Learning in Combinatorial Optimization: What and How to Explore[*]

Sajad Modaresi
Duke University

Denis Sauré
University of Chile

Juan Pablo Vielma
MIT Sloan School

August 25, 2014

Abstract

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

1 Introduction

Motivation. Traditional solution approaches to many operational problems are based on combinatorial optimization problems, and typically involve instantiating a deterministic mathematical program, whose solution is implemented repeatedly through time: nevertheless, in practice, instances are not usually known in advance. When possible, parameters characterizing said instances are estimated *off-line*, either by using historical data or from direct observation of the (idle) system.

[*] This research was supported by the National Science Foundation [Grant CMMI-1233441]. Correspondence: sajad.modaresi@duke.edu, dsaure@dii.uchile.cl, jvielma@mit.edu.

Unfortunately, off-line estimation is not always possible as, for example, historical data (if available) might only provide partial information pertaining previously implemented solutions. Consider, for instance, shortest path problems in network applications: repeated implementation of a given path might reveal cost information about arcs on such a path, but might provide no further information about costs of other arcs in the graph. Similar settings arise, for example, in other network applications (e.g., tomography and connectivity) in which feedback about cost follows from instantiating and solving combinatorial problems such as spanning and Steiner trees.

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

Unfortunately, the features above either prevent or discourage the use of known bandit algorithms. First, in the combinatorial setting, traditional algorithms might not be implementable as they would typically require solving an instance of the underlying combinatorial problem between decision epochs, for which, depending on the application, there might not be enough computational resources. Second, even with enough computational resources, such algorithms would typically call for implementing each feasible solution at least once, which in the settings of interest might take a prohibitively large number of periods, and thus 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. Thus, our objective can be seen as interpreting said principles and adapting them to the combinatorial setting with the goal of *developing efficient policies that are amenable to implementation.* In doing so, we also aim at understanding how the specifics of the underlying combinatorial problem affect achievable performance. For this, we consider settings in which an agent must implement solutions to a series of arriving instances of a given combinatorial problem (i.e. whose feasible solutions are structured subsets of a *ground* set), and there is initial uncertainty about said instances. In particular, we assume that instance uncertainty is restricted to cost-coefficients in the objective function. Hence, the feasible region is the same for each instance and known upfront by the agent.

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

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

i) **We develop policies that significantly outperform existing and concurrently developed policies.** Let A be the set of ground elements of the underlying combinatorial optimization problem (e.g. set of arcs in the shortest path problem), s be the size of the largest feasible solution to the combinatorial optimization problem (e.g. length of the longest path) and N be the time horizon. The first policy we develop is based on a solution cover of the ground set A. The cumulative regret of this *cover-based* policy admits an upper bound of either order $|\mathcal{E}|\,(\ln N)^{1+\varepsilon}$, for arbitrarily small $\varepsilon > 0$, or $|\mathcal{E}|\,s^2 \ln N$, where $\mathcal{E}$ is a *solution cover* of the ground set A (thus $|\mathcal{E}| \leq |A|$). This immediately compares favorably with existing bounds of order $|A|^4 \ln N$ and $s^2\,|A| \ln N$ for the policies in Gai et al. (2012) and Chen et al. (2013), respectively, and to that of order $|\mathcal{B}|\,(\ln N)^{1+\epsilon}$ for the policy in Liu et al. (2012)[1] ($\mathcal{B}$ is a set of solutions that can be thought of as a variation of a solution cover and it always satisfies $|\mathcal{E}| \leq |\mathcal{B}|$). The second policy we develop is based on an optimization problem which we denote the *Optimality Cover Problem* (OCP). The cumulative regret of this *OCP-based* policy admits an upper bound of order $G\,(\ln N)^{1+\varepsilon}$, for arbitrarily small $\varepsilon > 0$, where G is the size of an instance-dependent set of feasible solutions that always satisfies $G \leq |A|$ and normally satisfies $G \leq |\mathcal{E}|$ ($G \sim |A|$ only for trivial combinatorial optimization problems). We show that, for many families of instances, $|\mathcal{E}|$ is arbitrarily larger than G and hence the OCP-based policy can significantly outperform the cover-based policy. However, we also show that it is possible that $|\mathcal{E}| < G$ (with some probability) for somewhat pathological instances. In this regard, we delineate a family of hybrid cover/OCP-based policies that is guaranteed to outperform the cover-based policy. Nonetheless, extensive computational experiments show the dominance of the pure OCP-based policy over the cover-based policy in practice.

ii) **We show that the proposed policies are essentially optimal with respect to the combinatorial aspects of the problem.** To achieve this we show that no policy can achieve a regret lower than order $L \ln N$, where L is an instance-dependent constant that is strongly related to the solution to a relaxation of OCP, and thus to G. We show that such a lower bound, which is the first for the stochastic combinatorial setting, arises naturally as the limit of performance bounds for a sequence of policies within the proposed class of hybrid policies, thus establishing the tightness of the bound, and the efficiency of such policies.

[1] Such a policy was independently and concurrently developed after our initial submission.

iii) **We provide strong evidence that our policies are implementable in practice.** The proposed policies focus exploration (either fully or partially) on the solution to either OCP or a relaxation of it. These problems could be much harder than the underlying combinatorial optimization problem and are potentially intractable. However, unlike some of the benchmark, our policies only require solving these optimization problems (and the underlying combinatorial optimization problem) with an exponentially decreasing frequency. (Computationally, the cover-based policy can be thought of as a special case of the OCP-based policy.) Furthermore, we show that these optimization problems can be formulated as a Mixed Integer Programming (MIP) problem that can be effectively tackled by state of the art solvers. Finally, we also develop an oracle polynomial time heuristic that utilizes a solution oracle for the underlying combinatorial optimization problem to construct a reasonable solution to OCP and its relaxation. We show that while the performance of the OCP-based policies deteriorates when the heuristic is used, the resulting policies still outperform the cover-based policy as well as other long-term benchmarks, and remains competitive with the short-term benchmarks.

The optimal $\ln N$ scaling of the regret is well known in the bandit literature (Lai and Robbins 1985) and can even be achieved in the combinatorial setting by traditional algorithms. The regret of such algorithms, however, is proportional to the number of solutions, which for combinatorial settings is typically exponential in $|A|$, which suggests that the dependence on N might not be the major driver of performance in this setting. Thus, the focus should be on the accompanying constants: we show that modest modifications to bandit ideas suffice to take such a constant from being proportional to the size of the solution set, which one obtains from the direct application of bandit algorithms, to the size of a minimal solution cover of the whole ground set. In this regard, efficiency is achieved when only a "critical" subset of components of the ground set is covered: OCP solves for such a critical subset and covers it while incurring minimal regret. Our results speak of a fundamental principle in active learning, which is somewhat obscured in the traditional bandit setting: that of only exploring what is necessary to reconstruct the solution to the underlying problem, and doing so at the least possible cost.

The remainder of the paper. Next, we review related work. Section 3 formulates the problem, and advances a simple adaptation of classical bandit ideas (i.e. the cover-based policy). In Section 4 we present the OCP-based policy, develop performance guarantees, and assess asymptotic efficiency. In Section 5 we discuss practical policy implementation, and Section 6 illustrates the results in the paper by means of numerical experiments. Finally, Section 7 presents extensions and concluding remarks. Proofs and supporting material are relegated to Appendices A and B.

2 Literature Review

Classical bandit settings. Introduced in Thompson (1933) and Robbins (1952), the multi-armed bandit setting is a classical framework for dynamic decision making under uncertainty. In its basic formulation a gambler maximizes cumulative reward by pulling arms of a slot machine sequentially over time when limited prior information on reward distributions is available. The gambler faces the classical exploration vs. exploitation trade-off: either pulling the arm thought to be the "best" at the risk of failing to actually identify such an arm, or trying other arms which allows identifying the best arm but hampers reward maximization.

The seminal work of Gittins (1979) shows that, for the case of independent and discounted arm rewards, and infinite horizon, the optimal policy is of the index type. Unfortunately, index-based policies are not always optimal (see Berry and Fristedt (1985), and Whittle (1982)) or cannot be computed in closed form. In their seminal work, Lai and Robbins (1985) study asymptotically efficient policies for the undiscounted case. They establish a fundamental limit on achievable performance, which implies the (asymptotic) optimality of the order $\ln N$ dependence in the regret (see Kulkarni and Lugosi (1997) for a finite-sample minimax version of the result). Our proof of efficiency is based on the change of measure argument in this paper: see the discussion in Section 4.4. In the same setting, Auer et al. (2002) introduces the celebrated index-based UCB1 policy, which is both efficient and implementable. We revisit their results in the next section.

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

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

A first strain of literature considers settings with a continuous set of arms, where exploring all arms is not feasible. Agrawal (1995) studies a multi-armed bandit in which arms represent points in the real line and their expected rewards are continuous functions of the arms. Mersereau et al. (2009) and Rusmevichientong and Tsitsiklis (2010) study bandits with possibly infinite arms when expected rewards are linear functions of an (unknown) scalar and a vector, respectively. In a more general setting, Kleinberg et al. (2008) consider the case where arms form a metric space, and expected rewards satisfy a Lipschitz condition. See Bubeck et al. (2011) for a review of work in *continuum* bandits.

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

Gai et al. (2012) study combinatorial bandits when the underlying problem belongs to a restricted class, and extend the UCB1 policy to this setting. They establish what is essentially an order $|A|^4 \ln N$ performance bound, where $|A|$ denotes the size of the ground set A. Their policy applies to the more general setting we study, and is used as a benchmark in our numerical experiments. Concurrent to our work, two papers examine the combinatorial setting: Chen et al. (2013) provide a tighter order $s^2 |A| \ln N$ performance bound for the UCB1-type policy of Gai et al. (2012) applied to the general combinatorial case (here s denotes a bound on the size of a solution); also, Liu et al. (2012) develop a version of the cover-based policy for network optimization problems (their ideas can be adapted to the general case as well) but under a different form of feedback. Their policy collects information through implementation of solutions in a *barycentric spanner* of the solution set, which in our feedback setting could be set as a solution cover: see further discussion in Section 7. Probable performance of their policy is essentially that of a static cover-based policy, which is (asymptotically) always lower than or equal to that of its dynamic version, and might be arbitrarily worse than the OCP-based policy.

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

Online subset selection. Broadly speaking, our work contributes to the literature of online learning with combinatorial number of alternatives. There are several studies that focus on similar online learning problems, from the ranking and selection perspective. Ryzhov and Powell (2011) study information collection in settings where the decision maker selects individual arcs from a directed graph, and Ryzhov and Powell (2012) consider a more general setting where selection is made from a polyhedron. (See also Ryzhov et al. (2012).) The ideas in Harrison and Sunar (2013) regarding selection of efficient learning mechanisms relate to the insight derived from our work. Also, see Jones et al. (1998), and the references within, for related work in the global optimization literature.

3 Combinatorial Formulation vs. Traditional Bandits

3.1 Problem formulation

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

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

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

We assume that each element $b_{a,n} \in B_n$ is a random variable, independent and identically distributed across instances, and independent of other components in B_n. We let $F(\cdot)$ denote the common distribution of B_n for $n \in \mathbb{N}$, which we assume is initially *unknown*. We assume, however, that upper and lower bounds on its range are known upfront.[2] That is, it is known that $l_a \leq b_{a,n} \leq u_a$ a.s. (with $l_a < u_a$), for all $a \in A$ and $n \in \mathbb{N}$. Furthermore, while our general approach and some of our results hold in more general settings, we assume for simplicity that the distributions of $b_{a,n}$ are absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}$.

We assume that, in addition to not knowing $F(\cdot)$, the agent does not observe B_n prior to implementing a solution. Instead, we assume that B_n is only revealed *partially* and *after* a solution is implemented. More specifically, we assume that if solution $S_n \in \mathcal{S}$ is implemented, only cost-coefficients associated with ground elements in S_n (i.e., $\{b_{a,n} : a \in S_n\}$) are observed by the agent and after the corresponding cost is incurred. Finally, we assume that the agent is interested in minimizing the expected cumulative cost associated with implementing a sequence of solutions.

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

[2] Our results extend to the case of unbounded range provided that $F(\cdot)$ is light-tailed.

sequences.

In practice, the agent does not know F upfront, hence no admissible policy can incur costs below those incurred by the clairvoyant agent, in expectation. Thus, we measure the performance of a policy in terms of its expected *regret*: let $\pi := (S_n)_{n=1}^{\infty}$ denote a non-anticipative policy, where $S_n : \Omega \to \mathcal{S}$ is a $\mathcal{F}_n/2^{\mathcal{S}}$-measurable function that maps the available "history" at time n, $\mathcal{F}_n := \sigma(\{b_{a,m} : a \in S_m , m < n\})$, to a solution in $\mathcal{S}$; given F and N, the expected regret of policy π is

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

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

Remark 3.1. Although the regret also depends on the combinatorial optimization problem through A and $\mathcal{S}$, we omit this dependence to simplify the notation.

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

$$\Delta_S^F := \sum_{a \in S} \mathbb{E}_F \left\{ b_{a,n} \right\} - z^* \left(\mathbb{E}_F \left\{ B_n \right\} \right).$$

(Note that the expected optimality gap associated with $S^* \in \mathcal{S}^* \left(\mathbb{E}_F \left\{ B_n \right\} \right)$ is zero.) For $S \in \mathcal{S}$, let

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

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

$$T_n(\{a\}) := \left| \{ m < n \, : \, a \in S_m \} \right|$$

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

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

3.2 Known results for the non-combinatorial case.

Traditional multi-armed bandits correspond to settings where $\mathcal{S}$ is formed by ex-ante *identical* singleton subsets of A (i.e., $\mathcal{S} = \{\{a\} : a \in A\}$, l_a and u_a equal for all $a \in A$), thus the combinatorial structure is absent. In this setting, the seminal work of Lai and Robbins (1985) (henceforth, LR) establishes an asymptotic lower bound on the regret attainable by any *consistent* policy when F is regular,[3] and provides policies achieving asymptotic efficiency. LR show that consistent policies must explore (pull) each element (arm) in A at least order $\ln N$ times, hence, by (2), their regret must also be of at least order $\ln N$.

