Learning in Combinatorial Optimization: What and How to Explore

Sajad Modaresi
Kenan-Flagler Business School, University of North Carolina at Chapel Hill
Sajad_Modaresi@kenan-flagler.unc.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, a decision-maker implements a solution to a combinatorial optimization problem. The objective coefficient vectors of said problem, which are unobserved prior to implementation, vary from period to period. 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 versus 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. We establish a fundamental limit on the asymptotic performance of any admissible policy that is proportional to the optimal objective value of the LBP problem. We show that such a lower bound might be asymptotically attained by policies that adaptively reconstruct and solve LBP at an exponentially decreasing frequency. Because LBP is likely intractable in practice, we propose policies that instead reconstruct and solve a proxy for LBP, which we call the *Optimality Cover Problem* (OCP). We provide strong evidence of the practical tractability of OCP which implies that the proposed policies can be implemented in real-time. We test the performance of the proposed policies through extensive numerical experiments and show that they significantly outperform relevant benchmarks 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 and typically involve instantiating a deterministic mathematical program, whose solution is implemented repeatedly over 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 from 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 versus exploitation* trade-off which 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, the combinatorial setting has salient features that distinguish it from the traditional bandit. In particular, the combinatorial structure induces correlation between the cost of different solutions, thus raising the question of *how* to collect (i.e., by implementing what solutions) and combine information for parameter estimation. Also, because of such correlation, the underlying combinatorial problem might be invariant to changes in certain parameters, hence not all parameters might need to be estimated to solve said problem. Therefore, answering the question of *what* parameters to estimate is also crucial in the combinatorial setting.

Unfortunately, the features above either prevent or discourage the use of traditional bandit algorithms. First, in the combinatorial setting, traditional algorithms might not be implementable as they would typically require solving the underlying combinatorial problem at each period, 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 amount of time (i.e., 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*, and in the process, understanding how performance depends on the structure of the underlying combinatorial problem.

We consider a decision-maker that at each period must solve a combinatorial optimization problem with a linear objective function whose cost coefficients are random draws from a distribution that is identical in all periods and initially unknown (except for its range). We assume (without

loss of generality) that the underlying combinatorial problem is that of cost minimization, and that the feasible region consists of a time-invariant nonempty collection of nonempty subsets (e.g., paths on a graph) of a discrete finite *ground set* (e.g., arcs of a graph), which is known upfront by the decision-maker. By implementing a solution, the decision-maker observes the cost realizations for the ground elements contained in said solution. Following the bulk of the bandit literature, we measure performance in terms of the cumulative *regret*, which is the expected cumulative additional cost incurred relative to that of an oracle with prior knowledge of the cost distribution.

Main Contributions. Our contributions are as follows:

i) **We establish a fundamental limit on the asymptotic performance of any admissible policy and show that this lower bound is attainable:** We prove that no policy can achieve an asymptotic (on N, which denotes the total number of periods) regret lower than $z^*_{LBP} \ln N$, where z^*_{LBP} is the optimal objective value of an instance-dependent optimization problem, which we call the *Lower Bound Problem* (LBP). This problem simultaneously answers the questions of *what* to explore and *how* to do so. More specifically, we show that in the combinatorial setting it suffices to focus exploration on a subset of the ground set which we call a *critical set*. To the best of our knowledge, ours is the first lower bound for the stochastic combinatorial bandit setting. Then, we show that said lower bound might be asymptotically attained (up to a sub-logarithmic term) by near-optimal policies that adaptively reconstruct and solve LBP at an exponentially decreasing frequency.

ii) **We develop an efficient policy amenable for real-time implementation:** The near-optimal policies alluded above reconstruct LBP adaptively over time. However, their implementation is impractical mainly because LBP depends non-trivially on the cost distribution (and thus, is hard to reconstruct), and because LBP is often an exponentially-sized problem that is unlikely to be timely solvable in practice. Nonetheless, we develop an implementable policy, which we call the OCP-based policy, by means of replacing LBP in the near-optimal policies by a proxy that distills LBP's two main goals: determining what should be explored and how to do so. Said proxy, which we denote the *Optimality Cover Problem* (OCP), is a combinatorial optimization problem that is easier to reconstruct in practice as it depends solely on the vector of mean costs. 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, or via problem-specific heuristics. Finally, we show that a variant of the OCP-based policy admits an asymptotic performance guarantee that is similar to that of the near-optimal policy.

iii) **We numerically show that the OCP-based policy significantly outperforms existing benchmarks:** The key to the efficiency of the OCP-based policy is that it explores as dictated by OCP (i.e., focusing exploration on critical elements) and rarely explores every ground element, let alone every solution, of the combinatorial problem. Through extensive computational experiments we show that such a policy significantly outperforms existing upper-confidence-bound-type benchmarks (i.e., adaptations of the UCB1 policy of Auer et al. (2002) to the combinatorial setting), even when OCP is solved heuristically in a greedy way.

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 in combinatorial settings, is typically exponential. This suggests that the dependence on N might not be the major driver of performance in the combinatorial setting, especially in finite time. To this end, we aim at studying the optimal scaling of the regret with respect to the combinatorial aspects of the setting. In doing so, our performance bounds sacrifice the optimal dependence on N (by adding a sub-logarithmic term) for the sake of clarity in terms of their dependence on the underlying combinatorial aspects of the problem, thus facilitating their comparison to the fundamental performance limit. In this regard, our analysis shows that efficient exploration is achieved when it is focused on a critical set 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 optimal solution to the underlying problem, and doing so at the least possible cost.

The Remainder of the Paper. Section 2 reviews the related work. Section 3 formulates the problem and reviews ideas from the classic bandit setting. In Section 4 we establish a fundamental limit on the asymptotic performance and propose a near-optimal policy. Section 5 presents an efficient practical policy, amenable to implementation, whose performance is similar to that of the near-optimal policy. Section 6 discusses the computational aspects for solving OCP, and Section 7 illustrates the numerical experiments. Finally, Section 8 presents extensions and concluding remarks. All proofs and supporting material are relegated to Online Appendix A.

2. Literature Review

Traditional Bandit Settings. Introduced in Thompson (1933) and Robbins (1952), the multi-armed bandit setting is a classical framework for studying dynamic decision-making under uncertainty. In its traditional 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 versus exploitation trade-off: either pulling

the arm thought to be the "best" (exploitation) at the risk of failing to actually identify such an arm, or trying other arms (exploration) which allows identifying the best arm but hampers reward maximization.

The seminal work of Gittins (1979) shows that for the case of independent arm rewards and discounted 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 $\ln N$ (where N is the total number of periods) dependence in the regret (see Kulkarni and Lugosi (1997) for a finite-sample minimax version of the result). In the same setting, Auer et al. (2002) introduce the celebrated index-based UCB1 policy, which is both efficient and implementable.

Envisioning each feasible solution as an arm, the combinatorial bandit setting that we study 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)). Alternatively, envisioning each ground element (e.g., arcs of a graph) as an arm, the combinatorial 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 such a setting and propose efficient allocation rules; see also Agrawal et al. (1990). The setting we study 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 structure that is exploited to improve upon the performance of traditional bandit algorithms.

A first strand of (non-combinatorial) 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 number of 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. 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), Sauré and Zeevi (2013), and Bernstein et al.

(2018), for example, product assortments are implemented in sequence and (non-linear) rewards are driven by a choice model with initially unknown parameters. Unlike in these papers, we assume in our model that the random cost vector is independent of the implemented solution at each period – see Remark 1 for further details. Also, see Caro and Gallien (2007) for a similar assortment planning formulation with linear independent 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 is expressed in terms of a polynomial of the size of the ground set. 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), which they extend to the combinatorial setting we study – their bound is still expressed as a polynomial of the size of the ground set; also, Liu et al. (2012) develop a policy for network optimization problems (their ideas can be adapted to the setting we study as well) but in a different feedback setting. Their policy collects information through implementation of solutions in a "barycentric spanner" of the solution set, which in the feedback setting of this paper could be set as a solution-cover: see further discussion in Online Appendix A.6. Probable performance of their policy might be arbitrarily worse than that of the OCP-based policy that we propose.

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 develop efficient policies that are implementable in real-time.

3. Combinatorial Formulation versus Traditional Bandits

3.1. Problem Formulation

Model Primitives and Basic Assumptions. We consider a decision-maker that faces a combinatorial optimization problem with a linear objective function repeatedly over time. The feasible region of the combinatorial problem is time-invariant and consists of a nonempty collection $\mathcal{S}$ of nonempty subsets (e.g., paths on a graph) of a discrete finite *ground set* A (e.g., arcs in a graph). We assume that both A and $\mathcal{S}$ are known upfront by the decision-maker, and without loss of generality, that the problem is that of cost minimization.

The cost coefficient vector at each period is a vector of independent random variables. (We assume that all random variables are defined in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.) Furthermore, these random variables jointly form a sequence of i.i.d. random vectors across periods. We let $B_n(a)$ denote the random cost coefficient associated with element $a \in A$ in period $n \geq 1$, and define $B_n := (B_n(a) : a \in A)$ as the random cost coefficient vector in period n. (Throughout the paper, we use the notation $x(a)$ to refer to the a-th element of vector x.) Let F denote the (common) distribution of the cost coefficient vectors and $B \sim F$ with $B := (B(a) : a \in A)$ so that each B_n is an independent copy of B. We assume that F is initially unknown (by the decision-maker) except for its range: it is known that $l(a) < B(a) < u(a)$ a.s. for each $a \in A$ for given vectors $l := (l(a) : a \in A)$ and $u := (u(a) : a \in A)$ such that $l < u$ component-wise. (We also assume for simplicity that the marginal distributions of F are absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}$.)

At the beginning of period n, the decision-maker selects and implements a solution $S_n \in \mathcal{S}$. Then, the random cost vector B_n is realized and the cost associated with solution S_n is incurred by the decision-maker. Finally, the decision-maker observes the realized cost coefficients *only* for those ground elements *included* in the solution implemented, i.e., the decision-maker observes $(b_n(a) : a \in S_n)$, where $b_n(a)$ denotes the realization of $B_n(a)$, $a \in A$, $n \geq 1$.

The decision-maker is interested in minimizing the total expected cost incurred in N periods (N is not necessarily known upfront). Let $\pi := (S_n)_{n=1}^{\infty}$ denote a non-anticipating policy, where $S_n : \Omega \to \mathcal{S}$ is an $\mathcal{F}_n$-measurable function that maps the available "history" at period n to a solution in $\mathcal{S}$, where $\mathcal{F}_n := \sigma(\{B_m(a) : a \in S_m\,, m < n\}) \subseteq \mathcal{F}$ for $n \geq 1$, with $\mathcal{F}_0 := \sigma(\emptyset)$. Finally, note that the expected cumulative cost associated with a policy π is given by

$$J^{\pi}(F, N) := \sum_{n=1}^{N} \mathbb{E}\left\{ \sum_{a \in S_n} B(a) \right\}.$$

(Note that the right-hand-side above depends on the policy π through the sequence $(S_n)_{n=1}^{\infty}$).

REMARK 1. In our formulation, B_n is 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 models.

Full-Information Problem and Regret. Define $\mathcal{B} := \prod_{a \in A}(l(a), u(a))$. For a cost vector $\nu := (\nu(a) : a \in A) \in \mathcal{B}$, define the underlying combinatorial problem, denoted by $Comb(\nu)$, as follows:

$$z^*_{Comb}(\nu) := \min\left\{ \sum_{a \in S} \nu(a) : S \in \mathcal{S} \right\}, \tag{1}$$

where $z^*_{Comb}(\nu)$ denotes the optimal objective value of $Comb(\nu)$. Let $\mathcal{S}^*(\nu)$ denote the set of optimal solutions to $Comb(\nu)$, and define $\mu(a) := \mathbb{E}\{B(a)\}$ for each $a \in A$ and $\mu := (\mu(a) : a \in A)$.

Suppose for a moment that F is known upfront: it can be seen that always implementing an optimal solution to $Comb(\mu)$ is the best among all non-anticipating policies. That is, because of the linearity of the objective function, a clairvoyant decision-maker with prior knowledge of F would implement $S_n \in \mathcal{S}^*(\mu)$ for all $n \geq 1$, thus incurring an expected cumulative cost of

$$J^*(F,N) := N\, z^*_{Comb}(\mu).$$

(Note that the right-hand-side above depends on F through μ.) In practice, the decision-maker does not know F upfront, hence no admissible policy incurs an expected cumulative cost below that incurred by the clairvoyant decision-maker. Thus, we measure the performance of a policy π in terms of its expected *regret*, which for given F and N is defined as

$$R^\pi(F,N) := J^\pi(F,N) - J^*(F,N).$$

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

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

3.2. Known Results and Incorporating Combinatorial Aspects

We begin with two definitions and then discuss the existing results in the bandit literature.

Definition 1 (Regularity). The distribution F is regular if $\mu \in \mathcal{B}$ and the density of $B(a)$: (i) can be parameterized by its mean $\mu(a)$, and thus we denote it by $f_a(\cdot;\mu(a))$; (ii) $0 < I_a(\mu(a),\lambda(a)) < \infty$ for all $l(a) < \lambda(a) < \mu(a) < u(a)$, $a \in A$, where $I_a(\mu(a),\lambda(a))$ denotes the Kullback-Leibler divergence (see e.g., Cover and Thomas (2006)) between $f_a(\cdot;\mu(a))$ and $f_a(\cdot;\lambda(a))$; and (iii) $I_a(\mu(a),\lambda(a))$ is continuous in $\lambda(a) < \mu(a)$ for all $\mu(a) \in (l(a),u(a))$.

The assumption of parameterizing the density function f_a by its mean $\mu(a)$ is made for clarity of exposition and can be relaxed (see Lai and Robbins (1985)).

Definition 2 (Consistency). A policy π is said to be consistent if $R^\pi(F,N) = o(N^\alpha)$ for all $\alpha > 0$, for every regular F.

Traditional multi-armed bandits correspond to settings where $\mathcal{S}$ is formed by ex-ante *identical* singleton subsets of A, i.e., settings where $\mathcal{S} = \{\{a\} : a \in A\}$, and all marginal distributions of F are identical, thus the combinatorial structure is absent. In such settings, and under mild assumptions, the seminal work of Lai and Robbins (1985) establishes an asymptotic lower bound on the regret attainable by any *consistent* policy. Different policies, such as the celebrated index-based UCB1

algorithm (Auer et al. 2002), have been shown to (nearly) attain such asymptotic performance limit. Combining the results in Theorem 1 of Lai and Robbins (1985) and Theorem 1 in Auer et al. (2002), we have that

$$\sum_{a\in A:\mu(a)>\mu^*} (\mu(a)-\mu^*)\, K(a) \leq \liminf_{N\to\infty} \frac{R^{UCB1}(F,N)}{\ln N} \leq \sum_{a\in A:\mu(a)>\mu^*} \frac{8}{\mu(a)-\mu^*},$$

where $\mu^* := \min\{\mu(a) : a \in A\}$, and $K(a)$ denotes the inverse of the Kullback-Leibler divergence between F and an alternative distribution F_a under which $\mu^* = \mu(a)$. Lai and Robbins (1985) show that consistent policies must explore (pull) each element (arm) in A at least on order $\ln N$ times. Thus, balancing the exploration versus exploitation trade-off in the traditional setting narrows down to answering how frequently to explore each element $a \in A$. (The answer to this question is given by $\ln N/N$ exploration frequency in Lai and Robbins (1985)).

Note that the combinatorial setting can be seen as a traditional bandit with a combinatorial number of arms, where arm rewards are correlated. Thus, one might attempt to apply off-the-shelf index-based policies such as UCB1 envisioning each solution $S \in \mathcal{S}$ as an arm. However, this approach has two important disadvantages in our setting (consider that $|\mathcal{S}|$ is normally exponential in $|A|$): (i) computing an index for every solution in $\mathcal{S}$ is comparable to solving the underlying combinatorial problem by enumeration which, in most settings of interest, is impractical; and (ii) because 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, this might no longer be necessary in the combinatorial setting.

To illustrate the issues above, consider two examples in which, for simplicity of exposition, we ignore the exploration frequencies. That is, we assume that whatever elements in A are selected for exploration, they are selected persistently over time (irrespective of how), so that their mean cost estimates are accurate.

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

Consider constants $0 < \epsilon < c \ll M$ and let the distribution F be such that $\mu(e_i) = c$, $\mu(p_{i,j}) = \mu(q_{i,j}) = M$, for all $i \in \{1,\dots,k\}$, $i \leq j \leq k$, $n \in \mathbb{N}$, and $l(a) = \epsilon$ and $u(a) = \infty$ for every arc $a \in A$. The shortest (expected) path is $S^* = \{e_1, e_2, \dots, e_k\}$ with expected length (cost) $z^*_{Comb}(\mu) = kc$, $|A| = k(k+2)$, 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).

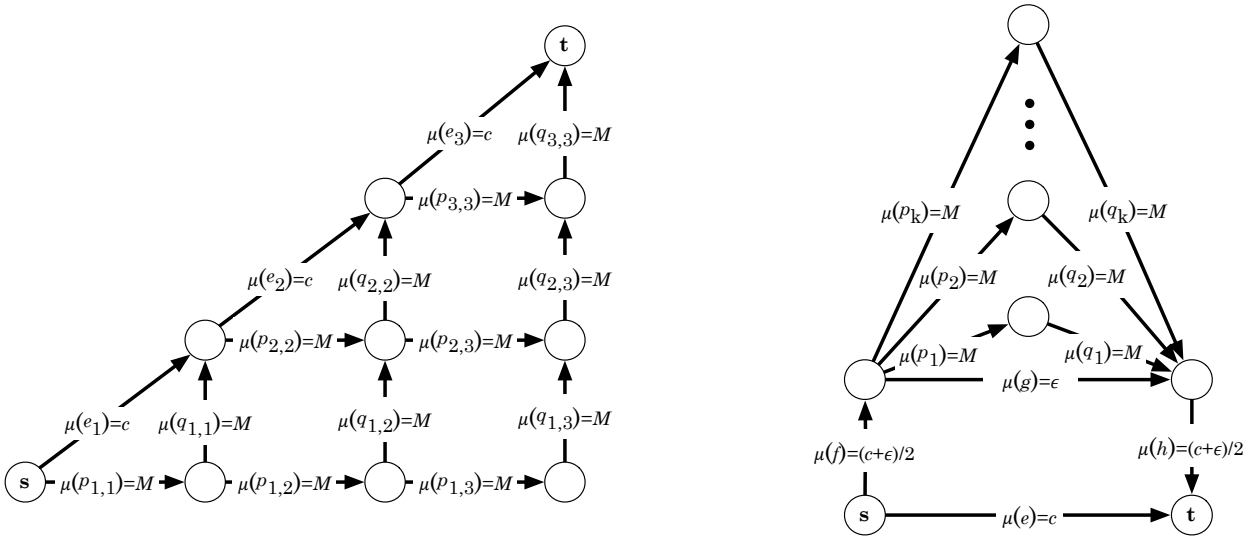

Figure 1 Graph for Example 1 (left), and Example 2 (right).

