$ when implemented. Thus, one has that $\sum_{S^*\in\mathcal{S}^*_\lambda} T_{N+1}(S) \leq \max\left\{\tilde{T}_{N+1}(a) : a \in D\right\}$, and therefore

$$\mathbb{E}_\lambda\left\{N - \max\left\{\tilde{T}_{N+1}(a) : a \in D\right\}\right\} \leq \mathbb{E}_\lambda\left\{N - \sum_{S^*\in\mathcal{S}^*_\lambda} T_{N+1}(S)\right\} = o(N^\alpha). \tag{A-1}$$

Take $\epsilon > \alpha$. Define $I(D,\lambda) := |D| \max\{I_a(\theta_a,\lambda_a) : a \in D\}$, $D \in \mathcal{D}(\theta)$. We then have that

$$\begin{aligned}\mathbb{P}_\lambda\left\{\max\left\{\tilde{T}_{N+1}(a) : a \in D\right\} < \frac{(1-\epsilon)\ln N}{I(D,\lambda)}\right\} &= \mathbb{P}_\lambda\left\{N - \max\left\{\tilde{T}_{N+1}(a) : a \in D\right\} > N - \frac{(1-\epsilon)\ln N}{I(D,\lambda)}\right\} \\ &\overset{(a)}{\leq} \frac{\mathbb{E}_\lambda\left\{N - \max\left\{\tilde{T}_{N+1}(a) : a \in D\right\}\right\}}{N - \frac{(1-\epsilon)\ln N}{I(D,\lambda)}},\end{aligned}$$

where (a) follows from Markov's inequality. Note that for N large enough, we have that $N - ((1-\epsilon)\ln N/I(D,\lambda)) > 0$, and because $(1-\epsilon)\ln N/I(D,\lambda) = O(\ln N)$, from (A-1) we have that

$$(N - O(\ln N))\ \mathbb{P}_\lambda\left\{\max\left\{\tilde{T}_{N+1}(a) : a \in D\right\} < \frac{(1-\epsilon)\ln N}{I(D,\lambda)}\right\} = o(N^\alpha).$$

The above can be re-written as

$$\mathbb{P}_\lambda\left\{\max\left\{\tilde{T}_{N+1}(a) : a \in D\right\} < \frac{(1-\epsilon)\ln N}{I(D,\lambda)}\right\} = o(N^{\alpha-1}). \tag{A-2}$$

For $a \in D$ and $n \in \mathbb{N}$ define

$$L_n(a) := \sum_{k=1}^{n} \ln\left(f_a(\hat{b}_a^k;\theta_a)/f_a(\hat{b}_a^k;\lambda_a)\right),$$

where $\hat{b}_a^k$ denotes the k-th cost observation for $a \in D$ when policy π is implemented. Also, define the event

$$\Xi(N) := \left\{L_{\tilde{T}_{N+1}(a)}(a) \le \frac{(1-\alpha)\ln N}{|D|} \text{ for all } a \in D\,,\, \max\left\{\tilde{T}_{N+1}(a) : a \in D\right\} < \frac{(1-\epsilon)\ln N}{I(D,\lambda)}\right\},$$

and note that

$$\mathbb{P}_\lambda\{\Xi(N)\} \le \mathbb{P}_\lambda\left\{\max\left\{\tilde{T}_{N+1}(a) : a \in D\right\} < \frac{(1-\epsilon)\ln N}{I(D,\lambda)}\right\}.$$

Next, we relate the probability of the event $\Xi(N)$ under the two parameter configurations.

$$\begin{aligned}
\mathbb{P}_\lambda\{\Xi(N)\} &= \int_{\omega\in\Xi(N)} d\mathbb{P}_\lambda(\omega) \\
&\overset{(a)}{=} \int_{\omega\in\Xi(N)} \prod_{a\in D} \exp(-L_{\tilde{T}_{N+1}(a)}(a))\ d\mathbb{P}_\theta(\omega) \\
&\overset{(b)}{\ge} \int_{\omega\in\Xi(N)} \exp(-(1-\alpha)\ln N)\ d\mathbb{P}_\theta(\omega) \\
&= N^{\alpha-1}\mathbb{P}_\theta\{\Xi(N)\},
\end{aligned}$$

where (a) follows from noting that probabilities under λ and θ differ only in that cost observations in D have different probabilities under λ and θ, and (b) follows from noting that $L_{\tilde{T}_{N+1}(a)}(a) \le (1-\alpha)\ln N/|D|$ for all $\omega \in \Xi(N)$.

From above and (A-2) we have that

$$\lim_{N\to\infty} \mathbb{P}_\theta\{\Xi(N)\} \le \lim_{N\to\infty} N^{1-\alpha}\ \mathbb{P}_\lambda\{\Xi(N)\} = 0. \tag{A-3}$$

Now, fix $a \in D$. By the Strong Law of Large Numbers we have that

$$\lim_{n\to\infty} \max_{m\le n} L_m(a)/n = I_a(\theta_a, \lambda_a), \quad \text{a.s.}[\mathbb{P}_\theta], \quad \forall a \in D.$$

Because $\alpha < \epsilon$, we have that

$$\lim_{N\to\infty} \mathbb{P}_\theta \left\{ L_m(a) > \frac{(1-\alpha)\ln N}{|D|} \text{ for some } m < \frac{(1-\epsilon)\ln N}{|D|\, I_a(\theta_a,\lambda_a)} \right\} = 0 \quad \forall a \in D,$$

and because $I(D,\lambda) \ge |D|\, I_a(\theta_a,\lambda_a)$, we further have that

$$\lim_{N\to\infty} \mathbb{P}_\theta \left\{ L_m(a) > \frac{(1-\alpha)\ln N}{|D|} \text{ for some } m < \frac{(1-\epsilon)\ln N}{I(D,\lambda)} \right\} = 0 \quad \forall a \in D.$$

Then, in particular by taking $m = \tilde{T}_{N+1}(a)$ we have that

$$\lim_{N\to\infty} \mathbb{P}_\theta \left\{ L_{\tilde{T}_{N+1}(a)}(a) > \frac{(1-\alpha)\ln N}{|D|}, \quad \tilde{T}_{N+1}(a) < \frac{(1-\epsilon)\ln N}{I(D,\lambda)} \right\} = 0 \quad \forall a \in D,$$

which in turn implies

$$\lim_{N\to\infty} \mathbb{P}_\theta \left\{ L_{\tilde{T}_{N+1}(a)}(a) > \frac{(1-\alpha)\ln N}{|D|}, \quad \max\left\{\tilde{T}_{N+1}(a) : a \in D\right\} < \frac{(1-\epsilon)\ln N}{I(D,\lambda)} \right\} = 0 \quad \forall a \in D.$$

Finally, by taking the union of events over $a \in D$ we have that

$$\lim_{N\to\infty} \mathbb{P}_\theta \left\{ L_{\tilde{T}_{N+1}(a)}(a) > \frac{(1-\alpha)\ln N}{|D|} \text{ for some } a \in D\,, \max\left\{\tilde{T}_{N+1}(a) : a \in D\right\} < \frac{(1-\epsilon)\ln N}{I(D,\lambda)} \right\} = 0. \tag{A-4}$$

Thus, by (A-3), (A-4), and the definition of $\Xi(N)$ we have that

$$\lim_{N\to\infty} \mathbb{P}_\theta \left\{ \max\left\{\tilde{T}_{N+1}(a) : a \in D\right\} < \frac{(1-\epsilon)\ln N}{I(D,\lambda)} \right\} = 0.$$

The result follows from letting ϵ approach zero, and taking $K_D := I(D,\lambda)^{-1}$.

Proposition 2. *For any consistent policy π and regular F we have that*

$$\lim_{N\to\infty} \mathbb{P}_F\Big(\zeta^\pi(F,N) \ge z_L^*\left(\mathbb{E}_F\left\{B_n\right\}\right)\ln N\Big) = 1.$$

Define $\theta := \mathbb{E}\left\{B_n\right\}$ and $\Upsilon_N := \bigcap_{D\in\mathcal{D}(\theta)} \left\{\max\left\{\tilde{T}_{N+1}(a) : a \in D\right\} \ge K_D \ln N\right\}$. Note that $\zeta^\pi(F,N) \ge z_L^*(\theta)\ln N$ when Υ_N occurs, because $\left(x_a = \frac{\tilde{T}_{N+1}(a)}{\ln N}, a \in A\right)$ and $\left(y_S = \frac{T_{N+1}(S)}{\ln N}, S \in \mathcal{S}\right)$

are feasible to LBP. Thus, one has that

$$\begin{aligned}\mathbb{P}_F\left\{\frac{\zeta^\pi(F,N)}{\ln N} < z_L^*(\theta)\right\} &= \mathbb{P}_F\left\{\frac{\zeta^\pi(F,N)}{\ln N} < z_L^*(\theta)\,,\, \Upsilon_N\right\} + \mathbb{P}_F\left\{\frac{\zeta^\pi(F,N)}{\ln N} < z_L^*(\theta)\,,\, \Upsilon_N^c\right\} \\ &\leq \mathbb{P}_F\left\{\Upsilon_N^c\right\}. \end{aligned} \tag{A-5}$$

From Proposition 1 and the union bound, we have that

$$\lim_{N\to\infty} \mathbb{P}_F\left\{\Upsilon_N^c\right\} \leq \sum_{D\in\mathcal{D}(\theta)} \lim_{N\to\infty} \mathbb{P}_F\left\{\max\left\{\tilde{T}_{N+1}(a) : a\in D\right\} < K_D \ln N\right\} = 0,$$

because $|\mathcal{D}(\theta)| < \infty$. Thus, taking the limit in (A-5) we have that

$$\lim_{N\to\infty} \mathbb{P}_F\left\{\zeta^\pi(F,N) < z_L^*(\theta)\ln N\right\} = 0.$$

This concludes the proof.

PROPOSITION 3. *If $f(B)$ corresponds to a shortest path, minimum-cost spanning tree, minimum-cost perfect matching, generalized Steiner tree or knapsack problem, then there exists a family of instances where $z_L^*(B) = 0$ while the minimum-sized cover of A is arbitrarily large.*

The family for the shortest path problem is that based on Example 2 (which is parametrized by an integer k), and described at the end of Section 4.1.

For minimum-cost spanning tree, consider a complete graph $G = (V, A)$ with $|V| = k$ nodes, $b_a = \epsilon$ and $l_a = 0$ for all $a \in \{(i, i+1) : i < k\}$, and $l_a = M > 0$ for all $a \notin \{(i, i+1) : i < k\}$ with $k\,\epsilon < M$. One can check that any cover of A is of size at least $(k-2)/2$. In contrast, $\mathcal{D}(B) = \emptyset$, independent of k, thus $z_L^*(B) = 0$. Note that the Steiner tree problem generalizes the minimum-cost spanning tree problem, thus this instance covers the Steiner tree case as well.

For minimum-cost perfect matching consider a complete graph $G = (V, A)$ with $|V| = 2k$ nodes, $b_a = \epsilon$ and $l_a = 0$ for all $a \in \{(2i+1, 2i+2) : i < k\}$, and $l_a = M > 0$ for all $a \notin \{(2i+1, 2i+2) : i < k\}$ with $k\,\epsilon < M$. One can check that any cover of A is of size at least $2(k-1)$. In contrast, $\mathcal{D}(B) = \emptyset$, independent of k, thus $z_L^*(B) = 0$.

Finally, for the knapsack problem, consider the items $A := \{0, 1, \ldots, Ck\}$, where $C \in \mathbb{N}$ denotes the knapsack capacity, and weights $w \in \mathbb{R}^{Ck+1}$ so that $w_0 = C$, and $w_i = 1$ for $i > 0$. In addition, set $u_0 = 0$ and $b_0 = \epsilon$ and $u_i = -M < 0$ for $i > 0$ (where u_a denotes the upper bound on the range of the "utility" distribution of ground element a), with $\epsilon < M$. Note that in this case the problem is of utility maximization. One can check that any cover of A is of size at least $k+1$. In contrast, $\mathcal{D}(B) = \emptyset$, independent of k, thus $z_L^*(B) = 0$.

A.1.2. A near-optimal policy

THEOREM 4. *Consider $\gamma \in (0,1)$ and $\varepsilon > 0$ arbitrary. Suppose that F is regular and Assumption 1 holds, and let $\pi^*(\gamma)$ denote the LBP policy when we choose $n_i := \max\{\lfloor e^{i^{1/(1+\varepsilon)}} \rfloor, n_{i-1}+1\}$ for all $i \geq 2$, then*

$$\lim_{N\to\infty} \frac{R^{\pi^*(\gamma)}(F,N)}{(\ln N)^{1+\varepsilon}} \leq z_L^*\left(\mathbb{E}_F\left\{B_n\right\}\right) + \gamma z_C^*\left(\mathbb{E}_F\left\{B_n\right\}\right). \tag{11}$$

The regret of the policy π^* (we abuse the notation and ignore the dependence of π^* on γ) stems from two sources: exploration efforts and exploitation errors. That is,

$$R^{\pi^*}(F,N) = R_1(F,N) + R_2(F,N), \tag{A-6}$$

where $R_1(F,N)$ is the exploration-based regret, i.e., that incurred at instance n during cycle i if $\tilde{T}_n(a) < \gamma\ i$ for some $a \in A$, or alternatively when sampling a solution, picking $S_n \neq S^*$, and $R_2(F,N)$ is the exploitation-based regret, i.e., that incurred when $\tilde{T}_n(a) \geq \gamma\ i$ for all $a \in A$ and we sample $S_n = S^*$. We prove the result by bounding each term above separately. (We dropped the dependence of $R_1(F,N)$ and $R_2(F,N)$ on the policy π^* to simplify notation.)

In the remainder of this proof, $\mathbb{E}$ and $\mathbb{P}$ denote expectation and probability when costs are distributed according to F and policy π^* is implemented.

Step 1 (Exploitation-based regret). Exploitation-based regret during cycle i is due to implementing suboptimal solutions when minimum cover-based exploration requirements are met.

Let i' denote a finite upper bound on the first cycle in which one is sure to randomize a solution on at least one instance, e.g., $i' := 1 + \inf\{i \in \mathbb{N}, i \geq 2 : n_i \geq i\,|A|\,, n_{i+1} - n_i > |A|\}$. (Note that i' does not depend on N).

Fix $i \geq i'$ and note that when cover-based exploration requirements are met for $n \in [n_i, n_{i+1}-1]$, one may exploit, that is, one may implement $S_n = S^*$ for some $S^* \in \mathcal{S}^*(\bar{B}_{n_i}))$. We use the event $\left\{S_n \in \mathcal{S}^*(\bar{B}_{n_i})\right\}$ to denote exploitation. Thus,

$$\begin{aligned} R_2(F,N) &\leq n_{i'}\Delta_{max}^F + \sum_{i=i'}^{\lceil (\ln N)^{1+\varepsilon}\rceil} \sum_{n=n_i}^{n_{i+1}-1} \mathbb{E}\left\{\mathbf{1}\left\{\tilde{T}_n(a) \geq \gamma(i-1)\,, \forall a \in A\,, S_n \in \mathcal{S}^*(\bar{B}_{n_i})\right\} \Delta_{S_n}^F\right\} \\ &\leq n_{i'}\Delta_{max}^F + \sum_{i=i'}^{\infty} (n_{i+1}-n_i)\,\mathbb{P}\left\{\mathcal{S}^*(\bar{B}_{n_i}) \not\subseteq \mathcal{S}^*(\mathbb{E}\left\{B_n\right\}), \tilde{T}_{n_i}(a) \geq \gamma(i-1), \forall a \in A\right\} \Delta_{max}^F \end{aligned} \tag{A-7}$$

where $\Delta_{max}^F := \max_{S\in\mathcal{S}}\left\{\Delta_S^{\mathbb{E}\{B_n\}}\right\}$.

Next, we find an upper bound for the probability inside the sum in (A-7). For this, note that

$$\left\{\mathcal{S}^*(\bar{B}_{n_i}) \not\subseteq \mathcal{S}^*(\mathbb{E}\left\{B_n\right\})\right\} \subseteq \bigcup_{a\in A}\left\{\left|\bar{b}_{a,n_i} - \mathbb{E}\left\{b_{a,n}\right\}\right| \geq \frac{\Delta_{min}^F}{2s}\right\}, \tag{A-8}$$

where $s := \max\{|S| : S \in \mathcal{S}\}$ and $\Delta^F_{min} := \min\left\{\Delta_S^{\mathbb{E}\{B_n\}} : S \in \mathcal{S} \setminus \mathcal{S}^*\left(\mathbb{E}\{B_n\}\right)\right\}$ denote the maximum solution size and minimum optimality gap for the full-information instance, respectively. (We assume, without loss of generality, that Δ^F_{max} and Δ^F_{min} are both positive, since otherwise, the problem is trivial.) Indeed, note that

$$\left\{\left|\bar{b}_{a,n_i} - \mathbb{E}\{b_{a,n}\}\right| < \frac{\Delta^F_{min}}{2s}, \forall a \in A\right\} \subseteq \left\{\sum_{a\in S^*} \bar{b}_{a,n_i} < \sum_{a\in S} \bar{b}_{a,n_i}, \forall S^* \in \mathcal{S}^*(\mathbb{E}\{B_n\}), S \in \mathcal{S} \setminus \mathcal{S}^*\left(\mathbb{E}\{B_n\}\right)\right\}.$$

The next proposition, whose proof can be found in Appendix B, allows us to bound (A-7) using the observation above.

PROPOSITION 5. *For any fixed $a \in A$, $n \in \mathbb{N}$, $k \in \mathbb{N}$, and $\epsilon > 0$ we have that*

$$\mathbb{P}\left\{\left|\bar{b}_{a,n} - \mathbb{E}\{b_{a,n}\}\right| \geq \epsilon, \tilde{T}_n(a) \geq k\right\} \leq 2\exp\left\{-\frac{2\epsilon^2 k}{\mathcal{L}^2}\right\},$$

where $\mathcal{L} := \max\{u_a - l_a : a \in A\}$.