Theorem 3.2 (Lai and Robbins 1985). *Let $\mathcal{S} = \{\{a\} : a \in A\}$, then for any consistent policy π and regular distribution F, we have that*

$$\liminf_{N\to\infty} \frac{R^{\pi}(F,N)}{\ln N} \geq \sum_{a\in A} \Delta^{F}_{\{a\}} K_a, \tag{3}$$

where K_a is a positive finite constant depending on F, for all $a \in A$.

In the above, K_a is the inverse of Kullback-Leibler distance (see e.g., Cover and Thomas (2006)) between the original distribution F and a distribution F_a that makes a optimal (which always exists because arms are ex-ante identical). The argument behind the result above is the following: in order to distinguish F from a distribution F_a, consistent policies cannot restrict the exploration of any given arm to a finite number of times (independent of N), and must explore all arms periodically. Thus, broadly speaking, balancing the exploration vs. exploitation trade-off in the traditional setting narrows down to answering *when* (or how frequently) to explore each element $a \in A$. (The answer to this question is given by LR's $\ln N/N$ exploration frequency).

Different policies have been shown to attain the logarithmic dependence on N in (3), and in general, there is a trade-off between computational complexity and larger leading constants. For instance, the index-based UCB1 algorithm introduced by Auer et al. (2002) is simple to compute and provides a finite-time theoretical performance guarantee.

Theorem 3.3 (Auer et al. 2002). *Let $\mathcal{S} = \{\{a\} : a \in A\}$ and for each $a \in A$, let $\tilde{K}_a := 8/(\Delta^{F}_{\{a\}})^2$. Then the expected regret of policy UCB1 after N plays is such that*

$$\frac{R^{\pi}(F,N)}{\ln N} \leq \sum_{a\in A} \Delta^{F}_{\{a\}} \tilde{K}_a + O(1/\ln N). \tag{4}$$

The left-hand sides of (3) and (4) admit asymptotic lower and upper bounds of the form $C_F\,|A|$, respectively, where C_F is a finite constant depending on F. Informally, such bounds imply that the regret $R^{\pi}(F,N)$ grows like $C_F\,|A| \ln N$ where the mismatch between the bounds (3) and (4) is primarily due to the difference in C_F (for (3), $C_F = \min_{a\in A} K_a$, and for (4), $C_F = \max_{a\in A} \tilde{K}_a$). We can then identify three components of the regret:

(i) Component $\ln N$ that is exclusively dependent on time,

[3] A policy π is said to be *consistent* if $R^{\pi}(F,N) = o(N^{\alpha})$ for all $\alpha > 0$, for every F on a class of *regular* distributions satisfying certain *indistinguishability* condition: see proof of Proposition 4.9. This avoids considering policies that perform well in a particular setting at the expense of performing poorly in others.

(ii) component $|A|$ that is exclusively dependent on the combinatorial structure of $\mathcal{S}$ (absent in the traditional setting), and

(iii) component C_F that is almost exclusively dependent on the distribution F.[4]

Giving simple comparisons between the distribution dependent constants C_F can be extremely challenging (e.g. above, only $\tilde{K}_a$ has a simple formula[5]). For the most part, we concentrate on the combinatorial and temporal components of the regret, and examine components (ii) and (iii) jointly only when deemed necessary. Thus, informally, we refer to the regret of policy UCB1 and the LR lower bound *being proportional* to $|A| \ln N$ and declare them *matching up to a distribution-dependent constant.*

As mentioned in the introduction, the combinatorial setting can be seen as a traditional bandit with a combinatorial number of arms, and where arm rewards are correlated. Implementing off-the-shelf traditional bandit policies to this setting would result in a regret proportional to $|\mathcal{S}| \ln N$: unfortunately, for most combinatorial problems of interest, $|\mathcal{S}|$ is exponential in $|A|$, hence traditional bandit algorithms will exhibit regrets scaling rather unfavorably with the size of A. (Note, however, that the fundamental bound in (3) does not apply to the combinatorial setting, thus at this point it is not clear what would constitute a favorable scaling.) Moreover, from a practical standpoint, implementing index-based policies such as UCB1 involves computing an exponential number of indices, and the complexity of selecting an arm is comparable to that of solving the underlying problem by enumeration, which in most cases of interest is impractical.

3.3 Setting comparison and a first simple approach

The results in the traditional bandit neither apply nor are likely to lead to efficiency in the combinatorial setting. However, their underlying principles are still applicable. For example, all bandit algorithms, in one way or another, impose a $\ln N/N$ exploration frequency on each arm, as this allows to estimate mean performance of each arm with a precisely increasing confidence. In our setting, the same goal can be achieved while leveraging the combinatorial structure of the solution set to expedite estimation: a key observation is that one might conduct mean cost estimation for elements in the ground set, and then aggregate those to produce estimates for all solutions.

A natural way of incorporating the observation above into most algorithms is to select as an exploration set 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). We can then alternate between exploring the elements of $\mathcal{E}$ and exploiting an optimal solution according to current mean cost estimates.

[4] While C_F can depend on the combinatorial structure (e.g. through $\min_{a \in A}$), F has a significantly stronger impact on it (e.g. changing F we can keep C_F constant for a changing combinatorial structure or make C_F change for a fixed combinatorial structure).

[5] Auer et al. (2002) show that $K_a \leq \tilde{K}_a/16$ and propose UCB2, that can be tuned so that $\tilde{K}_a \approx 1/(2\left(\Delta_a^F\right)^2)$.

To induce the appropriate exploration frequency we use an idea known as the *doubling trick* (Cesa-Bianchi and Lugosi 2006, Chapter 2.3). This approach allows us to impose the traditional $\ln N/N$ exploration frequency and more importantly, it allows us to minimize the number of times that the underlying combinatorial problem needs 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 $n_{i+2} - n_{i+1} > n_{i+1} - n_i$ for all $i \in \mathbb{N}$. Within each cycle we first implement every element of the exploration set (at most once, in the long-run) and only after that consider the implementation of exploitation solutions. The frequency of exploration can then be controlled by varying the increment in length of the cycles (e.g. to achieve the $\ln N/N$ exploration frequency we can use cycles of exponentially increasing lengths).

Combining the minimal cover exploration set with the doubling trick we obtain a simple *static cover-based* policy $\pi_s(\mathcal{E})$ (which we hereafter refer to as the *simple* policy) that proceeds as follows:

- Starting cycle i, implement solutions in $\mathcal{E}$ until $T_n(\{a\}) \geq i$ for all $a \in A$, or until the end of the cycle. Then, if there is still time left in the cycle, implement $S \in \mathcal{S}^*\left(\bar{B}_{n_i}\right)$ for the rest of it, where $\bar{B}_n := \left(\bar{b}_{a,n}, a \in A\right)$ and

$$\bar{b}_{a,n} := \frac{1}{T_n(\{a\})} \sum_{m<n\,:\,a\in S_m} b_{a,m} \quad a \in A,\; n \in \mathbb{N}. \tag{5}$$

In the following section we introduce a significantly improved policy, so we relegate a detailed description of the simple policy to Appendix A. Next, we show that this simple adaptation of traditional algorithms can significantly improve performance. We begin with the following regret bound results whose proofs are also relegated to Appendix A.

Theorem 3.4. *For any cover $\mathcal{E}$, let $\pi_s(\mathcal{E})$ denote static cover-based policy and for an arbitrary $\delta > 1$ let $H := (1+\delta)\left(s/\Delta^F_{\min}\right)^2$, where $s := \max\{|S| : S \in \mathcal{S}\}$ and $\Delta^F_{min} := \min\left\{\Delta^F_S \,:\, \Delta^F_S > 0\,,\, S \in \mathcal{S}\right\}$. If we choose $n_i := \max\left\{\lfloor e^{i/H}\rfloor, n_{i-1}+1\right\}$, for all $i \geq 2$, then* [6]

$$\frac{R^{\pi_s(\mathcal{E})}(F,N)}{\ln N} \leq (1+\delta)\frac{C}{\left(\Delta^F_{\min}\right)^2}s^2 + O(1/\ln N) \leq (1+\delta)\frac{\Delta^F_{max}}{\left(\Delta^F_{\min}\right)^2}|\mathcal{E}|\,s^2 + O(1/\ln N),$$

where $\Delta^F_{max} := \max\left\{\Delta^F_S \,:\, S \in \mathcal{S}\right\}$, and $C := \sum_{S\in\mathcal{E}} \Delta^F_S$. If instead we choose $n_i := \max\{\lfloor e^{i^{1/(1+\varepsilon)}}\rfloor, n_{i-1}+1\}$, with $\varepsilon > 0$ arbitrary, for all $i \geq 2$, then

$$\frac{R^{\pi_s(\mathcal{E})}(F,N)}{(\ln N)^{1+\varepsilon}} \leq C + O(1/(\ln N)^{1+\varepsilon}) \leq \Delta^F_{max}|\mathcal{E}| + O\left(1/(\ln N)^{1+\varepsilon}\right).$$

The two variants of this simple policy have regrets proportional to $|\mathcal{E}|\,s^2\ln N$ and $|\mathcal{E}|\,(\ln N)^{1+\epsilon}$, with distribution-dependent constants $C_F = (1+\delta)s^2\Delta^F_{max}/\left(\Delta^F_{\min}\right)^2$ and $C_F = \Delta^F_{max}$, respectively.

[6] For simplicity of exposition, the above assumes without loss of generality that $u_a - l_a < 1$. The proof of the result, however, keeps track of the dependence on $u_a - l_a$, which only affects the distribution-dependent constant C_F.

(Note that $\Delta^F_{\min}$ plays the role of $\Delta^F_{\{a\}}$ in the traditional bandit setting.) The second of these bounds clarifies the fact that, for this policy, the regret follows from the cost of suboptimal exploration (at the cost of an arbitrarily small suboptimal scaling with the horizon).

Remarkably, incorporating this simple combinatorial aspect of the problem in the design of algorithms results in performance that compares favorably to relevant benchmarks. To illustrate this point, consider the following example:

Example 3.5. *Consider the digraph $G = (V, A)$ for $V = \{v_{i,j} : i, j \in \{1, \ldots, k+1\}, i \leq j\}$ 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 Figure 1 for $k = 3$. Let $\mathcal{S}$ be composed of all paths from node $s := v_{1,1}$ to node $t := v_{k+1,k+1}$.*

Let $\varepsilon < c \ll M$ and set $l_a = \varepsilon$ and $u_a = \infty$ for every arc $a \in A$. Define F to be such that $\mathbb{E}_F\{b_{e_i,n}\} = c$, and $\mathbb{E}_F\{b_{p_{i,j},n}\} = \mathbb{E}_F\{b_{q_{i,j},n}\} = M$, for all $i \in \{1, \ldots, k\}$ and $i \leq j \leq k$, $n \in \mathbb{N}$. The shortest (expected) path is $S^*(\mathbb{E}_F\{B_n\}) = (e_1, e_2, \ldots, e_k)$ with expected length (cost) $z^*(\mathbb{E}_F\{B_n\}) = kc$, $|A| = (k+3)(k+2)/2$, 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).

In Example 3.5, the regret of traditional algorithms is proportional to $\frac{4^{k+1}}{(k+1)^{3/2}\sqrt{\pi}} \ln N$. Also, the regret of the policy in Gai et al. (2012) and Chen et al. (2013), which is designed to operate in the combinatorial setting, admits a bound proportional to $(k+3)^4(k+2)^4 \ln N/16$ and $2(k+3)(k+2)k^2 \ln N$, respectively. Compare this to $\pi_s(\cdot)$: one can easily construct a cover $\mathcal{E}$ of size $k+1$, in which case its regret is proportional to $4(k+1)k^2 \ln N$ or, alternatively, to $(k+1)(\ln N)^{1+\epsilon}$. Clearly, for moderate k

$$(k+1) \ll 4(k+1)k^2 \ll 2(k+3)(k+2)k^2 \ll \frac{4^{k+1}}{(k+1)^{3/2}\sqrt{\pi}}.$$

This example is representative of settings with combinatorial structure, where $|\mathcal{E}| \ll |A| \ll |\mathcal{S}|$.

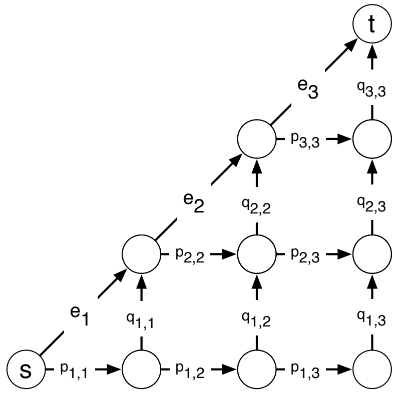

Figure 1: Graph for Example 3.5.

Can one do better? The combinatorial structure of the problem allows significant improvement when applied to policy design. It turns out that additional improvement follows from exploiting the ideas in the lower bound result in LR as well. To see this, note that, unlike in the traditional setting, solutions are not ex-ante identical in the combinatorial one, thus, for some $a \in A$, there might not exist a distribution F_a such that $a \in S$ for some $S \in \mathcal{S}^*(\mathbb{E}_{F_a}\{B\})$. Moreover, even if such a distribution exists, one might be able to distinguish F from F_a without implementing solutions containing a. This opens the possibility that information collection in some ground elements might be stopped after a finite number of instances, independent of N, without affecting asymptotic efficiency. We show next that this is indeed the case.