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 cost 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 Lai and Robbins (1985). To see this, note that, unlike in the traditional setting, solutions are not ex-ante identical in the combinatorial case. This opens up the possibility that information collection on some ground elements might be stopped after a finite number of periods, independent of N, without affecting asymptotic efficiency. This is illustrated in the following example.

EXAMPLE 2. Let $G = (V, A)$ be the digraph depicted in the right panel of Figure 1 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 let F be such that $\mu(e) = c$, $\mu(g) = \epsilon$, $\mu(f) = \mu(h) = \frac{c+\epsilon}{2}$, $\mu(p_i) = \mu(q_i) = M$ for $n \in \mathbb{N}$ and for all $i \in \{1, \ldots, k\}$ 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* scenarios. However, because the combinatorial problem is that of cost minimization, 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 note 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 arcs f and h is sufficient to guarantee the optimality of $\{e\}$. We can explore arcs 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 efficient 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 questions of *what* elements of A to explore (e.g., arcs f and h) and *how* to explore them (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* (i.e., with what frequency) to explore. To achieve this, we extend the fundamental performance limit of Lai and Robbins (1985) from the traditional multi-armed bandits to the combinatorial setting.

4. Bounds on Achievable Asymptotic Performance

4.1. A Limit on Achievable Performance

Following the arguments in the traditional bandit setting, consistent policies must explore those subsets of suboptimal ground elements that have a chance of becoming part of any optimal solution, i.e., those subsets for which there exists an alternative cost distribution F' such that said subset belongs to each optimal solution in $\mathcal{S}^*(\mu')$, where μ' denotes the vector of mean costs under distribution F'. Because the range of F is known, for a given set $D \subseteq A$, it is only necessary to check whether D belongs to each optimal solution in $\mathcal{S}^*\left((\mu \wedge l)(D)\right)$, where

$$(\mu \wedge l)(D) := \left(\mu(a)\mathbf{1}\{a \notin D\} + l(a)\mathbf{1}\{a \in D\} : a \in A\right),$$

and $\mathbf{1}\{\}$ denotes the indicator function of a set. We let $\mathcal{D}(\mu)$ denote the collection of all nonempty subsets of suboptimal ground elements satisfying the condition alluded above, that are minimal with respect to inclusion. We have that $D \in \mathcal{D}(\mu)$ if and only if

(a) $D \subseteq A$ and $D \neq \emptyset$,

(b) $D \cap S^* = \emptyset$ for all $S^* \in \mathcal{S}^*(\mu)$,

(c) $D \subseteq S$ for all $S \in \mathcal{S}^*\left((\mu \wedge l)(D)\right)$,

(d) There is no subset $D' \subset D$ for which $(a) - (c)$ hold.

In other words, we take a pessimistic approach and define $\mathcal{D}(\mu)$ as the collection of nonempty subsets of suboptimal ground elements that become part of any optimal solution if their mean costs are set to their lowest possible values.

As an illustration, consider the examples in the previous section. In Example 1 we have that

$$\mathcal{D}(\mu) = \{\{p_{1,1}, q_{1,1}\}, \{p_{2,2}, q_{2,2}\}, \{p_{3,3}, q_{3,3}\}, \{p_{1,1}, p_{1,2}, q_{1,2}, q_{2,2}\}, \{p_{2,2}, p_{2,3}, q_{2,3}, q_{3,3}\}, \\ \{p_{1,1}, p_{1,2}, p_{1,3}, q_{1,3}, q_{2,3}, q_{3,3}\}, \{p_{1,1}, p_{1,2}, q_{1,2}, p_{2,3}, q_{2,3}, q_{3,3}\}\}$$

and in Example 2 we have that $\mathcal{D}(\mu) = \{\{f\}, \{h\}\}$.

We conclude that for any $D \in \mathcal{D}(\mu)$, there exists an alternative distribution F' under which D is included in every optimal solution. Because said elements are suboptimal under distribution F (condition (b) above), consistent policies must distinguish F from F' to attain asymptotic optimality. The following proposition, whose proof can be found in Online Appendix A.1.1, shows that this can be accomplished by selecting *at least* one element in each set $D \in \mathcal{D}(\mu)$ at a minimum frequency. For $n \geq 1$ and $a \in A$, define the random variable $T_n(a)$ as the number of times that the decision-maker has selected ground element a prior to period n, that is $T_n(a) := |\{m < n : a \in S_m\}|$.

PROPOSITION 1. *For any consistent policy π and $D \in \mathcal{D}(\mu)$ we have that*

$$\lim_{N\to\infty} \mathbb{P}_F \left\{ \frac{\max\{T_{N+1}(a) : a \in D\}}{\ln N} \geq K_D(\mu) \right\} = 1, \tag{2}$$

for a positive finite constant $K_D(\mu)$.

Similar to the traditional bandit setting, $K_D(\mu)$ represents the inverse of the Kullback-Leibler divergence between F and the alternative distribution F' alluded above.

Proposition 1 characterizes *what* needs to be explored by a consistent policy by imposing a lower bound on the number of times that certain subsets of A ought to be explored. To obtain 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 implement solutions in $\mathcal{S}$ that include those ground elements in the most efficient manner. To assess the regret associated with implementing a solution $S \in \mathcal{S}$ given a mean cost vector $\nu \in \mathcal{B}$, we define

$$\Delta_S^\nu := \sum_{a\in S} \nu(a) - z^*_{Comb}(\nu).$$

The following *Lower Bound Problem* (henceforth, *LBP*) 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 3 (LBP). For a given cost vector $\nu \in \mathcal{B}$, define the lower bound problem $LBP(\nu)$ as

$$z^*_{LBP}(\nu) := \min \sum_{S\in\mathcal{S}} \Delta^\nu_S \; y(S) \tag{3a}$$

$$s.t. \qquad \max\{x(a) : a \in D\} \geq K_D(\nu), \quad D \in \mathcal{D}(\nu) \tag{3b}$$

$$x(a) \leq \sum_{S\in\mathcal{S}: a\in S} y(S), \quad a \in A \tag{3c}$$

$$x(a),\, y(S) \in \mathbb{R}_+, \quad a \in A, S \in \mathcal{S}, \tag{3d}$$

where $z^*_{LBP}(\nu)$ denotes the optimal objective value of $LBP(\nu)$. Also, define $\Gamma_{LBP}(\nu)$ as the set of optimal solutions to $LBP(\nu)$

Consider a solution (x, y) to $LBP(\mu)$ where $x = (x(a) : a \in A)$ and $y = (y(S) : S \in \mathcal{S})$. 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 $T_{N+1}(a)/N$. 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 the values $y(S)$ represent the exploration frequencies of the solutions in the cover. Indeed, constraints (3b) enforce exploration conditions (2) and constraints (3c) enforce the cover of the elements of A selected by (3b). The next result establishes a lower bound on the asymptotic regret of any consistent policy in the combinatorial setting which is proportional to $z^*_{LBP}(\mu)$.

THEOREM 1. *The regret of any consistent policy π is such that*

$$\liminf_{N\to\infty} \frac{R^\pi(F, N)}{\ln N} \geq z^*_{LBP}(\mu). \tag{4}$$

From Theorem 1 we see that the fundamental limit on performance is deeply connected to both the combinatorial structure of the problem, as well as the range and mean of distribution F.

REMARK 3. A value of zero for $z^*_{LBP}(\mu)$ suggests that the regret may not necessarily grow as a function of N. To see how this indeed can be the case, 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, $\mathcal{D}(\mu) = \emptyset$ as any suboptimal solution includes arcs f and h, whose cost lower bounds already ensure the optimality of solution $\{e\}$. Thus, in this case, $z^*_{LBP}(\mu) = 0$ and a finite regret (independent of N) might be attainable. Indeed, this setting is such that active learning is not necessary, and information from implementing optimal solutions in $\mathcal{S}^*(\mu)$ suffices to guarantee the optimality of said solutions. (This is not restricted to the case of shortest path problems: in Online Appendix A.1.2 we discuss settings in which $z^*_{LBP}(\mu) = 0$ and the underlying combinatorial problem is minimum-cost spanning tree, minimum-cost perfect matching, generalized Steiner tree, or knapsack.)

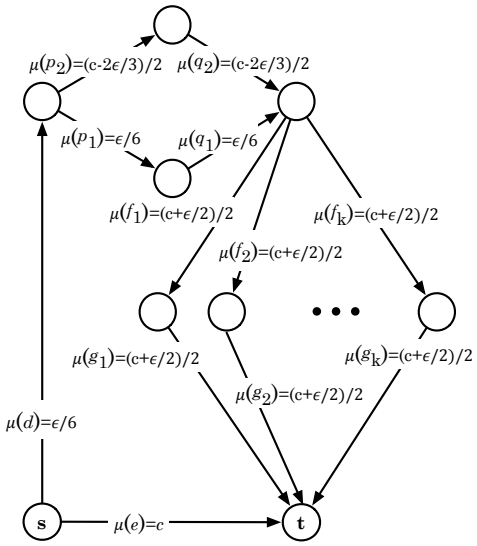

Figure 2 Graph for Example 3.

4.2. An Asymptotically Near-Optimal Policy

For $n \geq 1$, define $\widehat{\mu}_n := (\widehat{\mu}_n(a) : a \in A)$, where

$$\widehat{\mu}_n(a) := \frac{\sum_{m<n:a\in S_m} b_m(a)}{|m < n : a \in S_m|}, \quad a \in A,$$

denotes the sample mean of cost realizations for ground element a prior to period n. (Initial estimates are either collected from implementing a solution-cover or from expert knowledge.)

To match the lower bound of Theorem 1, given the construction of $LBP(\mu)$, it is natural to try allocating exploration efforts only to the solutions prescribed by $LBP(\mu)$ (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 focus exploration on the solution to $LBP(\widehat{\mu}_n)$ hoping that said solution converges to that of $LBP(\mu)$. While this is indeed the case when $\widehat{\mu}_n \to \mu$, collecting information only on solutions prescribed by $LBP(\widehat{\mu}_n)$ does not suffice (in general) to accurately estimate the full mean cost vector, as the following example illustrates.

EXAMPLE 3. 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 $\mu(e) = c$, $\mu(d) = \mu(p_1) = \mu(q_1) = \epsilon/6$, $\mu(p_2) = \mu(q_2) = \frac{c-2\epsilon/3}{2}$, and $\mu(f_i) = \mu(g_i) = \frac{c+\epsilon/2}{2}$ 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 $\mathcal{D}(\mu) = \{\{f_1\}, \{f_2\}, \ldots, \{f_k\}, \{g_1\}, \{g_2\}, \ldots, \{g_k\}\}$. This, in turn, implies that the

minimum-regret solution-cover (i.e., exploration set) induced by the optimal solution to $LBP(\mu)$ 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 $LBP(\widehat{\mu}_n)$ consists of solutions $\{\tilde{S}_i\}_{i=1}^k$. There on, focusing exploration on the solution to $LBP(\widehat{\mu}_n)$ might imply that no further samples of p_1 and q_1 are needed, thus $\widehat{\mu}_n \to \nu' = (\nu'(a) : a \in A)$, with $\nu'(a) = \mu(a)$ for all a in A except $a \in \{p_1, q_1\}$. One can see that in such a case, the exploration set (solution-cover) that $LBP(\widehat{\mu}_n)$ could converge to is $\{\tilde{S}_i\}_{i=1}^k$ with a regret of ck which is not an optimal solution to $LBP(\mu)$.

Example 3 shows that convergence of $LBP(\widehat{\mu}_n)$ to $LBP(\mu)$ (and even $z^*_{LBP}(\widehat{\mu}_n)$ to $z^*_{LBP}(\mu)$) is not guaranteed if exploration is restricted to the solution to $LBP(\widehat{\mu}_n)$. Thus, to assure convergence of $z^*_{LBP}(\widehat{\mu}_n)$ to $z^*_{LBP}(\mu)$ (so as to attain the asymptotic performance in the lower bound result in Theorem 1), one must collect samples on a subset of A that might contain more elements than those explored by the solution to $LBP(\mu)$, 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 $LBP(\widehat{\mu}_n)$ to $LBP(\mu)$, 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 $LBP(\widehat{\mu}_n)$.

Following the discussion above, next we propose a policy that focuses exploration on the solution to $LBP(\widehat{\mu}_n)$, 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 $LBP(\mu)$ by guaranteeing that $\widehat{\mu}_n \to \mu$ (see below for a more detailed discussion). To simplify the reconstruction of the LBP (and the exposition), we make the following technical assumption, needed for proving a performance guarantee.

ASSUMPTION 1. *F is regular and the density function $f_a(\cdot;\cdot)$ is known by the decision-maker for all $a \in A$, and there exists a known finite constant K such that $K_D(\mu) \leq K$ for all $D \in \mathcal{D}(\mu)$. In addition, there is no set $D \subseteq A$ such that $z^*_{Comb}(\mu) = z^*_{Comb}((\mu \wedge l)(D))$ with $\mathcal{S}^*(\mu) \neq \mathcal{S}^*((\mu \wedge l)(D))$.*

Knowing the parametric form of the cost density function for each $a \in A$ reduces the burden of estimating $K_D(\mu)$ as this can be performed by simply estimating μ (as is also the case for Δ^μ_S and the set $\mathcal{D}(\mu)$). The last part of Assumption 1 is necessary to correctly reconstruct the set of constraints (3b), 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 $z^*_{LBP}(\widehat{\mu}_n)$ to $z^*_{LBP}(\mu)$ is assured if $\widehat{\mu}_n$ converges to μ. As discussed in Example 1, this can be achieved by exploring a cover of A. We formalize the problem of finding a minimum-regret cover of A in the following definition.

DEFINITION 4 (*Cover* PROBLEM). For a given cost vector $\nu \in \mathcal{B}$, define the cover problem $Cover(\nu)$ as

$$z^*_{Cover}(\nu) := \min \sum_{S\in\mathcal{S}} \Delta^{\nu}_S \; y(S) \tag{5a}$$

$$s.t. \qquad 1 \le \sum_{S\in\mathcal{S}: a\in S} y(S), \quad a \in A \tag{5b}$$

$$y(S) \in \{0,1\}, S \in \mathcal{S}, \tag{5c}$$

where $z^*_{Cover}(\nu)$ denotes the optimal objective value of the $Cover(\nu)$ problem. Also, define $\Gamma_{Cover}(\nu)$ as the set of optimal solutions to $Cover(\nu)$.

The proposed policy, which we refer to as the *LBP-based* policy and denote by π^*, is described by Algorithm 1. The LBP-based policy formulates and solves $LBP(\widehat{\mu}_n)$ and $Cover(\widehat{\mu}_n)$, and focuses exploration efforts (at different degrees) on the solutions to said problems. To enforce the logarithmic exploration frequency found in Theorem 1, we use 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 $Comb(\widehat{\mu}_n)$ and auxiliary exploration problems $LBP(\widehat{\mu}_n)$ and $Cover(\widehat{\mu}_n)$ 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 $n_{i+2} - n_{i+1} \ge n_{i+1} - n_i$ for all $i \in \mathbb{N}$. Within each cycle, we first solve $Comb(\widehat{\mu}_n)$, $LBP(\widehat{\mu}_n)$ and $Cover(\widehat{\mu}_n)$, 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). In Algorithm 1, we choose $n_i := \max\{\lfloor e^{i^{1/(1+\varepsilon)}} \rfloor, n_{i-1}+1\}$, $i \ge 2$, given a tuning parameter $\varepsilon > 0$. For $S \in \mathcal{S} \setminus \mathcal{S}^*(\widehat{\mu}_n)$, we define

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

where $y(S)$ (in the definition of p_S) refers to the solution to the LBP (see Algorithm 1). We also define $p_{S^*} := (1 - \sum_{S\in\mathcal{S}\setminus\mathcal{S}^*(\widehat{\mu}_n)} p_S)/|\mathcal{S}^*(\widehat{\mu}_n)|$ for $S^* \in \mathcal{S}^*(\widehat{\mu}_n)$. Note that p_S is a probability measure that enforces the right exploration frequency (as prescribed by LBP) for solution $S \in \mathcal{S}$. Also, in Algorithm 1, γ is a tuning parameter that controls the cover-based exploration frequency.

The LBP-based policy admits the following performance guarantee which we prove in Online Appendix A.1.3.

Algorithm 1 LBP-based policy $\pi^*(\gamma,\varepsilon)$

Set $i=0$, and draw $(b_1(a): a\in A)$ randomly from $\mathcal{B}$
for $n=1$ to N **do**
  if $n=n_i$ **then**
    Set $i=i+1$
    Set $S^*\in\mathcal{S}^*(\widehat{\mu}_n)$ [Update exploitation set]
    Set $\mathcal{E}\in\Gamma_{Cover}(\widehat{\mu}_n)$ [Update Cover-exploration set]
    Set $(x,y)\in\Gamma_{LBP}(\widehat{\mu}_n)$ [Update LBP-exploration set]
  end if
  if $T_n(a)<\gamma\, i$ for some $a\in A$ **then**
    Set $S_n=S$ for any $S\in\mathcal{E}$ such that $a\in S$ [Cover-based exploration]
  else
    Set $S_n=S$ with probability p_S, $S\in\mathcal{S}$ [LBP-based exploration/Exploitation]
  end if
end for

THEOREM 2. *Consider $\gamma\in(0,1)$ and $\varepsilon>0$ arbitrary. The LBP-based policy $\pi^*(\gamma,\varepsilon)$ is such that*

$$\lim_{N\to\infty}\frac{R^{\pi^*(\gamma,\varepsilon)}(F,N)}{(\ln N)^{1+\varepsilon}}\leq z^*_{LBP}(\mu)+\gamma\, z^*_{Cover}(\mu). \tag{6}$$

4.3. Performance Gap Analysis

We observe that the constants accompanying the $\ln N$ term in the lower bound and upper bound results in Theorems 1 and 2 do not match exactly. In this section we provide a discussion on this gap.

Optimal Scaling with Respect to N. While it is possible to achieve the optimal $\ln N$ dependence in the upper bound in Theorem 2 (through a different definition of cycles n_i and introduction of additional tunable parameters), this comes at the price of additional constants in front of the second term in the right-hand side of (6). We introduce an additional sub-logarithmic term to the optimal $\ln N$ 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 the near-optimal policy (i.e., the LBP-based policy) is distributed between the solution to LBP and, at a lower frequency, the solution to $Cover$.