Using the above, the union bound, and (A-8), we have that

$$\mathbb{P}\left\{\mathcal{S}^*(\bar{B}_{n_i}) \not\subseteq \mathcal{S}^*(\mathbb{E}\{B_n\}), \tilde{T}_{n_i}(a) \geq \gamma(i-1), \forall a \in A\right\} \leq$$
$$\sum_{a\in A} \mathbb{P}\left\{\left|\bar{b}_{a,n_i} - \mathbb{E}\{b_{a,n}\}\right| \geq \frac{\Delta^F_{min}}{2s}, \tilde{T}_{n_i}(a) \geq \gamma(i-1)\right\} \leq 2|A|\exp\left\{-\frac{\left(\Delta^F_{min}\right)^2\gamma(i-1)}{2s^2\mathcal{L}^2}\right\}. \quad \text{(A-9)}$$

Now, for $i \geq i'$, one has that $n_{i+1} \leq e^{(i+1)^{1/(1+\varepsilon)}}$ and $n_i \geq e^{(i-1)^{1/(1+\varepsilon)}}$. Hence, $n_{i+1} - n_i \leq e^{(i+1)^{1/(1+\varepsilon)}}$. Using this, (A-7) and (A-9) we conclude that

$$R_2(F,N) \leq \Delta^F_{max}\left(n_{i'} + \sum_{i=i'}^{\infty} 2|A|\exp\left\{(i+1)^{1/(1+\varepsilon)} - \frac{\left(\Delta^F_{min}\right)^2\gamma(i-1)}{2s^2\mathcal{L}^2}\right\}\right).$$

Because $(i+1)^{1/(1+\varepsilon)} < i\frac{\left(\Delta^F_{min}\right)^2\gamma}{2s^2\mathcal{L}^2}$ for i large enough, we conclude that $R_2(F,N) \leq C_1$, for a positive finite constant C_1, independent of N.

Step 2 (Exploration-based regret). We separate the exploration-based regret into cover-based and LBP-based regrets. The former arises at instance n when there exists $a \in A$ such that $\tilde{T}_n(a) < \gamma\, i$. The latter arises when the cover-based exploration requirements are met and one samples $S_n \neq S^*$ for $S^* \in \mathcal{S}^*(\bar{B}_{n_i})$. Let $R_1^C(F,N)$ and $R_1^{LBP}(F,N)$ denote the cover-based and LBP-based exploration regrets, respectively, so that

$$R_1(F,N) := R_1^C(F,N) + R_1^{LBP}(F,N).$$

Step 2.1 (Cover-based exploration regret). We first bound the cover-based exploration regret. Let $\mathbf{C}$ denote the set of minimal covers of A, and Δ^C_{min} denote the minimum optimality gap for Cover($\mathbb{E}\{B_n\}$) problem in (10), i.e.,

$$\Delta^C_{min} := \min\left\{\left(\sum_{S\in\mathcal{E}} \Delta_S^{\mathbb{E}\{B_n\}}\right) - z^*_C(\mathbb{E}\{B_n\}) : \mathcal{E}\in\mathbf{C}\setminus\Gamma_C(\mathbb{E}\{B_n\})\right\}.$$

We assume that $\Delta^C_{min} > 0$, since otherwise, the cover problem is trivial. Consider $i > i'$ and let $\mathcal{E}_i \in \Gamma_C\left(\bar{B}_{n_i}\right)$ denote the cover-based exploration set for any instance $n \in [n_i, n_{i+1} - 1]$. Define $c := \max\{|\mathcal{E}| : \mathcal{E}\in\mathbf{C}\}$ as the maximum size of a minimal cover of A and let $I := \left\{i \le (\ln N)^{1+\varepsilon} : i > i',\ \lceil\gamma(i-1)\rceil < \lceil\gamma\ i\rceil\right\}$ denote the set of cycles in which cover-based exploration requirements are increased. Noting that $\tilde{T}_{n_i}(a) \ge \gamma(i-1)$ for all $a\in A$ when $i > i'$, we have that

$$\begin{aligned}
R^C_1(F,N) &\le c\,i'\Delta^F_{max} + \sum_{i\in I}\mathbb{E}\left\{\mathbf{1}\left\{\tilde{T}_{n_i}(a)\ge\gamma(i-1)\ \forall a\in A\,,\,\mathcal{E}_i\in\Gamma_C(\mathbb{E}\{B_n\})\right\}\sum_{S\in\mathcal{E}_i}\Delta_S^{\mathbb{E}\{B_n\}}\right\}\\
&\qquad + \sum_{i\in I}\mathbb{E}\left\{\mathbf{1}\left\{\tilde{T}_{n_i}(a)\ge\gamma(i-1)\ \forall a\in A\,,\,\mathcal{E}_i\notin\Gamma_C(\mathbb{E}\{B_n\})\right\}\sum_{S\in\mathcal{E}_i}\Delta_S^{\mathbb{E}\{B_n\}}\right\}\\
&\le c\,i'\Delta^F_{max} + \left(\gamma(\ln N)^{1+\varepsilon}+1\right) z^*_C(\mathbb{E}\{B_n\})\\
&\qquad + \Delta^F_{max}\,c\sum_{i\in I}\mathbb{P}\left\{\tilde{T}_{n_i}(a)\ge\gamma(i-1)\ \forall a\in A\,,\,\mathcal{E}_i\notin\Gamma_C(\mathbb{E}\{B_n\})\right\} \qquad \text{(A-10)}
\end{aligned}$$

Next, we bound the probability inside the sum in (A-10). For that, observe

$$\left\{\Gamma_C(\bar{B}_{n_i})\not\subseteq\Gamma_C(\mathbb{E}\{B_n\})\right\}\subseteq\bigcup_{a\in A}\left\{\left|\bar{b}_{a,n_i}-\mathbb{E}\{b_{a,n}\}\right|\ge\frac{\Delta_1}{4cs}\right\}, \qquad \text{(A-11)}$$

where $\Delta_1 := \min\{\Delta^C_{min},\Delta^F_{min}\}$. Indeed, note that

$$\begin{aligned}
\left\{\left|\bar{b}_{a,n_i}-\mathbb{E}\{b_{a,n}\}\right|<\frac{\Delta_1}{4cs}\,,\,\forall a\in A\right\} &\subseteq \left\{\left|\Delta_S^{\bar{B}_{n_i}}-\Delta_S^{\mathbb{E}\{B_n\}}\right|<\frac{\Delta_1}{2c}\,,\,\forall S\in\mathcal{S}\right\}\\
&\subseteq \left\{\left|\sum_{S\in\mathcal{E}}\left(\Delta_S^{\bar{B}_{n_i}}-\Delta_S^{\mathbb{E}\{B_n\}}\right)\right|<\frac{\Delta_1}{2}\,,\,\forall\mathcal{E}\in\mathbf{C}\right\}\\
&\subseteq \left\{\sum_{S\in\mathcal{E}}\Delta_S^{\bar{B}_{n_i}} > \sum_{S\in\mathcal{E}^*}\Delta_S^{\bar{B}_{n_i}}\,,\,\forall\mathcal{E}^*\in\Gamma_C(\mathbb{E}\{B_n\})\,,\,\mathcal{E}\in\mathbf{C}\setminus\Gamma_C(\mathbb{E}\{B_n\})\right\},
\end{aligned}$$

where $\Delta_S^{\bar{B}_{n_i}}$ is as defined after LBP formulation in (8). We note that as discussed in (A-8) in Step 1, by taking $\Delta_1\le\Delta^F_{min}$, we also ensure that $\left\{\mathcal{S}^*(\bar{B}_{n_i})\subseteq\mathcal{S}^*(\mathbb{E}\{B_n\})\right\}$.
Using Proposition 5, the union bound, and (A-11), we have that

$$\mathbb{P}\left\{\mathcal{E}_i \notin \Gamma_C(\mathbb{E}\{B_n\}), \tilde{T}_{n_i}(a) \geq \gamma(i-1), \forall a \in A\right\} \leq$$
$$\sum_{a\in A} \mathbb{P}\left\{\left|\bar{b}_{a,n_i} - \mathbb{E}\{b_{a,n}\}\right| \geq \frac{\Delta_1}{4sc}, \tilde{T}_{n_i}(a) \geq \gamma(i-1)\right\} \leq 2|A| \exp\left\{-\frac{(\Delta_1)^2 \gamma(i-1)}{8s^2c^2\mathcal{L}^2}\right\}. \quad \text{(A-12)}$$

Using the above and (A-10) we obtain that

$$R_1^C(F,N) \leq \gamma (\ln N)^{1+\varepsilon} z_C^*(\mathbb{E}\{B_n\}) + C_2,$$

for a positive finite constant C_2, independent of N.

Step 2.2 (LBP-based exploration regret). Consider now the LBP-based exploration regret $R_1^{LBP}(F,N)$. Let $\Delta^{\mathcal{D}}$ denote a uniform upper bound on the precision of each mean cost estimate necessary to *approximately* reconstruct the set $\mathcal{D}(\mathbb{E}\{B_n\})$. That is, $\Delta^{\mathcal{D}} := \min\{\Delta_{min}^F, \Delta_2^{\mathcal{D}}, \Delta_3^{\mathcal{D}}\}/(2s)$, where

$$\Delta_2^{\mathcal{D}} := \min\left\{\min\left\{\Delta_S^{\mathbb{E}\{B_n\}_D} : S \notin \mathcal{S}^*(\mathbb{E}\{B_n\}_D)\right\} : D \subseteq A \setminus H, \mathcal{S}^*(\mathbb{E}\{B_n\}) = \mathcal{S}^*(\mathbb{E}\{B_n\}_D)\right\},$$
$$\Delta_3^{\mathcal{D}} := \min\{z^*(\mathbb{E}\{B_n\}) - z^*(\mathbb{E}\{B_n\}_D) : D \subseteq A \setminus H, \mathcal{S}^*(\mathbb{E}\{B_n\}) \neq \mathcal{S}^*(\mathbb{E}\{B_n\}_D)\},$$

Δ_{min}^F is as defined in Step 1, $H := \bigcup_{S^* \in \mathcal{S}^*(\mathbb{E}\{B_n\})} \bigcup_{a\in S^*} \{a\}$, and $\mathbb{E}\{B_n\}_D = (\mathbb{E}\{b_{a,n}\}, a \in A \setminus D) \cup (l_a : a \in D)$. The first precision threshold Δ_{min}^F ensures that $\mathcal{S}^*(\bar{B}_n) \subseteq \mathcal{S}^*(\mathbb{E}\{B_n\})$. The second threshold ensures that

$$\mathcal{D}(\bar{B}_n) \subseteq \mathcal{D}(\mathbb{E}\{B_n\}) \cup 2^H,$$

and the third one ensures that $\mathcal{D}(\mathbb{E}\{B_n\}) \subseteq \mathcal{D}(\bar{B}_n)$. The first statement is supported by Step 1 (see (A-8)). The second follows from noting that: (i) for $D \notin \mathcal{D}(\mathbb{E}\{B_n\})$,

$$\bigcap_{a\in A} \left\{\left|\bar{b}_{a,n} - \mathbb{E}\{b_{a,n}\}\right| < \Delta^{\mathcal{D}}\right\} \subseteq \left\{z^*(\bar{B}_n) = z^*\left(\left(\bar{B}_n\right)_D\right)\right\},$$

implying that $D \notin \mathcal{D}\left(\bar{B}_n\right)$; and (ii) not all solutions in $\mathcal{S}^*(\mathbb{E}\{B_n\})$ are necessarily optimal in the approximate problem (i.e., using the average costs), therefore, some of their ground elements might belong to $\mathcal{D}(\bar{B}_n)$. Similarly, the third statement follows from noting that for $D \in \mathcal{D}(\mathbb{E}\{B_n\})$,

$$\bigcap_{a\in A} \left\{\left|\bar{b}_{a,n} - \mathbb{E}\{b_{a,n}\}\right| < \Delta^{\mathcal{D}}\right\} \subseteq \left\{z^*(\bar{B}_n) > z^*\left(\left(\bar{B}_n\right)_D\right)\right\},$$

implying that $D \in \mathcal{D}\left(\bar{B}_n\right)$. We conclude that

$$\bigcap_{a\in A} \left\{\left|\bar{b}_{a,n} - \mathbb{E}\{b_{a,n}\}\right| < \Delta^{\mathcal{D}}\right\} \subseteq \left\{\mathcal{D}\left(\bar{B}_n\right) = \mathcal{D}(\mathbb{E}\{B_n\}) \cup H_o\right\},$$

form some $H_o \in 2^H$. While we assume, without loss of generality, that Δ^F_{min} and $\Delta^{\mathcal{D}}_2$ are positive (since otherwise, the problem is trivial), Assumption 1 implies that $\Delta^{\mathcal{D}}_3 > 0$. Thus, we have that $\Delta^{\mathcal{D}} > 0$.

Consider now the issue of approximating the K_D constants. We denote such estimates by $\hat{K}_D$. By the continuity of $I_a(\cdot,\cdot)$ for all $a \in A$, we have that $K_D(B)$ is also continuous for all $D \in \mathcal{D}(\mathbb{E}\{B_n\})$. In addition, because it is known that $K_D(B) \leq K$, there exists a finite constant $\kappa > 0$ such that

$$\left|\hat{K}_D(\bar{B}_n) - K_D(\mathbb{E}\{B_n\})\right| \leq \kappa \sum_{a\in A} \left|\bar{b}_{a,n} - \mathbb{E}\{b_{a,n}\}\right|,$$

for $\bar{B}_n$ in a neighborhood of $\mathbb{E}\{B_n\}$ (specifically, we consider a ball -using infinite norm- of radious lower than $\varrho/(|A|\kappa)$ centered at $\mathbb{E}\{B_n\}$ for $\varrho > 0$ arbitrary). Note that we make use of the uniform bound and use the approximation

$$\hat{K}_D(B) := K_D(B) \wedge K.$$

This, in turn, implies that $\hat{K}_D(B) \leq K$.

Define $\Delta^K := \varrho/(|A|\kappa)$ for $\varrho > 0$ arbitrary. We conclude that

$$\bigcap_{a\in A} \left\{\left|\bar{b}_{a,n} - \mathbb{E}\{b_{a,n}\}\right| < \Delta^K\right\} \subseteq \left\{\left|\hat{K}_D(\bar{B}_n) - K_D(\mathbb{E}\{B_n\})\right| < \varrho, \quad D \in \mathcal{D}(\mathbb{E}\{B\}_n)\right\}.$$

Let $(x^n, y^n) \in \Gamma_L(\bar{B}_n)$, and consider $(x^*, y^*) \in \Gamma_L(\mathbb{E}\{B_n\})$, augmented so that $y^*_{S^*} = K$ for all $S^* \in \mathcal{S}^*(\mathbb{E}\{B_n\})$ (note that because $\Delta^F_{S^*} = 0$ for all $S^* \in \mathcal{S}^*(\mathbb{E}\{B_n\})$, one can make this augmentation without affecting the objective value of $LBP(\mathbb{E}\{B_n\})$). Suppose that $\|\bar{B}_n - \mathbb{E}\{B_n\}\|_\infty < \delta/(2s)$ for some $0 < \delta < \min\{\Delta^K, \Delta^{\mathcal{D}}, \varrho\}$, then we have that

$$\max\{x^n_a + \delta : a \in D\} \geq K_D(\mathbb{E}\{B_n\}),\ D \in \mathcal{D}(\mathbb{E}\{B_n\}) \tag{A-13}$$

$$\max\{x^*_a + \delta : a \in D\} \geq \hat{K}_D(\bar{B}_n),\ D \in \mathcal{D}(\bar{B}_n). \tag{A-14}$$

For $z \in \mathbb{R}^k$ and $\delta > 0$, we define z^δ so that $z^\delta_j := z_j + \delta \mathbf{1}\{z_j > 0\}$, $j \leq k$, where z_j is the j-th element of z. From (A-14) we conclude that $(x^{*,\delta}, y^{*,\delta})$ is feasible to $LBP(\bar{B}_n)$. Seeing that $\|\bar{B}_n - \mathbb{E}\{B_n\}\|_\infty < \delta/(2s)$, we have $\left|\Delta^{\mathbb{E}\{B_n\}}_S - \Delta^{\bar{B}_n}_S\right| < \delta$ for all $S \in \mathcal{S}$. Therefore, we have that

$$\begin{aligned}\sum_{S\in\mathcal{S}} y^n_S \Delta^{\mathbb{E}\{B_n\}}_S &\overset{(a)}{\leq} \sum_{S\in\mathcal{S}} y^n_S \Delta^{\bar{B}_n}_S + |\mathcal{S}| K\delta \\ &\overset{(b)}{\leq} \sum_{S\in\mathcal{S}} y^{*,\delta}_S \Delta^{\bar{B}_n}_S + |\mathcal{S}| K\delta\end{aligned}$$

$$\begin{aligned}
&\overset{(c)}{\leq} \sum_{S\in\mathcal{S}} y_S^* \Delta_S^{\mathbb{E}\{B_n\}} + \delta\,|\mathcal{S}|\left(\delta + \Delta^F_{\max} + K\right) + |\mathcal{S}|\,K\delta \\
&= z_L^*(\mathbb{E}\{B_n\}) + \delta\,|\mathcal{S}|\left(\delta + \Delta^F_{\max} + 2K\right).
\end{aligned}$$

where (a) follows from the fact that $y_S^n \leq K$ for all $S \in \mathcal{S}$ (this because $\hat{K}_D(\bar{B}_n) \leq K$), (b) comes from that $(x^{*,\delta}, y^{*,\delta})$ is feasible to $LBP(\bar{B}_n)$, and (c) follows from that $\left|\Delta_S^{\mathbb{E}\{B_n\}} - \Delta_S^{\bar{B}_n}\right| < \delta$ and $y_S^{*,\delta} \leq y_S^* + \delta$ for all $S \in \mathcal{S}$, and $y_S^* \leq K$ for all $S \in \mathcal{S}$. Seeing that $\delta < \Delta^{\mathcal{D}} < \Delta^F_{min}$, taking $\delta \leq \varrho z_L^*(\mathbb{E}\{B_n\})/\left(|\mathcal{S}|\left(\Delta^F_{min} + \Delta^F_{\max} + 2K\right)\right)$, we have that

$$\sum_{S\in\mathcal{S}} y_S^n \Delta_S^{\mathbb{E}\{B_n\}} \leq (1+\varrho) z_L^*(\mathbb{E}\{B_n\}). \tag{A-15}$$