4 The Optimality Cover Problem

4.1 What and How to Explore?

A fundamental principle in the traditional bandit setting is that feedback from periodically pulling all arms is necessary to find an optimal arm, and to guarantee its optimality through time. However, in the combinatorial bandit setting, exploring all solutions is not always necessary to achieve such objectives. From the previous section, we know that it suffices to collect information on a cover $\mathcal{E}$. Is such information necessary as well? To answer this, consider the setting in Example 3.5 for $k=3$: feedback from solutions $S_1 := (e_1, e_2, e_3)$ and $S_2 := (p_{1,1}, q_{1,1}, p_{2,2}, q_{2,2}, p_{3,3}, q_{3,3})$ allows to consistently estimate the mean cost of these arcs; but because every suboptimal path travels along at least two of the arcs in S_2 (and the fact that costs are non-negative a.s.), such a feedback suffices to deduce and guarantee the optimality of S_1. Thus, we see that a cover $\mathcal{E}$ might contain superfluous information. While one can check that $\{S_1, S_2\}$ is a minimally (w/r to inclusion) sufficient exploration set, it is by no means the only possible set of solutions with such a feature: consider for example the alternative minimally sufficient exploration set $\left\{S_1, \tilde{S}_2, \tilde{S}_3\right\}$ where $\tilde{S}_2 = (p_{1,1}, p_{1,2}, p_{1,3}, q_{1,3}, q_{2,3}, q_{3,3})$ and $\tilde{S}_3 = (e_1, p_{2,2}, p_{2,3}, q_{2,3}, q_{3,3})$. A crucial observation here is that the latter exploration set neither directly explores S_2, nor indirectly explores three of its arcs: $q_{1,1}$, $q_{2,2}$, and $p_{3,3}$. Thus, no set of paths or arcs is necessary to deduce and guarantee the optimally of a solution. Nonetheless, minimally sufficient exploration sets differ in the regret they incur (e.g. $\{S_1, S_2\}$ does result in a smaller regret). To find a minimally sufficient exploration set with minimum regret we first characterize the set of ground elements that are sufficient to guarantee optimality and then find a way to explore them with the smallest possible regret.

What to explore? Suppose that only a subset of ground elements $C \subseteq A$ is selected persistently over time (irrespective of how), so that their mean cost estimates are accurate. To check the sufficiency of the feedback from this selection, one should, in principle, check whether $z^*(\mathbb{E}_F(B_n))$ is robust with respect to *all* plausible changes to mean costs of ground elements in $A \setminus C$. However, because $f(\cdot)$ minimizes cost, it suffices to check only *one* possibility: that in which $\mathbb{E}\{b_{a,n}\} \downarrow l_a$ for all $a \in A \setminus C$. This leads to the following definition for our target set of ground elements to be explored.

Definition 4.1 (Critical Set). *A subset $C \subseteq A$ is a sufficient ground exploration set (or simply sufficient set) for a cost vector $B \in \mathbb{R}^{|A|}$ if and only if*

$$z^*(B) \leq z^*(B_C),$$

where $B_C := (b_a : a \in C) \cup (l_a : a \in A \setminus C)$. A subset $C \subseteq A$ is a critical set if and only if it is a *sufficient set that is minimal with respect to inclusion.*

How to explore? While minimality of the exploration set seems desirable, regret is ultimately driven by the cost of the solutions implemented and not their number in isolation. Hence, it is natural to look for a cheapest (in terms of regret) solution cover of a critical set. The following formulation finds such an exploration set.

Definition 4.2. *For a given cost vector B, we let the Optimality Cover Problem (henceforth, OCP) be the optimization problem given by*

$$\begin{aligned} OCP(B): \quad z^*_{OCP}(B) := \min \quad & \sum_{S\in\mathcal{S}} \Delta^B_S \; y_S && \text{(6a)}\\ s.t. \quad & x_a \leq \sum_{S\in\mathcal{S}:a\in S} y_S, \quad a\in A && \text{(6b)}\\ & \sum_{a\in S} \left(l_a(1-x_a)+b_a x_a\right) \geq z^*(B), \quad S\in\mathcal{S} && \text{(6c)}\\ & x_a,\, y_S \in \{0,1\}\,, \quad a\in A, S\in\mathcal{S}, && \text{(6d)} \end{aligned}$$

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

By construction, a feasible solution (x, y) to OCP corresponds to incidence vectors of a sufficient exploration set $C \subseteq A$ and a solution cover $\mathcal{E}$ of such a set.[7] In what follows we refer to a solution (x, y) to OCP and the induced pair of sets $(C, \mathcal{E})$ interchangeably.

Constraints (6c) guarantee the optimality of $\mathcal{S}^*(B)$ even if costs of elements outside C are set to their lowest possible values (i.e., $b_a = l_a$ for all $a \notin C$), thus ensuring sufficiency of the feedback, and constraints (6b) guarantee that $\mathcal{E}$ covers C (i.e., $a \in S$ for some $S \in \mathcal{E}$, for all $a \in C$). Finally, (6a) ensures that the regret associated with implementing the solutions in $\mathcal{E}$ is minimized. Note that when solving (6), one can impose $y_S = 1$ for all $S \in \mathcal{S}^*(B)$ without affecting the objective function, thus one can restrict attention to solutions that cover optimal elements of A. Furthermore, while the critical subset C in a solution $(C, \mathcal{E})$ to OCP may not be minimal, any minimal subset C' of C leads to a feasible solution $(C', \mathcal{E})$ with the same objective value. To avoid any issues arising from this potential lack of minimality we concentrate on optimal solutions $(C, \mathcal{E})$ to $OCP(B)$ for which both C and $\mathcal{E}$ are minimal. We denote such set of optimal solutions $\Gamma^*(B)$, while noting that any optimal solution to OCP can be efficiently converted to a solution in $\Gamma^*(B)$.

Next, we construct a policy that focuses information collection on the solution to OCP.

4.2 An improved adaptive policy

As in Section 3.3, we can use the doubling trick to impose the appropriate exploration frequency on a critical subset C by implementing the solutions in $\mathcal{E}$, where $(C, \mathcal{E})$ denotes a solution to an

[7] That is, $(x, y) := \left(x^C, y^{\mathcal{E}}\right)$ where $x^C_a = 1$ if $a \in C$ and zero otherwise and $y^{\mathcal{E}}_S = 1$ if $S \in \mathcal{E}$ and zero otherwise.

instance of $OCP(B)$. Ideally we would solve $OCP(B)$ for $B = \mathbb{E}_F\left\{B_n\right\}$, but this value is of course unknown. For this reason we instead use the best current estimate given by $\bar{B}_{n_i}$ and construct a new solution for an updated estimate at the beginning of each cycle. This leads to an adaptive policy π_a, which we refer to as the *OCP-based policy*, that proceeds as follows:

- Starting cycle i, find a solution $(C, \mathcal{E}) \in \Gamma^*\left(\bar{B}_{n_i}\right)$.
- Implement solutions in $\mathcal{E}$ until $T_n(\{a\}) \geq i$, for all $a \in C$ or until the end of the cycle. Then, if there is still time left in the cycle, implement $S \in \mathcal{S}^*\left(\bar{B}_{n_i}\right)$ for the rest of it.

Because $\mathbb{E}_F\left\{B_n\right\}$ is initially unknown, the implementation above solves a proxy of $OCP(\mathbb{E}_F\left\{B_n\right\})$, using the estimate $\bar{B}_{n_i}$. For this reason, $(C, \mathcal{E})$ is updated in an adaptive fashion at the beginning of each cycle, as more information is incorporated by the estimate $\bar{B}_{n_i}$. The details of the OCP-based policy are displayed in Algorithm 1. There, we let $\Phi := \{n_i : i \in \mathbb{N}\}$ be the set of starting points of all cycles, and $\Gamma(\cdot)$ is defined below.

Algorithm 1 Adaptive policy π_a

```
Set $i = 0$, $C = A$, and $\mathcal{E}$ a minimal cover of $A$
for $n = 1$ to $N$ do
    if $n \in \Phi$ then
        Set $i = i + 1$
        Set $S^* \in \mathcal{S}^*\left(\bar{B}_n\right)$                                  [Update exploitation set]
        if $(C, \mathcal{E}) \notin \Gamma\left(\bar{B}_n\right)$ then
            Set $(C, \mathcal{E}) \in \Gamma^*\left(\bar{B}_n\right)$                     [Update exploration set]
        end if
    end if
    if $T_n(\{a\}) < i$ for some $a \in \cup_{S \in \mathcal{E}} S$ then
        Try such an element, i.e., set $S_n = S$ with $S \in \mathcal{E}$ such that $a \in S$    [Exploration]
    else
        Implement $S_n = S^*$                                                  [Exploitation]
    end if
end for
```

To analyze the performance of this adaptive policy we first need to understand the effect of using estimate $\bar{B}_{n_i}$ instead of the exact value $\mathbb{E}_F\left\{B_n\right\}$. Suppose that the policy stabilizes, and eventually implements the same exploration set $(C_\infty, \mathcal{E}_\infty)$ each cycle, so that mean cost estimates converge to a vector $\bar{B}_\infty$. While it should be that $\bar{b}_{a,\infty} \approx \mathbb{E}_F\left\{b_{a,n}\right\}$ for all $a \in \bigcup_{S \in \mathcal{E}_\infty} S$, we would have no guarantee on the quality of the estimates for ground elements $a \in A \setminus \bigcup_{S \in \mathcal{E}_\infty} S$. Hence, we cannot determine if the solution $(C_\infty, \mathcal{E}_\infty)$ is optimal for $OCP(\mathbb{E}_F\left\{B_n\right\})$ (we can only guarantee it will be optimal for a cost vector $\bar{B}_\infty$). We refer to solutions with such limited optimality guarantees as *feedback-consistent* solutions and group them in the set given by

$$\Gamma\left(B\right) := \left\{ (C,\mathcal{E}) \text{ feasible to } OCP(B) : (C,\mathcal{E}) \in \Gamma^*\left(\tilde{B}\right) \text{ for } \tilde{b}_a = b_a, a \in \bigcup_{S\in\mathcal{E}} S,\ \tilde{b}_a = u_a, a \in A \setminus \bigcup_{S\in\mathcal{E}} S \right\}.$$

Note that Algorithm 1 solves OCP only after checking that the solution from the previous cycle is not in $\Gamma\left(\bar{B}_{n_i}\right)$, and ensures minimum information is collected in all arcs in the exploration set (not just in a critical subset). This contributes to establishing convergence of the estimates: in the numerical experiments, we solve OCP each cycle and ensure information collection in a critical subset, as it has a practical advantage. Using the definition above we can derive the following performance guarantee for the OCP-based policy.

Theorem 4.3. *Let π_a denote the policy in Algorithm 1 and set $\varepsilon > 0$ arbitrary. If we choose $n_i := \max\left\{\lfloor e^{i^{1/(1+\varepsilon)}}\rfloor, n_{i-1}+1\right\}$, for all $i \geq 2$, then $(C_{n_i},\mathcal{E}_{n_i})$ converges to $(C_\infty,\mathcal{E}_\infty) \in \Gamma\left(\mathbb{E}_F\left\{B_n\right\}\right)$. Moreover,*

$$\frac{R^{\pi_a(\mathcal{E})}(F,N)}{\left(\ln N\right)^{1+\varepsilon}} \leq \mathbb{E}_F\left\{z^*_{OCP}\left(\bar{B}_\infty\right)\right\} + O\left(1/\left(\ln N\right)^{1+\varepsilon}\right) \leq \Delta^F_{\max}\, G + O\left(1/\left(\ln N\right)^{1+\varepsilon}\right),$$

where $G := \max\left\{|\mathcal{E}| : (C,\mathcal{E}) \in \Gamma\left(\mathbb{E}_F\left\{B_n\right\}\right)\right\}$, and $\bar{B}_\infty$ is a random vector that coincides with $\mathbb{E}_F\left\{B_n\right\}$ for $a \in \bigcup_{S\in\mathcal{E}_\infty} S$.

As in the case of simple policy in Section 3.3, it is possible obtain a bound that scales optimally with N (i.e. with regret proportional to $\ln N$ instead of $\left(\ln N\right)^{1+\varepsilon}$) at the expense of a larger accompanying constant (we account for this possibility in the proof of this result in Appendix B).

Remark 4.4. Consider a version of Algorithm 1, where we impose that $x_a = 1$ for all $a \in A$ when solving OCP, i.e. one solves a proxy for the minimum-regret solution cover of A. The proof of Theorem 4.3 simplifies to show that the performance of the underlying policy π_d, which we refer to as the *dynamic cover-based* policy, admits an upper bound of the form

$$\frac{R^{\pi_d}(F,N)}{\left(\ln N\right)^{1+\varepsilon}} \leq C^* + O\left(1/\left(\ln N\right)^{1+\varepsilon}\right),$$

for $\varepsilon > 0$ arbitrary, where C^* denotes the regret of a minimum-regret solution cover of A (when costs are given by $\mathbb{E}_F\left\{B_n\right\}$). It follows that, asymptotically, the performance bound for the static cover-based policy π_s is at most as good as that of its dynamic counterpart.

4.3 Performance Bound Comparisons

From the previous section we see that comparing the performance of OCP-based and cover-based policies amounts to comparing $\mathbb{E}_F\left\{z^*_{OCP}\left(\bar{B}_\infty\right)\right\}$ and C^*. Unfortunately, while it is always the case that $z^*_{OCP}\left(\mathbb{E}_F\left\{B_n\right\}\right) \leq C^*$, in general we only have that $z^*_{OCP}\left(\mathbb{E}_F\left\{B_n\right\}\right) \leq \sum_{S\in\mathcal{E}} \Delta^F_S$ for

$(C, \mathcal{E}) \in \Gamma(\mathbb{E}_F\{B_n\})$. Thus, it is possible that neither bound dominates the other across all settings. First, it is possible for OCP-based policy to significantly outperform cover-based policies. To see this, consider the following example.