Improved Upper Bounds. By setting γ arbitrarily close to zero, one can set the leading constant in the right-hand side of (6) arbitrarily close to that in Theorem 1 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 1 and 2. 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 (2) 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

$$\lim_{N\to\infty} \frac{R^{\pi^*(\gamma,\varepsilon)}(F,N)}{(\ln N)^{1+\varepsilon}} \leq z_R^*(\mu,\gamma),$$

for $\gamma > 0$, where $z_R^*(\nu,\gamma)$ is the optimal solution to a "robust" variation of $LBP(\nu)$ for a given cost vector $\nu \in \mathcal{B}$ (this formulation is presented in Online Appendix A.1.4), such that

$$z_{LBP}^*(\nu) \leq z_R^*(\nu,\gamma) \leq z_{LBP}^*(\nu) + \gamma\, z_{Cover}^*(\nu).$$

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.

Improved Lower Bounds. As shown above, in general it is not possible to improve the leading constant in (6) as finding and validating an optimal solution to $LBP(\mu)$ might require knowledge of the mean costs of ground elements that are not explored by said solution. Hence, to find an optimal solution of $LBP(\mu)$ we may need complementary exploration through a cover or a robust version of $LBP(\mu)$. 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 (4) and (6). 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. An Efficient Practical Policy

A significant obstacle for the implementation of the LBP-based policy is the ability to reconstruct and solve formulation (3) repeatedly over time. Indeed, the right-hand-side of (3b) depends non-trivially on the distribution F, and while LBP is a continuous optimization problem, it has an exponential number of constraints (3b) that do not have a clear separation procedure. In addition, the maximum in constraint (3b) 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 1 and 2. In particular, we propose a policy that follows closely the near-optimal policy of Theorem 2, but replaces formulation (3) by a proxy that: (i) depends on the distribution F only through the vector of mean costs (and thus is easier to reconstruct); and (ii) can be solved effectively with modern optimization techniques. To achieve this, we distill the core combinatorial aspects of the LBP by focusing the questions of *what* ground elements to explore and *how* to do so (i.e., through implementation of which solutions), while somewhat ignoring the question of when to explore (e.g., the precise exploration frequencies).

5.1. The Optimality Cover Problem

With regard to the first question above (what to explore), from Proposition 1 we know that consistent policies must try at least one element in each $D \in \mathcal{D}(\mu)$ at a *specific* minimum frequency, so as to distinguish F from an alternative distribution that makes D part of any optimal solution. (Note that mean cost estimates for these elements should converge to their true values, and that ought to suffice to guarantee the optimality of the solutions in $\mathcal{S}^*(\mu)$.) Here, we consider an alternative, more direct mechanism which, in a nutshell, imposes the same exploration frequency on a set that contains at least one element from each set in $\mathcal{D}(\mu)$.

Suppose that exploration is focused on a subset $C \subseteq A$ and that elements outside C would not be permanently sampled: in the long-run, a consistent mean cost vector estimate $\nu \in \mathcal{B}$ will essentially be such that $\nu(a) \approx \mu(a)$ for $a \in C$, but not much can be said about $\nu(a)$ for $a \notin C$. If persistent exploration on the subset C is to guarantee the optimality of the solutions in $\mathcal{S}^*(\mu)$, independent of $(\mu(a) : a \notin C)$, then (taking a pessimistic approach) C must be such that

$$z^*_{Comb}(\mu) \leq z^*_{Comb}((\mu \wedge l)(A \setminus C)), \tag{7}$$

where we recall that $(\mu \wedge l)(A \setminus C) = (l(a)\mathbf{1}\{a \notin C\} + \mu(a)\mathbf{1}\{a \in C\} : a \in A)$. One can check that $D \cap C \neq \emptyset$ for any $D \in \mathcal{D}(\mu)$ for such a subset C. This, in turn, implies that setting $x(a) = K$ for all $a \in C$, for a large enough positive constant K should lead to a feasible solution to $LBP(\mu)$. This motivates the following definition.

Definition 5 (Critical Set). A subset $C \subseteq A$ is a *sufficient set* if and only if (7) holds. A sufficient set $C \subseteq A$ is a *critical set* if it does not contain any sufficient set $C' \subset C$.

We may use condition (7) to simplify LBP by just enforcing the exploration of a critical set (i.e., what to explore). Once the critical set is identified, we can explore it efficiently (in terms of regret) by implementing a minimum-regret solution-cover (exploration set) of it (i.e., how to explore). Both the selection of the critical set and its minimum-regret solution-cover can be achieved simultaneously through the following combinatorial optimization problem.

Definition 6 (OCP). For a given cost vector $\nu \in \mathcal{B}$, we let the *Optimality Cover Problem* (henceforth, $OCP(\nu)$) be the optimization problem given by

$$z^*_{OCP}(\nu) := \min \sum_{S \in \mathcal{S}} \Delta^\nu_S \; y(S) \tag{8a}$$

$$s.t. \qquad x(a) \leq \sum_{S \in \mathcal{S}: a \in S} y(S), \quad a \in A \tag{8b}$$

$$\sum_{a \in S} (l(a)(1 - x(a)) + \nu(a)x(a)) \geq z^*_{Comb}(\nu), \quad S \in \mathcal{S} \tag{8c}$$

$$x(a),\, y(S) \in \{0, 1\}, \quad a \in A, S \in \mathcal{S}, \tag{8d}$$

where $z^*_{OCP}(\nu)$ denotes the optimal objective value of the $OCP(\nu)$ problem. Also, define $\Gamma_{OCP}(\nu)$ as the set of optimal solutions to $OCP(\nu)$.

By construction, a feasible solution (x,y) to $OCP(\mu)$ 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 (8c) guarantee the optimality of solutions in $\mathcal{S}^*(\nu)$ even if costs of elements outside C are set to their lowest possible values (i.e., $\nu(a) = l(a)$ for all $a \notin C$), and constraints (8b) guarantee that $\mathcal{G}$ covers C (i.e., $a \in S$ for some $S \in \mathcal{G}$, for all $a \in C$). Finally, (8a) ensures that the regret associated with implementing the solutions in $\mathcal{G}$ is minimized. Note that when solving (8), one can impose $y(S^*) = 1$ for all $S^* \in \mathcal{S}^*(\nu)$ without affecting the objective function, thus one can restrict attention to solutions that cover optimal elements of A.

There is a clear connection between $LBP(\mu)$ and $OCP(\mu)$. This is formalized in the next Lemma, whose proof can be found in Online Appendix A.2.

LEMMA 1. *An optimal solution to a linear relaxation of $OCP(\mu)$ when one relaxes the integrality constraints over $y(S)$ variables is also optimal to formulation $LBP(\mu)$ when one replaces $K_D(\mu)$ by 1 for all $D \in \mathcal{D}(\mu)$.*

Proof of Lemma 1 shows that a feasible solution to $LBP(\mu)$ can be mapped to a feasible solution to a linear relaxation of $OCP(\mu)$ (via proper augmentation), and vice versa. The above elucidates that $OCP(\mu)$ is a version of $LBP(\mu)$ that imposes equal exploration frequencies across all solutions. In this regard, the formulations are essentially equivalent up to a minor difference: optimal solutions to $OCP(\mu)$ *must* cover all optimal ground elements; this, however, can be done without affecting performance in both formulations and hence it is inconsequential. In what follows we discuss our practical policy which periodically solves the OCP problem.

5.2. OCP-based Policy

We propose a practical policy, called the *OCP-based* policy, which closely follows the structure of the LBP-based policy, except for a few qualitative differences. The OCP-based policy: (i) solves OCP problem instead of LBP; (ii) does not complement exploration on the solution to the *Cover* problem; and (iii) enforces the logarithmic exploration frequency using the cycle definition $n_1 = 1$ and $n_i := \max\left\{\lfloor e^{i/H} \rfloor, n_{i-1} + 1\right\}$ for all $i \geq 2$, given a fixed tuning parameter $H > 0$. Note that the changes in (ii) and (iii) above ought to eliminate additional suboptimal exploration and induce the proper exploration frequency, respectively, and should result in improved practical performance (we test this policy in our numerical experiments)

The OCP-based policy, which we denote by $\pi_{OCP}(H)$, is depicted in Algorithm 2. At the beginning of each cycle, the OCP-based policy solves for $S^* \in \mathcal{S}^*(\widehat{\mu}_n)$, updates $\Gamma_{OCP}(\widehat{\mu}_n)$, and ensures that all elements in the critical set have been explored with sufficient frequency. If there is time remaining in the cycle, the policy implements (exploits) an optimal solution $S^* \in \mathcal{S}^*(\widehat{\mu}_n)$.

Algorithm 2 OCP-based policy $\pi_{OCP}(H)$

Set $i=0$, $C=A$, $\mathcal{G}$ a minimal cover of A, and draw $(b_1(a): a \in A)$ randomly from $\mathcal{B}$
for $n=1$ to N **do**
  if $n=n_i$ **then**
    Set $i=i+1$
    Set $S^* \in \mathcal{S}^*(\widehat{\mu}_n)$ [Update exploitation set]
    Set $(C,\mathcal{G}) \in \Gamma_{OCP}(\widehat{\mu}_n)$ [Update OCP-exploration set]
  end if
  if $T_n(a) < i$ for some $a \in C$ **then**
    Set $S_n = S$ for any $S \in \mathcal{G}$ such that $a \in S$ [OCP-based exploration]
  else
    Set $S_n = S^*$ [Exploitation]
  end if
end for

Proving a meaningful theoretical performance bound under the modifications $(i)-(iii)$ above is rather challenging. For this reason, we instead consider a variant of the OCP-based policy that simply ignores the changes (ii) and (iii). In addition, such a policy solves for a ϱ-optimal solution, instead of an optimal solution, to OCP, for a tuning parameter $\varrho > 0$. The parameter ϱ allows the policy to converge to an optimal solution to $OCP(\mu)$ – because there might exist multiple optimal solutions to $OCP(\mu)$, solving for a ϱ-optimal solution ensures that the policy settles on one of them. The resulting policy, which we refer to as the *modified OCP-based* policy and denote by π'_{OCP}, can be found in Algorithm 3 in Online Appendix A.2.2.

To prove a performance bound, we need a relaxed version of Assumption 1.

Assumption 2. *There is no set $D \subseteq A$ such that $z^*_{Comb}(\mu) = z^*_{Comb}((\mu \wedge l)(D))$ with $\mathcal{S}^*(\mu) \neq \mathcal{S}^*((\mu \wedge l)(D))$.*

Note that Assumption 2 ensures that Constraint (8c) is not active for any $S \notin \mathcal{S}^*(\mu)$ and any vectors x and y satisfying (8b) and (8d). As discussed in Section 4.2, this 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 is anticipated.

Under Assumption 2, we obtain the following performance bound for the modified OCP-based policy $\pi'_{OCP}(\gamma, \varepsilon, \varrho)$. We note that as in Algorithm 1, ε is a tuning parameter used in the definition of cycles, and γ is a tuning parameter that controls the cover-based exploration frequency.

THEOREM 3. *Consider $\gamma \in (0,1)$, $\varrho > 0$, and $\varepsilon > 0$ arbitrary. We then have that for ϱ sufficiently small*

$$\lim_{N\to\infty} \frac{R^{\pi'_{OCP}(\gamma,\varepsilon,\varrho)}(F,N)}{(\ln N)^{1+\varepsilon}} \leq z^*_{OCP}(\mu) + \gamma\, z^*_{Cover}(\mu).$$

The proof of Theorem 3 follows a similar line of arguments as that of Theorem 1 for the near-optimal LBP-based policy: we refer the reader to Online Appendix A.2.2 for details.

6. Computational Aspects for Solving OCP and Policy Implementation

In this section we address the computational aspects for the practical implementation of the OCP-based policy. 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, which our policies solve repeatedly over time. Note that $Comb(\cdot)$ and $OCP(\cdot)$ have generic combinatorial structures and hence are, a priori, theoretically hard to solve. Hence, practical tractability of said problems is essential for implementation.

Note that the OCP-based policy solves OCP at an exponentially decreasing frequency, thus ensuring its timely solvability in the long-run. In the short-run, a time-asynchronous version of the policy, that uses the incumbent solution to OCP until the new solution becomes available, can be implemented (see Online Appendix A.3.6).

As mentioned above, in general OCP might be theoretically intractable. Nonetheless, in Online Appendix A.3.7 we present a greedy oracle polynomial-time heuristic for OCP. The greedy heuristic requires a polynomial number of calls to an oracle for solving $Comb$. It therefore runs in polynomial time when $Comb$ is polynomially solvable. Furthermore, it provides a practical solution method for OCP when $Comb$ 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)).

An advantage of the greedy heuristic described in Online Appendix A.3.7 is that it only requires an oracle for solving $Comb$ and hence does not require any knowledge of the specific structure of $Comb$. In Section 7 we implement a variant of the OCP-based policy that uses this greedy heuristic to solve OCP. We show that even such a myopic approach can already provide a reasonable

approximation of the OCP-based policy and can significantly outperform alternative approaches. However, we would expect much better performance from heuristics or approximations that exploit the particular structure of $Comb$ for a specific class of problems. Such focus on a specific class of problems is, however, beyond the scope of this paper, thus we instead use mixed-integer programming (MIP) to exploit structure in a generic way.

Over 50 years of theoretical and computational developments in MIP (Jünger et al. 2010) have led to state-of-the-art MIP solvers with machine-independent speeds that nearly double every year (Achterberg and Wunderling 2013, Bixby 2012). One key to this speed is a wide range of highly effective generic primal heuristics (e.g., see Fischetti and Lodi (2011) and the "Primal Heuristic" sections of Gamrath et al. (2016), Maher et al. (2017), and Gleixner et al. (2017)). Hence, formulating OCP as a MIP opens up a wide range of exact and heuristic algorithms to solve it. However, the effectiveness of this approach is strongly contingent on constructing a formulation with favorable properties (Vielma 2015). In what follows we focus our attention on constructing such formulations for the case in which $Comb$ 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 MIP formulations of OCP, which can be effectively tackled by state-of-the-art solvers.

6.1. MIP Formulations of OCP for Polynomially-Solvable Problems

In this section we assume that $Comb$ is polynomially solvable. However, this does not imply that OCP is tractable or practically solvable, as it might contain an exponential (in $|A|$) number of variables and constraints. The following theorem, whose proof can be found in Online Appendix A.3.1, ensures that OCP remains in NP, the class of non-deterministic polynomially-solvable problems (see e.g., Cook et al. (1998)).

THEOREM 4. *If $Comb$ is in P, then OCP is in NP.*

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 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 $Comb$ admits a linear programming (LP) formulation. Then, we address the case when $Comb$ admits a polynomial-sized extended LP formulation, and finally, the case when $Comb$ does not admit such an extended formulation.

Problems with LP Formulations. We present a polynomial-sized formulation of OCP when $Comb$ admits a polynomial-sized LP formulation. To describe this formulation in simple matrix notation we assume that $A := \{1, \dots, |A|\}$. Moreover, for $v \in \mathbb{R}^r$, let $\text{diag}(v)$ be the $r \times r$ diagonal

matrix with v as its diagonal. Also, we remember that $l = (l(a) : a \in A)$ is the vector of lower bounds on the range of F.

PROPOSITION 2. *Let $y^S \in \{0,1\}^{|A|}$ 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 $\operatorname{conv}\left(\{y^S\}_{S\in\mathcal{S}}\right) = \left\{y \in [0,1]^{|A|} : My \le d\right\}$. Then a MIP formulation of $OCP(\nu)$ is given by*

$$\min \quad \sum_{i\in A}\left(\sum_{a\in A}\nu(a)y^i(a) - z^*_{Comb}(\nu)\right) \tag{9a}$$

$$s.t. \qquad x(a) \le \sum_{i\in A} y^i(a), \qquad a \in A \tag{9b}$$

$$My^i \le d, \qquad i \in A \tag{9c}$$

$$M^T w \le \operatorname{diag}(l)(\mathbf{1} - x) + \operatorname{diag}(\nu)x \tag{9d}$$

$$d^T w \ge z^*_{Comb}(\nu) \tag{9e}$$

$$x(a),\, y^i(a) \in \{0,1\},\, w \in \mathbb{R}^m, \qquad a,\, i \in A, \tag{9f}$$

where $x = (x(a) : a \in A)$, $y^i = (y^i(a) : a \in A)$, and $\mathbf{1}$ is a vector of ones.

In the above, x represents the incidence vector of a critical set. Such a condition is imposed via LP duality, using constraints (9d) and (9e), and eliminates the necessity of introducing constraint (8c) for each solution in $\mathcal{S}$. Similarly, each y^i represents the incidence vector of a solution $S \in \mathcal{S}$. A formal proof of the validity of this formulation is included in Online Appendix A.3.3.

Formulation (9) has $O(|A|^2)$ variables and $O(m|A|)$ constraints. If m is polynomial in the size of the input of *Comb*, then we should be able to solve (9) 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 (9c) 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 *Comb* 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 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 2 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 $\text{conv}\left(\{y^S\}_{S \in \mathcal{S}}\right) = P$. Then, a MIP formulation with a polynomial number of variables and constraints of OCP for the spanning tree problem is obtained by replacing (9c) with $Cy^i + Dz^i \leq d$, replacing (9d) with $C^T w \leq \text{diag}(l)(\mathbf{1} - x) + \text{diag}(\nu)x$ 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 2017). Hence, we cannot use the techniques in the previous paragraph to construct polynomial-sized IP formulations of OCP for matching. Fortunately, a simple linear programming observation and a result by Ventura and Eisenbrand (2003) allow constructing a version of (9) with a polynomial number of variables. The observation is that a solution y^* is optimal for $\max\{\nu^T y : My \leq d\}$ if and only if it is optimal for $\max\{\nu^T y : M_i^T y \leq d_i \quad \forall i \in I(y^*)\}$ 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 (9d) analog to that for spanning trees to obtain a version of (9) with a polynomial number of variables. We would still have an exponential number of constraints in (9c), but these can be separated in polynomial time for matching, so OCP 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) shows that there must exist a family of matroid problems without a polynomial-sized extended LP formulation. Fortunately, it can be shown (e.g., see Lemma 4 in Online Appendix A.3.2) that for matroid problems, there exists a unique critical set $C \subseteq A$ that can be found in polynomial time. Once this set is obtained, we can simply replace (9b) by $1 \leq \sum_{i \in A} y^i(a)$ for all $a \in C$ and remove (9d)–(9e). We are not aware of any other polynomially-solvable combinatorial optimization problem which requires non-trivial results to formulate OCP with a polynomial number of variables.

Remark 4. Further improvements and extensions to (9) can be achieved. We give two such examples in Online Appendices A.3.4 and A.3.5. The first example shows how (9) for OCP can be

extended to the case when *Comb* is not in P, but admits a compact IP formulation. The second example gives a linear-sized formulation of OCP for shortest path problems. We finally note that Online Appendix A.3 comments on similar results for the *Cover* problem.

7. Numerical Experiments

In this section we study the finite-time performance of the OCP-based policy from Section 5. In particular, we consider two policies: the OCP-based policy (as defined in Algorithm 2), and a variant that solves OCP heuristically in a greedy way (using Algorithm 5 presented in Online Appendix A.3.7). We refer to this latter policy as the *Greedy-Heuristic* policy. We divide the numerical experiments into two classes: long-term and short-term experiments. We discuss the long-term experiments in Section 7.1 and refer the reader to Online Appendix A.5 for the short-term experiments. In what follows, we first describe the benchmark policies and then discuss the studied settings and results.