Consider $i > i'$ and let $(x_i, y_i) \in \Gamma_L(\bar{B}_{n_i})$ be the solution used for LBP-based exploration for $n \in [n_i, n_{i+1} - 1]$. In what follows, with abuse of notation, we use the event $\left\{S_n \in \Gamma_L\left(\bar{B}_n\right)\right\}$ to denote the LBP-based exploration. We have that

$$\begin{aligned}
R_1^{LBP}(F,N) &\leq n_{i'}\Delta^F_{max} + \sum_{i=i'}^{\lceil (\ln N)^{1+\varepsilon}\rceil} \sum_{n=n_i}^{n_{i+1}-1} \mathbb{E}\left\{\mathbf{1}\left\{\tilde{T}_{n_i}(a) \geq \gamma(i-1)\ \forall a \in A\,,\, S_n \in \Gamma_L\left(\bar{B}_n\right)\right\}\Delta_{S_n}^{\mathbb{E}\{B_n\}}\right\} \\
&\leq n_{i'}\Delta^F_{max} + \sum_{i=i'}^{\lceil (\ln N)^{1+\varepsilon}\rceil} \sum_{n=n_i}^{n_{i+1}-1} \mathbb{E}\left\{\mathbf{1}\left\{\tilde{T}_{n_i}(a) \geq \gamma(i-1), \left|\bar{b}_{a,n} - \mathbb{E}\{b_{a,n}\}\right| < \delta/(2s), \forall a \in A, S_n \in \Gamma_L\left(\bar{B}_n\right)\right\}\Delta_{S_n}^{\mathbb{E}\{B_n\}}\right\} \\
&\quad + \sum_{i=i'}^{\lceil (\ln N)^{1+\varepsilon}\rceil} \sum_{n=n_i}^{n_{i+1}-1} \mathbb{E}\left\{\mathbf{1}\left\{\tilde{T}_{n_i}(a) \geq \gamma(i-1), \forall a \in A, \cup_{a\in A}\left\{\left|\bar{b}_{a,n} - \mathbb{E}\{b_{a,n}\}\right| \geq \delta/(2s)\right\}, S_n \in \Gamma_L\left(\bar{B}_n\right)\right\}\Delta_{S_n}^{\mathbb{E}\{B_n\}}\right\} \\
&\leq n_{i'}\Delta^F_{max} + \sum_{i=i'}^{\lceil (\ln N)^{1+\varepsilon}\rceil} (1+\varrho)\ z_L^*(\mathbb{E}\{B_n\}) \\
&\quad + \sum_{i=i'}^{\infty} (n_{i+1} - n_i)\,\Delta^F_{\max} \sum_{a\in A} \mathbb{P}\left\{\left|\bar{b}_{a,n} - \mathbb{E}\{b_{a,n}\}\right| \geq \delta/(2s), \tilde{T}_{n_i}(a) \geq \gamma(i-1)\right\}.
\end{aligned} \tag{A-16}$$

Using Proposition 5 to bound the probability in (A-16), we have that

$$R_1^{LBP}(F,N) \leq n_{i'}\Delta^F_{max} + (\ln N)^{1+\varepsilon}(1+\varrho)\ z_L^*(\mathbb{E}\{B_n\}) + \sum_{i=i'}^{\infty} e^{(i+1)^{1/(\varepsilon+1)}} \Delta^F_{\max} 2\,|A| \exp\left\{-\frac{\delta^2\ \gamma(i-1)}{2s^2\mathcal{L}^2}\right\}.$$

Because $(i+1)^{1/(1+\varepsilon)} < i\frac{\delta^2\gamma}{2s^2\mathcal{L}^2}$ for i large enough, we conclude that

$$R_1^{LBP}(F,N) \leq (\ln N)^{1+\varepsilon}(1+\varrho)\ z_L^*(\mathbb{E}\{B_n\}) + C_3$$

for a positive finite constant C_3, independent of N. Putting all the above together, we conclude that

$$R^{\pi^*}(F,N) \leq \left((1+\varrho)\ z_L^*(\mathbb{E}\{B_n\}) + \gamma z_C^*(\mathbb{E}\{B_n\})\right)(\ln N)^{1+\varepsilon} + C_4,$$

for a finite positive constant C_4, independent of N.

We finally note that the solution to the estimated *Cover* and LBP problems converge a.s. to optimal and ϱ-optimal solutions, respectively. For this, note that Proposition 5, (A-12) and (A-16) imply (via Borel-Cantelli) that $\mathbb{P}\{\mathcal{E}_i \in \Gamma_C(\mathbb{E}\{B_n\}) \text{ e.v.}\} = 1$ and

$$\mathbb{P}\{(x_i^\varrho, y_i^\varrho) \text{ is a } \varrho\text{-optimal solution to } LBP(\mathbb{E}\{B_n\}) \text{ e.v.}\} = 1.$$

The result follows from noting that one can choose ϱ arbitrarily small.

A.2. Ommitted Proofs from Section 5

LEMMA 1. *Consider a cost vector $B \in \mathcal{B}$. An optimal solution to the linear relaxation of $OCP(B)$ is also optimal to formulation $LBP(B)$ when one replaces $K_D(B)$ by $K > 0$ for all $D \in \mathcal{D}(B)$.*

We prove the result by contradiction. Without loss of generality, we assume that $K = 1$. Let (x,y) be a feasible solution to *R-OCP*$(\mathbb{E}_F\{B_n\})$, and suppose that there is a $D \in \mathcal{D}$ such that $\max\{x_a : a \in D\} = 0$. By the definition of $\mathcal{D}$, one has that $z^*(\mathbb{E}_F\{B_n\}_D) < z^*(\mathbb{E}_F\{B_n\})$, thus

$$\begin{aligned} z^*(\mathbb{E}_F\{B_n\}_D) &= \sum_{a \in S^* \setminus D} \mathbb{E}_F\{b_{a,n}\} + \sum_{a \in D} l_a \\ &\overset{(a)}{\geq} \sum_{a \in S^*} \left(l_a(1-x_a) + \mathbb{E}_F\{b_{a,n}\}\, x_a\right) \\ &\overset{(b)}{\geq} z^*(\mathbb{E}_F\{B_n\}), \end{aligned}$$

for $S^* \in \mathcal{S}^*(\mathbb{E}_F\{B_n\}_D)$, where (a) follows from the fact that $l_a = (l_a(1-x_a) + \mathbb{E}_F\{b_{a,n}\}\, x_a)$, for $a \in D$, and $\mathbb{E}_F\{b_{a,n}\} \geq (l_a(1-x_a) + \mathbb{E}_F\{b_{a,n}\}\, x_a)$, for $a \notin D$, and (b) follows from the fact that (x,y) satisfies constraints (13c) (because it is feasible to *R-OCP*$(\mathbb{E}_F\{B_n\})$). The last inequality above contradicts $z^*(\mathbb{E}_F\{B_n\}_D) < z^*(\mathbb{E}_F\{B_n\})$, thus we have that $\max\{x_a : a \in D\} = 1$ for all $D \in \mathcal{D}$, therefore (x,y) is feasible to $LBP(\mathbb{E}_F\{B_n\})$.

Now, let (x,y) be a feasible solution to $LBP(\mathbb{E}_F\{B_n\})$ such that $x_a \in \{0,1\}$ for all $a \in A$, and that $x_a = 1$ and $y_{S^*} = 1$ for $a \in S^*$ and $S^* \in \mathcal{S}^*(\mathbb{E}_F\{B_n\})$ (note that one can restrict attention only to feasible solutions to $LBP(\mathbb{E}_F\{B_n\})$ with x integral, and $\Delta_{S^*}^F = 0$ for all $S^* \in \mathcal{S}^*(\mathbb{E}_F\{B_n\})$,

thus this extra requirement does not affect the solution to $LBP(\mathbb{E}_F\{B_n\})$). Suppose (x,y) is not feasible to *R-OCP*$(\mathbb{E}_F\{B_n\})$, i.e., there exists some $S\in\mathcal{S}$ such that

$$\sum_{a\in S}(l_a(1-x_a)+\mathbb{E}_F\{b_{a,n}\}x_a)<z^*(\mathbb{E}_F\{B_n\}). \tag{A-17}$$

Let S_0 be one such S that additionally minimizes the left-hand side in (A-17) (in case of ties we pick any minimizing solution S_0 with smallest value of $|\{a\in S_0: x_a=0\}|$). Then $D=\{a\in S_0: x_a=0\}$ belongs to $\mathcal{D}$ (or a subset of it does), This contradicts the feasibility of (x,y) to $LBP(\mathbb{E}_F\{B_n\})$, because if (x,y) is feasible to $LBP(\mathbb{E}_F\{B_n\})$, then we must have $\max\{x_a: a\in D\}\geq 1$ for all $D\in\mathcal{D}$. Thus, we conclude that (x,y) is feasible to *R-OCP*$(\mathbb{E}_F\{B_n\})$.

Summarizing, feasible solutions to *R-OCP*$(\mathbb{E}_F\{B_n\})$ are feasible to $LBP(\mathbb{E}_F\{B_n\})$, and feasible solutions to $LBP(\mathbb{E}_F\{B_n\})$ that cover all optimal elements in A are feasible to *R-OCP*$(\mathbb{E}_F\{B_n\})$. The result follows from noting that there is always an optimal solution to $LBP(\mathbb{E}_F\{B_n\})$ such that x is integral, and $x_a=1$ and $y_{S^*}=1$ for $a\in S^*$ for all $S^*\in\mathcal{S}^*(\mathbb{E}_F\{B_n\})$.

THEOREM 5. *Let $\pi_h(\gamma,\varrho)$ denote the hybrid policy and suppose that Assumption 2 holds. If we choose $n_i:=\max\{\lfloor e^{i^{1/(1+\varepsilon)}}\rfloor, n_{i-1}+1\}$ with $\varepsilon>0$ arbitrary, for all $i\geq 2$, and we select ϱ to be smaller than the minimum optimality gap for OCP $(\mathbb{E}_F\{B_n\})$, then for $\gamma\in(0,1)$*

$$\lim_{N\to\infty}\frac{R^{\pi_h(\gamma,\varrho)}(F,N)}{(\ln N)^{1+\varepsilon}}\leq z^*_{OCP}(\mathbb{E}_F\{B_n\})+\gamma z^*_C(\mathbb{E}_F\{B_n\}).$$

As in the case of the LBP policy π^*, the regret of policy π_h (we again ignore the dependence of the policy on γ and ϱ) stems from three sources: cover-based and OCP-based exploration efforts, and exploitation errors. That is,

$$R^{\pi_h}(F,N)=R_1^C(F,N)+R_1^{OCP}(F,N)+R_2(F,N), \tag{A-18}$$

where $R_1^C(F,N)$ is the cover-based exploration regret, i.e., that incurred at instance n during cycle i if $\tilde{T}_n(a)<\gamma\, i$ for some $a\in A$, $R_1^{OCP}(F,N)$ is the OCP-based exploration regret, i.e., that incurred at instance n during cycle i if $\tilde{T}_n(a)<i$ for some $a\in C$, and $R_2(F,N)$ is the exploitation-based regret, i.e., that incurred when exploration conditions are met and one implements solution $S_n=S^*$.

We prove the result by bounding each term in (A-18) separately. It turns out that the bounds for $R_1^C(F,N)$ and $R_2(F,N)$ in Step 1 and Step 2.1 in the proof of Theorem 4 apply to this setting unmodified, thus we omit them here. Next, we bound the OCP-based exploration regret $R_1^{OCP}(F,N)$.

As in the proof of the LBP policy, in the remainder of this proof, $\mathbb{E}$ and $\mathbb{P}$ denote expectation and probability when costs are distributed according to F and policy π_h is implemented.

Step 2.2' (OCP-based exploration regret).

Following the arguments in Step 2.2 of the proof of Theorem 4, we first define the minimum precision threshold on the accuracy of mean cost estimates necessary to reconstruct the solution to $OCP(\mathbb{E}\{B_n\})$. For that, we define $\Delta^{\mathcal{D}} := \min\{\Delta^F_{min}, \Delta^{\mathcal{D}}_2, \Delta^{\mathcal{D}}_3, \Delta^{\mathcal{D}}_4\}/(8sc)$, where

$$\begin{aligned}
\Delta^{\mathcal{D}}_2 &:= \min\left\{\min\left\{\Delta^{\mathbb{E}\{B_n\}_D}_S : S \notin \mathcal{S}^*\left(\mathbb{E}\{B_n\}_D\right)\right\} : D \subseteq A \setminus H,\, \mathcal{S}^*\left(\mathbb{E}\{B_n\}\right) = \mathcal{S}^*\left(\mathbb{E}\{B_n\}_D\right)\right\},\\
\Delta^{\mathcal{D}}_3 &:= \min\left\{z^*(\mathbb{E}\{B_n\}) - z^*\left(\mathbb{E}\{B_n\}_D\right) :\, D \subseteq A \setminus H,\, \mathcal{S}^*\left(\mathbb{E}\{B_n\}\right) \neq \mathcal{S}^*\left(\mathbb{E}\{B_n\}_D\right)\right\},\\
\Delta^{\mathcal{D}}_4 &:= \min\left\{\left(\sum_{S\in\mathcal{G}} \Delta^{\mathbb{E}\{B_n\}}_S\right) - z^*_{OCP}(\mathbb{E}\{B_n\}) : (C,\mathcal{G}) \in \mathbf{G} \setminus \Gamma_{OCP}(\mathbb{E}\{B_n\})\right\},
\end{aligned}$$

and $\mathbf{G}$ denotes the set of all feasible solutions $(C,\mathcal{G})$ to $OCP(\mathbb{E}\{B_n\})$ problem. Note that as in the proof of Theorem 4, $\Delta^F_{min} = \min\left\{\Delta^{\mathbb{E}\{B_n\}}_S : S \in \mathcal{S} \setminus \mathcal{S}^*\left(\mathbb{E}\{B_n\}\right)\right\}$, $s = \max\{|S| : S \in \mathcal{S}\}$, $c = \max\{|\mathcal{E}| : \mathcal{E} \in \mathbf{C}\}$, i.e., the maximum size of a minimal cover of A, and $H = \bigcup_{S^*\in\mathcal{S}^*(\mathbb{E}\{B_n\})} \bigcup_{a\in S^*} \{a\}$. Also note that $\Delta^{\mathcal{D}}_4$ denote the minimum optimality gap of problem $OCP\left(\mathbb{E}_F\{B_n\}\right)$. Note that thresholds Δ^F_{min}, $\Delta^{\mathcal{D}}_2$ and $\Delta^{\mathcal{D}}_4$ are always positive, while $\Delta^{\mathcal{D}}_3 > 0$ by Assumption 2.

We now check that having mean cost estimates with enough precision allows to reconstruct the feasible set $\mathbf{G}$. Consider (x,y) satisfying (13b) and (13d). We first note that as discussed in Step 1 of the proof of Theorem 4, $\left\{\|\hat{B}_n - \mathbb{E}\{B_n\}\|_\infty < \Delta^F_{min}/(2s)\right\}$ ensures that $\left\{\mathcal{S}^*(\bar{B}_n) \subseteq \mathcal{S}^*(\mathbb{E}\{B_n\})\right\}$. One then has that

$$\begin{aligned}
\left\{\|\hat{B}_n - \mathbb{E}\{B_n\}\|_\infty < \Delta^{\mathcal{D}}\right\} \subseteq \left\{\left|\sum_{a\in S} x_a\left(\bar{b}_{a,n} - \mathbb{E}\{b_{a,n}\}\right)\right| < \Delta^{\mathcal{D}} s,\, \forall S \in \mathcal{S}\right\}\\
\cap\ \left\{\left|z^*(\bar{B}_n) - z^*(\mathbb{E}\{B_n\})\right| < \Delta^{\mathcal{D}} s\right\}\\
\subseteq \left\{\left|\left(\sum_{a\in S}\left(x_a\bar{b}_{a,n} + (1-x_a)l_a\right) - z^*(\bar{B}_n)\right) - \left(\sum_{a\in S}\left(x_a\mathbb{E}\{b_{a,n}\} + (1-x_a)l_a\right) - z^*(\mathbb{E}\{B_n\})\right)\right| < 2\Delta^{\mathcal{D}} s,\, \forall S \in \mathcal{S}\right\}.
\end{aligned}$$

We conclude that, because $2\Delta^{\mathcal{D}} s < \Delta^{\mathcal{D}}_2 \wedge \Delta^{\mathcal{D}}_3$,

$$\sum_{a\in S}\left(x_a\mathbb{E}\{b_{a,n}\} + (1-x_a)l_a\right) \geq z^*(\mathbb{E}\{B_n\}) \text{ iff } \sum_{a\in S}\left(x_a\bar{b}_{a,n} + (1-x_a)l_a\right) \geq z^*(\bar{B}_n).$$

Having the same feasible region for both $OCP(\mathbb{E}\{B_n\})$ and $OCP(\bar{B}_n)$ problems, we now show that ϱ-optimal solutions to the latter problem corresponds to an optimal solution to the former. Indeed, we have that

$$\left\{\|\hat{B}_n - \mathbb{E}\{B_n\}\|_\infty < \Delta^{\mathcal{D}}\right\} \subseteq \left\{\left|\Delta^{\bar{B}_n}_S - \Delta^{\mathbb{E}\{B_n\}}_S\right| < \frac{\Delta^{\mathcal{D}}_4}{4c},\, \forall S \in \mathcal{S}\right\}$$

$$\subseteq \left\{ \left| \sum_{S\in\mathcal{G}} \left(\Delta_S^{\bar{B}_n} - \Delta_S^{\mathbb{E}\{B_n\}} \right) \right| < \frac{\Delta_4^{\mathcal{D}}}{4}, \forall (C,\mathcal{G}) \in \mathbf{G} \right\}$$

$$\subseteq \left\{ \sum_{S\in\mathcal{G}} \Delta_S^{\bar{B}_n} > \Delta_4^{\mathcal{D}}/2 + \sum_{S\in\mathcal{G}^*} \Delta_S^{\bar{B}_n}, \forall (C^*,\mathcal{G}^*) \in \Gamma_{OCP}\left(\mathbb{E}\{B_n\}\right), (C,\mathcal{G}) \in \mathbf{G} \setminus \Gamma_{OCP}\left(\mathbb{E}\{B_n\}\right) \right\}.$$

The above not only implies that $\Gamma_{OCP}(\bar{B}_n) \subseteq \Gamma_{OCP}(\mathbb{E}\{B_n\})$, but also that ϱ-optimal solutions to $OCP(\bar{B}_n)$ are also optimal to $OCP(\mathbb{E}\{B_n\})$, as long as $\varrho < \Delta_4^{\mathcal{D}}/2$. Letting $\Gamma_{OCP}^{\varrho}(B)$ denote the set of ϱ-optimal solutions to $OCP(B)$, the above implies that for $\varrho < \Delta_4^{\mathcal{D}}/2$,

$$\left\{ \|\hat{B}_n - \mathbb{E}\{B_n\}\|_\infty < \Delta^{\mathcal{D}} \right\} \subseteq \left\{ \Gamma_{OCP}^{\varrho}(\bar{B}_n) \subseteq \Gamma_{OCP}(\mathbb{E}\{B_n\}) \right\}. \tag{A-19}$$