Example 4.5. *Let $G = (V, A)$ be the digraph depicted in Figure 2 and let $\mathcal{S}$ be composed of all paths from node s to node t. Set $l_a = 0$ and $u_a = \infty$ for every arc $a \in A$, and F be such that $\mathbb{E}_F\{b_{e,n}\} = c$, $\mathbb{E}_F\{b_{g,n}\} = 0$, $\mathbb{E}_F\{b_{f,n}\} = \mathbb{E}_F\{b_{h,n}\} = \frac{c+\varepsilon}{2}$, $\mathbb{E}_F\{b_{p_i,n}\} = \mathbb{E}_F\{b_{q_i,n}\} = M$ and $\mathbb{E}_F\{b_{v_i,n}\} = \mathbb{E}_F\{b_{w_i,n}\} = \frac{c+\varepsilon}{2}$ for $n \in \mathbb{N}$ and for all $i \in \{1, \ldots, k\}$ where $0 < \varepsilon \ll c \ll M$. The shortest (expected) path in this digraph is (e).*

In Example 4.5, $|\mathcal{S}| = (k+2)$, the only cover of A is $\mathcal{E} = \mathcal{S}$ and $s = 4$. Thus, the regret of cover-based policies is proportional to either $16(k+2)\ln N$ or $(k+2)(\ln N)^{1+\varepsilon}$ and hence does not seem to improve upon traditional algorithms, or other benchmark (although $|A| > |\mathcal{E}|$). In contrast, we can check that $G = 2$, thus the regret of π_a is proportional to $2(\ln N)^{1+\varepsilon}$, which is independent of k. The following proposition, whose proof can be found in Appendix A.2, shows that constructing similar examples where π_a has an arbitrarily better performance than the simple policy in Section 3.3 can be done for a wide variety of combinatorial optimization problems.

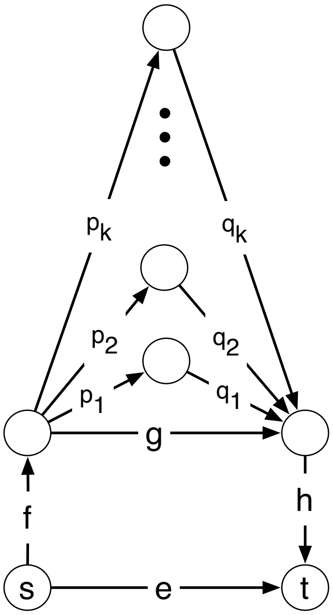

Figure 2: Graph for Example 4.5.

Proposition 4.6. *If $f(B)$ corresponds to a shortest path, minimum cost spanning tree, minimum cost perfect matching, generalized Steiner tree or knapsack problem, then there exists a family of instances where G is arbitrarily smaller than a minimum size cover of A.*

The construction of examples where the cover-based policy outperforms the OCP-based policy is more subtle and we illustrate it with the following example.

Example 4.7. *Let $G = (V, A)$ be the digraph depicted in Figure 3 and let $\mathcal{S}$ be composed of all paths from node s to node t. Set $l_a = 0$ and $u_a = \infty$ for every arc $a \in A$, and F be such that for all $n \in \mathbb{N}$ we have $\mathbb{E}_F\{b_{e,n}\} = c$, $\mathbb{E}_F\{b_{d,n}\} = \mathbb{E}_F\{b_{p_1,n}\} = \mathbb{E}_F\{b_{q_1,n}\} \approx 0$, $\mathbb{E}_F\{b_{p_2,n}\} = \mathbb{E}_F\{b_{q_2,n}\} = \frac{c-\varepsilon}{2}$ and $\mathbb{E}_F\{b_{f_i,n}\} = \mathbb{E}_F\{b_{g_i,n}\} = \frac{c+\varepsilon}{2}$ for all $i \in \{1, \ldots, n\}$ where $0 < \varepsilon \ll c$. The shortest (expected) path in this digraph is (e).*

For every $i \in \{1, \ldots, k\}$ let $S_i = (d, p_1, q_1, f_i, g_i)$ and $\tilde{S}_i = (d, p_2, q_2, f_i, g_i)$. Then, in Example 4.7 a minimum-regret cover of the digraph is given by $\{S_i\}_{i=1}^{k-1} \cup \left\{\tilde{S}_k\right\}$, which has a regret of $c+\varepsilon(k-1)$. In contrast, a minimum-regret feedback-consistent solution to $OCP(\mathbb{E}_F\{B_n\})$ is given by $\{S_i\}_{i=1}^{k}$, which has a regret of $z^*_{OCP}\left(\mathbb{E}_F\left\{B_n\right\}\right) = k\varepsilon$. Hence, an OCP-based policy that consistently implements this feedback-consistent solution as an exploration set will perform significantly better than a minimum-regret cover-based policy. Unfortunately, an alternative feedback-consistent solution to $OCP\left(\mathbb{E}_F\left\{B_n\right\}\right)$ with regret equal to kc is given by $\left\{\tilde{S}_i\right\}_{i=1}^{k}$ and we cannot guarantee that exploration set in the OCP-based policy will not converge to this set. The issue here is that initially the OCP-based policy could draw samples of the costs of p_1 and q_1 that are extremely high. This event could then permanently bias the policy towards $\left\{\tilde{S}_i\right\}_{i=1}^{k}$ as its feedback suffices to guarantee optimality of (e). One can see that the expected regret of the exploration set used by the OCP-based policy is such that $\mathbb{E}_F\left\{z^*_{OCP}\left(\bar{B}_\infty\right)\right\} \in (k\varepsilon, kc)$. If the distribution F is such that $\mathbb{E}_F\left\{z^*_{OCP}\left(\bar{B}_\infty\right)\right\} > c + \varepsilon(k-1)$, then the dynamic cover-based policy would outperform the OCP-based policy, on average.

Generally speaking, a sufficient condition for the OCP-based policy to outperform cover-based policies is that feedback-consistent solutions to OCP are also optimal, i.e. $\Gamma^*(\mathbb{E}_F\left\{B_n\right\}) = \Gamma(\mathbb{E}_F\left\{B_n\right\})$. Whether this condition holds depends on the specific setting at hand, however, the next lemma, which we prove in Appendix A.2, establishes that this is the case for an important class of combinatorial problems.

Lemma 4.8. *Let $f(\cdot)$ be a weighted basis or independent set matroid minimization problem. Then, for $B \in \mathbb{R}^{|A|}$ in the range of F, $\bigcup_{S\in\mathcal{E}} S \subseteq C$ for all $(C,\mathcal{E}) \in \Gamma^*(B)$.*

The above implies that one can always corroborate the optimality of a solution to OCP based solely on the feedback it generates. Thus, for the case of matroids, the OCP-based accepts a performance bound of

$$z^*_{OCP}(\mathbb{E}_F\left\{B_n\right\})\left(\ln N\right)^{1+\varepsilon} + O\left(1/\left(\ln N\right)^{1+\varepsilon}\right)$$

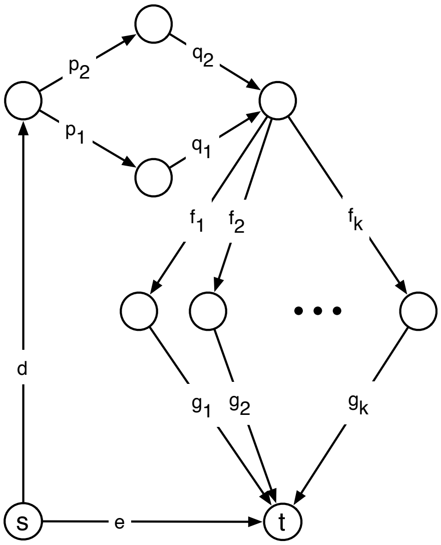

Figure 3: Graph for Example 4.7.

and hence always outperforms cover-based policies (asymptotically). However, the analysis also indicates that the OCP-based policy might not improve upon cover-based policies when feedback-consistent solutions to OCP are suboptimal.

Can one do better? From above, we see that further improvement in performance might be achieved if one is to supplement OCP-based feedback with that from a solution-cover of A, so as to provide the optimality guarantee that feedback-consistent solutions lack. Indeed, consider a policy that during cycle i conducts exploration such that it guarantees $T_n(\{a\}) \geq \gamma\, i$ for all $a \in A$, for $\gamma \in (0,1)$ arbitrary, and that $T_n(\{a\}) \geq i$ for all a in a critical subset C; this while incurring the lowest possible regret (according to current cost estimates). The proof techniques in this paper allow to show that the performance of such a hybrid cover/OCP-based policy admits an upper bound of order[8]

$$((1-\gamma)z^*_{OCP}(\mathbb{E}_F\{B_n\}) + \gamma\ C^*)\,(\ln N)^{1+\varepsilon} + O\left(1/\,(\ln N)^{1+\varepsilon}\right).$$

(We provide further details on this policy in the next section.) Note that the asymptotic performance of this policy is better than that of the dynamic cover-based policy, and better than the OCP-based policy (on average) provided that

$$(1-\gamma)z^*_{OCP}(\mathbb{E}_F\{B_n\}) + \gamma\ C^* < \mathbb{E}_F\left\{z^*_{OCP}\left(\bar{B}_\infty\right)\right\},$$

which is possible when $\Gamma^*\left(\mathbb{E}_F\{B_n\}\right) \subset \Gamma\left(\mathbb{E}_F\{B_n\}\right)$. Note that γ must be positive if one is to recover the optimal solution to $OCP(\mathbb{E}_F\{B_n\})$. In this regard, we see that $z^*_{OCP}(\mathbb{E}_F\{B_n\})\ln N$ arises as a natural lower bound on the probable performance of policies in this class.[9] This raises the question of whether this fundamental limit is shared by all admissible policies. To answer this question, we must first establish a theoretical limit on achievable performance.

4.4 A limit on Achievable Performance

In this section we establish that any consistent policy must explore all elements in some critical subset, at the frequency prescribed in LR. We then use this fact to establish a fundamental limit on performance.

Consistent policies must explore critical sets. Let $\mathcal{D}$ contain all subsets D of suboptimal ground elements such that they become part of every optimal solution if their costs are the lowest possible, and that are minimal with respect to inclusion. That is, $\mathcal{D} := \{D \in \mathcal{D}' : D' \notin \mathcal{D}'\ \forall\, D' \subset D\}$, where

$$\mathcal{D}' := \left\{ D \subseteq A : D \subseteq \bigcap_{S\in\mathcal{S}^*(\mathbb{E}_F\{B_n\})} (A\setminus S)\ ,\ D \subseteq \bigcap_{S\in\mathcal{S}^*((\mathbb{E}_F\{B_n\})_{A\setminus D})} S\ ,\ \right\},$$

with B_D defined as in Definition 4.1 for $B \in \mathbb{R}^{|A|}$. By construction, for any $D \in \mathcal{D}$ there exists an alternative distribution F_D under which all elements in D are part of any optimal solution. Because said elements are suboptimal, a consistent policy must distinguish F from F_D to attain asymptotic optimality. This can be accomplished by selecting *at least* one element in each set $D \in \mathcal{D}$ at a minimum frequency. The following proposition, which we prove in Appendix A.2, formalizes this.

Proposition 4.9. *For any consistent policy π, regular F, and $D \in \mathcal{D}$ we have that*

$$\lim_{N\to\infty} \mathbb{P}_F\left\{\frac{\max\{T_{N+1}(\{a\}) : a \in D\}}{\ln N} \geq K_D\right\} = 1, \tag{7}$$

[8] The proof of such a result, which is omitted, follows from those of Theorems 3.4 and 4.3, along the lines of the extension mentioned in Remark 4.4.

[9] Note that, for a given setting, one can get arbitrarily "close" (in terms of asymptotic performance) to such a bound by suitably choosing ε and γ.

where K_D is a positive finite constant depending on F.

In this proposition, K_D corresponds to the inverse of Kullback-Leibler distance between F and the alternative distribution F_D. Note that while Theorem 3.2 imposes lower bounds on the number of times that a solution (a singleton) is implemented, Proposition 4.9 imposes similar bounds, but on the number of times that certain subsets of A are selected.

From Proposition 4.9, we have that consistent policies try at least one element in each $D \in \mathcal{D}$ at a minimum frequency. Thus, such policies must *at least* explore "frequently" all elements on a set $C \in \mathcal{C}'$, where

$$\mathcal{C}' := \{C \subseteq A : \forall D \in \mathcal{D}, \exists\, a \in C \text{ s.t. } a \in D\}.$$

Note that by construction, any set $C \in \mathcal{C}'$ is such that $z^*\left(\mathbb{E}_F\{B_n\}\right) \leq z^*((\mathbb{E}_F\{B_n\})_C)$, thus C is a sufficient exploration set. Similarly, because elements in $\mathcal{D}$ are minimal with respect to inclusion, one has that $\min\{z^*((\mathbb{E}_F\{B_n\})_{C'}) : C' \subset C\} < z^*\left(\mathbb{E}_F\{B_n\}\right)$ for all $C \in \mathcal{C}'$ that are minimal with respect to inclusion. Thus, by Definition 4.1, we conclude that $\mathcal{C} := \{C \in \mathcal{C}' : C' \notin \mathcal{C}', \forall C' \subset C\}$ essentially corresponds to the family of all critical subsets.[10] This confirms the intuition that consistent policies must actively explore *at least* all elements of a critical set.