7.1. Long-Term Experiments

7.1.1. Benchmark Policies and Implementation Details

Benchmark Policies. Our benchmark policies are versions of UCB1 (Auer et al. 2002), adapted to the combinatorial setting. The UCB1 policy implements solution S_n in period n, where

$$S_n \in \arg\min_{S\in\mathcal{S}} \left\{ \frac{\sum_{m<n:S_m=S}\sum_{a\in S} b_m(a)}{|m<n:S_m=S|} - \sqrt{\frac{2\ln(n-1)}{|m<n:S_m=S|}} \right\}.$$

Note that the estimate cost for solution S is based solely on past implementations of that solution. We improve performance of UCB1 by: (i) conducting parameter estimation at the ground element level to reduce variance of estimation; (ii) using $\min\{T_n(a):a\in S\}$ instead of $|m<n:S_m=S|$ to adjust 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 – this could improve performance by preventing UCB1 to implement solutions that are clearly suboptimal. The resulting policy, which we denote by UCB1+, implements solution S_n in period n, where

$$S_n \in \arg\min_{S\in\mathcal{S}} \left\{ \max\left\{ \sum_{a\in S}\widehat{\mu}_n(a) - \sqrt{\frac{2\ln(n-1)}{\min\{T_n(a):a\in S\}}}, \sum_{a\in S} l(a) \right\} \right\}.$$

In a similar setting, Gai et al. (2012) propose an alternative adaptation of UCB1: a modified version of such a policy in period n implements

$$S_n \in \arg\min_{S\in\mathcal{S}} \left\{ \sum_{a\in S} \max\left\{ \widehat{\mu}_n(a) - \sqrt{\frac{(\mathcal{K}+1)\ln(n-1)}{T_n(a)}}, l(a) \right\} \right\},$$

for a tuning parameter $\mathcal{K}>0$. We denote this policy as Extended UCB1+.

REMARK 5. Note that computing S_n in Extended UCB1+ can be accomplished by solving an instance of $Comb(\cdot)$. Implementing UCB1+ in contrast, requires solving for S_n via explicit enumeration.

Implementation Details. We report results when the marginal distributions of F are exponential (we normalize the mean cost vector so that the maximum solution cost is at most one): we tested many cost distributions and observed consistent performance. For the OCP-based and Greedy-Heuristic policies, we report the results for $H = 5$: preliminary tests using $H \in \{5, 10, 20\}$ always resulted in logarithmic regrets. When choosing a solution from the exploration set to implement, in case of a tie, our proposed 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. 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. 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. Finally, all policies start by implementing each solution in a common minimum-size cover of A.

All 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, the optimization problem must be solved by enumeration). For Steiner tree and knapsack problems, we solved standard IP formulations using GUROBI 5.0 Optimizer. The OCP-based policy solves formulation (8) 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 was around 30 seconds for the UCB1+, Extended UCB1+ and Greedy-Heuristic policies, and around 1.5 minutes for the OCP-based policy. (Note, however, that while the running time of the OCP-based policy grows (roughly) logarithmically with the horizon, those of UCB1+ and Extended UCB1+ grow linearly.)

7.1.2. Settings and Results

We present settings complementary to those in Examples 1 and 2 in the sense that critical sets are large, thus the OCP-based and Greedy-Heuristic policies do not have an immediate advantage. (See Online Appendix A.4 for numerical experiments on Examples 1 and 2.) The settings are comprised of the shortest path, Steiner tree and knapsack problems. We observed consistent performance of our policies across these settings: here we only show a representative setting from each class.

Shortest Path Problem. We consider a shortest path problem on a randomly generated layered graph – see panel (a) of Figure 2 in Ryzhov and Powell (2011) for an example of a layered graph.

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-cover to $OCP(\mu)$ is of size 16 with an implied critical set of size 40.

Knapsack Problem. Here the set A represents items that might go into a 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-cover to $OCP(\mu)$ is of size 8 with an implied critical set of size 17.

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 $|A| = 18$, $|\mathcal{S}| = 10651$, and the minimum-size cover is of size 2. We consider two settings: one where the lower bound vector is $l = 0$ (the solution-cover to $OCP(\mu)$ is of size 7 and the critical set is of size 17) and one where lower bounds are positive numbers that are selected randomly (the solution-cover to $OCP(\mu)$ is of size 5 and the critical set is of size 12).

Results. The left and right panel in Figure 3 depict the average performance of different policies for the shortest path and knapsack settings, respectively. We see that in both settings, the OCP-based and Greedy-Heuristic policies significantly outperform the benchmarks. The left panel in Figure 4 depicts the average performance of different policies for the Steiner tree setting when all cost lower bounds are set to zero. In this case, all arcs (but those trivially suboptimal) are critical, however, the OCP-based and Greedy-Heuristic policies still outperform the benchmarks. The right panel in Figure 4 depicts average performance in the setting where lower bounds are positive numbers. Note that the OCP-based policy significantly outperforms the benchmarks 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. So far, the results in this section show that the average performance of our policies is significantly better than that for the benchmarks. It turns out that our policies outperform the benchmarks not only in terms of average, but also in terms of

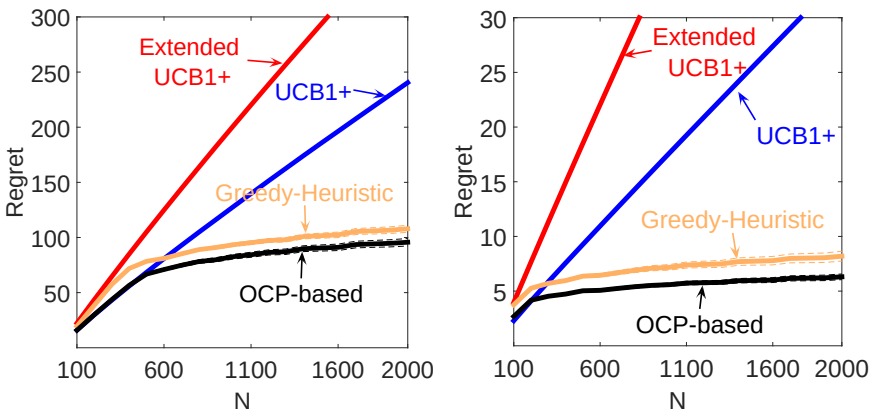

Figure 3 Average performance of different policies on the representative setting for the shortest path (left) and knapsack (right) problems.

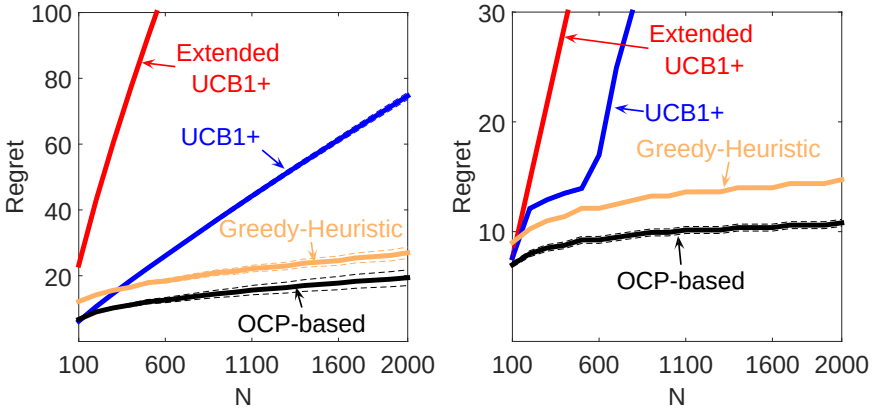

Figure 4 Average performance of different policies on the representative setting for the Steiner tree problem with zero (left) and positive (right) lower bounds.

worst-case regret: we compared the sample path final regrets (i.e., at time period $N = 2000$) of OCP-based policy with those of UCB1+ and Extended UCB1+ policies: out of 700 sample paths in the numerical experiments in Section 7.1.2 (and including those in the Online Appendix A.4), the OCP-based policy outperforms the UCB1+ and Extended UCB1+ policies in all 700 (i.e., 100% of sample paths) and 697 (i.e., 99.6% of sample paths), respectively.

7.2. Experiment with Size of the Ground Set

At the end of Section 1, we argued that, in the combinatorial setting, it is the constant accompanying the $\ln N$ term in a performance guarantee that is worth characterizing. However, prior work (see Section 2), lacking a fundamental performance limit, instead writes such an accompanying constant as a function of the size of the ground set (i.e., $|A|$). However, following Theorem 1, we know that such a constant is not a trivial function of $|A|$. Thus, the question of how said constant

scales *in practice* with the size of the underlying combinatorial problem is of much relevance. For this reason, we next explore how performance of various policies varies with the size of the ground set.

We experiment with the shortest path problem on a layered graph (see Section 7.1.2 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 $N = 20,000$ and 40 replications.

For the OCP-based policy, we solve the OCP problem using the linear-sized formulation (A-20) presented in the Online Appendix A.3.5. We observe a behavior similar to the graph on the left panel of Figure 6 for all choices of $\mathfrak{L}$. That is, the cumulative regret of all three policies grow similar to a function $\mathfrak{K}\ln(n)$ for some policy-dependent constant $\mathfrak{K}$. We consider two estimates for such a constant: (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 $n = 100, 200, \ldots, 20000$ (by minimizing the sum of squared errors). We present the value of both constants for the three policies and varying $|A|$ in Figure 5. 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 6 as we observed similar behavior in other instances. As can be seen in the left panel of Figure 6 (and also from Figure 5), the OCP-based 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-based policy than those for the benchmark policies. In addition, such constants grow with a much smaller rate for the OCP-based policy than the benchmarks. Moreover, as illustrated by the right panel of Figure 6, the computation time of the OCP-based policy grows logarithmically with N. Furthermore, there is a significant variation if we consider computation times. This is shown in Table 1, 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 the OCP-based policy, it includes

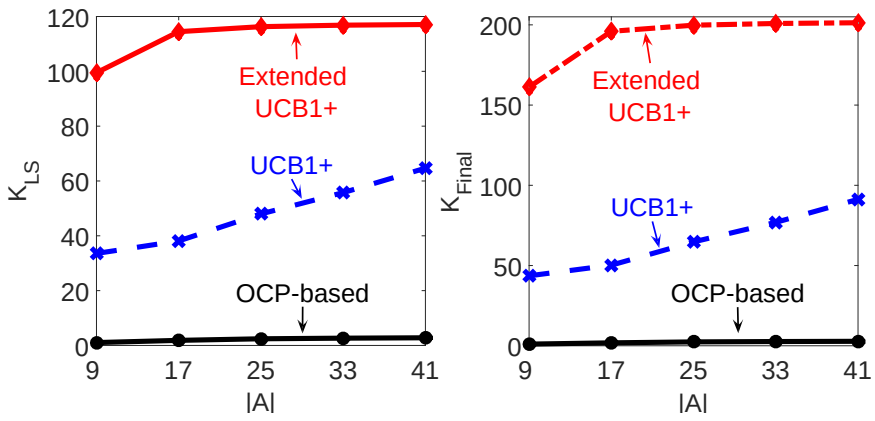

Figure 5 Constants K_{LS} (left) and K_{Final} (right) when increasing the size of the ground set.

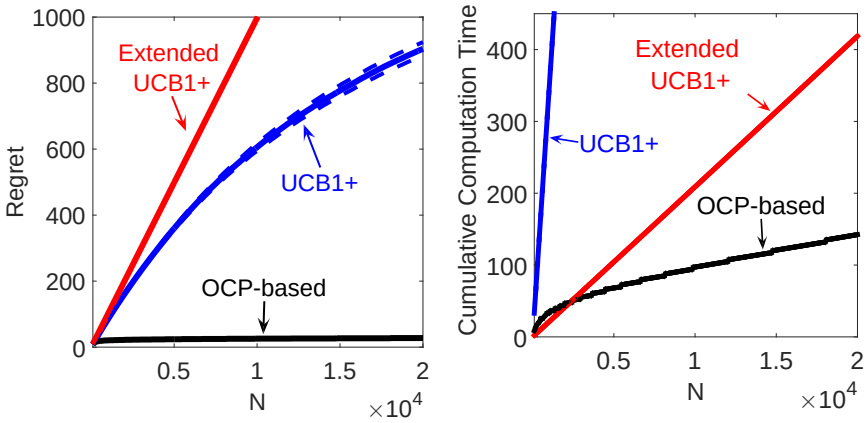

Figure 6 Average performance (left) and computation time (right) as a function of N for the instance with $\mathfrak{L} = 10$, $|A| = 41$, and $|S| = 1025$.

	$\|A\|$				
	9	17	25	33	41
OCP-based	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 1 Average total computation time (in seconds) for each replication of $N = 20,000$.

the solution time of all instances of *OCP* and *Comb* as dictated by Algorithm 2). We can see that the OCP-based 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-based policy, the Extended UCB1+ performs very poorly, in terms of regret, 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-based policy. These observations

further pronounce the practical advantage of the OCP-based policy both in terms of performance (i.e., regret) and computation time.

8. Conclusion

In this paper we study 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 adapt key ideas behind results in the classical bandit setting to develop efficient practical policies. We show that in addition to answering the question of *when* (i.e., with what frequency) to explore, which is key in the traditional bandit setting, in the combinatorial setting the key questions to answer are *what* and *how* to explore. We answer such questions by solving an optimization problem which we call the Lower Bound Problem (LBP). We establish a fundamental limit on the asymptotic performance of any admissible policy that is proportional to the optimal objective value of the LBP problem. We show that such a lower bound might be asymptotically attained by near-optimal policies that adaptively reconstruct and solve LBP at an exponentially decreasing frequency. Because LBP is likely intractable in practice, we propose a simpler and more practical policy, namely the OCP-based policy, that instead reconstructs and solves a proxy for LBP, which we call the Optimaltiy Cover Problem (OCP). This proxy explicitly solves for the cheapest optimality guarantee for the optimal solution to the underlying combinatorial problem. We prove a performance guarantee for a variant of the OCP-based policy, which is proportional to the optimal objective value of the OCP and can be compared to that of LBP. We also provide strong evidence of the practical tractability of OCP which in turn implies that the proposed OCP-based policies are scalable and implementable in real-time. Moreover, we test performance of the proposed policies through extensive numerical experiments and show that they significantly outperform relevant benchmarks in the long-term and are competitive in the short-term.

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. Indeed, it can be shown that in such feedback settings, efficient exploration amounts to focusing exploration on the solution to a variant of OCP. Moreover, the performance guarantee in Theorem 3 would remain valid with the constants associated with this alternative formulation. See Online Appendix A.6 for further details.

From our results, we observe a performance gap between the fundamental limit on (asymptotic) performance (Theorem 1) and the upper bound on the regret associated with near-optimal policies

(Theorem 2). Although we provide a detailed discussion of this gap in Section 4.3, future research can further investigate the possibility of closing this gap. Moreover, studying combinatorial bandit settings with non-linear objective functions is another direction for future research.

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 this work in various fronts. This research is supported in part by the National Science Foundation [Grant CMMI-1233441], and the Complex Engineering Systems Institute, ISCI (CONICYT: PIA FB0816).

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.

Achterberg, T. and Wunderling, R. (2013), Mixed integer programming: Analyzing 12 years of progress, *in* M. Jünger and G. Reinelt, eds, 'Facets of Combinatorial Optimization: Festschrift for Martin Grötschel', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 449–481.

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.

Bernstein, F., Modaresi, S. and Sauré, D. (2018), 'A dynamic clustering approach to data-driven assortment personalization', *To appear in Management Science* . DOI:10.1287/mnsc.2018.3031.

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

Bixby, R. E. (2012), 'A brief history of linear and mixed-integer programming computation', *Documenta Mathematica* pp. 107–121.

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.

Fischetti, M. and Lodi, A. (2011), Heuristics in mixed integer programming, *in* J. Cochran, ed., 'Wiley Encyclopedia of Operations Research and Management Science', Vol. 3, Wiley.

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.

Gamrath, G., Fischer, T., Gally, T., Gleixner, A. M., Hendel, G., Koch, T., Maher, S. J., Miltenberger, M., Müller, B., Pfetsch, M. E., Puchert, C., Rehfeldt, D., Schenker, S., Schwarz, R., Serrano, F., Shinano, Y., Vigerske, S., Weninger, D., Winkler, M., Witt, J. T. and Witzig, J. (2016), The scip optimization suite 3.2, Technical Report 15-60, ZIB, Takustr.7, 14195 Berlin.

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

Gleixner, A., Eifler, L., Gally, T., Gamrath, G., Gemander, P., Gottwald, R. L., Hendel, G., Hojny, C., Koch, T., Miltenberger, M., Müller, B., Pfetsch, M. E., Puchert, C., Rehfeldt, D., Schlösser, F., Serrano, F., Shinano, Y., Viernickel, J. M., Vigerske, S., Weninger, D., Witt, J. T. and Witzig, J. (2017), The scip optimization suite 5.0, Technical Report 17-61, ZIB, Takustr.7, 14195 Berlin.

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

Jünger, M., Liebling, T., Naddef, D., Nemhauser, G., Pulleyblank, W., Reinelt, G., Rinaldi, G. and Wolsey, L. (2010), *50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art*, Springer-Verlag, New York.

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.

Maher, S. J., Fischer, T., Gally, T., Gamrath, G., Gleixner, A., Gottwald, R. L., Hendel, G., Koch, T., Lübbecke, M. E., Miltenberger, M., Müller, B., Pfetsch, M. E., Puchert, C., Rehfeldt, D., Schenker, S., Schwarz, R., Serrano, F., Shinano, Y., Weninger, D., Witt, J. T. and Witzig, J. (2017), The scip optimization suite 4.0, Technical Report 17-12, ZIB, Takustr.7, 14195 Berlin.

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), 'Some 0/1 polytopes need exponential size extended formulations', *Mathematical Programming* **142**, 255–268.

Rothvoß, T. (2017), 'The matching polytope has exponential extension complexity', *Journal of the ACM (JACM)* **64**(6), 41.

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., 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.

Vielma, J. P. (2015), 'Mixed integer linear programming formulation techniques', *SIAM Review* **57**, 3–57.

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.