We are now ready to provide a bound on $R_1^{OCP}(F,N)$. Similar to the proof of Theorem 4, let i' be a finite upper bound on a cycle in which one is sure to conduct all OCP-based explorations (e.g., $i' := 1 + \inf\{i \in \mathbb{N}, i \geq 2 : n_{i+1} - n_i > i\,|A|\}$). Fix $i > i'$ and let $(C_i, \mathcal{G}_i)$ denote the OCP-based exploration set for any instance $n \in [n_i, n_{i+1}-1]$. Define the events $\Xi_i^1 := \{(C_i,\mathcal{G}_i) \in \Gamma_{OCP}\left(\mathbb{E}\{B_n\}\right)\}$ and $\Xi_i^2 := \{\mathcal{G}_i = \mathcal{G}_{i-1}\}$. We then have that

$$\begin{aligned}
R_1^{OCP}(F,N) &\leq n_{i'+1}\Delta_{max}^F + \sum_{i=i'+1}^{\lceil (\ln N)^{1+\varepsilon} \rceil} \mathbb{E}\left\{ \mathbf{1}\left\{ \tilde{T}_{n_{i-1}}(a) \geq \gamma(i-2), \forall a \in A, \left(\Xi_i^1 \cap \Xi_i^2\right) \right\} \sum_{S\in\mathcal{G}_i} \Delta_S^{\mathbb{E}\{B_n\}} \right\} \\
&\quad + \sum_{i=i'+1}^{\lceil (\ln N)^{1+\varepsilon} \rceil} i\ \mathbb{E}\left\{ \mathbf{1}\left\{ \tilde{T}_{n_{i-1}}(a) \geq \gamma(i-2), \forall a \in A, \left(\Xi_i^1 \cap \Xi_i^2\right)^c \right\} \sum_{S\in\mathcal{G}_i} \Delta_S^{\mathbb{E}\{B_n\}} \right\} \\
&\leq n_{i'+1}\Delta_{max}^F + \left((\ln N)^{1+\varepsilon} + 1 \right) z_{OCP}^*\left(\mathbb{E}\{B_n\}\right) \\
&\quad + \Delta_{max}^F\, c \sum_{i=i'+1}^{\infty} i\ \mathbb{P}\left\{ \tilde{T}_{n_{i-1}}(a) \geq \gamma(i-2), \forall a \in A, \left(\Xi_i^1 \cap \Xi_i^2\right)^c \right\}
\end{aligned} \tag{A-20}$$

Next, we bound the probability inside the sum in (A-20). For that, observe that

$$\begin{aligned}
\left\{ \|\bar{B}_{n_{i-1}} - \mathbb{E}\{B_n\}\|_\infty \vee \|\bar{B}_{n_i} - \mathbb{E}\{B_n\}\|_\infty < \Delta^{\mathcal{D}} \right\} \subseteq\ & \left\{ \Gamma_{OCP}^{\varrho}(\bar{B}_{n_{i-1}}) \subseteq \Gamma_{OCP}(\mathbb{E}\{B_n\}) \right\} \\
& \cap \left\{ \Gamma_{OCP}(\mathbb{E}\{B_n\}) \subseteq \Gamma_{OCP}^{\varrho}(\bar{B}_{n_i}) \right\} \\
& \subseteq \left(\Xi_i^1 \cap \Xi_i^2\right).
\end{aligned}$$

Using above and Proposition 5, we conclude that

$$\mathbb{P}\left\{ \tilde{T}_{n_{i-1}}(a) \geq \gamma(i-2), \forall a \in A, \left(\Xi_i^1 \cap \Xi_i^2\right)^c \right\} \leq 4\,|A| \exp\left\{ -\frac{2(\Delta^{\mathcal{D}})^2\, \gamma(i-2)}{\mathcal{L}^2} \right\}. \tag{A-21}$$

Using the above and (A-20), we have that

$$R_1^{OCP}(F,N) \leq (\ln N)^{1+\varepsilon} z^*_{OCP}(\mathbb{E}\{B_n\}) + C_5,$$

for a finite positive constant C_5, independent of N. Putting the results from Steps 1, 2.1 (from the proof of Theorem 4) and Step 2.2' together, we conclude that

$$R^{\pi_h}(F,N) \leq (z^*_{OCP}(\mathbb{E}\{B_n\}) + \gamma z^*_C(\mathbb{E}\{B_n\}))(\ln N)^{1+\varepsilon} + C_6,$$

for a finite positive constant C_6, independent of N.

We finally note that the solution to the estimated *Cover* and *OCP* problems converge a.s. to optimal solutions. For this, note that Proposition 5, (A-12) and (A-21) imply (via Borel-Cantelli) that $\mathbb{P}\{\mathcal{E}_i \in \Gamma_C(\mathbb{E}\{B_n\}) \text{ e.v.}\} = 1$ and

$$\mathbb{P}\{(C_i, \mathcal{G}_i) \in \Gamma_{OCP}(\mathbb{E}\{B_n\}) \text{ e.v.}\} = 1.$$

A.3. Appendix for Section 6

A.3.1. Asynchronous Policy from Section 6.1 The time-constrained Asynchronous Policy in Section 6.1 is depicted in Algorithm 4.

A.3.2. General Complexity of OCP To prove Theorem 6 and Proposition 4 we will use the following lemma.

Lemma 2. *We may restrict OCP or Cover to have at most $|A|$ non-zero y_S variables without changing the problems.*

For OCP it follows from noting that any critical subset can be covered by at most $|A|$ solutions. Hence, if an optimal solution for OCP has $|\mathcal{E}| > |A|$ we may remove one solution from it while preserving feasibility. If the removed solution is sub-optimal for $f(B)$ we would obtain a solution with lower objective value contradicting the optimality for OCP. If the removed solution is optimal for $f(B)$ we obtain an alternate optimal solution for OCP.

For *Cover* the result follows by noting that A can be covered by at most $|A|$ solutions.

Theorem 6. *If $f(B)$ is in P, then $OCP(B)$ and $Cover(B)$ are in NP.*

By Lemma 2, optimal solutions to OCP and *Cover* have sizes that are polynomial in $|A|$ and their objective function can be evaluated in polynomial time. Checking the feasibility of these solutions for OCP can be achieved in polynomial time, because checking (13c) can be achieved by solving $f(B_x)$ where $B_x := (b_{a,x} : a \in A)$ for $b_{a,x} := l_a(1-x_a) + b_a x_a$. This problem is polynomially solvable by assumption.

Algorithm 4 Basic Time-Constrained Asynchronous hybrid policy $\pi_h^A(\gamma, \varrho)$

Set $i = 0$, $C = A$, $\mathcal{E}$ a minimal cover of A, and $\mathcal{G} = \mathcal{E}$
Let $S^* \in \mathcal{S}$ be an arbitrary solution and $B_f = B_{OCP} = B_{Cover}$ be an initial cost estimate
Asynchronously begin solving $f(B_f)$, $OCP(B_{OCP})$ and $Cover(B_{Cover})$
for $n = 1$ to N **do**
  if $n \in \Phi$ **then**
    Set $i = i + 1$
    if Asynchronous solution to $f(B_f)$ has finished **then**
      Set $S^* \in \mathcal{S}^*(B_f)$ and $B_f = \bar{B}_n$ [Update exploitation set]
      Asynchronously begin solving $f(B_f)$
    end if
    if Asynchronous solution to $OCP(B_{OCP})$ has finished **then**
      if $(C, \mathcal{G})$ is not a ϱ-optimal solution to $OCP(B_{OCP})$ **then**
        Set $(C, \mathcal{G}) \in \Gamma_{OCP}(B_{OCP})$ [Update OCP-exploration set]
      end if
      Set $B_{OCP} = \bar{B}_n$
      Asynchronously begin solving $OCP(B_{OCP})$
    end if
    if Asynchronous solution to $Cover(B_{Cover})$ has finished **then**
      Set $\mathcal{E} \in \Gamma_C(\bar{B}_n)$ and $B_{Cover} = \bar{B}_n$ [Update Cover-exploration set]
      Asynchronously begin solving $Cover(B_{Cover})$
    end if
  end if
  if $\tilde{T}_n(a) < \gamma\ i$ for some $a \in A$ **then**
    Set $S_n = S$ with $S \in \mathcal{E}$ such that $a \in S$ [Cover-based exploration]
  else if $\gamma < 1$ and $\tilde{T}_n(a) < i$ for some $a \in C$ **then**
    Set $S_n = S$ with $S \in \mathcal{G}$ such that $a \in S$ [OCP-based exploration]
  else
    Implement $S_n = S^*$ [Exploitation]
  end if
end for

A.3.3. Complexity of OCP for Matroids To prove Theorem 7 we need the following lemma.

LEMMA 3. *Let $f(\cdot)$ be a weighted basis or independent set matroid minimization problem. Then there exist a unique critical set.*

To simplify the exposition, we assume that $\mathcal{S}^*(B) = \{S^*\}$ is a singleton. Also, for $S \in \mathcal{S}$, we let e^S denote the incidence vector associated with S (i.e., $e^S_a \in \{0,1\}$, $a \in A$, is such that $e^S_a = 1$ if $a \in S$ and $e^S_a = 0$ otherwise).

Let $P := \operatorname{conv}\{e^S\}_{S\in\mathcal{S}} \subseteq \mathbb{R}^n$ be the independent set (base) polytope of $\mathcal{S}$. Then, for B feasible, $S^* \in \mathcal{S}^*(B)$ if and only if $\sum_{a\in S^*} b_a \leq \sum_{a\in S} b_a$ for any $S \in \mathcal{S}$ such that e^{S^*} and e^S are adjacent vertices in P. Furthermore, each adjacent vertex to e^{S^*} can be obtained from S^* by: removing (R), adding (A) or exchanging (E) a single element of S^* (Schrijver 2003, Theorem 40.6). Thus, we construct the set C so that S^* is always optimal if and only if the cost of all elements of C are at their expected value. The construction procedure starts with $C = S^*$. In some steps we distinguish between $\mathcal{S}$ corresponding to independent sets or bases.

R. (for the independent set case) From the optimality of S^* removing an element never leads to optimality.

A. (for the independent set case) For each $a \in A \setminus S^*$ such that $S^* \cup \{a\}$ is an independent set; if $l_a < 0$, then add a to C.

E. (for both cases) For each $a \in A \setminus S^*$, add a to C if

$$l_a < \max\{b_{a'} : a' \in S^*,\, S^* \cup \{a\} \setminus \{a'\} \text{ is an indep. set (base)}\}.$$

By construction, covering all elements in C guarantees optimality of S^*, and not covering some guarantees S^* is no longer optimal. Note that the set C is unique. For the case of multiple optimal solutions we simply repeat this procedure for each one.

We also need the following well-known Lemma (Schrijver 2003) and its simple corollary.

LEMMA 4 **(Uncrossing Technique)**. *Let*

$$P = \left\{ x \in \mathbb{R}^{|A|} : x_a \geq 0 \quad \forall a \in A, \quad \sum_{a\in S} x_a \leq R(S) \quad \forall S \subseteq A \right\}$$

be the independent set polytope of a matroid with rank function $R(\cdot)$, $x \in P$ and $W_1 \subset \ldots \subset W_k$ be an inclusion-wise maximal chain of subsets of A such that $\sum_{a\in W_l} x_a = R(W_l)$ for all $l \leq k$. Then, for any set $S \subseteq A$ such that $\sum_{a\in S} x_a = R(S)$ we have that

$$e^S \in \operatorname{span}\left(\left\{e^{W_l}\right\}, l \leq k\right)$$

We use the following corollary of Lemma 4.

COROLLARY 1. *Let P be the independent set or base polytope of a matroid and let $x \in P$. If $x_a \in (0,1)$, then there exist $\varepsilon > 0$ and $a' \in A \setminus \{a\}$ such that $x_{a'} \in (0,1)$ and $x \in \operatorname{conv}\{\overline{x}, \underline{x}\}$, for $\overline{x}, \underline{x} \in P \setminus \{x\}$ defined by*

$$\overline{x}(\varepsilon, a, a') := x + \varepsilon \left(e^a - e^{a'}\right); \quad \underline{x}(\varepsilon, a, a') := x + \varepsilon \left(e^{a'} - e^a\right). \tag{A-22}$$

Let $W_0 \subset W_1 \subset W_2 \subset \ldots \subset W_k$ be the maximal chain from Lemma 4 (with $W_0 = \emptyset$). If $k = 0$ then the result follows trivially ($x \in \operatorname{int}(P)$), so we will assume that $k \geq 1$. Let l_0 be the smallest $l \in \{1, \ldots, k\}$ such that $a \in W_l$. There exists $a' \in W_{l_0} \setminus \{a\}$ such that $x_{a'} \in (0,1)$: to see this, note that $R(W_{l_0-1}) \in \mathbb{Z}_+$, and that

$$R(W_{l_0}) = \left(R(W_{l_0-1}) + x_a + \sum_{h \in W_{l_0} \setminus \left(W_{l_0-1} \cup \{a\}\right)} x_h \right) \in \mathbb{Z}_+,$$

thus one can always find such an a' in $W_{l_0} \setminus (W_{l_0-1} \cup \{a\})$. For any choice of a' we have that $y \in \{\overline{x}, \underline{x}\}$ defined in (A-22) satisfies $\sum_{h \in W_l} y_h = r(W_l)$ for all $l \leq k$, thus by Lemma 4 $y \in P$ for $\varepsilon < \min\{x_a, x_{a'}\}$ (so that $y \geq 0$ and $\sum_{h \in S} y_h \leq R(S)$ for constraints not active at x). The result follows since $x \in \operatorname{conv}\{\overline{x}, \underline{x}\}$ by construction.

Finlay we need the following proposition.

PROPOSITION 6. *Consider the Linear Programming (LP) problems given by*

$$MC(C): \quad \min \quad \sum_{l=1}^{r} \sum_{a \in A} b_a x_a^l \tag{A-23a}$$

$$s.t. \qquad 1 \leq \sum_{l=1}^{r} x_a^l, \quad a \in C \tag{A-23b}$$

$$\sum_{a \in S} x_a^l \leq R(S), \quad S \subseteq A, l \in \{1, \ldots, r\} \tag{A-23c}$$

$$0 \leq x_a^l \leq 1, \quad a \in A, l \in \{1, \ldots, r\}, \tag{A-23d}$$

for $r = 1$ to $r = |A|$ (for the basis problem we also add $\sum_{a \in S} x_a^l = R(S)$ for all l). Then, (A-23) *has integral extreme points.*

Because the feasible region of the basis problem is a face of the independent set problem it suffices to prove this result for the latter one. For this, we need a few auxiliary results. Let x be a fractional extreme point of (A-23). Without loss of generality x^1 has a fractional component $x_{i_1}^1 \in (0,1)$: we will reach a contradiction by constructing a set of solutions whose convex hull contains

x. Corollary 1 implies that there exist $\varepsilon_1 > 0$, j_1 such that $x^1 \in \text{conv}\left\{\overline{x}^1(\varepsilon_1, i_1, j_1), \underline{x}^1(\varepsilon_1, i_1, j_1)\right\}$, with $\overline{x}^1(\varepsilon_1, i_1, j_1), \underline{x}^1(\varepsilon_1, i_1, j_1) \in P$ defined by (A-22).

Define $\tilde{C} := \left\{h \in C : 1 = \sum_{l=1}^r x_h^l\right\}$. If $\{i_1, j_1\} \cap \tilde{C} \neq \emptyset$, by symmetry we may assume w.l.o.g. that $j_1 \in \tilde{C}$ (if not rename i_1 and j_1). Because $x_{j_1}^1 \in (0,1)$, $j_1 \in \tilde{C}$ and the definition of $\tilde{C}$, there exists $l \in \{2, \ldots, r\}$ such that $x_{j_1}^l \in (0,1)$. We assume w.l.o.g. that $l = 2$ and let $i_2 = j_1$. By Corollary 1 applied to x^2 we have that there exist $\varepsilon_2 > 0$, j_2 and $\overline{x}^2(\varepsilon_2, i_2, j_2), \underline{x}^2(\varepsilon_2, i_2, j_2) \in P$ defined by (A-22) such that $x^2 \in \text{conv}\left\{\overline{x}^2(\varepsilon_2, i_2, j_2), \underline{x}^2(\varepsilon_2, i_2, j_2)\right\}$. If $\{i_1, j_2\} \cap \tilde{C} \neq \emptyset$, again by symmetry we can assume that $j_2 \in \tilde{C}$ and repeat this construction and continue until we obtain a sequence $i_1, j_1 = i_2, \ldots, j_{k-1} = i_k, j_k$ and $\varepsilon_1, \ldots, \varepsilon_k$ for $k \geq 1$ which satisfies one of the following conditions:

1. $\{i_1, j_k\} \cap \tilde{C} = \emptyset$.
2. $j_k = i_l$ for some $l \in \{1, \ldots, k-1\}$, in which case we may assume that $l = 1$.

For case 1. let $\varepsilon := \min\left\{\min\{\varepsilon_l : l = 1, \ldots, k\}, 1 - \sum_{l=2}^r x_{i_1}^l, 1 - \sum_{l=1}^{k-1} x_{j_k}^l - \sum_{l=k+1}^r x_{j_k}^l\right\}$ and for case 2. let $\varepsilon := \min\{\varepsilon_l : l = 1, \ldots, k\}$. For both cases define $\hat{X} := (\hat{x}^l)_{l=1}^r$, $\check{X} := (\check{x}^l)_{l=1}^r$ so that $\hat{x}^l = \overline{x}^l(\varepsilon, i_l, j_l)$, $\check{x}^l = \underline{x}^l(\varepsilon, i_l, j_l)$ for $l \in \{1, \ldots, k\}$ and $\hat{x}^l = \check{x}^l = x^l$ for all $l \in \{k+1, \ldots, r\}$. We then have that $\hat{X}, \check{X} \subseteq Q$, $x \notin \hat{X} \cup \check{X}$ and $x \in \text{conv}\left\{\hat{X}, \check{X}\right\}$, which contradicts x being an extreme point.

THEOREM 7. *$OCP(B)$ and $Cover(B)$ are in P for weighted basis or independent set matroid minimization problems.*

From Lemma 3 we know that there exists a unique critical set. Moreover, such a set can be found in polynomial time (e.g., by solving $|A|$ instances of $f(\cdot)$). Let C^* denote the unique critical set and $R : 2^{\mathcal{N}} \to \mathbb{Z}_+$ be the rank function of the matroid.

We claim that OCP can be solved through $MC(C^*)$ and $Cover$ can be solved by $MC(A)$. Indeed, formulation (A-23) is the fractional covering of C with at most r solutions of the matroid and if we change $0 \leq x_a^l \leq 1$ to $x_a^l \in \{0,1\}$ we have the standard covering with exactly r solutions of the matroid. By Proposition 6 the problems with $0 \leq x_a^l \leq 1$ and $x_a^l \in \{0,1\}$ are the same. By Lemma 2, for both OCP and $Cover$ it suffices to consider cases $r \in \{1, \ldots, |A|\}$; we just need to evaluate the regret for each case and pick the best. The result follows by noting that (A-23) can be solved in polynomial time because (A-23c) can be separated in polynomial time (Schrijver 2003, Corollary 40.4c).