A limit on achievable performance. The above establishes that any consistent solution must generate feedback sufficient to recover the optimal solution to the underlying problem. Efficiency of such feedback, on the other hand, follows from minimizing the regret associated with collecting such information. In that regard, transforming the bounds in Proposition 4.9 into a valid performance bound might be accomplished by solving the following Lower Bound Problem (LPB)

$$\begin{aligned}
LBP: \quad L(F) := \min \quad & \sum_{S \in \mathcal{S}} \Delta_S^F \, y_S && \text{(8a)}\\
s.t. \quad & \max\{x_a : a \in D\} \geq K_D, \quad D \in \mathcal{D} && \text{(8b)}\\
& x_a \leq \sum_{S \in \mathcal{S}: a \in S} y_S, \quad a \in A && \text{(8c)}\\
& x_a,\, y_S \in \mathbb{R}_+, \quad a \in A, S \in \mathcal{S}. && \text{(8d)}
\end{aligned}$$

In this formulation, y_S and x_a are meant to represent $T_{N+1}(S)/\ln N$ and $T_{N+1}(\{a\})/\ln N$, respectively. Note that, without loss of generality, one can restrict attention to solutions that set $x_a = 0$ when this does not affect the objective value. Because of this, the a's with non-zero x_a correspond to a critical subset, and the S's with non-zero y_S correspond to the cover of the critical subset. Indeed, constraints (8b) enforce exploration conditions (7) on the critical subset and constraints (8c) enforce the cover of the critical subset.

For any consistent policy π, define $\zeta^\pi(F,N) := \sum_{S \in \mathcal{S}} \Delta_S^F \, T_{N+1}(S)$ to be the total additional cost (relative to an oracle agent) associated with that policy. Note that $\mathbb{E}_F\{\zeta^\pi(F,N)\} = R^\pi(F,N)$.

[10] Unlike $\mathcal{C}$, Definition 4.1 requires maintaining the optimal cost, thus requiring the inclusion of ground elements in all optimal solutions. Nonetheless, such elements can be explored without incurring in regret.

The next proposition, which we prove in Appendix A.2, ties the asymptotic bounds in (7) to the solution to LBP to establish an asymptotic bound on the regret incurred by any consistent policy.

Proposition 4.10. *For any consistent policy π and any regular F we have that*

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

Note that the result above establishes convergence in probability (hence it can be used to bound $\zeta^{\pi}(F,N)$, rather than just its expectation, which is the regret). The next result, whose proof we omit as it follows directly from Proposition 4.10 and Markov's inequality, shows that the regret of any consistent policy in the combinatorial setting is (at least) proportional to $L(F)\ln N$.

Theorem 4.11. *The regret of any consistent policy π is such that for any regular F we have*

$$\liminf_{N\to\infty} \frac{R^{\pi}(F,N)}{\ln N} \geq L(F),$$

where $L(F)$ is the optimal objective value of formulation LBP in (8).

Theorem 4.11 not only provides a fundamental limit on performance, it also supports the principle behind the OCP-based policy: frequent exploration might be restricted to a solution cover of a critical subset. Note, however, that while both OCP and LBP aim to minimize the regret associated with covering a critical subset, LBP is more flexible in terms of the frequency and means by which such a critical subset is covered. In particular, differences in the K_D values might translate into different values of the x_a variables, thus signaling different requirements for feedback collection. In addition, since the y_S variables are continuous in LBP, it is possible to cover a critical element with multiple solutions (implemented a different frequencies implied by the y_S's).

Define R-OCP as the relaxation of OCP where the integrality constraints (6d) over the y_S variables are replaced by those of non-negativity. The next lemma, which we prove in Appendix A.2.2, establishes the connection between LBP and R-OCP

Lemma 4.12. *Let $K_D = K$ for some constant $K > 0$, for all $D \in \mathcal{D}$ in formulation LBP. An optimal solution (x,y) to R-OCP$(\mathbb{E}_F\{B_n\})$ is also optimal to LBP.*

The connection between these formulations goes beyond that indicated in Lemma 4.12. When $K_D = K$ for all $D \in \mathcal{D}$, one can always select a feasible solution to (8) and map it into a feasible solution to R-OCP (via proper augmentation), and the opposite holds true as well. Thus, for the equal K_D case, one can argue that these formulations are essentially equivalent up to a minor difference: optimal solutions to R-OCP *must* cover all optimal ground elements; this, however, can be done without affecting performance and hence is inconsequential.

Back to bound comparison. Using the results above we can compare the performance of the hybrid policy discussed at the end of Section 4.3 with that of an idealized optimal policy. In particular, ignoring the order $(1/\ln N)$ terms assuming that N is large enough, we have that

$$L(F)\ \ln N \overset{(a)}{\leq} K_{\max}\, z^*_{\textit{R-OCP}}\left(\mathbb{E}_F\left\{B_n\right\}\right)\ \ln N \overset{(b)}{\leq} \left((1-\gamma) z^*_{OCP}\left(\mathbb{E}_F\left\{B_N\right\}\right) + \gamma\, C^*\right)\ \left(\ln N\right)^{1+\varepsilon}, \quad (9)$$

where $K_{\max} := \max\left\{K_D : D \in \mathcal{D}\right\}$. We ignore $K_{\max}$ as it is a distribution-dependent constant (we assume $K_{\max} = 1$). The theoretical gap in attainable performance for the proposed policies arises from (a) and (b) above. This raises the question of whether any of these sources of gap can be closed.

Lemma 4.12 establishes that (a) arises because both R-OCP and OCP impose equal exploration frequencies on all elements on a critical subset. However, it is possible to eliminate (a) by modifying the proposed policy such that it adjusts exploration frequencies adaptively over time. Such is the spirit of the more complex UCB2 and UCB1-normal (Auer et al. 2002), that modify UCB1. Nonetheless, improvement from such a modification might be shadowed by the increase in computational complexity from approximating the coefficients K_D (which depend on F) and from solving the resulting proxy formulation (LBP admits an integer programming formulation, but constraint (8b) is notoriously difficult to handle (Toriello and Vielma 2012)). We do not pursue such direction here, as R-OCP already captures the combinatorial nature of LBP while still providing enough tractability in practice.

With regard to the gap in (b), it can be reduced for a given setting by: (i) focusing information collection on the solution to R-OCP instead of that to OCP; and (ii) making both ε and γ tend to zero. Regarding (i), in Appendix A we present a hybrid policy that collects feedback from both a solution cover of A and from the solution to R-OCP. The proof techniques in this paper apply to establish a performance bound of

$$\left((1-\gamma+\rho) z^*_{\textit{R-OCP}}\left(\mathbb{E}_F\left\{B_N\right\}\right) + \gamma\, C^*\right)\ \left(\ln N\right)^{1+\varepsilon},$$

for $\rho > 0$ arbitrary. With respect to (ii) above, while we can get a sequence of policies whose regret tend to the left hand side of (b), it is not clear if it is possible to construct a single policy that achieves this regret (at least not uniformly over all possible settings). Note that making ε and γ tend to zero increases the constants accompanying order $(1/\left(\ln N\right)^{1+\varepsilon})$ terms. Hence, achieving the optimal asymptotic performance for identical constants K_D may come at the expense of a significant deterioration of the practical finite-time performance. For this reason we also do not pursue such direction here. Nonetheless, one can get arbitrarily close to the optimal performance within the class of hybrid policies while maintaining practical implementability, thus we claim that such a class essentially matches the fundamental limit on performance up to a distribution-dependent constant.

5 Practical Policy Implementation

In this section, we address the practical implementation of the proposed policies. We provide strong evidence that, at least for a large class of combinatorial problems, the proposed policies scale reasonably well and should be implementable for real-size instances. For this, we focus our attention on the practical solvability of OCP and $R\text{-}OCP$, which our policies solve repeatedly during the horizon, for many inputs B. Note that $f(B)$, $OCP(B)$ and $R\text{-}OCP(B)$ have generic combinatorial structures and hence could be extremely hard to solve. Thus, practical tractability of said problems is essential for implementation. For simplicity we concentrate on the solution of OCP and comment on adaptations for $R\text{-}OCP$ when needed.

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

We also present an oracle polynomial time heuristic for OCP. This heuristic requires a polynomial number of calls to an oracle for solving $f(B)$. It then runs in polynomial time when $f(B)$ is polynomialy solvable. Furthermore, it provides a practical solution method for OCP when $f(B)$ is not expected to be solvable in polynomial time, but is frequently tractable in practice (e.g. medium size instances of NP-complete problems such as the traveling salesman (Applegate et al. 2011), Steiner tree (Magnanti and Wolsey 1995, Koch and Martin 1998, Carvajal et al. 2013), and set cover problems (Etcheberry 1977, Hoffman and Padberg 1993, Balas and Carrera 1996)).

5.1 A Time-constrained Asynchronous Policy

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

Nonetheless, as described, for example, in Algorithm 1, the OCP-based policy cannot proceed until the corresponding problems are solved. However, one can easily modify the policy so that it begins solving $f(B)$ and/or $OCP(B)$ at the beginning of a cycle, but continues to implement solutions while these problems are solved (such solutions might be computed either upfront or in previous cycles). Solution to these problems update incumbent solutions as they become available, which for long cycles would be at the beginning of the next one.

Algorithm 5, which can be found in Appendix A.3 presents one such possible modification. It essentially applies the static cover-based policy in the transient period and eventually implements the OCP-based policy with one cycle delay (once the cycles are long enough, the policy essentially implements the exploration and exploitation solutions that the OCP-based policy would have implemented in the previous cycle). Note that this one cycle delay does not affect the asymptotic analysis of the policy and hence the performance guarantee of the OCP-based policy is preserved. In addition, the short-term performance of this policy is that of the static cover-based policy.

5.2 MIP formulations for OCP for Polynomially Solvable Problems

In this section we assume $f(B)$ is polynomially solvable. However, this does not imply that neither $OCP\,(B)$ nor $R\text{-}OCP(B)$ are tractable or practically solvable, as they might contain an exponential (in $|A|$) number of variables and constraints.[11] Nonetheless, the following theorem, whose proof can be found in Appendix A.3.1, ensures that both $OCP\,(B)$ and $R\text{-}OCP\,(B)$ remain in NP, the class of non-deterministic polynomially solvable problems (see e.g., Cook et al. (1998)).

Theorem 5.1. *If $f\,(B)$ is in P, then $OCP\,(B)$ and $R\text{-}OCP\,(B)$ are in NP.*

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

Theorem 5.2. *OCP and R-OCP are in P for weighted basis or independent set matroid minimization problems.*

While it is possible to establish a non-trivial jump in theoretical complexity for problems within P, we deem the study of the theoretical complexity of OCP/R-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 $f(B)$ admits a linear programming (LP) formulation. Then, we address the case when $f(B)$ admits a polynomial-sized extended LP formulation, and finally, the case when $f(B)$ does not admit such an extended formulation.

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

[11] In most cases the natural size of $f(B)$ is $O(|A| + \sum_{a\in A:|l_a|<\infty} \ln_2 l_a + \sum_{a\in A:|u_a|<\infty} \ln_2 u_a)$, and l_a and u_a are usually bounded. For instance, in the class of shortest path problems in Example 3.5, the natural size of the considered graph is the number of arcs $|A| = O(k^2)$ (not the number of paths) and the bounds of all arc lengths are constant. If we instead had non-constant finite upper bounds u_a (with $u_a > 0.03$ for the analysis to remain valid), we would also include the number of bits needed to encode them, which is equal to $\sum_{a\in A} \ln_2 u_a$. We refer the interested reader to Schrijver (2003) for more information on the input sizes of combinatorial optimization problems.

Proposition 5.3. *Let y^S be the incidence vector of $S \in \mathcal{S}$, $M \in \mathbb{R}^{m\times|A|}$, and $d \in \mathbb{R}^m$ be such that $\left\{y^S\right\}_{S\in\mathcal{S}} = \left\{y \in \{0,1\}^{|A|} : My \le d\right\}$ and $\operatorname{conv}\left(\left\{y^S\right\}_{S\in\mathcal{S}}\right) = \left\{y \in [0,1]^{|A|} : My \le d\right\}$. Then an MIP formulation of $OCP(B)$ is given by*

$$\min \quad \sum_{i=1}^{|A|}\left(\sum_{a\in A} b_a y_a^i - z^*(B)\right) \tag{10a}$$

$$s.t. \quad x_a \le \sum_{i=1}^{|A|} y_a^i, \qquad a \in A \tag{10b}$$

$$My^i \le d, \qquad i \in \{1,\dots,|A|\} \tag{10c}$$

$$M^T w \le \operatorname{diag}(l)(\mathbf{1}-x) + \operatorname{diag}(b)x \tag{10d}$$

$$d^T w \ge z^*(B) \tag{10e}$$

$$x_a,\, y_a^i \in \{0,1\}, w \in \mathbb{R}^m, \qquad a \in A, i \in \{1,\dots,|A|\}, \tag{10f}$$

where for $v \in \mathbb{R}^r$, $\operatorname{diag}(v)$ is the $r \times r$ diagonal matrix with v as its diagonal. A formulation for R-$OCP(B)$ is obtained by replacing $y_a^i \in \{0,1\}$ with $0 \le y_a^i \le 1$.

In the above, x represents the incidence vector of a critical set. Such a condition is imposed via LP duality, using constraints (10d) and (10e), and eliminates the necessity of introducing constraint (6c) for each solution in $\mathcal{S}$. Similarly, each y^i represents the incidence vector of a solution $S \in \mathcal{S}$ for OCP and fractions of solutions for R-OCP[12].

Formulation (10) has $O(|A|^2)$ variables and $O(m|A|)$ constraints. If m is polynomial in the size of the input of $f(B)$, then we should be able to solve (10) directly with a state of the art 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 (10c) within a Branch-and-Cut algorithm. However, in such a case one would have an exponential number of w variables, which would force us to use a more intricate, and potentially less effective, branch-and-cut-and-price procedure. Nonetheless, when $f(B)$ does not admit a polynomial-sized LP formulation, one can still provide formulations with a polynomial number of variables, many of them also having a polynomial number of constraints. We discuss such cases next.