Online Appendix Companion to "Learning in Combinatorial Optimization: What and How to Explore"

Appendix A: Omitted Proofs and Complementary Material

A.1. Omitted Proofs and Material from Section 4

A.1.1. A Limit on Achievable Performance

In this section we prove Proposition 1 and Theorem 1. We begin with some preliminaries. Define $\Theta_a := (l(a), u(a))$. For $\lambda(a) \in \Theta_a$, the Kullback-Leibler divergence between $f_a(\cdot;\mu(a))$ and $f_a(\cdot;\lambda(a))$ is defined as

$$I_a(\mu(a),\lambda(a)) := \int_{-\infty}^{\infty} [\ln\left(f_a(x_a;\mu(a))/f_a(x_a;\lambda(a))\right)] f_a(x_a;\mu(a))\ dx_a.$$

Define $\lambda := (\lambda(a) : a \in A)$ and let $\mathbb{E}_\lambda$ and P_λ denote the expectation and probability induced when each f_a receives the parameter $\lambda(a) \in \Theta_a$ for all $a \in A$.

Define $\tilde{T}_{N+1}(S)$ as the number of times that the decision-maker has implemented solution $S \in \mathcal{S}$ prior to period $N+1$, that is, $\tilde{T}_{N+1}(S) := |m < N+1 : S_m = S|$. We can then rewrite the regret function as

$$R^\pi(F,N) = \sum_{S\in\mathcal{S}} \Delta_S^\mu\ \mathbb{E}_F\left\{\tilde{T}_{N+1}(S)\right\}.$$

Next, we prove Proposition 1.

PROPOSITION 1. *For any consistent policy π and $D \in \mathcal{D}(\mu)$ we have that*

$$\lim_{N\to\infty} \mathbb{P}_F\left\{\frac{\max\{T_{N+1}(a) : a \in D\}}{\ln N} \geq K_D(\mu)\right\} = 1, \tag{2}$$

for a positive finite constant $K_D(\mu)$.

Proof of Proposition 1. For simplicity, we denote $\mathcal{D}(\mu)$ by $\mathcal{D}$. Consider $D \in \mathcal{D}$ as defined in Section 4, and take $\lambda \in \mathcal{B} = \prod_{a\in A}(l(a),u(a))$ so that $\lambda(a) = \mu(a)$ for $a \notin D$, and that $D \subseteq S^*$ for all $S^* \in \mathcal{S}^*(\lambda)$. By the consistency of π, one has that

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

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

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

We focus on $0 < \alpha < 1$ and take ϵ such that $0 < \alpha < \epsilon < 1$. Define $I(D,\lambda) := |D| \max\{I_a(\mu(a), \lambda(a)) : a \in D\}$, $D \in \mathcal{D}$. We then have that

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

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

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

where in above, $(N - O(\ln N))$ refers to $N - ((1-\epsilon)\ln N / I(D,\lambda))$. The above can be re-written as

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

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

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

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

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

and note that

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

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

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

where (a) follows from noting that probabilities under λ and μ differ only in that cost realizations in D have different probabilities under λ and μ, and (b) follows from noting that $L_{T_{N+1}(a)}(a) \leq (1-\alpha)\ln N/|D|$ for all $\omega\in\Xi(N)$.

From above and (A-2) we have that

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

Now, fix $a\in D$. By the Strong Law of Large Numbers (see page 8 of Lai and Robbins (1985)) we have that

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

From above, we have that

$$\lim_{N\to\infty} \max\left\{ \frac{L_m(a)}{\frac{(1-\epsilon)\ln N}{|D|\, I_a(\mu(a),\lambda(a))}} : m < \frac{(1-\epsilon)\ln N}{|D|\, I_a(\mu(a),\lambda(a))} \right\} = I_a(\mu(a),\lambda(a)), \quad \text{a.s.}[\mathbb{P}_\mu], \quad \forall a\in D.$$

From above and seeing that $1-\alpha > 1-\epsilon$, we have for all $a\in D$ that

$$\begin{gathered}\lim_{N\to\infty} \mathbb{P}_\mu\left\{ L_m(a) > \tfrac{(1-\alpha)\ln N}{|D|} \text{ for some } m < \tfrac{(1-\epsilon)\ln N}{|D| I_a(\mu(a),\lambda(a))} \right\} =\\ \lim_{N\to\infty} \mathbb{P}_\mu\left\{ \max\left\{ \tfrac{L_m(a)}{\frac{(1-\epsilon)\ln N}{|D| I_a(\mu(a),\lambda(a))}} : m < \tfrac{(1-\epsilon)\ln N}{|D| I_a(\mu(a),\lambda(a))} \right\} > \left(\tfrac{1-\alpha}{1-\epsilon}\right) I_a(\mu(a),\lambda(a)) \right\} = 0.\end{gathered}$$

Because $I(D,\lambda) \geq |D|\, I_a(\mu(a),\lambda(a))$, we further have that

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

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

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

which in turn implies

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

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

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

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

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

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

□

THEOREM 1. *The regret of any consistent policy π is such that*

$$\liminf_{N\to\infty} \frac{R^\pi(F,N)}{\ln N} \geq z^*_{LBP}(\mu). \tag{4}$$

Proof of Theorem 1. For any consistent policy π, define $\zeta^\pi(F,N) := \sum_{S\in\mathcal{S}} \Delta^\mu_S \ \tilde{T}_{N+1}(S)$ to be the total additional cost (relative to an oracle) associated with that policy. Note that $\mathbb{E}_F\{\zeta^\pi(F,N)\} = R^\pi(F,N)$. The next lemma ties the asymptotic bounds in (2) to the solution to $LBP(\mu)$ and establishes an asymptotic bound on the regret of any consistent policy.

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

$$\lim_{N\to\infty} \mathbb{P}_F\Big(\zeta^\pi(F,N) \geq z^*_{LBP}(\mu)\ln N\Big) = 1.$$

Proof of Lemma 2. Define the event $\Upsilon_N := \bigcap_{D\in\mathcal{D}(\mu)} \{\max\{T_{N+1}(a) : a \in D\} \geq K_D(\mu)\ln N\}$ and let Υ^c_N denote the complement of the event Υ_N. Note that $\zeta^\pi(F,N) \geq z^*_{LBP}(\mu)\ln N$ when Υ_N occurs, because $\Big(x(a) = \frac{T_{N+1}(a)}{\ln N}, a \in A\Big)$ and $\Big(y(S) = \frac{\tilde{T}_{N+1}(S)}{\ln N}, S \in \mathcal{S}\Big)$ are feasible to $LBP(\mu)$. Thus, one has that

$$\begin{aligned} \mathbb{P}_F\left\{\frac{\zeta^\pi(F,N)}{\ln N} < z^*_{LBP}(\mu)\right\} &= \mathbb{P}_F\left\{\frac{\zeta^\pi(F,N)}{\ln N} < z^*_{LBP}(\mu)\,,\, \Upsilon_N\right\} + \mathbb{P}_F\left\{\frac{\zeta^\pi(F,N)}{\ln N} < z^*_{LBP}(\mu)\,,\, \Upsilon^c_N\right\} \\ &\leq \mathbb{P}_F\{\Upsilon^c_N\}. \end{aligned} \tag{A-5}$$

From Proposition 1 and the union bound, we have that

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

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

$$\lim_{N\to\infty} \mathbb{P}_F\{\zeta^\pi(F,N) < z^*_{LBP}(\mu)\ln N\} = 0.$$

□

Note that Lemma 2 establishes convergence in probability (hence it can be used to bound $\zeta^\pi(F,N)$, rather than just its expectation, which is the regret). Theorem 1 then follows directly from Lemma 2 and Markov's inequality.

A.1.2. Family of Instances with Finite Regret

PROPOSITION 3. *If the combinatorial problem $Comb(\nu)$ in (1) 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 $z^*_{LBP}(\mu) = 0$ while the minimum-size cover of A is arbitrarily large.*

Proof of Proposition 3. The family for the shortest path problem is that based on Example 2 (which is parametrized by an integer k), and described after Theorem 1 in Section 4.

For minimum-cost spanning tree, consider a complete graph $G = (V, A)$ with $|V| = k$ nodes, $\mu(a) = \epsilon$ 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, $\mathcal{D}(\mu) = \emptyset$, independent of k, thus $z^*_{LBP}(\mu) = 0$. 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, $\mu(a) = \epsilon$ 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, $\mathcal{D}(\mu) = \emptyset$, independent of k, thus $z^*_{LBP}(\mu) = 0$.

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 $\mu(0) = \epsilon$ 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, $\mathcal{D}(\mu) = \emptyset$, independent of k, thus $z^*_{LBP}(\mu) = 0$.

A.1.3. Performance Guarantee of the LBP-based policy

Suppose that Assumption 1 holds. The following result provides a performance guarantee for the LBP-based policy.

THEOREM 2. *Consider $\gamma \in (0,1)$ and $\varepsilon > 0$ arbitrary. The LBP-based policy $\pi^*(\gamma,\varepsilon)$ is such that*

$$\lim_{N\to\infty} \frac{R^{\pi^*(\gamma,\varepsilon)}(F,N)}{(\ln N)^{1+\varepsilon}} \leq z^*_{LBP}(\mu) + \gamma\, z^*_{Cover}(\mu). \tag{6}$$

Proof of Theorem 2. The regret of the policy π^* (we drop the dependence of π^* on γ and ε for simplicity) stems from two sources: exploration efforts and exploitation errors. That is,

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

where $R_1(F,N)$ is the exploration-based regret, i.e., that incurred at period n during cycle i if $T_n(a) < \gamma\, i$ for some $a \in A$, or alternatively when sampling a solution, picking $S_n \neq S^*$ with $S^* \in \mathcal{S}^*(\widehat{\mu}_{n_i})$, and $R_2(F,N)$ is the exploitation-based regret, i.e., that incurred when $T_n(a) \geq \gamma\, i$ 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 period, 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 $S^* \in \mathcal{S}^*(\widehat{\mu}_{n_i})$. We use the event $\{S_n \in \mathcal{S}^*(\widehat{\mu}_{n_i})\}$ to denote exploitation. We also define $\Delta^\mu_{max} := \max_{S\in\mathcal{S}}\{\Delta^\mu_S\}$. We then have that

$$\begin{aligned} R_2(F,N) &\leq n_{i'}\Delta^\mu_{max} + \sum_{i=i'}^{\lceil (\ln N)^{1+\varepsilon}\rceil} \sum_{n=n_i}^{n_{i+1}-1} \mathbb{E}\left\{\mathbf{1}\left\{T_n(a) \geq \gamma(i-1)\,, \forall a \in A\,, S_n \in \mathcal{S}^*(\widehat{\mu}_{n_i})\right\}\Delta^\mu_{S_n}\right\} \\ &\leq n_{i'}\Delta^\mu_{max} + \sum_{i=i'}^{\infty} (n_{i+1}-n_i)\,\mathbb{P}\left\{\mathcal{S}^*(\widehat{\mu}_{n_i}) \not\subseteq \mathcal{S}^*(\mu), T_{n_i}(a) \geq \gamma(i-1), \forall a \in A\right\}\Delta^\mu_{max}. \end{aligned} \tag{A-6}$$

Next, we find an upper bound for the probability inside the sum in (A-6). For this, note that

$$\{\mathcal{S}^*(\widehat{\mu}_{n_i}) \not\subseteq \mathcal{S}^*(\mu)\} \subseteq \bigcup_{a\in A}\left\{|\widehat{\mu}_{n_i}(a) - \mu(a)| \geq \frac{\Delta^\mu_{min}}{2s}\right\}, \tag{A-7}$$

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

$$\left\{|\widehat{\mu}_{n_i}(a) - \mu(a)| < \frac{\Delta^{\mu}_{min}}{2s}\,, \forall a \in A\right\} \subseteq \left\{\sum_{a\in S^*} \widehat{\mu}_{n_i}(a) < \sum_{a\in S} \widehat{\mu}_{n_i}(a)\,, \forall S^* \in \mathcal{S}^*(\mu)\,, S \in \mathcal{S} \setminus \mathcal{S}^*(\mu)\right\}.$$

The next proposition, whose proof can be found in Online Appendix A.7, allows us to bound (A-6) using the observation above.

PROPOSITION 4. *For any fixed $a \in A$, $n \in \mathbb{N}$, $k \in \mathbb{N}$, and $\epsilon > 0$ we have that*

$$\mathbb{P}\{|\widehat{\mu}_n(a) - \mu(a)| \geq \epsilon, T_n(a) \geq k\} \leq 2\exp\left\{-\frac{2\epsilon^2 k}{\mathcal{L}^2}\right\},$$

where $\mathcal{L} := \max\{u(a) - l(a) : a \in A\}$.

Using the above, the union bound, and (A-7), we have that

$$\mathbb{P}\{\mathcal{S}^*(\widehat{\mu}_{n_i}) \not\subseteq \mathcal{S}^*(\mu)\,, T_{n_i}(a) \geq \gamma(i-1)\,, \forall a \in A\} \leq$$
$$\sum_{a\in A} \mathbb{P}\left\{|\widehat{\mu}_{n_i}(a) - \mu(a)| \geq \frac{\Delta^{\mu}_{min}}{2s}, T_{n_i}(a) \geq \gamma(i-1)\right\} \leq 2\,|A| \exp\left\{-\frac{\left(\Delta^{\mu}_{min}\right)^2 \gamma(i-1)}{2s^2\mathcal{L}^2}\right\}. \quad \text{(A-8)}$$

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-6) and (A-8) we conclude that

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

Because $(i+1)^{1/(1+\varepsilon)} < i\frac{\left(\Delta^{\mu}_{min}\right)^2 \gamma}{2s^2\mathcal{L}^2}$ 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 period n when there exists $a \in A$ such that $T_n(a) < \gamma\, i$. The latter arises when the cover-based exploration requirements are met and one samples $S_n \neq S^*$ for $S^* \in \mathcal{S}^*(\widehat{\mu}_{n_i})$. Let $R_1^{Cover}(F,N)$ and $R_1^{LBP}(F,N)$ denote the cover-based and LBP-based exploration regrets, respectively, so that

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

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 the $Cover\,(\mu)$ problem in (5), i.e.,

$$\Delta^C_{min} := \min\left\{\left(\sum_{S\in\mathcal{E}}\Delta^\mu_S\right) - z^*_{Cover}(\mu) : \mathcal{E}\in\mathbf{C}\setminus\Gamma_{Cover}\,(\mu)\right\}.$$

We assume that $\Delta^C_{min}>0$, since otherwise, the cover problem is trivial. Consider $i>i'$ and let $\mathcal{E}_i\in\Gamma_{Cover}\,(\widehat{\mu}_{n_i})$ denote the cover-based exploration set for any period $n\in[n_i,n_{i+1}-1]$. 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 $T_{n_i}(a)\ge\gamma(i-1)$ for all $a\in A$ when $i>i'$, we have that

$$\begin{aligned}
R_1^{Cover}(F,N) &\le c\,i'\Delta^\mu_{max} + \sum_{i\in I}\mathbb{E}\left\{\mathbf{1}\left\{T_{n_i}(a)\ge\gamma(i-1)\;\forall a\in A\,,\,\mathcal{E}_i\in\Gamma_{Cover}\,(\mu)\right\}\sum_{S\in\mathcal{E}_i}\Delta^\mu_S\right\}\\
&\qquad + \sum_{i\in I}\mathbb{E}\left\{\mathbf{1}\left\{T_{n_i}(a)\ge\gamma(i-1)\;\forall a\in A\,,\,\mathcal{E}_i\notin\Gamma_{Cover}\,(\mu)\right\}\sum_{S\in\mathcal{E}_i}\Delta^\mu_S\right\}\\
&\le c\,i'\Delta^\mu_{max} + \left(\gamma\,(\ln N)^{1+\varepsilon}+1\right)z^*_{Cover}(\mu)\\
&\qquad + \Delta^\mu_{max}\,c\sum_{i\in I}\mathbb{P}\left\{T_{n_i}(a)\ge\gamma(i-1)\;\forall a\in A\,,\,\mathcal{E}_i\notin\Gamma_{Cover}\,(\mu)\right\}. \qquad \text{(A-9)}
\end{aligned}$$

Next, we bound the probability inside the sum in (A-9). For that, observe

$$\{\Gamma_{Cover}(\widehat{\mu}_{n_i})\not\subseteq\Gamma_{Cover}(\mu)\}\subseteq\bigcup_{a\in A}\left\{|\widehat{\mu}_{n_i}(a)-\mu(a)|\ge\frac{\Delta_1}{4cs}\right\}, \qquad \text{(A-10)}$$

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

$$\begin{aligned}
\left\{|\widehat{\mu}_{n_i}(a)-\mu(a)|<\frac{\Delta_1}{4cs}\,,\,\forall a\in A\right\} &\subseteq \left\{\left|\Delta^{\widehat{\mu}_{n_i}}_S-\Delta^\mu_S\right|<\frac{\Delta_1}{2c}\,,\,\forall S\in\mathcal{S}\right\}\\
&\subseteq \left\{\left|\sum_{S\in\mathcal{E}}\left(\Delta^{\widehat{\mu}_{n_i}}_S-\Delta^\mu_S\right)\right|<\frac{\Delta_1}{2}\,,\,\forall\mathcal{E}\in\mathbf{C}\right\}\\
&\subseteq \left\{\sum_{S\in\mathcal{E}}\Delta^{\widehat{\mu}_{n_i}}_S>\sum_{S\in\mathcal{E}^*}\Delta^{\widehat{\mu}_{n_i}}_S\,,\,\forall\mathcal{E}^*\in\Gamma_{Cover}\,(\mu)\,,\,\mathcal{E}\in\mathbf{C}\setminus\Gamma_{Cover}\,(\mu)\right\},
\end{aligned}$$

where we remember that for a cost vector $\nu\in\mathcal{B}$, $\Delta^\nu_S=\sum_{a\in S}\nu(a)-z^*_{Comb}(\nu)$. We note that as discussed in (A-7) in Step 1, by taking $\Delta_1\le\Delta^\mu_{min}$, we also ensure that $\{\mathcal{S}^*(\widehat{\mu}_{n_i})\subseteq\mathcal{S}^*(\mu)\}$.
Using Proposition 4, the union bound, and (A-10), we have that

$$\mathbb{P}\left\{\mathcal{E}_i \notin \Gamma_{Cover}(\mu)\,, T_{n_i}(a) \geq \gamma(i-1)\,, \forall\, a \in A\right\} \leq$$
$$\sum_{a\in A} \mathbb{P}\left\{|\widehat{\mu}_{n_i}(a) - \mu(a)| \geq \frac{\Delta_1}{4sc}, T_{n_i}(a) \geq \gamma(i-1)\right\} \leq 2\,|A| \exp\left\{-\frac{(\Delta_1)^2 \gamma(i-1)}{8s^2c^2\mathcal{L}^2}\right\}. \quad \text{(A-11)}$$

Using the above and (A-9) we obtain that

$$R_1^{Cover}(F,N) \leq \gamma\,(\ln N)^{1+\varepsilon}\, z^*_{Cover}(\mu) + C_2,$$