A.3.4. Basic MIP Formulation

PROPOSITION 4. *Let y^S be the incidence vector of $S \in \mathcal{S}$, $M \in \mathbb{R}^{m \times |A|}$, and $d \in \mathbb{R}^m$ be such that $\{y^S\}_{S \in \mathcal{S}} = \left\{y \in \{0,1\}^{|A|} : My \leq d\right\}$ and $\text{conv}\left(\{y^S\}_{S \in \mathcal{S}}\right) = \left\{y \in [0,1]^{|A|} : My \leq d\right\}$. Then a MIP formulation of $OCP(B)$ is given by*

$$\min \quad \sum_{i=1}^{|A|} \left(\sum_{a \in A} b_a y_a^i - z^*(B)\right) \tag{14a}$$

$$s.t. \quad x_a \leq \sum_{i=1}^{|A|} y_a^i, \qquad a \in A \tag{14b}$$

$$My^i \leq d, \qquad i \in \{1,\dots,|A|\} \tag{14c}$$

$$M^T w \leq \operatorname{diag}(l)(\mathbf{1}-x) + \operatorname{diag}(b)x \tag{14d}$$

$$d^T w \geq z^*(B) \tag{14e}$$

$$x_a,\, y_a^i \in \{0,1\},\, w \in \mathbb{R}^m, \qquad a \in A, i \in \{1,\dots,|A|\}, \tag{14f}$$

where for $v \in \mathbb{R}^r$, $\operatorname{diag}(v)$ *is the* $r \times r$ *diagonal matrix with* v *as its diagonal, and* $\mathbf{1}$ *is a vector of ones. A formulation for* $Cover(B)$ *is obtained by setting* $x_a = 1$ *for all* $a \in A$ *and removing* (14d)–(14e).

We begin with the result for OCP. For any feasible solution (x,y) to (14) we have that x is the incidence vector of a critical subset. This, because (14d) enforces dual feasibility of w when elements with $x=0$ are not covered, and (14e) forces the objective value of the dual of $f(B')$ to be greater than or equal to $z^*(B)$, where $B' = \operatorname{diag}(l)(\mathbf{1}-x) + \operatorname{diag}(b)x$. With this, the optimal objective value of $f(B')$ is greater than or equal to $z^*(B)$. On the other hand, any y_a^i is the incidence vector of some $S \in \mathcal{S}$ because of (14c) and the assumptions on M and d. Finally, (14b) ensures that $\mathcal{E} = \{\operatorname{supp}(y_a^i)\}_{i,a\in A}$ covers the critical subset. Lemma 2 ensures that the $|A|$ variables in y are sufficient for an optimal solution. If less than $|A|$ elements are needed for the cover, then the optimization can pick the additional y variables to be the incidence vector of an optimal solution to $f(B)$ so that they do not increase the objective function value. The extension for *Cover* is straightforward.

A.3.5. IP formulation for OCP when $f(B)$ admits a compact IP formulation Suppose $f(B)$ admits a compact IP formulation such that $\{y^S\}_{S\in\mathcal{S}} = \left\{y \in \{0,1\}^{|A|} : My \leq d\right\}$ for some $M \in \mathbb{R}^{m\times|A|}$ and $d \in \mathbb{R}^m$, where y^S denotes the incidence vector of $S \in \mathcal{S}$. Then an IP formulation of $OCP(B)$ is given by

$$\min \quad \sum_{i=1}^{|A|} \left(\sum_{a\in A} b_a y_a^i - z^*(B) \right) \tag{A-24a}$$

$$s.t. \quad x_a \leq \sum_{i=1}^{|A|} y_a^i, \quad a \in A \tag{A-24b}$$

$$My^i \leq d, \qquad i \in \{1,\dots,|A|\} \tag{A-24c}$$

$$\sum_{a\in S} (l_a(1-x_a) + b_a x_a) \geq z^*(B), \quad S \in \mathcal{S} \tag{A-24d}$$

$$x_a,\, y_a^i \in \{0,1\}, \quad a \in A, i \in \{1,\dots,|A|\}. \tag{A-24e}$$

Like in formulation (14), a feasible solution (x,y) to (A-24) is such that x is the incidence vector of a critical subset (this is enforced by (A-24d)), and the y^i's are a cover of such set, due to (A-24c) and the assumptions on M and d. Note that an efficient cover includes at most $|A|$ elements (the optimization can pick the additional y^i to be the incidence vector of an optimal solution).

Formulation (A-24) has a polynomial number of variables, but the number of constraints described by (A-24d) is in general exponential. However, the computational burden of separating these constraints is the same as solving $f(B)$ (finding a violated inequality (A-24d) or showing that it satisfies all these inequalities can be done by solving $f(B')$ for $b'_a = l_a(1-x_a)+b_a x_a$). Hence, if we can solve $f(B)$ sufficiently fast (e.g., when the problem is in P, or it is a practically solvable NP-hard problem) we should be able to effectively solve (A-24) with a Branch-and-Cut algorithm that dynamically adds constraints (A-24d) as needed. Finally, note that a formulation for *Cover* is obtained by setting $x_a = 1$ for all $a \in A$ and removing (A-24d).

A.3.6. Linear-sized formulation for OCP for the shortest path problem Let $f(B)$ correspond to a shortest $s-t$ path problem in a digraph $G=(V,A)$. Define $\hat{A} = A \cup \{(t,s)\}$ and let $\hat{\delta}_{out}$ and $\hat{\delta}_{in}$ denote the outbound and inbound arcs in digraph $\hat{G} = (V,\hat{A})$. An optimal solution (x,p,w) to

$$\min \quad \left(\sum_{a\in A} b_a p_a\right) - z^*(B)\; p_{(t,s)} \tag{A-25a}$$

$$s.t. \qquad x_a \le p_a, \qquad a \in A \tag{A-25b}$$

$$\sum_{a\in\hat{\delta}_{out}(v)} p_a - \sum_{a\in\hat{\delta}_{in}(v)} p_a = 0, \qquad v \in V \tag{A-25c}$$

$$w_u - w_v \le l_{(u,v)}(1-x_{(u,v)}) + b_{(u,v)}x_{(u,v)}, \quad (u,v)\in A \tag{A-25d}$$

$$w_s - w_t \ge z^*(B) \tag{A-25e}$$

$$p_a \in \mathbb{Z}_+, \qquad a \in \hat{A} \tag{A-25f}$$

$$x_a \in \{0,1\},\; w_v \in \mathbb{R}, \qquad a \in A,\; v \in V, \tag{A-25g}$$

is such that $(C,\mathcal{E})$ is an optimal solution to $OCP(B)$, where $C = \{a \in A : x_a = 1\}$ and $\mathcal{E} \subseteq \mathcal{S}$ is a set of paths for which $p_a = |\{S \in \mathcal{E} : a \in S\}|$. Such a set $\mathcal{E}$ can be constructed from p in time $O(|A||V|)$.

The first difference between formulations (A-25) and (14) is the specialization of the LP duality constraints to the shortest path setting. The second one is the fact that the paths in cover $\mathcal{E}$ are aggregated into an integer circulation in augmented graph $\hat{G}$, which is encoded in variables p. Indeed, using known properties of circulations (Schrijver 2003, pp. 170-171), we have that $p = \sum_{S\in\mathcal{E}} y^{\hat{S}}$, where $y^{\hat{S}}$ is the incidence vector of the circulation obtained by adding (t,s) to path

S. Furthermore, given a feasible p we can recover the paths in $\mathcal{E}$ in time $O(|A||V|)$. To obtain a formulation for *Cover* we simply set $x_a = 1$ for all $a \in A$ and remove (A-25d)–(A-25e).

It is possible to construct similar formulations for other problems with the well-known integer decomposition property Schrijver (2003).

A.4. Adjoint Formulation for Tighter Upper Bound

The following formulation is a variation of LBP that is robust with respect to changes in the mean cost vector of elements that are not "covered" by its optimal solution. For that, we introduce an additional variable w_a indicating whether one would impose additional exploration (beyond that required in the lower bound result - the parameter γ indicates the frequency of such exploration) on a ground element $a \in A$, and variable r_a indicates the degree at which element $a \in A$ is covered in a solution. The variable $\underline{z}$ computes the minimum cost attainable if one were to change the mean cost of an unexplored ground element. Finally, the formulation imposes that the optimal cost is no lower than such an alternative minimum cost.

$$
\begin{aligned}
z_R^*(B,\gamma) = \min \; & \sum_{S\in\mathcal{S}} \Delta_S^B \; y_S \\
s.t. \quad & \underline{z} = \min_{y'\in\mathbb{R}_+^{|\mathcal{S}|}} \left\{ \sum_{S\in\mathcal{S}} \Delta_S^{B_{\{a\in A; r_a=0\}}} \; y'_S : r_a \le \sum_{S\in\mathcal{S}:a\in S} y'_S,\, a\in A \right\} \\
& \sum_{S\in\mathcal{S}} \Delta_S^B \; y_S \le \underline{z} \\
& r_a \le \sum_{S\in\mathcal{S}:a\in S} y_S,\, a\in A \\
& r_a = x_a + \gamma\, w_a,\, a\in A \\
& \max\{x_a : a\in D\} \ge K_D(B), \quad D\in\mathcal{D}(B) \\
& w_a \in \{0,1\},\, x_a,\, r_a, y_S \in \mathbb{R}_+, \quad a\in A,\, S\in\mathcal{S}.
\end{aligned}
$$

Appendix B: Auxiliary Results and Omitted Proofs.

B.1. Auxiliary Result for the Proof of Theorem 4 and Theorem 5

PROPOSITION 5. *For any fixed $a \in A$, $n \in \mathbb{N}$, $k \in \mathbb{N}$, and $\epsilon > 0$ we have that*

$$\mathbb{P}\left\{\left|\bar{b}_{a,n} - \mathbb{E}\left\{b_{a,n}\right\}\right| \geq \epsilon, \tilde{T}_n(a) \geq k\right\} \leq 2\exp\left\{-\frac{2\epsilon^2 k}{\mathcal{L}^2}\right\},$$

where $\mathcal{L} := \max\{u_a - l_a : a \in A\}$.

For $m \in \mathbb{N}$ define $t_m(a) := \inf\left\{n \in \mathbb{N} : \tilde{T}_n(a) = m\right\} - 1$. Indexed by m, one has that $b_{a,t_m(a)} - \mathbb{E}\left\{b_{a,n}\right\} = b_{a,t_m(a)} - \mathbb{E}\left\{b_{a,t_m(a)}\right\}$ is a bounded martingale difference sequence, thus one has that

$$\begin{aligned}
\mathbb{P}\left\{\left|\bar{b}_{a,n} - \mathbb{E}\left\{b_{a,n}\right\}\right| \geq \epsilon\,, \tilde{T}_n(a) \geq k\right\} &= \mathbb{P}\left\{\left|\sum_{m=1}^{\tilde{T}_n(a)} \left(b_{a,t_m(a)} - \mathbb{E}\left\{b_{a,n}\right\}\right)\right| \geq \epsilon\ \tilde{T}_n(a)\,, \tilde{T}_n(a) \geq k\right\} \\
&\leq \sum_{h=k}^{\infty} \mathbb{P}\left\{\left|\sum_{m=1}^{h} \left(b_{a,t_m(a)} - \mathbb{E}\left\{b_{a,n}\right\}\right)\right| \geq \epsilon\ h\,, \tilde{T}_n(a) = h\right\} \\
&\overset{(a)}{\leq} 2\ \sum_{h=k}^{\infty} \exp\left\{\frac{-2\ h\epsilon^2}{\mathcal{L}^2}\right\} \mathbb{P}\left\{\tilde{T}_n(a) = h\right\} \\
&\leq 2\exp\left\{\frac{-2\ k\epsilon^2}{\mathcal{L}^2}\right\},
\end{aligned}$$

where (a) follows from the Hoeffding-Azuma Inequality (see, for example, (Cesa-Bianchi and Lugosi 2006, Lemma A.7)).

Letter to the Area Editor and Associate Editor and Point-by-Point Response to Referees' Comments: "Learning in Combinatorial Optimization: What and How to Explore" (OPRE-2014-08-500)

October 24, 2016

Overview

We begin by sincerely expressing our gratitude to the review team for their continued feedback throughout the review process. As put by the associate editor, the two main issues to be addressed in this revision are: first, we should address the issue of the gap between our lower and upper bounds; and second, we should provide a clear characterization on how all accompanying *leading* constants (in these bounds) depend on the problem primitives.

With respect to the first of these issues, the manuscript now provides a policy that while computationally impractical, has a regret bound that scales optimally with the horizon (up to sub-logarithmic terms) and has the same leading constant as that in the lower bound (up to terms that can be made arbitrarily small but never zero in general). (The impracticality of such a policy is akin to that of similar policies in the traditional bandit setting; e.g., the optimal policy in Lai and Robbins (1985).) The inclusion of sub-logarithmic terms is a deliberate choice that allows us to present our results in a much clearer way (such a choice is not unprecedented in the literature and can be avoided, as we clarify in the manuscript). On the other hand, we show that the additional terms in the leading constant cannot be eliminated in the general setting (see Example 3). Nonetheless, we fully explain the source of these additional terms and give a family of settings in which it can be eliminated. The associated discussion is very theoretical in nature, but provides the insight that informs about our more practical policy.

Providing matching bounds in general is a very challenging task, especially if one is concerned with the implementability of the underlying policies. Similar to many valuable work published in *Operations Research*, our work does not fully close the gap in performance in general settings. However, unlike some of said work, we do establish a lower bound on performance (the first in the combinatorial setting), and the analysis in Section 4.3 suggests that our upper bounds are the best possible and that closing the gap would amount to developing new techniques to improve the lower bound result. Moreover, we again note that as shown in Example 3, closing the gap may not be possible in the general setting. Overall, we hope that the review team appreciates the degree at which we have advanced on closing the performance gap.

Regarding the issue of providing a clear characterization of the accompanying constants in our bounds, let us advance that in this revision we have significantly changed the structure and presentation of the manuscript to improve its readability. In particular, we now directly state that the optimal constants accompanying the $\ln N$ terms in the lower and upper bounds for combinatorial bandits are the solution of an optimization problem and do not have simple formulas in terms of the problem primitives. Up to this point we had attempted to write these bounds in terms of the model primitives in ways that could not capture the underlying combinatorial aspects of the setting, mostly to facilitate the theoretical comparison with the benchmark (in terms of upper bounds). Because the leading constants in our bounds are the solution of optimization problems and therefore have a non-trivial dependence on the problem primitives, we believe that one should not look for simplified upper bounds in terms of problem primitives in the combinatorial bandit setting. Also, because our lower and upper bounds are equal up to tunable constants and sub-logarithmic terms, the *theoretical* comparison with the benchmark (beyond the extensive empirical comparison in our numerical experiments) seems no longer critical. We hope that the review team respects our choice of exposition style in this regard, but we would be happy to add to the appendix any material that might help comparing performance bounds striped from the combinatorial setting, should the review team think it is necessary. We believe that our discussion about the gap in performance and the settings in which it can be closed have resulted in a significant improvement in the clarity of exposition of the paper, as it allows us to address the issue of the accompanying constants from a fresh new perspective. We are again grateful that the review team pushed us to achieve it.

Below we summarize the main changes to the paper in this revision. Considering that there has been some changes in the review team, we then summarize the history of revisions of the paper by discussing the main issues raised by the review team in each round of revision and our responses. Then, we produce detailed responses to all comments raised by the AE and the referees (we reproduce the AE and referees' reports in full in brown text).

Summary of Main Changes

i) We propose a new policy, called the LBP policy, and prove an upper bound on its asymptotic regret that scales optimally with the horizon (up to sub-logarithmic terms) and whose leading constant is that of the lower bound (plus tunable constants). More specifically, we show that any consistent policy must incur a regret whose asymptotic growth is at least $L \ln N$ where N is the time horizon and L is a precisely defined constant that depends on the complete optimization problem (its structure and cost distribution). In addition, we present a family of policies that incur a regret whose asymptotic growth is upper bounded by $(L + \gamma\, C)\, (\ln N)^{1+\varepsilon}$

for positive parameters γ, ε, and C which is a precisely defined constant that depends on the complete optimization problem. We show, in Example 3, that γ cannot be made zero in general settings, but identify a family of settings in which this is possible.

ii) To simplify the exposition and comparison of bounds we now present all upper bounds as a constant (independent of N) multiplied by $(\ln N)^{1+\varepsilon}$ where $\varepsilon > 0$ can be taken arbitrarily close to zero. The use of such a presentation is not novel, and ensures clarity of the bounds. Indeed, avoiding it (which we have done in the past) would result in additional complicated leading constants that would make it significantly harder (if at all possible) to compare the upper bound to the lower bound. As the leading constant in the lower and upper bounds present the same structure (thus facilitating their comparison, an issue raised by the review team), we hope the reviewers respect this stylistic choice. Please refer to the more detailed discussion regarding the comparison of upper and lower bounds after Theorem 4.

iii) In relation to (*ii*) above, we no longer approximate upper bounds via simplified formulas in terms of the problem primitives (e.g., the size of the ground set). In previous rounds we opted for giving such simplified formulas to facilitate bound comparisons with competing heuristics in the literature. However, this resulted in the appearance of additional leading constants (which we had denoted as distribution-dependent constants) that primarily depend on the cost distribution, but may also depend on the instance size. Such simplifications are also done by other papers describing competitive heuristics in the literature (although they lacked the insight from our lower bound result). Doing so obscured the true dependence of the constants on the combinatorial problem and resulted in upper bounds that were quite rough and loose (i.e., worst-case bounds). Indeed, as our lower and upper bounds show, this dependence is through the solution to a rather complicated auxiliary optimization problem that defines L. This auxiliary optimization problem precisely depends on the original combinatorial problem, but, as expected, its solution does not have a simple structure in terms of problem primitives. We believe that this fact is one of the main messages of the paper and hence adding the simplifications (i.e., using the worst-case bounds) would only obscure it and reduce clarity. Nonetheless, we would be happy to provide a comparison with benchmark heuristics in terms of simplified (worst-case) performance upper bounds in the online appendix, should the review team think it is necessary.

iv) Following the changes outlined above, we have restructured the paper exposition into three parts. In the first part we present the lower bound and propose a new policy (i.e., LBP policy) that nearly matches this lower bound (we provide a precise discussion in the main text). However, this policy depends on an optimization problem (i.e., LBP) that will likely not be solvable in practice. For this reason, the second part of the paper deals with a modification

of the optimization problem (i.e., OCP) that can be solved in practice. We support this practical solvability with several non-trivial combinatorial optimization results. This modified problem leads to a policy (i.e., hybrid policy) that can be implemented in practice and has a performance guarantee. The third part of the paper deals with the practical performance of a variant of the practical policy (i.e., OCP policy) that is significantly less conservative than the hybrid policy. A variant of this policy (i.e., a simplified version of the "adaptive" policy that we presented in the previous version of the manuscript which we no longer present) also admits a performance guarantee, but its proof (which was presented in the previous version of the manuscript) is significantly more technical and the resulting bound is hard to interpret. For this reason, we decided to omit this performance guarantee and focus on the policy's computational performance. By means of extensive computational experiments in Section 7 we show that our policy significantly outperforms all known alternative policies in the long-term and is competitive in the short-term. We also show that this policy requires significantly less computational resources to be implemented.