Problems with Polynomial-sized Extended Formulations. The first way to construct polynomial-sized IP formulations of $OCP(B)$ and R-$OCP(B)$ 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 5.3 is exponential in the number of nodes of the underlying graph. However, in the case of spanning trees, we can additionally use a known

[12] Because of assumption $\operatorname{conv}\left(\left\{y^S\right\}_{S\in\mathcal{S}}\right) = \left\{y \in [0,1]^{|A|} : My \le d\right\}$, y^i is a convex combination of incidence vectors, so it might correspond to fractions of more than one solution.

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 $\left\{y^S\right\}_{S \in \mathcal{S}} = P \cap \{0,1\}^{|A|}$ and $\operatorname{conv}\left(\left\{y^S\right\}_{S \in \mathcal{S}}\right) = P$. Then, a MIP formulation with a polynomial number of variables and constraints of $OCP(B)$ for the spanning tree problem is obtained by replacing (10c) with $Cy^i + Dz^i \leq d$, replacing (10d) with $C^T w \leq \operatorname{diag}(l)(\mathbf{1} - x) + \operatorname{diag}(b)x$ and $D^T w \leq 0$, and adding the polynomial number of variables z^i for $i \in \{1, \ldots, |A|\}$. Similar techniques can be used to construct polynomial-sized formulations for other problems with polynomial-sized extended LP formulations.

Problems without Polynomial-sized Extended Formulations. It has recently been shown that there is no polynomial-sized extended LP formulations for the non-bipartite perfect matching problem (Rothvoß 2013a). Hence, we cannot use the techniques in the previous paragraph to construct polynomial-sized IP formulations of $OCP\,(B)$ and $R\text{-}OCP\,(B)$ for matching. Fortunately, a simple Linear Programming observation and a result by Ventura and Eisenbrand (2003) allow constructing a version of (10) with a polynomial number of variables. The observation is that a solution y^* is optimal for $\max\left\{b^T y : My \leq d\right\}$ if and only if it is optimal for $\max\left\{b^T y : M_i^T y \leq d_i \quad \forall i \in I\,(y^*)\right\}$ where $I\,(y^*) := \left\{i \in \{1, \ldots, m\} : M_i^T y^* = d_i\right\}$ 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 (10d) analog to that for spanning trees to obtain a version of (10) with a polynomial number of variables. We would still have an exponential number of constraints in (10c), but these can be separated in polynomial time for matching, so $OCP(B)$ and $R\text{-}OCP\,(B)$ 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 (2013b) shows that there must exist a family of matroid problems without a polynomial-sized extended LP formulation. Fortunately, Theorem 5.2 shows that OCP/R-OCP for matroids can be solved in polynomial time. We are not aware of any other polynomially solvable combinatorial optimization problem which require non-trivial results to formulate $OCP(B)$ or $R\text{-}OCP\,(B)$ with a polynomial number of variables.

We end this subsection by noting that further improvements and extensions to (10) can be achieved. We give two such examples in Appendices A.3.2 and A.3.3. The first example shows how (10) for $OCP(B)$ can be extended to the case when $f(B)$ is not in P, but admits a compact IP formulation. The second example gives a linear-sized formulation of $OCP(B)$ for shortest path problems.

5.3 Oracle Polynomial-time Heuristic

One practical advantage of using MIP formulations of $OCP(B)$ or $R\text{-}OCP(B)$ is that the build-in heuristics of state of the art MIP solvers usually find good quality solutions early on. One can use such solutions within the asynchronous policy in Algorithm 5 to update the set of solutions used to collect information whenever these heuristics encounter a better solution. Of course, such heuristics might not have runtime or quality guarantees, but should improve the practical performance of our algorithm, relative to the simple policy of Section 3.3. To illustrate this, we develop one such heuristic, which only requires a polynomial number of queries to an oracle for $f(B)$ (plus a polynomial number of additional operations), and that returns a solution that is equal and possibly arbitrarily better than a minimal cover of A.

We begin by describing the heuristic for $OCP(B)$ in Algorithm 2. This heuristic first sets all costs to their lowest possible values, and successively solves instances of $f(B)$, each time incorporating the incumbent solution to the cover $\mathcal{E}$, adding its ground elements to C, and updating the cost vector accordingly. The procedure stops when the feedback from C *suffices* to guarantee the optimality of a solution (i.e. when $z^*(\tilde{B}) \geq z^*(B)$). To achieve *efficiency* of such a feedback, the heuristic then prunes elements in C that are not required to guarantee sufficiency of the feedback.

Algorithm 2 Oracle Polynomial-time Heuristic

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

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

The performance of the heuristic ultimately depends on the specifics of a setting. For instance, in the setting of Example 3.5, the heuristic returns, in the worst case, a solution with $|\mathcal{E}| = k$, which is of the order of a cover of A. In the setting of Example 4.5 on the other hand, the heuristic returns a solution with $|\mathcal{E}| = 2$ (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. In fact, one can check such is the case for the settings presented in the proof of Proposition 4.6. We test the practical performance of the heuristic, when embedded in the OCP-based policy, in the next section.

Finally, a heuristic for *R-OCP* can be obtained by using the critical set C obtained from Algorithm 2. Using this critical set to fix the x_a variables in (6) for *R-OCP* (i.e. with relaxed y_S variables) which yields an LP than can be solved in polynomial time by column generation.

6 Numerical Experiments

We illustrate the performance of the proposed policies via numerical experiments in two settings. First, long-term experiments aim to illustrate the ability of the proposed policies to leverage the combinatorial structures to improve upon the performance of relevant benchmark. Then, we compare policy performance against benchmark especially tuned for the short-term (recall that our policy aims at asymptotic optimality). For each setting, we first describe the benchmark and then present numerical results for settings of the shortest path, Steiner tree and knapsack problems.

6.1 Long-term Experiments

6.1.1 Benchmark Policies and Implementation

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

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

and $\bar{b}_{S,n}$ denotes an estimate of the expected cost of solution $S \in \mathcal{S}$ at period n, computed at the solution level.[13] We improve performance of UCB1 by: (*i*) conducting parameter estimation at the ground element level; (*ii*) adjusting confidence interval length to reflect better the amount of information used in estimating parameters; (*iii*) adjusting said length so that confidence bounds remain within the bounds implied by the range of F; and (*iv*) reducing the solution set so that it only includes solutions that are minimal with respect to inclusion. The resulting policy, which we denote UCB1+, implements solution S_n for instance n, where[14]

$$S_n \in \operatorname*{argmin}_{S\in\mathcal{S}}\left\{\max\left\{\sum_{a\in S}\bar{b}_{a,n} - \sqrt{2\ln(n-1)/(\min_{a\in S}\{T_n(\{a\})\})}, \sum_{a\in S} l_a\right\}\right\}.$$

[13] This is the average cost incurred in previous implementations of solution S.

[14] As the effect of truncation in UCB1+ and Extended UCB1+ produces mixed results in practice, our experiments consider the point-wise minimum regret among the policies with and without truncation.

Note that the policy incorporates many of ideas in Section 3.3, thus its performance should be comparable to that of the simple policy. In a similar setting, Gai et al. (2012) propose another adaptation of UCB1: a modified version of such a policy implements

$$S_n \in \operatorname*{argmin}_{S \in \mathcal{S}} \left\{ \sum_{a \in S} \max \left\{ \bar{b}_{a,n} - \sqrt{(K+1)\ln(n-1)/T_n(\{a\})}, l_a \right\} \right\}$$

for instance n, for some positive finite constant K. We denote this policy as Extended UCB1+. Note that the performance bound in Chen et al. (2013) compares rather unfavorably to that in Theorem 3.4 in settings with "highly" combinatorial solution sets.[15] Our experiments test whether such an ordering is preserved in practice. We test the performance of the OCP-based (adaptive), static cover-based (simple) and dynamic cover-based (dynamic cover) policies, as well as that of the version of the OCP-based policy that solves OCP heuristically using Algorithm 2 (heuristic), in addition to Extended UCB1+ and UCB1+.

It is worth mentioning that all solutions to R-OCP are integral in our experiments. While we do not observe a gap between OCP and its relaxation in the settings of this Section, in general one can present counter-examples where that is not the case.

Implementation Details. We report results when the marginals of F are exponential (we normalize the mean costs of the ground elements so that the maximum solution cost is at most one): we tested many cost distributions and observed consistent performance. All policies start with an initialization phase in which each solution in a common minimum size cover of A is implemented. We report results where $H = 5$: preliminary tests using $H = \{5, 10, 20\}$ always resulted in logarithmic regrets. For the simple policy, we use an arbitrary minimum size cover of A to perform exploration. For the adaptive policy, we updated the exploration set on each cycle. We implemented UCB1+ and Extended UCB1+ with and without truncating indices at the implied lower bounds. Here, we present the point-wise minimum regret among both versions of each policy. Finally, we set $K = 1$ 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. The figures in this section report average performance for $N = 2000$ over 100 replications, and dotted lines represent 95% confidence intervals.

All policies were implemented in MATLAB R2011b. Shortest path problems were solved using Dijkstra's algorithm except when implementing UCB1+ (note that because of the index computation, $f(\cdot)$ must be solved by enumeration). For Steiner tree and knapsack problems, we solved standard IP formulations using GUROBI 5.0 Optimizer. The adaptive policy solves formulation (6) 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

[15] Such a $O(s^2 |A| \ln N)$ performance bound, however, does not apply directly to this modification. Nonetheless, it stands to reason that such bounds would remain valid in practice.

ranged from less than 5 seconds for simple policy to around 1.5 minutes for the adaptive policy.[16]

6.1.2 Settings and Results

The settings are comprised of the shortest path problems in Examples 3.5, 4.5 and 4.7 for $k = 3$ (as shown in Figure 1), $k = 20$ and $k = 20$, respectively, followed by randomly generated instances (structures and costs) of shortest path, Steiner tree and knapsack problems. We observed consistent performance of our policies across these settings: here we only show a representative from each class. The random settings are complementary to Examples 3.5 and 4.5 in that the optimal critical subsets are large and hence the OCP-based policy does not have an immediate advantage.

Examples 3.5, 4.5 and 4.7. Figure 4 depict the average performance of six different policies on Examples 3.5 (left), 4.5 (center) and 4.7 (right), respectively.

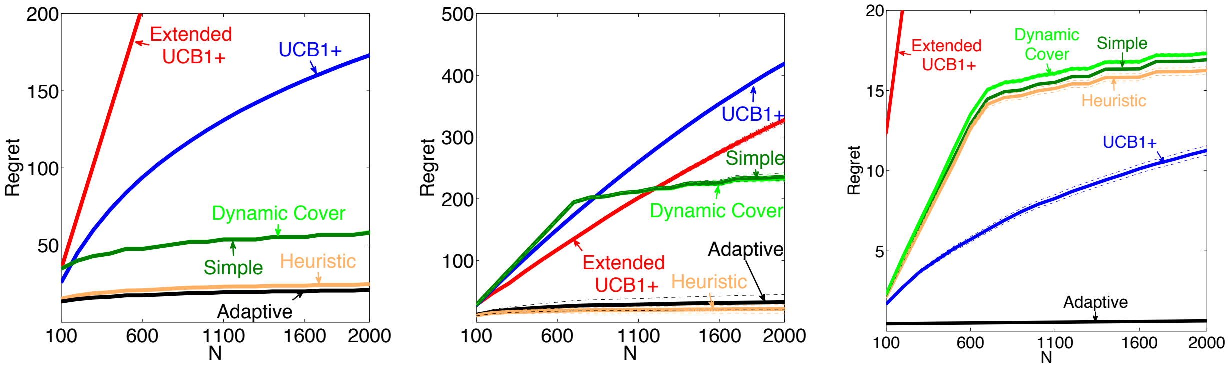

Figure 4: Average performance of different policies on Examples 3.5 (left), 4.5 (center) and 4.7 (right).

On Example 3.5, the adaptive and heuristic policies perform significantly better than the other policies, as they successfully limit exploration to feedback-consistent solution sets. The static cover policy provides a slightly better performance than its dynamic counterpart (the minimum size cover has 4 elements). The situation is essentially the same on Example 4.5, only that this time Extended UCB1+ initially outperforms the simple and dynamic cover policies (recall that in this setting, the minimum size cover is equal to $\mathcal{S}$, which has size 22). In contrast, minimum regret exploration set of the adaptive policy is only of size 2, which helps it achieve the best performance. (Note that for this setting, the heuristic solution to OCP tends to find the actual optimal solution, even with unreliable estimates.) On Example 4.7, the heuristic solution to OCP coincides with the minimum regret cover of $\mathcal{S}$, thus the performance of heuristic coincides with those of the cover-based policies, which in turn are outperformed by UCB1+ (note that this latter policy rarely uses the arcs p_2 and

[16] Note, however, that while the running times of simple and adaptive policies grow (roughly) logarithmically with the horizon, those of UCB1+ and Extended UCB1+ grow linearly.

q_2, since the costs of p_1 and q_1 close to 0).

In terms of efficient information collection, one can divide the set of ground elements (arcs) into three classes: those that are part of the optimal solution (Optimal arcs), those that are covered by at least one optimal solution to $OCP(\mathbb{E}_F\{B_n\})$ (exploration arcs), and the rest (*uninformative* arcs). Table 1 shows the average number of times each type of arc is tested up to horizon $N = 2000$ by each policy. Note that the adaptive and heuristic policies spend significantly less time exploring uninformative arcs.

	Example 3.5			Example 4.5			Example 4.7		
	Opt. Arcs	Exp. Arcs	Uninf. Arcs	Opt. Arcs	Exp. Arcs	Unin. Arcs	Opt. Arcs	Exp. Arcs	Unin. Arcs
Adaptive	1958.93	470.67	2.25	1858.25	548.12	4.55	140.03	214.50	1.00
Heuristic	1951.62	472.18	3.38	1918.43	524.20	3.32	106.83	215.94	35.71
Dyn. Cover	1885.88	482.03	38.00	1159.20	734.66	37.15	119.47	214.68	38.09
Simple	1886.52	481.91	37.81	1128.79	749.88	37.15	142.95	212.59	37.19
UCB1+	1660.75	533.35	42.12	474.31	929.80	66.61	92.45	217.75	24.61
Ext. UCB1+	791.31	684.36	364.72	870.88	795.78	53.76	14.87	219.02	151.79

Table 1: Average number of trials of different arcs up to horizon $N = 2000$ over different policies on Examples 3.5, 4.5 and 4.7.