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 $\mathcal{D}(\mu)$. That is, $\Delta^{\mathcal{D}} := \min\left\{\Delta^{\mu}_{min}, \Delta^{\mathcal{D}}_2, \Delta^{\mathcal{D}}_3\right\}/(2s)$, where

$$\Delta^{\mathcal{D}}_2 := \min\left\{\min\left\{\Delta_S^{(\mu\wedge l)(D)} : S \notin \mathcal{S}^*\left((\mu\wedge l)(D)\right)\right\} : D \subseteq A\setminus H,\, \mathcal{S}^*(\mu) = \mathcal{S}^*\left((\mu\wedge l)(D)\right)\right\},$$
$$\Delta^{\mathcal{D}}_3 := \min\left\{z^*_{Comb}(\mu) - z^*_{Comb}\left((\mu\wedge l)(D)\right) : D \subseteq A\setminus H,\, \mathcal{S}^*(\mu) \neq \mathcal{S}^*\left((\mu\wedge l)(D)\right)\right\},$$

Δ^{μ}_{min} is as defined in Step 1, $H := \bigcup_{S^*\in\mathcal{S}^*(\mu)}\bigcup_{a\in S^*}\{a\}$, and $(\mu\wedge l)(D) = (\mu(a), a\in A\setminus D) \cup (l(a) : a \in D)$. The first threshold Δ^{μ}_{min} ensures that $\mathcal{S}^*(\widehat{\mu}_n) \subseteq \mathcal{S}^*(\mu)$. This is supported by Step 1 (see (A-7)). The second threshold $\Delta^{\mathcal{D}}_2$ ensures that

$$\mathcal{D}(\widehat{\mu}_n) \subseteq \mathcal{D}(\mu) \cup 2^H.$$

This follows from noting that: (i) for $D \notin \mathcal{D}(\mu)$,

$$\bigcap_{a\in A}\left\{|\widehat{\mu}_n(a) - \mu(a)| < \Delta^{\mathcal{D}}\right\} \subseteq \left\{z^*_{Comb}(\widehat{\mu}_n) = z^*_{Comb}\left((\widehat{\mu}_n\wedge l)(D)\right)\right\},$$

implying that $D \notin \mathcal{D}(\widehat{\mu}_n)$; and (ii) not all solutions in $\mathcal{S}^*(\mu)$ are necessarily optimal in the approximate problem (i.e., using the average costs), therefore, some of their ground elements might belong to $\mathcal{D}(\widehat{\mu}_n)$. The third threshold $\Delta^{\mathcal{D}}_3$ ensures that $\mathcal{D}(\mu) \subseteq \mathcal{D}(\widehat{\mu}_n)$. This follows from noting that for $D \in \mathcal{D}(\mu)$,

$$\bigcap_{a\in A}\left\{|\widehat{\mu}_n(a) - \mu(a)| < \Delta^{\mathcal{D}}\right\} \subseteq \left\{z^*_{Comb}(\widehat{\mu}_n) > z^*_{Comb}\left((\widehat{\mu}_n\wedge l)(D)\right)\right\},$$

implying that $D \in \mathcal{D}(\widehat{\mu}_n)$. We conclude that

$$\bigcap_{a\in A}\left\{|\widehat{\mu}_n(a) - \mu(a)| < \Delta^{\mathcal{D}}\right\} \subseteq \left\{\mathcal{D}(\widehat{\mu}_n) = \mathcal{D}(\mu) \cup H_o\right\},$$

for some $H_o \in 2^H$. While we assume, without loss of generality, that Δ^{μ}_{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 $K_D(\nu)$ is also continuous for all $D \in \mathcal{D}(\mu)$. In addition, because it is known that $K_D(\mu) \leq K$, there exists a finite constant $\kappa > 0$ such that

$$\left|\hat{K}_D(\widehat{\mu}_n) - K_D(\mu)\right| \leq \kappa \sum_{a\in A} \left|\widehat{\mu}_n(a) - \mu(a)\right|,$$

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

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

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

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

$$\bigcap_{a\in A} \left\{\left|\widehat{\mu}_n(a) - \mu(a)\right| < \Delta^K\right\} \subseteq \left\{\left|\hat{K}_D(\widehat{\mu}_n) - K_D(\mu)\right| < \varrho,\ D \in \mathcal{D}(\mu)\right\}.$$

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

$$\begin{aligned} &\max\{x^n(a) + \delta : a \in D\} \geq K_D(\mu),\ D \in \mathcal{D}(\mu) \\ &\max\{x^*(a) + \delta : a \in D\} \geq \hat{K}_D(\widehat{\mu}_n),\ D \in \mathcal{D}(\widehat{\mu}_n). \end{aligned} \tag{A-12}$$

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-12) we conclude that $(x^{*,\delta}, y^{*,\delta})$ is feasible to $LBP(\widehat{\mu}_n)$. Seeing that $\|\widehat{\mu}_n - \mu\|_\infty < \delta/(2s)$, we have $\left|\Delta^{\mu}_S - \Delta^{\widehat{\mu}_n}_S\right| < \delta$ for all $S \in \mathcal{S}$. Therefore, we have that

$$\begin{aligned} \sum_{S\in\mathcal{S}} y^n(S)\Delta^{\mu}_S &\overset{(a)}{\leq} \sum_{S\in\mathcal{S}} y^n(S)\Delta^{\widehat{\mu}_n}_S + |\mathcal{S}|\,K\delta \\ &\overset{(b)}{\leq} \sum_{S\in\mathcal{S}} y^{*,\delta}(S)\Delta^{\widehat{\mu}_n}_S + |\mathcal{S}|\,K\delta \end{aligned}$$

$$\begin{aligned}&\overset{(c)}{\leq} \sum_{S\in\mathcal{S}} y^*(S)\Delta^\mu_S + \delta\,|\mathcal{S}|\,(\delta+\Delta^\mu_{max}+K) + |\mathcal{S}|\,K\delta\\ &= z^*_{LBP}(\mu) + \delta\,|\mathcal{S}|\,(\delta+\Delta^\mu_{max}+2K)\,,\end{aligned}$$

where (a) follows from the fact that $y^n(S)\leq K$ for all $S\in\mathcal{S}$ (this because $\hat{K}_D(\widehat{\mu}_n)\leq K$), (b) comes from that $(x^{*,\delta},y^{*,\delta})$ is feasible to $LBP(\widehat{\mu}_n)$, and (c) follows from that $\left|\Delta^\mu_S-\Delta^{\widehat{\mu}_n}_S\right|<\delta$ and $y^{*,\delta}(S)\leq y^*(S)+\delta$ for all $S\in\mathcal{S}$, and $y^*(S)\leq K$ for all $S\in\mathcal{S}$. Seeing that $\delta<\Delta^{\mathcal{D}}<\Delta^\mu_{min}$, taking $\delta\leq \varrho\, z^*_{LBP}(\mu)/\left(|\mathcal{S}|\,(\Delta^\mu_{min}+\Delta^\mu_{max}+2K)\right)$, we have that

$$\sum_{S\in\mathcal{S}} y^n(S)\Delta^\mu_S \leq (1+\varrho) z^*_{LBP}(\mu).$$

Consider $i>i'$ and let $(x_i,y_i)\in\Gamma_{LBP}(\widehat{\mu}_{n_i})$ 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 $\{S_n\in\Gamma_{LBP}\left(\widehat{\mu}_n\right)\}$ to denote the LBP-based exploration. We have that

$$\begin{aligned}R^{LBP}_1(F,N) &\leq n_{i'}\Delta^\mu_{max} + \sum_{i=i'}^{\lceil(\ln N)^{1+\varepsilon}\rceil}\sum_{n=n_i}^{n_{i+1}-1}\mathbb{E}\left\{\mathbf{1}\left\{T_{n_i}(a)\geq\gamma(i-1)\ \forall a\in A\,,\, S_n\in\Gamma_{LBP}\left(\widehat{\mu}_n\right)\right\}\Delta^\mu_{S_n}\right\}\\
&\leq n_{i'}\Delta^\mu_{max} + \sum_{i=i'}^{\lceil(\ln N)^{1+\varepsilon}\rceil}\sum_{n=n_i}^{n_{i+1}-1}\mathbb{E}\left\{\mathbf{1}\left\{T_{n_i}(a)\geq\gamma(i-1), |\widehat{\mu}_n(a)-\mu(a)|<\delta/(2s), \forall a\in A, S_n\in\Gamma_{LBP}\left(\widehat{\mu}_n\right)\right\}\Delta^\mu_{S_n}\right\}\\
&\quad + \sum_{i=i'}^{\lceil(\ln N)^{1+\varepsilon}\rceil}\sum_{n=n_i}^{n_{i+1}-1}\mathbb{E}\left\{\mathbf{1}\left\{T_{n_i}(a)\geq\gamma(i-1), \forall a\in A, \cup_{a\in A}\left\{|\widehat{\mu}_n(a)-\mu(a)|\geq\delta/(2s)\right\}, S_n\in\Gamma_{LBP}\left(\widehat{\mu}_n\right)\right\}\Delta^\mu_{S_n}\right\}\\
&\leq n_{i'}\Delta^\mu_{max} + \sum_{i=i'}^{\lceil(\ln N)^{1+\varepsilon}\rceil}(1+\varrho)\ z^*_{LBP}(\mu)\\
&\quad + \sum_{i=i'}^{\infty}(n_{i+1}-n_i)\,\Delta^\mu_{max}\sum_{a\in A}\mathbb{P}\left\{|\widehat{\mu}_n(a)-\mu(a)|\geq\delta/(2s), T_{n_i}(a)\geq\gamma(i-1)\right\}. \qquad \text{(A-13)}\end{aligned}$$

Using Proposition 4 to bound the probability in (A-13), we have that

$$R^{LBP}_1(F,N)\leq n_{i'}\Delta^\mu_{max} + (\ln N)^{1+\varepsilon}\,(1+\varrho)\ z^*_{LBP}(\mu) + \sum_{i=i'}^{\infty} e^{(i+1)^{1/(\varepsilon+1)}}\Delta^\mu_{max}2\,|A|\exp\left\{-\frac{\delta^2\ \gamma(i-1)}{2s^2\mathcal{L}^2}\right\}.$$

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

$$R^{LBP}_1(F,N)\leq (\ln N)^{1+\varepsilon}\,(1+\varrho)\ z^*_{LBP}(\mu)+C_3$$

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

$$R^{\pi^*}(F,N) \leq \left((1+\varrho)\, z^*_{LBP}(\mu) + \gamma\, z^*_{Cover}(\mu)\right)(\ln N)^{1+\varepsilon} + C_4,$$

for a finite positive constant C_4, independent of N.

We finally note that the optimal solutions to the $Cover(\widehat{\mu}_{n_i})$ and $LBP(\widehat{\mu}_{n_i})$ problems converge a.s. to optimal and ϱ-optimal solutions to $Cover(\mu)$ and $LBP(\mu)$, respectively. For this, note that Proposition 4, (A-11) and (A-13) imply (via Borel-Cantelli) that $\mathbb{P}\{\mathcal{E}_i \in \Gamma_{Cover}(\mu) \text{ eventually}\} = 1$ and

$$\mathbb{P}\{(x^\varrho_i, y^\varrho_i) \text{ is a } \varrho\text{-optimal solution to } LBP(\mu) \text{ eventually}\} = 1.$$

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

A.1.4. 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 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. For a vector $r := (r(a) : a \in A)$, we define

$$\underline{z}(r) := \min_{y' \in \mathbb{R}_+^{|\mathcal{S}|}} \left\{ \sum_{S \in \mathcal{S}} \Delta_S^{(\nu \wedge l)(\{a \in A:\, r(a)=0\})}\, y'(S) : r(a) \leq \sum_{S \in \mathcal{S}: a \in S} y'(S),\, a \in A \right\},$$

where we recall that for a set D, $(\nu \wedge l)(D) = (\nu(a)\mathbf{1}\{a \notin D\} + l(a)\mathbf{1}\{a \in D\} : a \in A)$. The variable $\underline{z}(r)$ computes the minimum cost attainable if one were to change the mean cost of an unexplored ground element. The following adjoint formulation imposes that the optimal cost is not greater than such an alternative minimum cost.

$$\begin{aligned}
z^*_R(\nu,\gamma) := \min\ & \sum_{S \in \mathcal{S}} \Delta_S^\nu\ y(S) \\
s.t.\quad & \sum_{S \in \mathcal{S}} \Delta_S^\nu\ y(S) \leq \underline{z}(r) \\
& r(a) \leq \sum_{S \in \mathcal{S}: a \in S} y(S),\, a \in A \\
& r(a) = x(a) + \gamma\, w(a),\, a \in A \\
& \max\{x(a) : a \in D\} \geq K_D(\nu), \quad D \in \mathcal{D}(\nu) \\
& x(a) = 1, \quad \forall\, a \in S, \forall\, S \in \mathcal{S}^*(\nu) \\
& w(a) \in \{0,1\},\, x(a),\, r(a), y(S) \in \mathbb{R}_+, \quad a \in A,\, S \in \mathcal{S}.
\end{aligned}$$

A.2. Omitted Proofs and Material from Section 5

A.2.1. Equivalence of LBP and OCP

LEMMA 1. *An optimal solution to a linear relaxation of $OCP(\mu)$ when one relaxes the integrality constraints over $y(S)$ variables is also optimal to formulation $LBP(\mu)$ when one replaces $K_D(\mu)$ by 1 for all $D \in \mathcal{D}(\mu)$.*

Proof of Lemma 1. Let *R-OCP*(μ) denote the linear relaxation of $OCP(\mu)$ where the integrality constraints over $y(S)$ variables are replaced by those of non-negativity. We prove Lemma 1 by showing that a feasible solution to *R-OCP*(μ) is also feasible to $LBP\,(\mu)$ and vice versa. We prove each feasibility result by contradiction.

We first note that when $K_D(\mu) = 1$ for all $D \in \mathcal{D}(\mu)$, one can restrict attention only to feasible solutions to $LBP\,(\mu)$ with binary x. Let (x, y) be a feasible solution to *R-OCP*(μ) and suppose that (x, y) is not feasible to $LBP\,(\mu)$, i.e., there exists a $D \in \mathcal{D}(\mu)$ such that $\max\{x(a) : a \in D\} = 0$ which implies that $x(a) = 0$ for all $a \in D$. Thus, for $S^* \in \mathcal{S}^*((\mu \wedge l)\,(D))$, we have that

$$\begin{aligned} z^*_{Comb}((\mu \wedge l)\,(D)) &= \sum_{a \in S^* \setminus D} \mu(a) + \sum_{a \in D} l(a) \\ &\overset{(a)}{\geq} \sum_{a \in S^*} \left(l(a)(1 - x(a)) + \mu(a)\, x(a)\right) \\ &\overset{(b)}{\geq} z^*_{Comb}(\mu), \end{aligned}$$

where (a) follows from the fact that $l(a) = (l(a)(1 - x(a)) + \mu(a)\, x(a))$ as $x(a) = 0$ for $a \in D$, and $\mu(a) \geq (l(a)(1 - x(a)) + \mu(a)\, x(a))$ for $a \notin D$, and (b) follows from the fact that (x, y) satisfies constraints (8c) (because it is feasible to *R-OCP*(μ)). However, by the definition of $\mathcal{D}(\mu)$, one has that $z^*_{Comb}((\mu \wedge l)\,(D)) < z^*_{Comb}(\mu)$, which is contradicted by the last inequality above, thus we have that $\max\{x(a) : a \in D\} = 1$ for all $D \in \mathcal{D}(\mu)$, therefore (x, y) is feasible to $LBP\,(\mu)$.

Now, let (x, y) be a feasible solution to $LBP\,(\mu)$ such that $x(a) \in \{0, 1\}$ for all $a \in A$, and that $x(a) = 1$ and $y(S^*) = 1$ for all $a \in S^*$ and $S^* \in \mathcal{S}^*(\mu)$ (because $\Delta^{\mu}_{S^*} = 0$ for all $S^* \in \mathcal{S}^*(\mu)$, this extra requirement does not affect the optimal solution to $LBP\,(\mu)$). Suppose (x, y) is not feasible to *R-OCP*(μ), i.e., there exists some $S \in \mathcal{S}$ such that

$$\sum_{a \in S} \left(l(a)(1 - x(a)) + \mu(a)\, x(a)\right) < z^*_{Comb}(\mu). \tag{A-15}$$

Let S_0 be one such S that additionally minimizes the left-hand side in (A-15) (in case of ties we pick any minimizing solution S_0 with smallest value of $|\{a \in S_0 : x(a) = 0\}|$). Then $D := \{a \in S_0 : x(a) = 0\}$ (or a subset of D) belongs to $\mathcal{D}(\mu)$. This contradicts the feasibility of (x, y) to

$LBP(\mu)$, because if (x,y) is feasible to $LBP(\mu)$, then we must have $\max\{x(a): a \in D\} \geq 1$ for all $D \in \mathcal{D}(\mu)$. Thus, we conclude that (x,y) is feasible to $R\text{-}OCP(\mu)$.

Summarizing, when $K_D(\mu) = 1$ for all $D \in \mathcal{D}(\mu)$, feasible solutions to $R\text{-}OCP(\mu)$ are feasible to $LBP(\mu)$, and feasible solutions to $LBP(\mu)$ that cover all optimal elements in A are feasible to $R\text{-}OCP(\mu)$. The result follows from noting that there always exists an optimal solution to $LBP(\mu)$ such that x is binary, and $x(a) = 1$ and $y(S^*) = 1$ for all $a \in S^*$ and $S^* \in \mathcal{S}^*(\mu)$.

A.2.2. Modified OCP-Based Policy

The modified OCP-based policy $\pi'_{OCP}(\gamma, \varepsilon, \varrho)$ is detailed in Algorithm 3. This policy closely follows the structure of the LBP-based policy in Algorithm 1, but solves the OCP problem instead of LBP. As in Algorithm 1, we define the cycles as $n_1 = 1$ and $n_i := \max\{\lfloor e^{i^{1/(1+\varepsilon)}} \rfloor, n_{i-1}+1\}$ for all $i \geq 2$, given a tuning parameter $\varepsilon > 0$. Moreover, as in Algorithm 1, γ is a tuning parameter that controls the cover-based exploration frequency. Finally, the parameter ϱ in Algorithm 3 allows the policy to converge to an optimal solution to $OCP(\mu)$ – because there might exist multiple optimal solutions, the "Update OCP-exploration set" step ensures that the policy settles on one of them.