Review and Revision History

- **First Review:**
  1. **Generality of the Proposed Approach:** The ideas on the paper have always been presented for general combinatorial optimization problems, but our initial submission only had computational results and details on practical applicability (i.e., practical solvability of OCP) for shortest path problem. Practical solvability of OCP is now covered for a wide range of problems in Section 6 and computational results for a wide range of problems are now presented in Section 7.
  2. **Theoretical Results:** We were asked to close the gap between the lower and upper bounds or identify its source. This is now achieved in Sections 4 and 5.
  3. **Numerical Experiments:** We were asked to include combinatorial optimization problems other than shortest path and add some competitive heuristics for short-term performance. All these are included in Section 7.
- **Second Review:**
  1. **Gap between upper and lower bounds:** We were asked to provide a more detailed analysis of the gap, closing it if possible.
  2. **Theoretical Results Supporting OCP and the Algorithm's Performance:** The main concerns were: 1) OCP may not be polynomially solvable; 2) polynomial-sized formulations of OCP for polynomially-solvable combinatorial optimization problems; and 3) size comparison of a cover and a solution to OCP. For 1) we argued that in the *Operations*

Research community there is a consensus that hardness results do not undermine the efforts for solving a problem in practice. However, we still provided an oracle polynomial-time heuristic for OCP. For 2) we included a description of how to use polynomial-sized extended formulations of combinatorial optimization problems to construct polynomial-sized formulations for OCP. We also showed that even though there is no polynomial-sized extended formulation for general matching problem, we can still construct a formulation for OCP with polynomial number of variables and an exponential number of constraints that can be separated in polynomial time. (Note that this is a non-trivial result.) Finally, for 3) we presented a wide range of problems where a solution to OCP is arbitrarily smaller than a cover.

3. **Extend the Results to General Observations:** The review team wondered how one would approach cases where one observe sufficient statistics (e.g., the total cost of a path) instead of element-level observations (e.g., the cost of each arc in a path). We were also asked to compare our upper bounds with those for other heuristics in the literature. In Section 8 we show that the extension to general observations is straightforward. Furthermore, in a previous version we showed that the performance of other benchmark heuristics in the literature is roughly that of our cover-based policy. As previously mentioned, this is no longer included to improve readability and emphasize one of the main messages of the paper: that the optimal constants accompanying the $\ln N$ terms for combinatorial bandits are the solution of an optimization problem and do not have simple formulas in terms of problem primitives.

- **Third Review:**
  1. **Gap between upper and lower bounds:** As we show in Example 3 and the discussion following that example, the gap cannot be closed in general, but it is possible in particular settings. We provide a clear characterization of the source of the gap and of the distance between the upper and lower bounds. Please see Sections 4 and 5 of the current manuscript.
  2. **Constants Should All be Verifiably Independent of Problem Size:** In previous versions we had divided constants into distribution-dependent constants that were primarily affected by the cost distribution, and combinatorial constants that primarily depended on the structure of the problem. The reason for this distinction was to give simple formulas for our upper bounds in terms of the problem primitives that could be compared to other bounds in the literature. As mentioned earlier in the summary of main changes, we now provide constants that have a precise (albeit complicated due to them being optimal solutions to an auxiliary optimization problem) dependence on the original combinatorial optimization problem.

Response to Comments of Associate Editor

The paper considers combinatorial optimization problems with unknown objective coefficients. The coefficients come from an unknown distribution. By implementing different solutions, the decision maker gets to see samples of some objective coefficients, but also experiences the costs of the implemented solutions. The objective is to balance the exploration/exploitation trade-off to efficiently learn the coefficients. The paper gives two classes of policies. Static cover-based policy partitions the objective function coefficients, and explores by implementing the solution corresponding to each partition over multiple cycles. The length of each cycle doubles the earlier one to facilitate the right exploration/exploitation trade-off. Optimality cover problem (OCP)-based policy acknowledges that not all objective coefficients have to be explored to find the optimal decision. To capitalize on the idea, it solves an optimization problem to choose the objective coefficients to explore. Regrets bounds for both policies are given. The size of the problem in the OCP-based policy can increase exponentially with the size of the original combinatorial optimization problem. The paper discusses when the problem in the OCP-based policy can be solved efficiently (with an LP or a MIP with polynomial size).

I have 3 reports for the paper. Referee 1's view of the paper is optimistic. The referee finds the contributions of the paper substantial and recommends minor revision. Main comment from Referee 2 is the paper doesn't give matching upper and lower bounds for the regret. Both upper and lower bounds depend on the horizon N logarithmically but the effect of N may be dominated by the constants in the regret bounds, and the upper and lower bounds do not have matching coefficients. Referee 2 recommends the paper be rejected. Referee 3 brings up a possible error in the definition of D that would make the lower bound vacuous for some instances. Referee 3 also mentions that the constants in the regret bounds do not match. Referee 3 recommends major revision.

A: In the new manuscript we present a new policy that is computationally impractical, but has a regret bound that scales optimally with the horizon (up to sub-logarithmic terms) and has the same leading constant as in the lower bound (up to terms that can be made arbitrarily small but never zero in general). Sacrificing the optimal dependence on N allows us to present our results in a much clearer way. This simplification is not unprecedented in the literature and can be avoided as we clarify in the manuscript. Regarding the leading constants accompanying the $\ln N$ term in the regret, we now directly state that these are the solution of an optimization problem and do not have simple formulas in terms of the problem primitives. Note that our lower and upper bounds are equal up to tunable constants and sub-logarithmic terms, which greatly facilitates their comparison. We believe that our discussion about the gap in performance and the settings in which it can be closed has resulted in a significant improvement in the clarity of exposition of the paper.

Regarding Referee 3's major comment #2, the definition of set $\mathcal{D}'$ should indeed exclude the empty set (it does so in this new version). However, the situation in Referee 3's major comment #3 is not an error: the fact that the solution to LBP is zero does not make the lower bound result vacuous; on the contrary, it signals that it is possible to achieve a finite regret independent of the horizon. (Moreover, for the case when $f(\cdot)$ is a matroid, an LBP-based policy does indeed attain a finite regret in such settings.) Please see the more detailed discussion in the response to Referee 3's major comment #3.

Even very naive policies get a bound of $O(\log N)$ (Theorem 3.2); hence the right thing to ask is how the constants in the regret bounds depend on the problem size $|A|$. This is acknowledged in the paper but the paper doesn't answer the question. It is not clear how Δ^F_{max} and $L\left(\mathbb{E}_F\{B_n\}\right)$ in Theorems 4.3 and 4.11 depend on $|A|$, and so we don't understand if the proposed policies are close to being optimal. Same happens in Theorem 3.4. s and Δ^F_{max} are likely to depend on $|A|$, and these dependencies are not explicit.

A: As discussed before, the optimal constants accompanying the $\ln N$ terms for combinatorial bandits are the solution of an optimization problem and do not have simple formulas in terms of the problem primitives. We believe that this fact is one of the main messages of the paper and hence adding the simplifications (i.e., expressing the upper bounds in terms of the problem primitives) would only obscure it and result in bounds that are quite rough and far from tight in terms of the problem primitives. Thus, we believe that one should not look for simplified upper bounds in terms of problem primitives in the combinatorial bandit setting. This is supported by the fact that both our lower and upper bounds are written in terms of the solution to the same optimization problem.

My assessment is that the paper didn't thoroughly address the main issue in the referee reports: "If the authors can close the gap between the upper and lower bounds..., then this can perhaps address the referees' concerns. In that case, I would welcome a re-submission..." The argument in page 22 is not precise, and it is not clear that the claims made here are possible. The paper made progress by giving polynomial-size MIP formulations of OCP but the main issue remains unresolved. I will recommend major revision but iterate that the main issue is still to be fixed: The revision must give upper and lower bounds that match, and the paper must clearly characterize how the bounds depend on the problem size (i.e., the constants should all be verifiably independent of problem size). This is a firm requirement in the revision, and the paper will not pass the bar without resolving this issue.

A: We sincerely appreciate the clarity of the AE's recommendation, as it gave us a clear path forward. In the current manuscript we propose a new policy that admits an upper bound whose leading constant matches that in the proved lower bound up to sub-logarithmic terms. We again note that, as we show in Example 3, obtaining matching lower and upper bounds may not be possible in general. In the new manuscript, we emphasize that the optimal constants accompanying the $\ln N$ terms for combinatorial bandits are the solution of an optimization problem and do not have simple formulas in terms of the problem primitives. On a side note, we are glad that the AE finds our polynomial-size MIP formulations of OCP satisfactory.

The technical issues raised by Referee 3 should also be settled. Please take a look major comment #1 and #3, and make sure that the regret lower bound doesn't become vacuous. (These technical issues may not be relevant anymore when you match the bounds.) I was also puzzled by $\Gamma(B)$ in Algorithm 1. Please justify why it's there. Preferably introduce $\Gamma(B)$ before the algorithm. Referee 2's comment about the size of the problems in numerical experiments should also be answered.

A: Regarding Referee 3's major comment #3, as mentioned above, the fact that the solution to LBP is zero does not make the lower bound result vacuous but quite the opposite, as it identifies settings where active *suboptimal* exploration is rather unnecessary. Regarding Referee 3's major comment #1, we now clearly discuss all the assumptions made in the paper. For example, Definition 1 states assumptions that were implicitly made when stating the lower bound result, and Assumption 2 explicitly states the technical assumption identified by Referee 3 in the major comment #1 (although equivalent, because of changes in the exposition, such an assumption looks different to the assumption that the Referee refers to).

The issue related to $\Gamma(B)$ (which we used to refer to as "feedback-consistent" solutions) no longer appears in the paper as we do not introduce such a set in the revised manuscript.

Finally, regarding Referee 2's comment about the size of the problem, we have added a new set of numerical experiments in Section 7.3 which evaluates the effect of problem size on both the performance and computation time of different policies.

Response to Comments of Referee 1

Review of: Learning in combinatorial optimization: what and how to explore

This is a revised version of a paper that studies learning in subset selection problems where information is collected about subsets of a set of ground elements, rather than about individual elements as in traditional bandit problems. The model assumes that beliefs about the ground elements are always independent, and that the values of those elements are bounded from below. The authors introduce the notion of an optimal cover, which is a set of elements that must be explored sufficiently often. Due to the structure of the subsets (e.g., a network structure), it may not be necessary to explore elements outside this set, because they may not provide any improvement over the current-best solution even when their values are equal to the lower bounds. The authors describe policies that explore such elements in an asymptotically optimal manner.

The technical content of the paper is substantial. While the authors prove various performance bounds, the real value of the paper, in my view, is that it characterizes the structure of learning in such combinatorial problems and points out important conceptual differences from the standard setting. This includes the optimal cover concept and the algorithmic developments for special problem classes in the later sections. These go a long way toward making the case that OCP is useful as a solution method. The authors also make the valuable point that, in this setting, practical performance is determined by factors other than the traditionally-studied dependence on N.

I think that the paper should be published after a minor revision to address the following comments (and a proofreading to fix remaining typos).

A: We thank you for your kind words. We hope to address all the remaining issues satisfactorily.

1. Page 7: It would be useful to give some discussion of the assumption that all observations are independent of the selection decision. This may not always be the case (for example, in an assortment planning problem, the performance of each product would depend on the other products offered at the same time).

A: We now discuss this assumption in Remark 1.

2. Page 8: What is the meaning of "$\mathcal{F}_n/2^{\mathcal{S}}$-measurable? (as opposed to $\mathcal{F}_n$-measurable)?

A: This is the notation used in the textbook "A Probability Path" by Resnick (2013). We have changed it back to the more standard $\mathcal{F}_n$-measurable, per referee's comment.

3. Page 8: There is still some abuse of notation in the definition of T_n. The quantity $T_n(\{a\})$ is not the same as $T_n(S)$ with $S = \{a\}$.

A: We now use the notation $\tilde{T}_n(a)$ instead of $T_n(\{a\})$ to avoid any confusion.

4. Page 11: Are the implemented solutions chosen uniformly at random from the given set? Is any ordering allowed as long as the conditions are satisfied?

A: There are rather minor details about our implementation of the OCP policy which we excluded from the description of the algorithm, for simplicity of exposition. For example, when choosing a solution from the exploration set to implement, in case of a tie, our implementation selects the solution that contains the most number of critical elements. In case of a second tie, our implementation selects a solution with the smallest average cost. We now discuss such implementation details in Section 7.1 per referee's comment.

5. Section 4: The optimality cover concept is one of the most interesting ideas in the paper. It may be useful to discuss similarities and differences with the concept of "optimistic reinforcement learning" of Wen & Van Roy (2013), "Efficient exploration and value function generalization in deterministic systems," perhaps in the conclusion. The similarity with optimistic methods is that, in this paper, an element remains a candidate for exploration as long as there is some chance that it may be optimal under the most optimistic value for its cost. Unlike Wen & Van Roy, here stochastic observations are allowed (and are the focus). But, the boundedness of the values is still crucial, as it seems that every element would be critical under e.g., normality assumptions on the observations.

A: We thank you for bringing this point to our attention. In Section 2 we now discuss the connection of our paper to Wen and Van Roy (2013). Regarding your comment on the criticality of the boundedness of the values, you are correct: every ground element would be critical when the support of the marginal arc distribution is the entire real line.

6. Page 17: In the paragraph following Example 4.7, should e be included in one of the sets in order to form a cover of the digraph?

A: The referee is correct: e should be included in a cover. In the revised manuscript, however, Example 3 is discussed in relation to the elements of set $\mathcal{D}$ (and the LBP problem) which only contain suboptimal arcs, thus excluding e.

7. Page 19 and following: If there is some a satisfying $a \notin D$ for all $D \in \mathcal{D}$, is it possible to never measure any set containing such a and preserve consistency?

A: Yes, it is possible. Moreover, Proposition 3 provides families of instances where many ground elements are not in any element of $\mathcal{D}$.

In general, however, one may need to explore ground elements that are not in any $D \in \mathcal{D}$. We discuss this point using Examples 2 and 3. In Example 2, we have that $\mathcal{D} = \{\{f\}, \{h\}\}$ and one can afford not to explore the arcs $\{p_1, \ldots, p_k, q_1, \ldots, q_k\}$ (i.e., only implement solutions in the exploration set $\{\{e\}, \{f, g, h\}\}$) in order to achieve asymptotic optimality. In Example 3, however, as we discuss in the paragraphs following that example, in order to achieve asymptotic optimality, one may need to explore more arcs than those in $\mathcal{D} = \{\{f_1\}, \{f_2\}, \ldots, \{f_k\}, \{g_1\}, \{g_2\}, \ldots, \{g_k\}\}$. (Please see the discussion following Example 3.)

8. Page 20: The meaning of "all elements on a set C? is not entirely clear. Does this mean that, for each $a \in C$, we have to select subsets S containing a sufficiently often (but the subsets can be different, as long as they include a)? Currently it sounds like C is the subset being explored, but I don't think this is what is intended. Also, please give more detail in the paragraph beginning "Note that by construction..." The definitions of $\mathcal{C}$ and $\mathcal{D}$ are quite dense, and it would be helpful to walk the reader through this argument.

A: We have modified the definitions of both $\mathcal{D}$ and $\mathcal{C}$. The latter definition is now preceded by a discussion that connects it to that of $\mathcal{D}$.

9. Section 6: In all of the graphs, performance is evaluated as a function of N, the number of subsets measured. It would be interesting to mention the sizes of the subsets chosen by different policies. Suppose that a and b are exploration elements, while c is uninformative. Is there any difference between implementing $\{a, b\}$ vs. $\{a, b, c\}$ if both are feasible? Which would OCP choose? Do different types of policies prefer bigger vs. smaller subsets? In Table 1, for example, it seems like UCB1+ is measuring smaller subsets overall.

A: We have added to Table 1 the average length of solutions (i.e., the average number of arcs in the solutions) implemented by different policies (the column called "Length"). Regarding the example mentioned by the referee, the OCP policy explores as dictated by the exploration set induced by the solution to the OCP problem. In this example, if arcs a and b are exploration arcs, they will be explored by implementing a solution in the exploration set that contains them.

10. Page 33: The authors explained in their response why the gamma-exponential KG formula cannot be applied. Would it be possible to apply the formula for multivariate normal priors? Note that independent normal beliefs on the ground elements (using the sample mean and sample variance as the parameters) induce correlations between subsets, given by the sums of the variances of their common elements. This policy would implicitly allow the elements to have negative values, but the formula could still be applied in a non-normal problem, and may work well as long as the sample means are not too close to the lower bounds.

A: One can possibly apply the formula for multivariate normal priors. Because the cost distribution in our numerical experiments is Exponential, we used the Exponential-Gamma conjugate prior for KG policy (this is, in fact, an advantage given to the KG policy as this policy, unlike OCP, requires a prior distribution for the cost).

Response to Comments of Referee 2

Review of "Learning in Combinatorial Optimization: What and How to Explore"

Summary

The authors study a class of combinatorial multi-armed bandit problems, that includes sequential shortest-path, knapsack, spanning-tree, and Steiner-tree problems with unknown costs coefficients. Three decision policies are developed, called static cover-based policy, OCP-based policy, and a hybrid cover/OCP-based policy. The static cover-based policy is based on the idea of separating the time horizon in exploration and exploitation phases; during exploration, solutions from a 'cover' are used that enable the decision maker to learn all unknown costs coefficients; during exploitation, the best solution according to available cost-estimates is used. The OCP-based policy is of similar nature, with the difference that the chosen cover is dynamically updated over time: in each exploration period an optimization problem (called OCP) is solved to determine which cover is used. For both these policies, an $O(\ln N)$ and an $O((\ln N)^{(1+epsilon)})$ upper bound on the regret is proven and with explicit constants before the leading asymptotic terms. For the hybrid policy, a regret upper bound is informally derived. The authors further provide a lower bound on the regret achievable by any (consistent) policy, and discuss the gap between lower and upper bounds.

The paper also contains an extensive discussion on practical implementation, focusing on problems for which the full-information problem, f(B), is polynomially solvable. It is shown that OCP and its relaxation R-OCP are in NP, but in P for a subclass of problems. The authors discuss a MIP formulation of OCP given that f(B) admits an LP formulation, a polynomial-sized extended LP formulation, or no such formulation. Numerical experiments compare the performance of the policies proposed in the paper to several policies from the literature; these results show that the 'adaptive' policy often outperforms other policies.