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

Knapsack problem. Here the set A represents items that might go into the knapsack. 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 minimum regret exploration set is of size 8 with an implied critical subset of size 17. The right panel in Figure 5 depicts the average performance of different policies on the representative for the knapsack setting. We see that the adaptive policy outperform the benchmark, with the heuristic and dynamic cover policies being close seconds.

Minimum Steiner tree problem. We consider a generalized version of the Steiner tree problem (Williamson and Shmoys 2011), where for a given undirected graph with non-negative edge costs

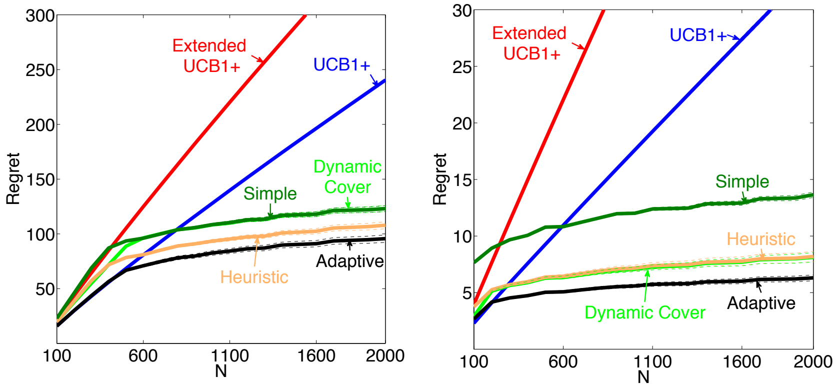

Figure 5: Average performance of different policies on the representative from the shortest path (left) and knapsack (right) settings.

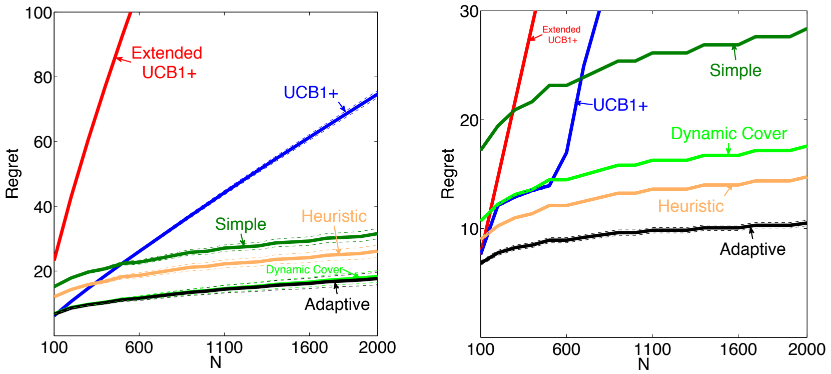

Figure 6: Average performance of different policies on the representative from the Steiner tree setting with zero (left) and positive (right) lower bounds.

and a set of pairs of vertices, the objective is to find a minimum cost subset of edges (tree) such that every given pair is connected in the set of selected edges. The graphs as well as the pairs of vertices are generated randomly, as well as the mean cost values. The representative setting is such that $|A| = 18$, $|\mathcal{S}| = 10651$, the minimum size cover is of size 2. The left panel in Figure 6 depicts average performance when all cost lower bounds are set to zero. In this representative setting, we have that the minimum regret exploration set is of size 7 with an implied critical subset of size 17. In this case, all arcs (but those trivially suboptimal) are critical, thus the dynamic cover policy is essentially equivalent to the adaptive policy, which is corroborated by our results. The right panel in Figure 6 depicts average performance when lower bounds are chosen randomly.[17] The representative setting is such that the minimum regret exploration set is of size 5 with an implied critical subset of size 12. Note that the adaptive policy outperforms the benchmark as it successfully limits exploration to a critical set.

[17] 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 $n = 1200$.

6.2 Short-term Experiments

6.2.1 Benchmark Policies and Implementation

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

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

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

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

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

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

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

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

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_n^{Gitt}, where

$$S_n^{Gitt} \in \underset{S \in \mathcal{S}}{\operatorname{argmin}} \left\{ \sum_{a \in S} g_{n,N}^{a}(\mu_{a,n}, \lambda_{a,n}, \alpha_{a,n}, \beta_{a,n}) \right\}.$$

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

6.2.2 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 adaptive policy begins outperforming the benchmark. In all settings, the benchmark policies initially provide a better performance compared to the adaptive policy, but the latter policy eventually surpasses the benchmarks for moderate values of N. The same holds true for the case of the heuristic policy.

Shortest path problem. The left panel at Figure 7 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 minimum regret exploration set is of size 10 with an implied critical subset of size 23.

Minimum Steiner tree problem. The central panel at Figure 7 depicts the average performances on a representative from the Steiner tree setting. The representative setting is such that $|A| = 9$, $|\mathcal{S}| = 50$, the minimum size cover is of size 2, and the minimum regret exploration set is of size 4 with an implied critical subset of size 8.

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

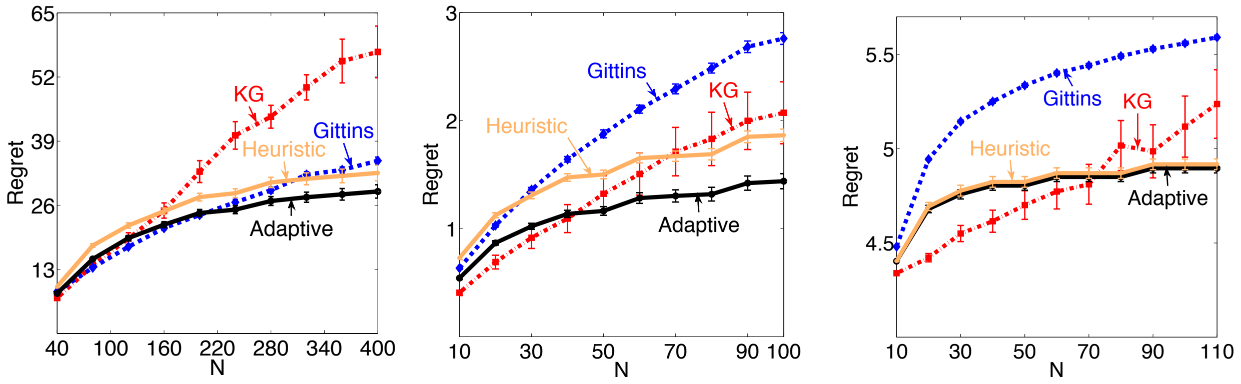

Figure 7: Average performance of different policies on the representative from the shortest path (left), Steiner tree (center) and knapsack (right) settings.

7 Final Remarks and Extensions

In this paper we have studied a class of sequential decision-making problems where the underlying single-period decision problem is a combinatorial optimization problem, and there is initial uncertainty about its objective coefficients. By framing the problem as a *combinatorial* multi-armed bandit, we have adapted key ideas behind results in the classical bandit setting to develop asymptotically efficient policies, and gave theoretical and practical evidence that its performance is *near-optimal* among practical policies. In doing so, we have shown that in addition to answering the question of *when* (with what frequency) to explore, which is key in the traditional setting, in the combinatorial setting one must also answer the questions of *what* and *how* to explore. We answer such questions by explicitly solving for the cheapest optimality guarantee for the optimal solution to the underlying combinatorial problem (i.e. by solving OCP). We have shown evidence that the proposed policies are scalable and implementable in practice, and our numerical experiments show they perform reasonably well relative to relevant benchmark, both in the short- and long-term.

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

$$\sum_{S \in \mathcal{E}} \gamma_S(a) I_S = I_{\{a\}}. \tag{11}$$

Without loss of generality, one can assume that each ground element is recovered by at least one solution set.[18] Let $\mathcal{E}$ be a solution set that recovers A, and let $\gamma := (\gamma(a), a \in A)$ be such that $\sum_{S\in\mathcal{E}} \gamma_S(a) I_S = I_{\{a\}}$, for all $a \in A$. One can implement a cover-based policy with $\mathcal{E}$ playing the role of a cover while replacing the estimate in (5) with

$$\bar{b}_{a,n} := \sum_{S\in\mathcal{E}} \frac{\gamma_S(a)}{T_n(S)} \sum_{m<n:S_m=S} \sum_{a\in S} b_{a,m}, \quad a \in A. \tag{12}$$

The estimate above reconstructs the expected cost of each solution in $\mathcal{E}$ and uses (11) to translate such estimates to the ground-element level. Implementing this modification requires precomputing a solution set $\mathcal{E}$ recovering A. Such a set can be selected so that $|\mathcal{E}| \leq |A|$, and computed by solving $O(|A|)$ instances of $f(\cdot)$ (see e.g., the algorithm in Awerbuch and Kleinberg (2004)). A close inspection to the proof of Theorem 3.4 reveals that its performance guarantee would remain valid (modulo changes to constants) after incorporating the new estimation procedure.

The idea above can also be used to extend the OCP-based policy to this new setting. In particular, Algorithm 1 would consider the estimates in (12) and $(C, \mathcal{E})$ to be solution to an alternative version of OCP where in addition to (6b)-(6d), one imposes that $\mathcal{E}$ recovers C, that is

$$\begin{aligned}
OCP'(B): \quad \min \quad & \sum_{S\in\mathcal{S}} \Delta_S^B \, y_S && \text{(13a)}\\
s.t. \quad & \sum_{S\in\mathcal{S}} \gamma_S(a) I_S = x_a I_{\{a\}}, \quad a \in A && \text{(13b)}\\
& \gamma_S(a) \leq Q \, y_S, \quad S \in \mathcal{S},\, a \in A && \text{(13c)}\\
& -\gamma_S(a) \leq Q \, y_S, \quad S \in \mathcal{S},\, a \in A && \text{(13d)}\\
& \sum_{a\in S} (l_a(1-x_a) + b_a x_a) \geq z^*(B), \quad S \in \mathcal{S} && \text{(13e)}\\
& x_a,\, y_S \in \{0,1\},\, \gamma_S(a) \in \mathbb{R}, \quad a \in A, S \in \mathcal{S}, && \text{(13f)}
\end{aligned}$$

where Q is an instance-dependent constant, whose size is polynomial in the size of the instance. The additional constraints(13b)-(13d) in OCP' ensure that the solution set $\mathcal{E}$ recovers the critical subset C. Like OCP, the formulation above can be specialized to accommodate the combinatorial structure of $f(\cdot)$ (as shown in Section 5.2). The performance guarantee in Theorem 4.3 would remain valid with the constants associated to OCP'. We anticipate that the challenge of solving OCP' effectively is comparable to that of solving OCP.

[18] If this is not the case, then it must be that a appears in a solution if and only if that solution also includes some ground element a'. Thus, one can (w.l.o.g.) combine such ground elements into a single element. Alternatively, one can assign costs only to one of such elements, so as to not modify the combinatorial structure of the solution set.

8 Acknowledgments

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

A Omitted Proofs and Complementary Material

A.1 Appendix for Section 3

The static cover-based policy is detailed in Algorithm 3. Next, we prove its performance bound.

Algorithm 3 Static cover-based policy $\pi_s(\mathcal{E})$

```
Set $i = 0$, and $\mathcal{E}$ a minimal cover of $A$
for $n = 1$ to $N$ do
  if $n \in \Phi$ then
    $i \leftarrow i + 1$, and set $S^* \in \mathcal{S}^*\left(\bar{B}_n\right)$        [Update exploitation set]
  end if
  if $T_n(\{a\}) < i$ for some $a \in S$, for some solution $S \in \mathcal{E}$ then
    Implement such a solution, i.e., set $S_n = S$        [Exploration]
  else
    Implement $S_n = S^*$        [Exploitation]
  end if
end for
```

Theorem 3.4. *For any cover $\mathcal{E}$, let $\pi_s(\mathcal{E})$ denote static cover-based policy and for an arbitrary $\delta > 1$ let $H := (1+\delta)\left(s/\Delta^F_{\min}\right)^2$, where $s := \max\{|S| : S \in \mathcal{S}\}$ and $\Delta^F_{min} := \min\left\{\Delta^F_S \,:\, \Delta^F_S > 0\,,\, S \in \mathcal{S}\right\}$. If we choose $n_i := \max\left\{\lfloor e^{i/H}\rfloor, n_{i-1}+1\right\}$, for all $i \geq 2$, then*

$$\frac{R^{\pi_s(\mathcal{E})}(F,N)}{\ln N} \leq (1+\delta)\frac{C}{\left(\Delta^F_{\min}\right)^2}s^2 + O(1/\ln N) \leq (1+\delta)\frac{\Delta^F_{max}}{\left(\Delta^F_{\min}\right)^2}|\mathcal{E}|\, s^2 + O(1/\ln N),$$

where $\Delta^F_{max} := \max\left\{\Delta^F_S \,:\, S \in \mathcal{S}\right\}$, and $C := \sum_{S\in\mathcal{E}} \Delta^F_S$. If instead we choose $n_i := \max\{\lfloor e^{i^{1/(1+\varepsilon)}}\rfloor, n_{i-1}+1\}$, with $\varepsilon > 0$ arbitrary, for all $i \geq 2$, then

$$\frac{R^{\pi_s(\mathcal{E})}(F,N)}{(\ln N)^{1+\varepsilon}} \leq C + O(1/(\ln N)^{1+\varepsilon}) \leq \Delta^F_{max}|\mathcal{E}| + O\left(1/(\ln N)^{1+\varepsilon}\right).$$

Proof. First, we prove the result for the case $n_i = \left\{\lfloor e^{i/H}\rfloor, n_{i-1+1}\right\}$. The regret of the simple policy $\pi_s(\mathcal{E})$ stems from two sources: exploration and errors during exploitation. That is,

$$R^{\pi_s(\mathcal{E})}(F,N) = \sum_{S\in\mathcal{S}} \Delta^F_S\ \mathbb{E}_F\left\{T_{N+1}(S)\right\} = R_1^{\pi_s(\mathcal{E})}(F,N) + R_2^{\pi_s(\mathcal{E})}(F,N), \tag{A-1}$$

where $R_1^{\pi_s(\mathcal{E})}(F,N)$ is the exploration-based regret, i.e., that incurred while $T_n(\{a\}) < i$ for some $a \in A$ at instance n in cycle i, and $R_2^{\pi_s(\mathcal{E})}(F,N)$ is the exploitation-based regret, i.e., that incurred when $T_n(\{a\}) \geq i$ for all $a \in A$. We prove the result by bounding each term above separately.