Algorithm 3 Modified OCP-based policy $\pi'_{OCP}(\gamma, \varepsilon, \varrho)$

Set $i = 0$, $C = A$, $\mathcal{E}$ a minimal cover of A, $\mathcal{G} = \mathcal{E}$, and draw $(b_1(a): a \in A)$ randomly from $\mathcal{B}$
for $n = 1$ to N **do**
  if $n = n_i$ **then**
    Set $i = i + 1$
    Set $S^* \in \mathcal{S}^*(\widehat{\mu}_n)$ [Update exploitation set]
    Set $\mathcal{E} \in \Gamma_{Cover}(\widehat{\mu}_n)$ [Update Cover-exploration set]
    if $(C, \mathcal{G})$ is not a ϱ-optimal solution to $OCP(\widehat{\mu}_n)$ **then**
      Set $(C, \mathcal{G}) \in \Gamma_{OCP}(\widehat{\mu}_n)$ [Update OCP-exploration set]
    end if
  end if
  if $T_n(a) < \gamma\, i$ for some $a \in A$ **then**
    Set $S_n = S$ for any $S \in \mathcal{E}$ such that $a \in S$ [Cover-based exploration]
  else if $\gamma < 1$ and $T_n(a) < i$ for some $a \in C$ **then**
    Set $S_n = S$ for any $S \in \mathcal{G}$ such that $a \in S$ [OCP-based exploration]
  else
    Set $S_n = S^*$ [Exploitation]
  end if
end for

Next, under Assumption 2, we prove a performance bound for the modified OCP-based policy.

THEOREM 3. *Consider $\gamma \in (0,1)$, $\varrho > 0$, and $\varepsilon > 0$ arbitrary. We then have that for ϱ sufficiently small*

$$\lim_{N\to\infty} \frac{R^{\pi'_{OCP}(\gamma,\varepsilon,\varrho)}(F,N)}{(\ln N)^{1+\varepsilon}} \leq z^*_{OCP}(\mu) + \gamma\, z^*_{Cover}(\mu).$$

Proof of Theorem 3. As in the case of the LBP-based policy π^*, the regret of policy π'_{OCP} (we again ignore the dependence of the policy on γ, ε, and ϱ to simplify the notation) stems from three sources: Cover-based and OCP-based exploration efforts, and exploitation errors. That is,

$$R^{\pi'_{OCP}}(F,N) = R_1^{Cover}(F,N) + R_1^{OCP}(F,N) + R_2(F,N), \tag{A-16}$$

where $R_1^{Cover}(F,N)$ is the Cover-based exploration regret, i.e., that incurred at period n during cycle i if $T_n(a) < \gamma\, i$ for some $a \in A$, $R_1^{OCP}(F,N)$ is the OCP-based exploration regret, i.e., that incurred at period n during cycle i if $T_n(a) < i$ 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 $S_n = S^*$ with $S^* \in \mathcal{S}^*(\widehat{\mu}_n)$.

We prove the result by bounding each term in (A-16) separately. It turns out that the bounds for $R_1^{Cover}(F,N)$ and $R_2(F,N)$ in Step 1 and Step 2.1 in the proof of Theorem 2 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-based 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'_{OCP} is implemented.

Step 2.2' (OCP-based exploration regret).

Following the arguments in Step 2.2 of the proof of Theorem 2, we first define the minimum precision threshold on the accuracy of mean cost estimates necessary to reconstruct the solution to $OCP(\mu)$. For that, we define $\Delta^{\mathcal{D}} := \min\{\Delta^{\mu}_{min}, \Delta^{\mathcal{D}}_2, \Delta^{\mathcal{D}}_3, \Delta^{\mathcal{D}}_4\}/(8sc)$, where

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

and $\mathbf{G}$ denotes the set of all feasible solutions $(C,\mathcal{G})$ to $OCP(\mu)$ problem. Note that as in the proof of Theorem 2, $\Delta^{\mu}_{min} = \min\{\Delta^{\mu}_S : S \in \mathcal{S}\setminus\mathcal{S}^*(\mu)\}$, $s = \max\{|S| : S \in \mathcal{S}\}$, $c = \max\{|\mathcal{E}| : \mathcal{E} \in \mathbf{C}\}$, i.e., the maximum size of a minimal cover of A, and $H = \bigcup_{S^*\in\mathcal{S}^*(\mu)}\bigcup_{a\in S^*}\{a\}$. Also note that $\Delta^{\mathcal{D}}_4$

denotes the minimum optimality gap of problem $OCP(\mu)$. Note that thresholds Δ^{μ}_{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 us to reconstruct the feasible set $\mathbf{G}$. Consider (x, y) satisfying (8b) and (8d). We first note that as discussed in Step 1 of the proof of Theorem 2, $\{\|\widehat{\mu}_n - \mu\|_\infty < \Delta^{\mu}_{min}/(2s)\}$ ensures that $\{\mathcal{S}^*(\widehat{\mu}_n) \subseteq \mathcal{S}^*(\mu)\}$. One then has that

$$\begin{aligned}\left\{\|\widehat{\mu}_n - \mu\|_\infty < \Delta^{\mathcal{D}}\right\} \subseteq & \left\{\left|\sum_{a\in S} x(a)\left(\widehat{\mu}_n(a) - \mu(a)\right)\right| < \Delta^{\mathcal{D}} s, \forall S \in \mathcal{S}\right\} \\ & \cap \left\{|z^*_{Comb}(\widehat{\mu}_n) - z^*_{Comb}(\mu)| < \Delta^{\mathcal{D}} s\right\} \\ \subseteq \left\{\left|\left(\sum_{a\in S}(x(a)\widehat{\mu}_n(a) + (1-x(a))l(a)) - z^*_{Comb}(\widehat{\mu}_n)\right) - \left(\sum_{a\in S}(x(a)\mu(a) + (1-x(a))l(a)) - z^*_{Comb}(\mu)\right)\right| < 2\Delta^{\mathcal{D}} s, \forall S \in \mathcal{S}\right\}&.\end{aligned}$$

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

$$\sum_{a\in S}(x(a)\mu(a) + (1-x(a))l(a)) \geq z^*_{Comb}(\mu) \text{ iff } \sum_{a\in S}(x(a)\widehat{\mu}_n(a) + (1-x(a))l(a)) \geq z^*_{Comb}(\widehat{\mu}_n).$$

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

$$\begin{aligned}\left\{\|\widehat{\mu}_n - \mu\|_\infty < \Delta^{\mathcal{D}}\right\} &\subseteq \left\{\left|\Delta^{\widehat{\mu}_n}_S - \Delta^{\mu}_S\right| < \frac{\Delta^{\mathcal{D}}_4}{4c}, \forall S \in \mathcal{S}\right\} \\ &\subseteq \left\{\left|\sum_{S\in\mathcal{G}}\left(\Delta^{\widehat{\mu}_n}_S - \Delta^{\mu}_S\right)\right| < \frac{\Delta^{\mathcal{D}}_4}{4}, \forall (C,\mathcal{G}) \in \mathbf{G}\right\} \\ \subseteq \left\{\sum_{S\in\mathcal{G}} \Delta^{\widehat{\mu}_n}_S > \Delta^{\mathcal{D}}_4/2 + \sum_{S\in\mathcal{G}^*} \Delta^{\widehat{\mu}_n}_S, \forall (C^*,\mathcal{G}^*) \in \Gamma_{OCP}(\mu), (C,\mathcal{G}) \in \mathbf{G}\setminus\Gamma_{OCP}(\mu)\right\}&.\end{aligned}$$

The above not only implies that $\Gamma_{OCP}(\widehat{\mu}_n) \subseteq \Gamma_{OCP}(\mu)$, but also that ϱ-optimal solutions to $OCP(\widehat{\mu}_n)$ are also optimal to $OCP(\mu)$, as long as $\varrho < \Delta^{\mathcal{D}}_4/2$. Letting $\Gamma^{\varrho}_{OCP}(\nu)$ denote the set of ϱ-optimal solutions to $OCP(\nu)$, the above implies that for $\varrho < \Delta^{\mathcal{D}}_4/2$,

$$\left\{\|\widehat{\mu}_n - \mu\|_\infty < \Delta^{\mathcal{D}}\right\} \subseteq \left\{\Gamma^{\varrho}_{OCP}(\widehat{\mu}_n) \subseteq \Gamma_{OCP}(\mu)\right\}.$$

We are now ready to provide a bound on $R^{OCP}_1(F, N)$. Similar to the proof of Theorem 2, 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 period $n \in [n_i, n_{i+1}-1]$. Define the events $\Xi_i^1 := \{(C_i, \mathcal{G}_i) \in \Gamma_{OCP}(\mu)\}$ and $\Xi_i^2 := \{\mathcal{G}_i = \mathcal{G}_{i-1}\}$. We then have that

$$\begin{aligned} R_1^{OCP}(F,N) &\le n_{i'+1}\Delta^{\mu}_{max} + \sum_{i=i'+1}^{\lceil (\ln N)^{1+\varepsilon}\rceil} \mathbb{E}\left\{ \mathbf{1}\left\{T_{n_{i-1}}(a) \ge \gamma(i-2)\,,\, \forall a \in A\,,\, \left(\Xi_i^1 \cap \Xi_i^2\right)\right\} \sum_{S\in\mathcal{G}_i} \Delta_S^{\mu} \right\} \\ &\quad + \sum_{i=i'+1}^{\lceil (\ln N)^{1+\varepsilon}\rceil} i\ \mathbb{E}\left\{ \mathbf{1}\left\{T_{n_{i-1}}(a) \ge \gamma(i-2)\,,\, \forall a \in A\,,\, \left(\Xi_i^1 \cap \Xi_i^2\right)^c\right\} \sum_{S\in\mathcal{G}_i} \Delta_S^{\mu} \right\} \\ &\le n_{i'+1}\Delta^{\mu}_{max} + \left((\ln N)^{1+\varepsilon} + 1\right) z^*_{OCP}(\mu) \\ &\quad + \Delta^{\mu}_{max}\, c \sum_{i=i'+1}^{\infty} i\ \mathbb{P}\left\{T_{n_{i-1}}(a) \ge \gamma(i-2)\,,\, \forall a \in A\,,\, \left(\Xi_i^1 \cap \Xi_i^2\right)^c\right\}, \end{aligned} \tag{A-17}$$

where $(\Xi_i^1 \cap \Xi_i^2)^c$ denotes the complement of the event $(\Xi_i^1 \cap \Xi_i^2)$. Next, we bound the probability inside the sum in (A-17). For that, observe that

$$\begin{aligned} \left\{\|\widehat{\mu}_{n_{i-1}} - \mu\|_\infty \vee \|\widehat{\mu}_{n_i} - \mu\|_\infty < \Delta^{\mathcal{D}}\right\} &\subseteq \left\{\Gamma^{\varrho}_{OCP}(\widehat{\mu}_{n_{i-1}}) \subseteq \Gamma_{OCP}(\mu)\right\} \\ &\quad \cap\ \left\{\Gamma_{OCP}(\mu) \subseteq \Gamma^{\varrho}_{OCP}(\widehat{\mu}_{n_i})\right\} \\ &\subseteq \left(\Xi_i^1 \cap \Xi_i^2\right). \end{aligned}$$

Using above and Proposition 4, we conclude that

$$\mathbb{P}\left\{T_{n_{i-1}}(a) \ge \gamma(i-2)\,,\, \forall a \in A\,,\, \left(\Xi_i^1 \cap \Xi_i^2\right)^c\right\} \le 4\,|A| \exp\left\{-\frac{2(\Delta^{\mathcal{D}})^2\ \gamma(i-2)}{\mathcal{L}^2}\right\}. \tag{A-18}$$

Using the above and (A-17), we have that

$$R_1^{OCP}(F,N) \le (\ln N)^{1+\varepsilon}\, z^*_{OCP}(\mu) + C_5,$$

for a finite positive constant C_5, independent of N. Putting the results from Steps 1 and 2.1 (from the proof of Theorem 2), and Step 2.2' together, we conclude that

$$R^{\pi'_{OCP}}(F,N) \le \left(z^*_{OCP}(\mu) + \gamma\, z^*_{Cover}(\mu)\right)(\ln N)^{1+\varepsilon} + C_6,$$

for a finite positive constant C_6, independent of N.

We finally note that the optimal solutions to the $Cover(\widehat{\mu}_{n_i})$ and $OCP(\widehat{\mu}_{n_i})$ problems converge a.s. to optimal solutions to $Cover(\mu)$ and $OCP(\mu)$, respectively. For this, note that Proposition 4,

(A-11) and (A-18) imply (via Borel-Cantelli) that $\mathbb{P}\{\mathcal{E}_i \in \Gamma_{Cover}(\mu) \text{ eventually}\} = 1$ and

$$\mathbb{P}\{(C_i, \mathcal{G}_i) \in \Gamma_{OCP}(\mu) \text{ eventually}\} = 1.$$

A.3. Appendix for Section 6

A.3.1. General Complexity of OCP

To prove Theorem 4 and Proposition 2, we will use the following lemma.

LEMMA 3. *We may restrict the OCP or Cover problems to have at most $|A|$ non-zero $y(S)$ variables without changing the problems.*

Proof of Lemma 3. For the OCP problem, the result follows from noting that any critical set C can be covered by at most $|A|$ solutions (i.e., by a solution-cover $\mathcal{G}$ of at most size $|A|$). Hence, if an optimal solution for OCP has $|\mathcal{G}| > |A|$, we may remove one solution from it while preserving feasibility. If the removed solution is suboptimal for $Comb$, we would obtain a solution with lower objective value contradicting the optimality for OCP. If the removed solution is optimal for $Comb$, we obtain an alternate optimal solution for OCP.

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

□

THEOREM 4. *If Comb is in P, then OCP is in NP.*

Proof of Theorem 4. By Lemma 3, 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 (8c) can be achieved by solving $Comb(\nu_x)$ where $\nu_x := (\nu_x(a) : a \in A)$ for $\nu_x(a) := l(a)(1 - x(a)) + \nu(a)x(a)$. This problem is polynomially solvable by assumption.

□

Note that the proof of Theorem 4 also shows that if $Comb$ is in P, then $Cover$ is in NP.

A.3.2. Critical Sets for Matroids

LEMMA 4. *Let $Comb(\nu)$ be a weighted basis or independent set matroid minimization problem. Then there exists a unique critical set that can be found in polynomial time.*

Proof of Lemma 4. To simplify the exposition, we assume that $\mathcal{S}^*(\nu) = \{S^*\}$ is a singleton. Also, for $S \in \mathcal{S}$, we let e^S denote the incidence vector associated with S (i.e., $e^S := (e^S(a) : a \in A)$ with $e^S(a) \in \{0, 1\}$, $a \in A$, such that $e^S(a) = 1$ if $a \in S$ and $e^S(a) = 0$ otherwise).

Let $P := \text{conv}\left(\{e^S\}_{S\in\mathcal{S}}\right) \subseteq \mathbb{R}^{|A|}$ be the independent set (base) polytope of $\mathcal{S}$. Then, for a feasible cost vector ν, we have that $S^* \in \mathcal{S}^*(\nu)$ if and only if $\sum_{a\in S^*}\nu(a) \leq \sum_{a\in S}\nu(a)$ 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 (denoted by "R"), adding (denoted by "A"), or exchanging (denoted by "E") a single element of S^* (Schrijver 2003, Theorem 40.6). Thus, we construct the critical 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

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

By construction, covering all elements in C guarantees optimality of S^*, and not covering some guarantees that 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. Finally, the only computationally non-trivial step in the construction of C is checking that this set is an independent set or a base, which can be done in polynomial time.

A.3.3. Basic MIP Formulation for OCP

Proposition 2. *Let $y^S \in \{0,1\}^{|A|}$ 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 $OCP(\nu)$ is given by*

$$\min \quad \sum_{i\in A}\left(\sum_{a\in A}\nu(a)y^i(a) - z^*_{Comb}(\nu)\right) \tag{9a}$$

$$s.t. \quad x(a) \leq \sum_{i\in A} y^i(a), \qquad a \in A \tag{9b}$$

$$My^i \leq d, \qquad i \in A \tag{9c}$$

$$M^T w \leq \text{diag}(l)(\mathbf{1} - x) + \text{diag}(\nu)x \tag{9d}$$

$$d^T w \geq z^*_{Comb}(\nu) \tag{9e}$$

$$x(a),\, y^i(a) \in \{0,1\},\, w \in \mathbb{R}^m, \qquad a,\, i \in A, \tag{9f}$$

where $x = (x(a) : a \in A)$, $y^i = (y^i(a) : a \in A)$, *and* $\mathbf{1}$ *is a vector of ones.*

Proof of Proposition 2. For any feasible solution (x, y) to (9), we have that x is the incidence vector of a critical set. This, because (9d) enforces dual feasibility of w when elements with $x(a) = 0$ are not covered, and (9e) forces the objective value of the dual of $Comb(\nu')$ to be greater than or equal to $z^*_{Comb}(\nu)$, where $\nu' = \text{diag}(l)(\mathbf{1} - x) + \text{diag}(\nu)x$. With this, the optimal objective value of $Comb(\nu')$ is greater than or equal to $z^*_{Comb}(\nu)$. On the other hand, any y^i is the incidence vector of some $S \in \mathcal{S}$ because of (9c) and the assumptions on M and d. Finally, (9b) ensures that the critical set is covered by the solution-cover (i.e., y^i's) induced by OCP. Lemma 3 ensures that the $|A|$ variables y^i are sufficient for an optimal solution to OCP. If less than $|A|$ solutions are needed for the cover, then the optimization problem can pick the additional y^i variables to be the incidence vector of an optimal solution to $Comb(\nu)$ so that they do not increase the objective function value. □

We note that Proposition 2 can be easily extended to obtain a formulation for $Cover(B)$ by setting $x_a = 1$ for all $a \in A$ and removing (9d)–(9e).

A.3.4. IP Formulation for OCP when $Comb(\nu)$ Admits a Compact IP Formulation

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

$$\min \quad \sum_{i\in A}\left(\sum_{a\in A}\nu(a)y^i(a) - z^*_{Comb}(\nu)\right) \tag{A-19a}$$

$$s.t. \qquad x(a) \le \sum_{i\in A} y^i(a), \quad a \in A \tag{A-19b}$$

$$My^i \le d, \qquad i \in A \tag{A-19c}$$

$$\sum_{a\in S}(l(a)(1 - x(a)) + \nu(a)x(a)) \ge z^*_{Comb}(\nu), \quad S \in \mathcal{S} \tag{A-19d}$$

$$x(a),\, y^i(a) \in \{0,1\}, \qquad a, i \in A. \tag{A-19e}$$

As in formulation (9), a feasible solution (x, y) to (A-19) is such that x is the incidence vector of a critical set (this is enforced by (A-19d)), and the y^i's are a cover of such set, due to (A-19b), (A-19c), and the assumptions on M and d. Note that an efficient cover includes at most $|A|$ solutions (the optimization can pick the additional y^i to be the incidence vector of an optimal solution).

Formulation (A-19) has a polynomial number of variables, but the number of constraints described by (A-19d) is in general exponential. However, the computational burden of separating these constraints is the same as solving $Comb(\nu)$ (finding a violated inequality (A-19d) or showing

that it satisfies all these inequalities can be done by solving $Comb(\nu')$ for $\nu'(a) = l(a)(1-x(a)) + \nu(a)x(a))$. Hence, if we can solve $Comb(\nu)$ 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-19) with a branch-and-cut algorithm that dynamically adds constraints (A-19d) as needed. Finally, note that a formulation for $Cover$ is obtained by setting $x(a) = 1$ for all $a \in A$ and removing (A-19d).