Assessment

Since the paper has already been thoroughly reviewed, I will focus on the general ideas and contributions of the paper, and neglect typos and small obscurities (although I have to say that the structure of the paper could be improved; I found some parts of the paper difficult to read, and difficult to follow the flow of ideas).

A: We sincerely appreciate your feedback, which helped us improve the paper in various fronts. We agree with your comment on the exposition; the paper has grown significantly since its initial submission. We have tried to streamline the exposition in this new version, and we had to make some tough choices with regard to what material to include, and what to relegate to the appendix. We hope this revision addresses your concerns satisfactorily.

Optimal regret

As remarked by the authors, traditional MAB policies have regret that grows as $K \ln N$, where N is the number of periods and K the number of arms. The relevant question for combinatorial MAB is to develop policies that decrease the constant before the $\ln N$ term; in particular because the number of arms in these problems is prohibitively large.

The authors prove in Theorem 4.11 a lower bound for the class of combinatorial bandits under consideration: $R(F, N) = \Omega(L(\mathbb{E}_F\{B_n\}) \ln N)$. The policies presented in the paper are not shown to have regret matching this lower bound. On page 22 the authors discuss the regret of a hybrid policy, in particular the 'gap' between the upper and lower bound on the regret. They argue that this gap can (largely) be removed, but they don't make this explicit because they believe that the computational complexity of such an optimal policy would make it intractable.

It would have been a very strong theoretical result if the authors had elaborated their intuition and had provided a policy with regret $O(L(\mathbb{E}_F\{B_n\}) \ln N)$, i.e., asymptotically optimal up to even the constant. Even if this policy would be computationally intractable for large instances, it would still have been very useful for smaller instances.

A: We now propose a new policy (i.e., LBP policy) and present an upper bound on its asymptotic regret that scales optimally with the horizon (up to sub-logarithmic terms) and whose leading constant is that of the lower bound (plus tunable constants). More specifically, we show that any consistent policy must incur a regret whose asymptotic growth is at least $L \ln N$ where N is the time horizon and L is a precisely defined constant that depends on the complete optimization problem (its structure and cost distribution). In addition we present a family of policies that incur a regret whose asymptotic growth is upper bounded by $(L + \gamma C)(\ln N)^{1+\varepsilon}$ for positive parameters γ, ε, and C which is a precisely defined constant that depends on the complete optimization problem. Sacrificing the optimal dependence on N allows us to present our results in a much clearer way. This simplification is not unprecedented in the literature and can be avoided as we clarify in the manuscript. The additional terms in the leading constant cannot be eliminated in the general setting. Said differently, as we show in Example 3, the gap between the lower and upper bounds may not be possible to be closed in some settings. We fully explain the source of this additional term and give a family of settings in which it can be vanished.

The proposed near-optimal policies (i.e., LBP policy) require advance knowledge of the parametric form of the cost distribution and, as predicted by the referee, are computationally intractable for large instances – they reconstruct an optimization problem (i.e., LBP) that might not be tractable in practice. For this reason, we propose a different class of policies that solves a proxy for such an optimization problem (i.e., OCP) that can be solved in practice. (We support this practical solvability with several non-trivial combinatorial optimization results and show in the numerical experiments of Section 7 that such policies significantly outperform the benchmark

policies.)

With asymptotic optimality up-to-the-constant being out-of-scope for general problems, it would still have been interesting to prove such a result for a subclass of combinatorial MAB (e.g., shortest path problems on a grid).

A: Please see our discussion above regarding closing the gap between the upper and lower bounds. Regarding the referee's idea of proving a sharper result for a special class of problem, in the discussion following Theorem 4 we state that such a result is possible when $f()$ is a matroid (in particular, we argue that in such cases it is possible to take $\gamma = 0$). This result follows from the fact that feedback from solutions to OCP suffices to guarantee the optimality of said solution (a result which we proved in the previous version of the paper – the extension to the case of LBP is rather straightforward), and from the performance guarantee proof of the OCP-based policy we presented in the previous version of the paper. Because of the length of the proof of these results, and to maintain the focus of the current manuscript, we have decided to omit this and other results pertaining the specific case of matroid problems, which we plan to include on a separate research note.

The constant $L\left(\mathbb{E}_F\left\{B_n\right\}\right)$ before the $\ln N$ term in the regret lower-bound may depend on the number of arms $|A|$, the size of the largest-path s, or the size of the largest cover $|Epsilon|$. The authors do not elaborate how the regret lower-bound depends on these quantities.

A: Being the optimal values of a combinatorial optimization problem, the dependence of $z_L^*\left(\mathbb{E}_F\left\{B_n\right\}\right)$ (which we used to refer to as $L\left(\mathbb{E}_F\left\{B_n\right\}\right)$ in the previous version) and $z_{OCP}^*\left(\mathbb{E}_F\left\{B_n\right\}\right)$ on the aforementioned parameters are quite non-trivial. In previous versions of the manuscript we expressed our upper bounds in terms of $|A|$ and s through rough approximations, this so we can compare with other upper bounds in the literature. These approximations are quite rough and hence far from tight. We believe that one should not look for simplified upper bounds in terms of problem primitives in the combinatorial bandit setting. The fact that $z_L^*\left(\mathbb{E}_F\left\{B_n\right\}\right)$ and $z_{OCP}^*\left(\mathbb{E}_F\left\{B_n\right\}\right)$ cannot be well-approximated by simple functions in terms of the problem size is a very important point which is essential for understanding and appreciating one of the main contributions of the paper. We have modified the paper so as to highlight the point above, and included the key takeaway of the discussion.

Additionally, because our lower and upper bounds are equal up to tunable constants and sub-logarithmic terms, the theoretical comparison with the benchmark (beyond the extensive empirical comparison in our numerical experiments) seems no longer critical. Alternatively, we can prove a simplified "worst-case" upper bound for a slightly modified version of our "adaptive" policy in the

previous version (which we no longer present) that is at least as good as the best bound in the literature. See the discussion following your next comment.

We greatly appreciate the referee's comment, as it has helped us improve the exposition of the paper and clarify its contribution.

They do show that their simple policy has regret $O(|Epsilon|\, s^2 \ln N)$, and that the OCP policy has regret $O(s^2|A|^2 \ln(N))$ (this result is hidden on page 91). It is hard to say how good these bounds are, compared to the best possible in terms of $|A|, s, |Epsilon|$. E.g., is there a policy with regret $O(|A|^2 \ln N)$? The fact that this lower bound is missing, makes it hard to evaluate the upper bounds on the regret. The authors do compare their results with concurrent literature; e.g., Gai et al. achieves regret $O(s^2|A| \ln N)$ and Chen et al. achieve $O(s^2|A| \ln N)$. These upper bounds are smaller than the upper bound $O(s^2|A|^2 \ln(N))$ proven for OCP. Because numerically OCP seems to outperform Gai and Chen, the theoretical regret upper bound on OCP seems not to be tight...
A: As argued above, we believe that one should not look for simplified upper bounds in terms of problem primitives in the combinatorial bandit setting.

Proof of this is that the performance bound for the LBP policy mimics the structure of the LBP problem and thus that of the lower bound result. In the previous version of the paper, we wrote our upper bounds in terms of the problem constants so that one can compare them to those in the literature and concurrent work. However, as referee also points out, the resulting bounds were not tight. It is important to note that these expressions were not the actual upper bounds but rather the worst-case bounds, as the true upper bounds depend on the solution of an auxiliary optimization problem, as explained above.

As a side note, we can prove a bound for a slightly modified version of our *adaptive* policy in the previous version of the paper (which we no longer present) that is $O(s^2 G \ln N)$ with $G \leq |A|$ which is at least as good as the bound of Chen et al. (2013). (We believe that the inclusion of such a result seems no longer relevant for this paper, however, we would be happy to include it in a future response document, should the review team think it is necessary.) It is of paramount importance to note that the regret of our policies are driven by suboptimal exploration efforts, which in the case of the policy in the previous manuscript was focused on a set of size of at most G; other constants in our upper bound, such as s, appear as a product of the probability bounding techniques used. This explains why the bound of our policies are better than those of the benchmark (or can be made better by modifying the probability bounding technique) and why it is crucial to avoid comparing such crude approximations of the theoretical upper bounds.

Apart from $O(\ln(N))$ results, the paper also contains various $O((\ln N)^{(1+epsilon)})$ regret-upper bounds. Although I understand that the regret in combinatorial MAB is mostly determined by the size of the problem and not by N, I still find it hard to evaluate the value of such $O((\ln N)^{(1+epsilon)})$ results. Given N, how should I choose epsilon to obtain the most sharp upper bound? And how does this compare to the $O(\ln(N))$ bound?

A: Following the feedback from the referee, in this version of the manuscript we focus mainly on the leading constants in the bounds, which do not depend on N. With this in mind, and in order to present our results in a clearer manner, we now adopt the idea of adding this suboptimal dependence on N, which is by no means ours (see, e.g., Sauré and Zeevi (2013), or Liu et al. (2012) which was brought to our attention by one of the referees in the second round of revision). We choose to add this additional sub-logarithmic term to the optimal scaling, so as to avoid introducing terms that emanate in part from the proof technique (similar results that adopt the optimal scaling with respect to N can be derived at the expense of introducing additional tuning variables - we discuss this in Section 4.3 of the manuscript), and so as to have a bound that reflects a fundamental insight about the result: asymptotic regret arises from suboptimal exploration, which in our LBP policy is distributed between the solution to LBP problem and, at a lower frequency, the solution to *Cover* problem. With regard to the choice of ε, because the lower bound result is asymptotic, we focus our theoretical results on asymptotic upper bounds, that is when N grows large. If N is known a priori, one can take advantage of that information for designing a policy that probably has a better finite-time performance. We, however, assume that N is not known upfront. We again note that we can prove a bound (which has the $\ln N$ dependence) for a slightly modified version of our *adaptive* policy in the previous version of the paper (which we no longer present) that is at least as good as the best bound in the literature. We also note that, while finite performance is not guaranteed to obey these asymptotic results, one may attempt to quantify the practical performance loss if N can be anticipated. We hope that the changes to the manuscript and the discussion above address the reviewer's concern.

Summarizing, I would like to see $\Omega(..)$ and $O(..)$ results in terms of $|A|, s$, and N; ideally the upper and lower bounds match; and if not, a proper discussion should be included on the theoretical gap. This is however not present in the paper, and apparently not the focus of the authors.

A: We now present upper and lower bounds that match up to sub-logarithmic terms which are not presented in terms of $|A|$ and s for the reasons described above.

Implementation and numerical performance .

A: The authors focus on developing policies with good 'practical' performance that are implementable for 'practical' problems. The Figures in the numerics-section look promising: with the adaptive policy the authors seem to have identified a well-performing policy.
What I miss are results/experiments/discussions on the influence of the *size of the problem*. If the policies developed by the authors are implementable in practice, does that mean that larger problems can be solved than current available alternative methods are capable of? How does computation time scale with problem size and with N, for different policies? (Numerical) result showing the computation time of the different policies could support the assertion that the paper develops 'practical' and 'implementable' policies; but such results are not presented.

A: Per referee's request, we have added another set of experiments to evaluate the influence of size of the problem on both the performance and computation time of different policies. The new experiments, presented in Section 7.3, show that, in a nutshell, the OCP policy significantly outperforms the benchmark policies regardless of the problem size. Moreover, the total computation time of the OCP policy is significantly smaller than that of the UCB1+ policy which is the more "competitive" benchmark in terms of performance. In addition, the computation time of the OCP policy grows logarithmically with N. These observations further pronounce the practical advantage of the OCP policy both in terms of performance (i.e., regret) and computation time.

Further, the figures report average performance. I would be interested to see worst-case values as well. I.e., does adaptive policy always outperform, say, UCB? Or can UCB be a lot better in some particular cases?

A: Per referee's request, we have investigated the sample path regret of different policies. Out of 700 sample paths in our numerical experiments in Section 7.2.2, the OCP policy outperforms the UCB1+ and Extended UCB1+ policies in 700 (i.e., 100% of sample paths) and 697 (i.e., 99.6% of sample paths), respectively. The manuscript presents these findings in Figure 5 via box plots of sample path regrets of different policies for three of the shortest path experiments.

Conclusion

I feel that the theoretical results of the paper in current form are somewhat unsatisfying; it is not really clear to me what these regret upper-bounds tell us about the structure of these problems and about the performance of these policies compared to existing literature. A big step could be made by investigating a regret lower-bound in terms of $|A|, s, |Epsilon|$, and N, and providing a policy that matches this lower bound.

A: The current manuscript presents bounds that are easy to compare; it identifies the source of the optimality gap when it exists, and presents one family of settings when this gap can be closed.

Following the feedback from the referee, we now focus on comparing the leading constants in our bounds, which we now characterize in terms of the solution to the same combinatorial problem. We hope to have presented a compelling case about why one should not approximate our bounds in terms of the problem primitives while striping them from the combinatorial setting they are embedded in, and also, why one should not compare the theoretical upper bounds of the benchmark policies to our upper bounds as we now present upper bounds that match our proved lower bound up to sub-logarithmic terms. We finally note that, as discussed before, our theoretical upper bounds are shown to be arbitrarily better that the benchmark in some settings, and never worse (even when wrote in terms of a worst-case approximation). Moreover, the numerical experiments of Section 7 show that the practical performance of our proposed policies is significantly better than the benchmark policies.

The practical performance of the proposed policies seems promising; especially the adaptive policy seems to perform very well in different problems and instances. I think that this paper could best be seen as a study that focuses on the practical, computational, performance of different policies. With such a focus, and with more attention paid to computation-time aspects in the numerical section, this could become a valuable paper. However, at this moment the paper spends many pages on theoretical results, but at the same time continuously stresses the focus on practical performance. Illustrative is the fact that several theoretically interesting results are expelled to the Appendix.

A: We agree with the referee's assessment: our focus is on practical performance and implementability, which arise as one of the main challenges when leveraging previous methods and concepts to the combinatorial bandit setting. We have indeed spent a significant amount of time on developing a lower bound on performance, which is a contribution of the paper, and proposing policies that admit regret upper bounds that match the lower bound up to sub-logarithmic terms. We have tried our best to discern what might be of interest to the reader and what not in terms of theoretical results; we gladly welcome any comment and advice in this regard, understanding that this is an issue of form and does not relate to the actual merits of the paper.

If this would be a first version, I would recommend major revision and encourage the authors to try to develop more refined theoretical regret bounds. Because the current manuscript is already a revision, I recommend rejecting the paper. I do believe that this paper contains many valuable ideas and results, but that the paper would need stronger theoretical results to justify publication in *Operations Research*.

A: We thank you for your comments about the value of the ideas and results in the paper. We hope that our response to your comments above as well as the modifications to the manuscript (which greatly benefited from your input) had made a solid case about the complex dependence of the lower and upper bounds on the model primitives. Also important, we hope to have shown the nature of our theoretical results: we have developed a first lower bound on performance; we have developed a policy whose upper bound on performance mimics the structure of the lower bound. This is of course not the case for other policies in the combinatorial bandit literature as they lack the lower bound result and consequently present upper bounds in terms of problem primitives. It is not clear whether one can re-write or improve said upper bounds from literature in light of the lower bound problem structure, although a quick review of the proof arguments and techniques reveals that such an extension is not straightforward.

Minor comments

As this paper has already been thoroughly refereed, I have refrained from making many detailed comments and remarks. However, I do want to point out the following:

- In Section 3.1, please clearly distinguish between a random variable and its realization, and please adapt your notation to conventional use. i.e., define a random variable $B = (B_a | a \in A)$, and let $b_1, \ldots, b_N$ be iid realizations, with $b_n = \{b_{a,n} | a \in A\}$. Currently $b_{a,n}$ is presented as a random variable, but since their values are used in the estimation of F, I assume they are meant to be realizations of a random variable.

A: We now make this clarification in Section 3.1 and denote the realization of the random variable $b_{a,n}$ by $\hat{b}_{a,n}$.

- Second, remove the n-dependency in $z^*(E_F(B_n))$. The optimal solution should be independent of n. Currently, in the definition of regret $R(\pi, N)$ on page 32/8, the dependency on n implies that $Nz^*(E_F(B_n))$ is inside the summation...? The same holds for the definition of $\Delta_S^{\mathbb{E}\{B_n\}}$; currently this depends on n, which is quite confusing.

A: Because the random vectors B_n are iid across instances (i.e., n), as the referee also points out, one could drop the dependence on n in $z^*(E_F(B_n))$ and also in the definition of $R(\pi, N)$. We, however, keep such dependence to emphasize the fact that the instances only differ in their costs. To avoid any confusion, we now clarify in the definition of $R(\pi, N)$ that the second term is outside the summation. Regarding the definition of $\Delta_S^{\mathbb{E}\{B_n\}}$, we now define the optimality gaps Δ_S^B for a general cost vector B to avoid any confusion.

- Third, I find the definition of $T_n(\{a\})$ a bit confusing. Suppose that $\{a\}$ is in $\mathcal{S}$. Then $T_n(a)$ is both defined as $|\{m < n : S_m = a\}|$ and as $|\{m < n : a \in S_m\}|$.

A: We now use the notation $\tilde{T}_n(a)$ instead of $T_n(\{a\})$ to avoid any confusion.

- Relating to the comments on LR on page 34/8: I recall that LR has a policy that even matches the constant asymptotically..?

A: This is correct, but those policies are likely not implementable in practice as they require specific knowledge of the parametric form of cost distribution (moreover, such policies must be tailored to each case). In this regard, our LBP policy attains asymptotic optimality using the same type of prior knowledge, although it circumvents the use of upper confidence bounds.

- I found section (i) of the main contributions difficult to understand without first reading Section 3.1.

A: We have entirely re-written the main contributions portion of the introduction. We hope this addresses your concern.

- Is $\mathcal{F}_n/2^{\mathcal{S}}$ conventional notation for a sigma-algebra? I have never seen this notation before.

A: This notation is used in the textbook "A Probability Path" by Resnick (2013), but we have changed it to $\mathcal{F}_n$ per referee's comment.

- Footnote 5: should it be $\{a\}$ instead of a, in Δ_a^F ?
- Perhaps first introduce Gamma(B) before stating algorithm 1.

A: These issues no longer appear in the revised version of the manuscript.

I further noticed that the authors do not state their implementation-results with much confidence. Section 5 starts with 'We provide strong *evidence* that, *at least* for a large class of combinatorial problems, the proposed policies scale *reasonably* well and *should be* implementable for real-size instances" (italics not in original). This statement gave me the impression that the results are rather weak. While I really appreciate avoiding overly-bold statements about the derived results, perhaps here the authors are a bit too modest?

A: We are glad that the referee finds the results satisfactory and strong. We have modified the statements to better reflect the results.

Response to Comments of Referee 3

Referee Report for OPRE-2014-08-500 **Learning in Combinatorial Optimization: What and How to Explore,** by Sajad, Saure, Vielma, submitted to Operations Research

Summary

This paper considers online combinatorial optimization, where the combinatorial optimization problem considered has a linear objective with unknown cost coefficients, corresponding to elements in some ground set. The feasible space of solutions to the combinatorial optimization problem is known. We solve this problem repeatedly, and wish to minimize cumulative regret.

The paper proposes two algorithms for solving this problem, which are both based on finding collections of solutions such that, if we sampled each solution in the collection infinitely often, we would learn which solution (among the much larger space of feasible solutions) is optimal. Upper bounds on asymptotic regret are shown for these algorithms, and a lower bound for regret is given for any consistent algorithm. Bounds are all of the form $c * log(N)$ plus terms that are slower in N, for some constant c.

Major Comments

1. If I am understanding correctly, there is an error in the proof of Lemma B.1, which is used to prove Theorem 4.3. Fixing this error does not seem to be trivial.

   In this lemma, in the second paragraph, we let $\Delta_{min}^{(C,\mathcal{E})}$ be the "minimum optimality gap (i.e., the absolute difference between the objective value of the best and second best solution) in $OCP(\tilde{B})$," where $\tilde{B}$ is obtained by replacing the true costs with upper bounds for elements of the ground set not sampled by the cover $\mathcal{E}$. Then, we let $\hat{\Delta}_{\mathcal{C}}$ be the infimum of these values where $(C,\mathcal{E})$ ranges over $\Gamma\left(E\left\{B_n\right\}\right)$.

   The proof then states that "We assume that $\hat{\Delta}_{\mathcal{C}}$ is positive and bounded". Moreover, later use of the lemma seems to indicate that "positive" means "strictly positive".

   Why is this assumption justified? I don't think that it is universally met. My gut feel is that it is actually a very special property, and many problems to which we would wish to apply this method won't satisfy it. I tried to work out whether Example 3.5 from the paper satisfies this assumption, and I believe that it does not. Let me show you my reasoning:

   For Example 3.5, as the paper points out at the top of page 13, there are actually several minimum cardinality solution covers that solve $OCP\left(E\left\{B_n\right\}\right)$. One of them is S_1, S_2, as described in that paragraph. Taking that as our $(C,\mathcal{E}) \in \Gamma^*\left(E\left\{B_n\right\}\right)$, which is also in $\Gamma\left(E\left\{B_n\right\}\right)$ and obtaining $\tilde{B}$, we see that there are actually several different solution covers that solve $OCP(\tilde{B})$ to optimality. One of them is the original S_1, S_2, and another is $(e_1, p_{2,2}, q_{2,2}, p_{3,3}, q_{3,3}), (p_{1,1}, q_{1,1}, e_2, e_3)$. So in this example, $\tilde{\Delta}_{\mathcal{C}}$ is 0.

Following where Lemma B.1 is used later in the proof of Theorem 4.3, it seems that this assumed strictly positive gap is used throughout the proof, and so fixing it would seem to require some care, and may actually be quite a bit of work.

A: Although we now provide a performance bound for an alternative policy, the issues brought up by the referee still apply: we refer to the proofs of the previous version of the manuscript, so as to avoid confusion.

Regarding the assumption that $\hat{\Delta}_{\mathcal{C}} > 0$, the referee is correct: this assumption may not hold in general and in retrospect, we should have been clearer in our statement. We thank the referee for pointing out this issue.

We now clearly state all the assumptions. In particular, Assumption 1 and its relaxation, Assumption 2, which fulfill the role of assuming that $\hat{\Delta}_{\mathcal{C}} > 0$, are now explicit and discussed in the paper. In particular, we argue that such an assumption holds when, for example, mean costs are randomly drawn from an absolutely continuous distribution. This suits most practical settings where mean costs are unknown and no particular structure for them is anticipated.

2. I believe that there is a simple error in the definition of $\mathcal{D}$ on page 19, line 35, regarding how the empty set is handled. First, observe that the empty set is always an element of $\mathcal{D}'$. (This is observed by examining the definition of $\mathcal{D}'$.) Second, for any $D \in \mathcal{D}'$ that is not the empty set, observe that $D' = \emptyset$ satisfies $D' \subset \mathcal{D}$ but fails to satisfy $D' \notin \mathcal{D}'$. So this tells us that by the current definition, $\mathcal{D} = \{\emptyset\}$ in all situations.

   Of course this is not what was meant. To correct the error, maybe the simplest thing would be to remove $\emptyset$ from $\mathcal{D}'$. The rest of my review will be written assuming this correction is made.

   If we don't remove $\emptyset$ from $\mathcal{D}$, then constraint 8b on page 20 in the formulation of LBP will have the max over the empty set, which I assume is $-\infty$. Then, for this formulation to make sense, we would require $K_{\emptyset} = -\infty$.

A: We thank the referee for highlighting this issue. We have eliminated the empty set in the definition of set $\mathcal{D}'$.

3. I believe there is another error in Section 4.4 ("A limit on achievable performance"), also regarding the collection of sets D, but that is more difficult to correct. I believe that in some examples, this collection of sets can be empty (or, can contain only the empty set if the definition of $\mathcal{D}'$ is not changed to remove the empty set as suggested above), which will cause $z_L^*(\mathbb{E}_F\{B_n\})$ to equal 0, and will cause Theorem 4.11 to be vacuous.

Consider the following example: We wish to solve a shortest path problem from one node (the "origin") to another node (the "destination") in a graph. There are two paths between origin and destination, call them "path a" and "path b", and each path contains only a single link. Let the lower bound l_a on all link costs be 0, and the upper bound u_a be 1. Under F, the expected link cost on path a is 0, and on path b is 1. Since each path contains a single link, I'll use a and b to indicate both paths and links, so $A = \{a, b\}$.

Now let's figure out what $\mathcal{D}'$ contains in this example, based on the definition on page 19.

The set of optimal paths under F, $\mathcal{S}^*(E_F\{B_n\})$, is $\{a\}$, so the intersection of $A \setminus S$ over S in $\mathcal{S}^*(E_F\{B_n\})$ is $\{b\}$. The only subsets of this are $\{b\}$ and $\emptyset$.

Let's see if $\{b\}$ is in $\mathcal{D}'$.

$(E_F\{B_n\})_{A\setminus\{b\}}$ is a B under which the expected cost of both paths a and b is 0. Thus, the set of optimal paths under this B is both paths, $\mathcal{S}^*(E_F\{B_n\})_{A\setminus\{b\}}) = \{a, b\}$. Thus, the intersection of all of these paths is $\cap_{S\in\mathcal{S}^*(E_F\{B_n\})_{A\setminus\{b\}})} S = \{a\} \cap \{b\} = \emptyset$. Thus, $\{b\}$ is not in $\mathcal{D}'$.

Thus $\mathcal{D}'$ is empty (or, contains only the empty set if its definition is not changed as suggested above). This implies $\mathcal{D}$ is empty, $z_L^*(\mathbb{E}_F\{B_n\})$ is 0, and theorem 4.11 is vacuous.

It is probably possible to fix this error, but the best route to doing it is not clear to me.

A: The referee is correct in that the set $\mathcal{D}$ might be empty, in which case $z_L^*(\mathbb{E}_F\{B_n\})$ will be zero. However, this by no means implies that the lower bound result is vacuous but quite the opposite as it identifies settings where active *suboptimal* exploration is rather unnecessary. For instance, in the example discussed by the referee, since the lower bound and the expected cost of arc (path) a are both zero, we conclude that the arc (path) a has a deterministic (and not random) cost of zero. This, in turn, implies that there is no need to explore the alternative arc (path) b indefinitely (e.g., on order $\ln N$) to conclude that the optimal arc (path) is a.

Summarizing, it is possible that the set $\mathcal{D}'$ is indeed *empty* which implies that the lower bound $z_L^*(\mathbb{E}_F\{B_n\})$ is in fact *zero*. This is not an error in the definition of set $\mathcal{D}'$ (except for the fact that the empty set should have been eliminated from the definition). When $z_L^*(\mathbb{E}_F\{B_n\})$ is in fact *zero*, the lower bound result implies that it is possible to achieve a finite regret independent of the horizon. (Moreover, for the case when $f(\cdot)$ is a matroid, a modified LBP-based policy would indeed attain a finite regret in such settings.)

4. There seem to be quite a number of other small errors or typos or steps that require more justification in the proof of Lemma B.1.
   (a) In equation B-16 and also line 36 of page 57, there seems to be an extra S as a superscript to $\mathcal{E}$ in the union over S.

(b) On line 41, of page 57, is $(C', \mathcal{E}')$ assumed to not be equal to $(C, \mathcal{E})$? Otherwise the following inequality (b) is not true.

(c) On line 51 of page 57, why is the first equality true? Do we know that $\tilde{b}_a = \tilde{b}'_a$ a for these a and do we know that $S^* = \overline{S}^*$?

(d) On page 58, line 12, we consider the case $\tilde{\Delta}_C > 0$, but we never consider the case where this is not true.

(e) On lines 22-24, what is f? Is this supposed to be S^*?

A: We thank the referee for bringing up these issues. We have corrected the errors in the new version of the manuscript. We note, however, that because of the new structure of the manuscript, some of these issues no longer arise.

5. The paper mentions [Mersereau et al., 2009] and [Rusmevichientong and Tsitsiklis, 2010], which are papers on linear bandits as part of a larger paragraph surveying literature on papers "with a continuous set of arms".

   However, the paper doesn't really bring out how relevant the well-known linear bandit model is to the model it studies, and doesn't point out that there is a larger literature on linear bandits. The two papers in this area that are probably most well-known are: [Abernethy et al., 2008, Dani et al., 2008]. These papers are not cited in the current paper. They represent probably 30 or more papers on similar models.

   In the most typical formulation of the linear bandit problem, there is a known subset $X \subset \mathbb{R}^d$ of arms, and an unknown vector $\theta \in \mathbb{R}^d$. At the start of each time period, the player chooses an arm $x \in X$ and observes a payoff whose expected value is $x \cdot \theta$. Most commonly, the assumption is that the player observes only the random payoff, but sometimes instead the player gets a noisy observation of $x_i \theta_i$ for each $i \in [d]$. The set X is usually taken to be finite (like the current paper does), and sometimes taken to be uncountable and compact. This problem can be posed in the adversarial setting, the Bayesian setting, or the non-Bayesian stochastic setting (this is the setting considered by the current paper).

   So the current paper is really studying the linear bandit problem, in the stochastic setting, with a finite number of arms, where we observe a noisy payoff for each component of the arm chosen. Now, I believe that the difference with most of the work on linear bandits is this assumption that we observe a noisy payoff for each component (each element of the ground set). I also believe that the authors have already pulled out the most relevant papers from linear bandits to the current work: [Gai et al., 2012] and [Chen et al., 2013].

   But the reader coming from the bandit literature is probably aware of the linear bandit literature, and sees that the model being considered is closely related, and would be made

more comfortable if the paper clearly described the relationship between this literature and the model studied.

A: We thank the referee for his/her comment which helped us make this connection clearer. We now discuss the connection to the linear bandits literature, and cite two of the papers mentioned above.

6. One main contribution of this work is on showing bounds on the performance of the proposed algorithms. These bounds are asymptotic, but it would be helpful to comment on (1) whether these bounds can be computed for finite values of N and (2) how tight these bounds are for the settings considered in the numerical experiments. If the bounds can be computed, the paper would be strengthened by plotting their upper and lower bounds versus actual achieved regret in the numerical experiments. This would help convey to the reader the tightness of the bounds for finite N in some examples, and help answer a more nebulous question: are the bounds meaningful in practice? That is, for values of N that one may have in one's particular problem, are the bounds tight enough to convey information that one could use to select which algorithm is best, or to offer a performance guarantee that would be meaningful in the application studied?

A: As the referee also points out, the lower bound in Theorem 3 is asymptotic, so it is not clear whether the lower bound is meaningful in the finite time. However, per referee's comment, we plot the lower bound for three of our shortest path examples in Figure 4. As can be noted from the graph, in Examples 1 and 2, the lower bound is in fact meaningful and the regret of the OCP and heuristic policies is much closer to the lower bound than the other benchmark policies. In Example 3, however, the lower bound is not meaningful, that is, the lower bound is larger than the regret of all policies as it only provides an asymptotic lower bound on regret. (We here note that solving the LBP problem is hard in general. Even though for these three instances, we were able to find the set $\mathcal{D}$ by enumeration and solve the LBP problem, we could not do so for other instances and as a result, we could not plot the lower bound.)

With regard to the upper bound, our numerical experiments use a variant of the hybrid policy that uses $\gamma = 0$ (i.e., the OCP policy). While this policy significantly outperforms all the benchmark policies in terms of the finite-time performance, the theoretical upper bound does not apply when $\gamma = 0$. For this reason, we decided not to plot the upper bound in our numerical experiments (see the discussion at the end of Section 5).

7. The abstract states, "In particular, we show how to construct policies whose asymptotic performance is arbitrarily close to the best possible." Also, on page 3 in boldface, the manuscript

states, "We show that the proposed policies are essentially optimal with respect to the combinatorial aspects of the problem." These are strong statements, and I am not sure if the analysis in the paper justifies them.

Equation (9) shows that there is a gap between the lower bound on $\liminf_N R^\pi(F,N)/\log(N)$ shown in Theorem 4.11 and the upper bound on the performance of the hybrid policy, which, if I am understanding correctly, says that the policies proposed in this paper do not have an optimal value for $\liminf_N R^\pi(F,N)/\log(N)$. Thus, if my understanding is correct, it doesn't seem well-justified to claim that the policies constructed in this paper have "asymptotic performance that is arbitrarily close to the best possible." The term "essentially optimal" is more vague, and so can be interpreted broadly, but the wording in the manuscript also does not mention that the results are asymptotic. Based on the proof, there seem to be large constants hiding in the terms that are slower in N than $\log(N)$, and so in addition to clarifying whether the constant associated with the $O(\log(N))$ rate is optimal or not, it would also be best to clarify that the performance analysis is asymptotic.

A: We thank the referee for pointing out this issue. In retrospect, we should have been clearer in our statements regarding the performance guarantees of our policies. In the revised manuscript, we have tried to be as clear as possible in our statements relating to such a result. In addition, we have restated our main theorems, so that it is clear that such statements are asymptotic in nature.

With respect to the gap in performance, we now present an upper bound on the asymptotic regret incurred by a new policy that scales optimally with the horizon (up to sub-logarithmic terms) and whose leading constant is that of the lower bound (plus tunable constants). More specifically, we show that any consistent policy must incur a regret whose asymptotic growth is at least $L \ln N$ where N is the time horizon and L is a precisely defined constant that depends on the complete optimization problem (its structure and cost distribution). In addition we present a family of policies that incur a regret whose asymptotic growth is upper bounded by $(L + \gamma C)(\ln N)^{1+\varepsilon}$ for positive parameters γ, ε, and C which is a precisely defined constant that depends on the complete optimization problem. Sacrificing the optimal dependence on N allows us to present our results in a much clearer way. This simplification is not unprecedented in the literature and can be avoided as we clarify in the manuscript. The additional terms in the leading constant cannot be eliminated in the general setting; we fully explain the source of this additional term and give a family of settings in which it can be eliminated. We hope that our response to your comments above as well as the modifications to the manuscript (which greatly benefited from your input) had made a solid case about the complex dependence of the lower and upper bounds on the model primitives.

Minor Comments

1. I found the definition of $\Gamma(B)$ at the top of page 16, and the preceding paragraph at the bottom of page 15, difficult to understand. For a long time, I read the algorithm and asked why $\Gamma(B)$ is there at all it seemed like the algorithm would both be cleaner and perform better if we simply got rid of the check on whether $(C, \mathcal{E})$ was in $\Gamma(\overline{B}_n)$. The check ensures that our cover is optimal if the elements we aren't sampling are much worse than we thought, but why not simply re-solve for $\Gamma^*(B)$?

   After reading the proof, my belief is that this check is here to avoid switching back and forth between different covers that are in $\Gamma^*(E_F[B_n])$, and instead to ensure that a single cover is used infinitely often. Is this correct? The text that surrounds this definition in the body of the paper does not discuss this at all. Please do what you can to explain in the body of the paper why this additional check is present in the algorithm.

A: The referee is correct; such a check aims at converging to a unique exploration set. Because the manuscript no longer introduces the set $\Gamma(B)$ of "feedback-consistent" solutions, we do not elaborate on this issue.

2. In the discussion of Example 4.7, at the bottom of page 17, the covers discussed do not include the solution (e). Shouldn't these be included in the solution covers?

A: The referee is correct, however, in the revised manuscript this example is discussed in relation to the elements of set $\mathcal{D}$ (and the LBP problem) which only contain suboptimal arcs.

3. Why does consistency of π imply line 36 on page 46 of the proof of Proposition 4.9? Can't I have a consistent policy π that tries some suboptimal solution linearly often? It wouldn't be a very good policy of course, but isn't that allowed?

A: According to the definition of consistent policies in (Lai and Robbins 1985) (which we introduce in Section 3.2), a policy π is said to be consistent if $R^\pi(F, N) = o(N^\alpha)$ for all $\alpha > 0$. Since the regret is proportional to the number of times that the policy implements suboptimal solutions (as shown in (2)), one can conclude that for a consistent policy π we have that $\mathbb{E}_\lambda \left\{ N - \sum_{S^* \in \mathcal{S}^*_\lambda} T_{N+1}(S) \right\} = o(N^\alpha)$ for any $\alpha > 0$, where $\mathcal{S}^*_\lambda$ denotes the set of optimal solutions under distribution λ (we note that the difference inside the expectation above is the total number of times, out of N instances, that policy π implements suboptimal solutions).

4. I did not understand how line 54 on page 46 was derived, in the proof of Proposition 4.9. Please add more details.

A: We have added the intermediate steps.

5. On page 8, there seems to be a conflict between the notations $T_n(S)$ and $T_n(\{a\})$. What if $\{a\}$ is also a solution? Then $T_n(S)$ and $T_n(\{a\})$ could indicate different things, even if $S = \{a\}$.

A: We now use the notation $\tilde{T}_n(a)$ instead of $T_n(\{a\})$ to avoid any confusion.

6. On page 15, in algorithm 1, what if $1 \in \Phi$ and we need to solve $\mathcal{S}^*(\overline{B}_1)$ before we have sampled anything? Is $\overline{B}_1$ well-defined?

A: The referee makes a good point: our explanation of the algorithm excluded some minor details (such as this one) for simplicity of exposition. As we mention at the beginning of the numerical experiments in Section 7, each policy starts with an initialization phase in which each solution in a common minimum-size cover of A is implemented. Per referee's comment, we now clarify this issue in the description of the algorithms.