A.3.5. Linear-sized Formulation for OCP for the Shortest Path Problem

Let $Comb(\nu)$ 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

$$\min \quad \left(\sum_{a\in A} \nu(a)\, p(a)\right) - z^*_{Comb}(\nu)\, p\left((t,s)\right) \tag{A-20a}$$

$$s.t. \quad x(a) \le p(a), \quad a \in A \tag{A-20b}$$

$$\sum_{a\in\hat{\delta}_{out}(v)} p(a) - \sum_{a\in\hat{\delta}_{in}(v)} p(a) = 0, \quad v \in V \tag{A-20c}$$

$$w(u) - w(v) \le l\left((u,v)\right)\left(1 - x\left((u,v)\right)\right) + \nu\left((u,v)\right) x\left((u,v)\right), \quad (u,v) \in A \tag{A-20d}$$

$$w(s) - w(t) \ge z^*_{Comb}(\nu) \tag{A-20e}$$

$$p(a) \in \mathbb{Z}_+, \quad a \in \hat{A} \tag{A-20f}$$

$$x(a) \in \{0,1\}, w(v) \in \mathbb{R}, \quad a \in A, v \in V, \tag{A-20g}$$

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

The first difference between formulations (A-20) and (9) is the specialization of the LP duality constraints to the shortest path setting. The second one is the fact that the paths in cover $\mathcal{G}$ 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 $p = \sum_{S\in\mathcal{G}} y^{\hat{S}}$, where $y^{\hat{S}}$ is the incidence vector of the circulation obtained by adding (t,s) to each path S. Furthermore, given a feasible p we can recover the paths in $\mathcal{G}$ 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-20d)–(A-20e).

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

A.3.6. 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 the cost vector B_n. However, the solution times for the problems $OCP(\cdot)$ or even $Comb(\cdot)$ could be longer than the time available to the executing policy. For example, most index-based policies must solve an instance of $Comb(\cdot)$ at each period, which might not be possible in practice. Fortunately, a key feature of our proposed OCP-based policies is that the frequency at which the problems $Comb(\cdot)$ and $OCP(\cdot)$ need to be solved decreases exponentially over time. Indeed, such problems are solved at the beginning of each cycle and the length of cycle i is $\Theta\left(\exp\left(i/H\right)\right)$ for a fixed tuning parameter $H > 0$. Hence, as cycles elapse, there will be eventually enough time to solve these problems.

Nonetheless, the policy cannot proceed until the problems $Comb(\cdot)$ and $OCP(\cdot)$ are solved. However, one can easily modify the policy so that it begins solving $Comb(\cdot)$ and $OCP(\cdot)$ at the beginning of a cycle, but continues to implement incumbent solutions while these problems are being solved (such solutions might be computed either upfront or in previous cycles). Solutions 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 presents one such possible modification for the OCP-based policy.

A.3.7. Greedy Oracle Polynomial-Time Heuristic

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

We begin by describing the heuristic for solving $OCP(\nu)$ in Algorithm 5. Given a cost vector ν, the heuristic first sets all costs to their lowest possible values, and successively solves instances of $Comb$, each time incorporating the incumbent solution into the solution-cover $\mathcal{G}$, adding its ground elements to the (critical) set 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 $z^*_{Comb}(\nu') \geq z^*_{Comb}(\nu)$). 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 5 calls an oracle for $Comb$ 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 $Comb$. In particular, if $Comb$ is in P, then the heuristic runs in *polynomial time*.

Algorithm 4 Basic Time-Constrained Asynchronous OCP-based policy $\pi^A_{OCP}(H)$

Set $i=0$, $C=A$, and $\mathcal{G}$ a minimal cover of A
Let $S^* \in \mathcal{S}$ be an arbitrary solution and $\widehat{\mu}_{Comb} = \widehat{\mu}_{OCP}$ be an initial cost estimate
Asynchronously begin solving $Comb(\widehat{\mu}_{Comb})$ and $OCP(\widehat{\mu}_{OCP})$
for $n=1$ to N **do**
  if $n = n_i$ **then**
    Set $i = i+1$
    if Asynchronous solution to $Comb(\widehat{\mu}_{Comb})$ has finished **then**
      Set $S^* \in \mathcal{S}^*(\widehat{\mu}_{Comb})$ [Update exploitation set]
      Set $\widehat{\mu}_{Comb} = \widehat{\mu}_n$
      Asynchronously begin solving $Comb(\widehat{\mu}_{Comb})$
    end if
    if Asynchronous solution to $OCP(\widehat{\mu}_{OCP})$ has finished **then**
      Set $(C, \mathcal{G}) \in \Gamma_{OCP}(\widehat{\mu}_{OCP})$ [Update OCP-exploration set]
      Set $\widehat{\mu}_{OCP} = \widehat{\mu}_n$
      Asynchronously begin solving $OCP(\widehat{\mu}_{OCP})$
    end if
  end if
  if $T_n(a) < i$ for some $a \in C$ **then**
    Set $S_n = S$ for any $S \in \mathcal{G}$ such that $a \in S$ [OCP-based exploration]
  else
    Set $S_n = S^*$ [Exploitation]
  end if
end for

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 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.

We finally note that the heuristic in Algorithm 5 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.

Algorithm 5 Oracle Polynomial-Time Heuristic

Set $\nu' := (\nu'(a) : a \in A) = (l(a) : a \in A)$, $\mathcal{G} = \emptyset$, $C = \emptyset$.
while $z^*_{Comb}(\nu') < z^*_{Comb}(\nu)$ **do**
  Select $S \in \mathcal{S}^*(\nu')$ and set $\nu'(a) = \nu(a)$ 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 $z^*_{Comb}((\nu' \wedge l)(\{a\})) \geq z^*_{Comb}(\nu)$ **then**
    $C \leftarrow C \setminus \{a\}$ and $\nu'(a) \leftarrow l(a)$
  end if
end for

A.4. Additional Computational Results

In this section we provide the computational results for Examples 1, 2 and 3. Figure 7 depicts the average performance of different policies on Examples 1 (left), 2 (center) and 3 (right), respectively.

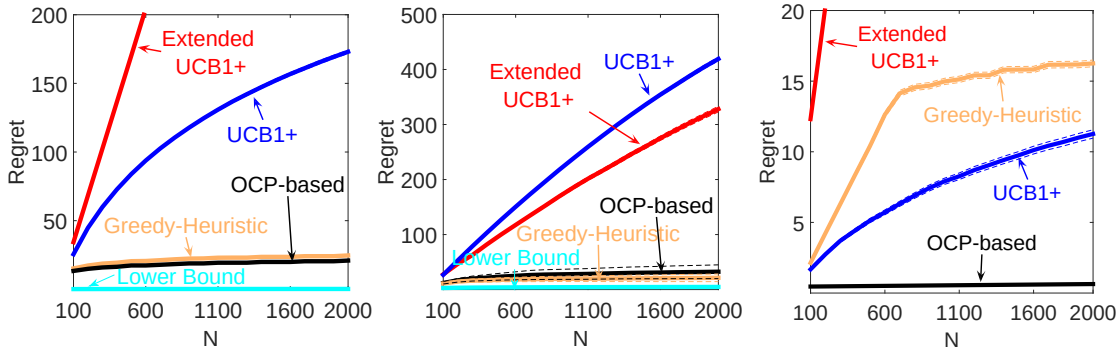

Figure 7 Average performance of different policies on Examples 1 (left), 2 (center) and 3 (right).

On Example 1, the OCP-based and Greedy-Heuristic policies perform significantly better than the benchmark policies. The situation is essentially the same on Example 2, only that this time Extended UCB1+ outperforms the UCB1+ policy. There, the solution to $OCP(\mu)$ is only of size 2, which helps our policies achieve the best performance. (Note that for this setting, the Greedy-Heuristic tends to find the actual optimal solution to $OCP(\mu)$ even with unreliable estimates.) On Example 3, the heuristic solution to OCP coincides with the minimum-regret cover of $\mathcal{S}$, thus the

Greedy-Heuristic policy is 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 discussed before, the lower bound in Theorem 1 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 7. 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-based and Greedy-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 (called the "Optimal arcs"), those that are covered by at least one optimal solution to $OCP(\mu)$ (called the "Exploration arcs"), and the rest (called the "*Uninformative* arcs"). Table 2 shows the average number of times that each type of arc (shown in columns called "Opt.", "Exp.", and "Uninf.", respectively) is tested up to period $N = 2000$ by each policy. Note that the OCP-based and Greedy-Heuristic policies spend significantly less time exploring uninformative arcs. Table 2 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").

	Example 1				Example 2				Example 3			
	Opt.	Exp.	Uninf.	Length	Opt.	Exp.	Unin.	Length	Opt.	Exp.	Unin.	Length
OCP-based	1958.93	470.67	2.25	3.06	1858.25	548.12	4.55	1.19	140.03	214.50	1.00	4.72
Greedy-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
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 2 Average number of trials of different arcs up to period $N = 2000$, and also average solution size for different policies on Examples 1, 2 and 3.

Figure 8 depicts box plots of the 100 different cumulative regrets at the final time period $N = 2000$ (i.e., sample path final regrets) for OCP-based, UCB1+ and Extended UCB1+ policies in Examples 1, 2 and 3. We observe that the OCP-based policy significantly outperforms UCB1+ and Extended UCB1+ not only in terms of average regret, but also for (almost) all sample path final regrets.

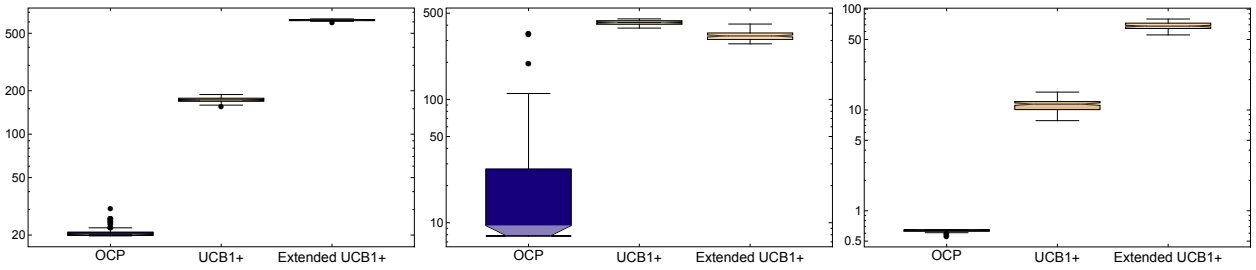

Figure 8 Box plots of sample path regrets for OCP-based and benchmark policies on Examples 1 (left), 2 (center) and 3 (right).

A.5. Short-Term Experiments

In this section we discuss the short-term experiments. In what follows, we first describe the benchmark policies and then discuss the studied settings and results.

A.5.1. 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 for each $a \in A$, $B(a)$ follows an exponential distribution with rate $\mu(a)$, but this rate itself is random, and initially distributed according to a Gamma distribution with parameters $\alpha_{a,0}$ and $\beta_{a,0}$. At period n, the posterior distribution of $\mu(a)$ is a Gamma with parameters

$$\alpha_{a,n} = \alpha_{a,0} + T_n(a), \quad \beta_{a,n} = \beta_{a,0} + \sum_{m<n: a \in S_m} b_m(a), \quad a \in A.$$

Thus at period 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_n. 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 indexes of the ground elements included in that solution. The policy requires a parametric representation of the uncertainty. To mimic a setting where the functional form of the cost distributions is unknown, we consider the approximation in Lai (1987) based on normally distributed costs 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 period 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 in Section 7.1. The average running time for a single replication ranged from around one second for the OCP-based policy to around 2 seconds for Gittins to less than 10 minutes for KG. We exclude the results for the UCB1+ and Extended UCB1+ policies, because they were consistently outperformed by the OCP-based policy.

A.5.2. 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-based policy begins outperforming the benchmarks. In all settings, the benchmark policies initially provide a better performance compared to the OCP-based policy, but the latter policy eventually surpasses the benchmarks for moderate values of N. The same holds true for the case of the Greedy-Heuristic policy.

Shortest Path Problem. The left panel of Figure 9 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-cover to $OCP(\mu)$ is of size 10 with an implied critical set of size 23.

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

Knapsack Problem. The right panel of Figure 9 depicts the average performances on a representative setting for the knapsack problem. (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-cover to $OCP(\mu)$ is of size 2 with an implied critical set of size 5.

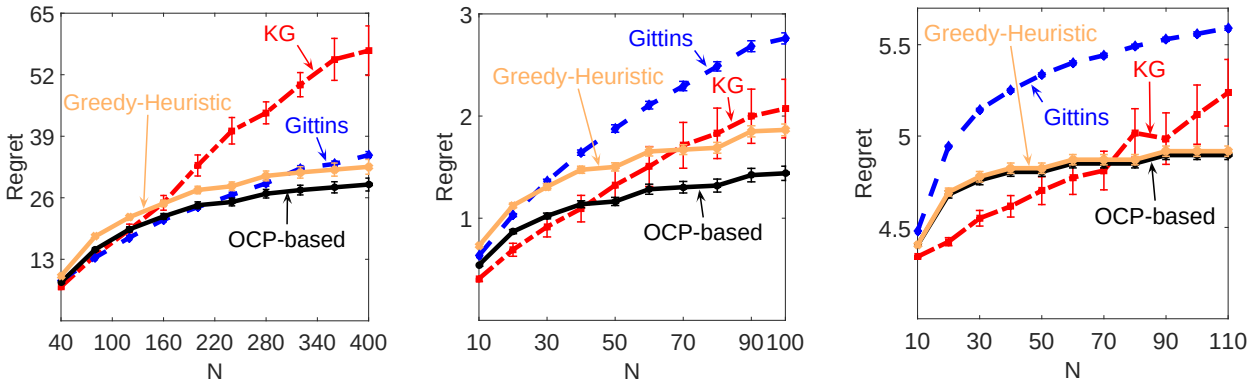

Figure 9 Average performance of different policies on the representative setting for the shortest path (left), Steiner tree (center) and knapsack (right) problems – the vertical lines show the 95% confidence intervals.

A.6. Alternative Feedback Setting

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 (i.e., a policy that conducts exploration by implementing solutions in a cover) can be adapted to this last setting. For a set of ground elements $S \subseteq A$, let $I_S := (I_S(a) : a \in A) \in \{0,1\}^{|A|}$ denote the incidence vector of the ground set (so that $S = \{a : I_S(a) = 1, a \in A\}$). We say that a

solution set $\mathcal{E}$ *recovers* a set $E \subseteq A$ if for each $a \in E$, there exists a vector $\gamma(a) := (\gamma_S(a), S \in \mathcal{E}) \in \mathbb{R}^{|\mathcal{E}|}$ such that

$$\sum_{S \in \mathcal{E}} \gamma_S(a) I_S = I_{\{a\}}. \tag{A-21}$$

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 using the estimate mean cost vector $\widehat{\mu}_n = (\widehat{\mu}_n(a) : a \in A)$, where

$$\widehat{\mu}_n(a) := \sum_{S \in \mathcal{E}} \frac{\gamma_S(a)}{|m < n : S_m = S|} \sum_{m < n : S_m = S} \sum_{a \in S} b_m(a), \quad a \in A. \tag{A-22}$$

The estimate above reconstructs the expected cost of each solution in $\mathcal{E}$ and uses (A-21) 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 $Comb(\cdot)$ (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 (A-22) and $(C, \mathcal{E})$ to be a solution to an alternative version of $OCP(\nu)$, denoted by $OCP'(\nu)$, where in addition to (8b)-(8d), one imposes that $\mathcal{E}$ recovers C, that is, $OCP'(\nu)$ is given by

$$\begin{aligned}
\min \quad & \sum_{S \in \mathcal{S}} \Delta^{\nu}_S \; y(S) && \text{(A-23a)}\\
s.t. \quad & \sum_{S \in \mathcal{S}} \gamma_S(a) I_S = x(a) I_{\{a\}}, \quad a \in A && \text{(A-23b)}\\
& \gamma_S(a) \leq Q \; y(S), \quad S \in \mathcal{S}, a \in A && \text{(A-23c)}\\
& -\gamma_S(a) \leq Q \; y(S), \quad S \in \mathcal{S}, a \in A && \text{(A-23d)}\\
& \sum_{a \in S} (l(a)(1 - x(a)) + b(a)x(a)) \geq z^*_{Comb}(\nu), \quad S \in \mathcal{S} && \text{(A-23e)}\\
& x(a), y(S) \in \{0, 1\}, \gamma_S(a) \in \mathbb{R}, \quad a \in A, S \in \mathcal{S}, && \text{(A-23f)}
\end{aligned}$$

where Q is an instance-dependent constant, whose size is polynomial in the size of the instance. The additional constraints(A-23b)-(A-23d) in $OCP'(\nu)$ ensure that the solution-cover $\mathcal{E}$ recovers the critical set C. Like OCP, the formulation above can be specialized to accommodate the combinatorial structure of $Comb$. The performance guarantee in Theorem 3 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.

A.7. Auxiliary Result for the Proof of Theorem 2 and Theorem 3

PROPOSITION 4. *For any fixed $a \in A$, $n \in \mathbb{N}$, $k \in \mathbb{N}$, and $\epsilon > 0$ we have that*

$$\mathbb{P}\left\{|\widehat{\mu}_n(a) - \mu(a)| \geq \epsilon, T_n(a) \geq k\right\} \leq 2\exp\left\{-\frac{2\epsilon^2 k}{\mathcal{L}^2}\right\},$$

where $\mathcal{L} := \max\{u(a) - l(a) : a \in A\}$.

Proof of Proposition 4. For $m \in \mathbb{N}$, define $t_m(a) := \inf\{n \in \mathbb{N} : T_n(a) = m\} - 1$. Indexed by m, one has that $B_{t_m(a)}(a) - \mu(a)$ is a bounded martingale difference sequence, thus one has that

$$\begin{aligned}
\mathbb{P}\left\{|\widehat{\mu}_n(a) - \mu(a)| \geq \epsilon\,, T_n(a) \geq k\right\} &= \mathbb{P}\left\{\left|\sum_{m=1}^{T_n(a)} \left(B_{t_m(a)}(a) - \mu(a)\right)\right| \geq \epsilon\ T_n(a)\,, T_n(a) \geq k\right\} \\
&\leq \sum_{h=k}^{\infty} \mathbb{P}\left\{\left|\sum_{m=1}^{h} \left(B_{t_m(a)}(a) - \mu(a)\right)\right| \geq \epsilon\ h\,, T_n(a) = h\right\} \\
&\overset{(a)}{\leq} 2\ \sum_{h=k}^{\infty} \exp\left\{\frac{-2\ h\epsilon^2}{\mathcal{L}^2}\right\} \mathbb{P}\left\{T_n(a) = h\right\} \\
&\leq 2\exp\left\{\frac{-2\ k\epsilon^2}{\mathcal{L}^2}\right\},
\end{aligned}$$

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