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

Step 1 (Exploration-based regret). By construction, $\pi_s(\mathcal{E})$ implements each solution $S \in \mathcal{E}$ at most $\lceil H \ln N \rceil$ while exploring. Therefore

$$R_1^{\pi_s(\mathcal{E})}(F,N) \leq C(H \ln N + 1) \leq |\mathcal{E}| \, \Delta_{max}^F (H \ln N + 1). \tag{A-2}$$

Step 2 (Exploitation-based regret). Exploitation-based regret during cycle i is due to implementing suboptimal solutions when all elements in A have been tried at least on i instances.

Let $i' := \inf \{ i \in \mathbb{N}, i \geq 2 : n_i \geq i \, |\mathcal{E}| \, , n_{i+1} - n_i > |\mathcal{E}| \}$ denote the first cycle in which one is sure to exploit on at least one instance. Note that i' does not depend on N, thus neither does the exploitation-based regret prior to cycle i'.

Fix $i \geq i'$. With some abuse of notation, for $n \in [n_i, n_{i+1} - 1]$, let $\bar{S}_n \in \mathcal{S}^*(\bar{B}_n)$ be any solution with minimum average cost at time n_i. We have that

$$\begin{aligned} R_2^{\pi_s(\mathcal{E})}(F,N) \;&\leq\; n_{i'} \Delta_{max}^F + \mathbb{E} \left\{ \sum_{i=i'}^{\lceil H \ln N \rceil} \sum_{n=n_i}^{n_{i+1}-1} \mathbb{P} \left\{ \bar{S}_n \notin \mathcal{S}^* \left(\mathbb{E} \left\{ B_n \right\} \right) \, , \, T_n(\{a\}) \geq i \, , \, \forall \, a \in A \right\} \Delta_{\bar{S}_n}^F \right\} \\ &\leq\; n_{i'} \Delta_{max}^F + \sum_{i=i'}^{\infty} \sum_{n=n_i}^{n_{i+1}-1} \mathbb{P} \left\{ \bar{S}_n \notin \mathcal{S}^* (\mathbb{E} \left\{ B_n \right\}) \, , \, T_n(\{a\}) \geq i \, , \, \forall \, a \in A \right\} \Delta_{max}^F . \end{aligned} \tag{A-3}$$

Next we find an upper bound for the probability inside the sum in (A-3). For this, note that

$$\left\{ \bar{S}_n \notin \mathcal{S}^* (\mathbb{E} \left\{ B_n \right\}) \right\} \subseteq \left\{ |\bar{z}_n^* - \mathbb{E} \left\{ \bar{z}_n^* \right\}| \geq \frac{\Delta_{min}^F}{2} \right\} \cup \left\{ |\bar{z}_n - \mathbb{E} \left\{ \bar{z}_n \right\}| \geq \frac{\Delta_{min}^F}{2} \right\}, \tag{A-4}$$

where $\bar{z}_n := \sum_{a \in \bar{S}_n} \bar{b}_{a,n}$, $\bar{z}_n^* := \sum_{a \in S^*} \bar{b}_{a,n}$ for some $S^* \in \mathcal{S}^*(\mathbb{E} \left\{ B_n \right\})$, and

$$\Delta_{min}^F := \min \left\{ \Delta_S^F \, : \, \Delta_S^F > 0 \, , \, S \in \mathcal{S} \right\},$$

is the minimum optimality gap. Indeed, note that $\left\{ |\bar{z}_n^* - \mathbb{E} \left\{ \bar{z}_n^* \right\}| < \frac{\Delta_{min}^F}{2} \right\}$ and $\left\{ |\bar{z}_n - \mathbb{E} \left\{ \bar{z}_n \right\}| < \frac{\Delta_{min}^F}{2} \right\}$ implies that $\bar{z}_n > \bar{z}_n^*$.

The next proposition, whose proof can be found in Appendix B, allows us to bound (A-3) using the observation above.

Proposition A.1. *For any fixed $S \subseteq A$, $n \in \mathbb{N}$, $i \in \mathbb{N}$, and $\epsilon > 0$ we have that*

$$\mathbb{P}\left\{ \left| \sum_{a \in S} \left(\bar{b}_{a,n} - \mathbb{E}\left\{ b_{a,n} \right\} \right) \right| \geq \epsilon, T_n(\{a\}) \geq i \,, \forall a \in S \right\} \leq 2 \; K(\epsilon) \left| S \right| \exp\left\{ -\frac{2\epsilon^2 i}{\left| S \right|^2 \mathcal{L}^2} \right\},$$

where $\mathcal{L} := \max\left\{ u_a - l_a : a \in A \right\}$, and $K(\epsilon)$ is a positive finite constant that only depends on ϵ.

Define $s := \max\left\{ |S| : S \in \mathcal{S} \right\}$: using the above, one has that

$$\mathbb{P}\left\{ \left| \bar{z}^*_n - \mathbb{E}\left\{ \bar{z}^*_n \right\} \right| \geq \frac{\Delta^F_{min}}{2} \,,\, T_n(\{a\}) \geq i \,, \forall\, a \in A \right\} \leq$$
$$\sum_{S \in \mathcal{S}} \mathbb{P}\left\{ \left| \sum_{a \in S} \left(\bar{b}_{a,n} - \mathbb{E}\left\{ b_{a,n} \right\} \right) \right| \geq \frac{\Delta^F_{min}}{2} \,,\, T_n(\{a\}) \geq i \,, \forall\, a \in S \right\} \overset{(a)}{\leq} 2 \; Ks \left| \mathcal{S} \right| \exp\left\{ -\frac{{\Delta^F_{min}}^2 i}{2s^2 \mathcal{L}^2} \right\},$$

where K is a positive finite constant, and (a) follows from noting that $|S| \leq s$ for all $S \in \mathcal{S}$. Consider (A-4): applying Proposition A.1 and the above to the first and second terms on its right-hand side, respectively, one obtains

$$\mathbb{P}\left\{ \bar{S}_n \notin \mathcal{S}^*(\mathbb{E}\left\{ B_n \right\}) \,,\, T_n(\{a\}) \geq i \,, \forall\, a \in A \right\} \leq 4 \; Ks \left| \mathcal{S} \right| \exp\left\{ -C_1 i \right\}, \tag{A-5}$$

where $C_1 := {\Delta^F_{min}}^2 / (2s^2 \mathcal{L}^2)$. Note that this final bound does not depend on n but rather on i. Now, for $i \geq i'$, one has that $n_{i+1} \leq e^{(i+1)/H}$ and $n_i \geq e^{(i-1)/H}$, hence

$$n_{i+1} - n_i \leq C_2 \; e^{\frac{i}{H}}, \quad i \geq i',$$

where $C_2 := e^{1/H} - e^{-1/H}$. Using this latter fact, (A-3) and (A-5) we conclude that

$$R_2^{\pi_s(\mathcal{E})}(F, N) \leq n_{i'} \; \Delta^F_{max} + \sum_{i=i'}^{\infty} C_3 \exp\left\{ i \left(\frac{1}{H} - C_1 \right) \right\},$$

where $C_3 := 4 \; Ks \left| \mathcal{S} \right| \Delta^F_{max} C_2$. Because $H > 1/C_1$ we have that

$$R_2^{\pi_s(\mathcal{E})}(F, N) \leq C_4, \tag{A-6}$$

where C_4 is a positive finite constant. Combining (A-1), (A-2) and (A-6) we conclude that

$$R^{\pi_s(\mathcal{E})}(F, N) \leq CH \ln N + C_5 \leq \left| \mathcal{E} \right| \Delta^F_{max} H \ln N + C_5,$$

where C_5 is a positive finite constant. The result follows from the definition of H.

Consider now the case when $n_i := \max\{\lfloor e^{i^{1/(1+\varepsilon)}} \rfloor, n_{i-1}+1\}$. By construction, $\pi_s(\mathcal{E})$ implements each solution $S \in \mathcal{E}$ at most $\left\lceil (\ln N)^{1+\varepsilon} \right\rceil$ while exploring, thus

$$R_1^{\pi_s(\mathcal{E})}(F,N) \leq C\left((\ln N)^{1+\varepsilon}+1\right) \leq |\mathcal{E}|\, \Delta^F_{\max}\left((\ln N)^{1+\varepsilon}+1\right).$$

To bound $R_2^{\pi_s(\mathcal{E})}(F,N)$ note that $n_{i+1}-n_i \leq e^{(i+1)^{1/(\varepsilon+1)}}$ for $i \geq i'$. Also, let $i'' \geq i'$ be such that $i\,C_1/2 > (i+1)^{1/(1+\varepsilon)}$ for $i \geq i''$. The arguments in Step 2 above lead to the bound

$$R_2^{\pi_s(\mathcal{E})}(F,N) \leq n_{i''}\ \Delta^F_{max} + \tilde{C}_3 \sum_{i=i''}^{\infty} e^{(i+1)^{1/(1+\varepsilon)}-i\,C_1} \leq n_{i''}\ \Delta^F_{max} + \tilde{C}_3 \sum_{i=i''}^{\infty} e^{-C_1\,i/2} = \tilde{C}_4,$$

where $\tilde{C}_3$ and $\tilde{C}_4$ are finite positive constants. Then, we have that

$$R^{\pi_s(\mathcal{E})}(F,N) \leq C\,(\ln N)^{1+\varepsilon} + \tilde{C}_5 \leq |\mathcal{E}|\,\Delta^F_{max}\,(\ln N)^{1+\varepsilon} + \tilde{C}_5,$$

where $\tilde{C}_5$ is a positive finite constant. □

A.2 Appendix for Section 4

A.2.1 Performance Bound Comparisons

Proposition 4.6. *If $f(B)$ corresponds to a shortest path, minimum cost spanning tree, minimum cost perfect matching, generalized Steiner tree or knapsack problem, then there exists a family of instances where G is arbitrarily smaller than a minimum size cover of A.*

Proof. For shortest path problems the family of instances is that from Example 4.5, which is parametrized by an integer k. For this family of instances we have that the unique cover of A is given by all $s-t$ paths which is of order k. In contrast, solutions in $\Gamma(B)$ include path $\{e\}$ and any other path, hence their size is 2, independent of k.

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

For minimum cost perfect matching consider a complete graph $G=(V,A)$ with $|V|=2k$ nodes, $b_a=0$ for all $a \in \{(2i+1,2i+2) : i<k\}$ and $l_a = M > 0$ for all $a \notin \{(2i+1,2i+2) : i<k\}$.

One can check that any cover of A is of size at least $2(k-1)$. In contrast, solutions in $\Gamma(B)$ are of size 1, independent of k.

Finally, for the knapsack problem, consider the items $A := \{0, \ldots, Ck\}$, 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 $b_0 := 0$ and $u_i := -M < 0$ for $i > 0$. Note that in this case the problem is of maximization. One can check that any cover of A is of size at least $k+1$. In contrast, solutions in $\Gamma(B)$ are of size 1, independent of k. □

Lemma A.2. *Consider a matroid on a set A. Let $R \subseteq A$, $\{a_1, \ldots, a_k\} \subseteq R$ and $b \in A \setminus R$ be such that for all $i \leq k$, $R \setminus \{a_1, \ldots, a_{i-1}\}$ has a circuit containing $a_i$and $R \cup \{b\}$ has a circuit containing b. Then $R \setminus \{a_1, \ldots, a_k\} \cup \{b\}$ has a circuit containing b.*

Proof. We show by induction on i that $R \setminus \{a_1, \ldots, a_{i-1}\} \cup \{b\}$ has a circuit containing b for all $i \leq k$. The base case is straightforward as it simply states that $R \cup \{b\}$ has a circuit containing b. For the inductive step assume for $i \leq k$ that there exists a circuit $C \subseteq R \setminus \{a_1, \ldots, a_{i-1}\} \cup \{b\}$ such that $b \in C$. If $a_i \notin C$ then $C \subseteq R \setminus \{a_1, \ldots, a_i\} \cup \{b\}$. If $a_i \in C$ note that by the lemma's assumptions there exists a circuit $C' \subseteq R \setminus \{a_1, \ldots, a_{i-1}\}$ such that $a_i \in C'$. Then $a_i \in C \cap C'$ and $b \in C \setminus C'$. Then, by Schrijver (2003, Theorem 39.7) there exists a circuit $\bar{C} \subseteq (C \cup C') \setminus \{a_i\}$ such that $b \in \bar{C}$. We conclude the induction hypothesis holds for $i+1$ by noting that $\bar{C} \subseteq R \setminus \{a_1, \ldots, a_i\} \cup \{b\}$. □

Lemma 4.8. *Let $f(\cdot)$ be a weighted basis or independent set matroid minimization problem. Then, for $B \in \mathbb{R}^{|A|}$ in the range of F, $\bigcup_{S \in \mathcal{E}} S \subseteq C$ for all $(C, \mathcal{E}) \in \Gamma^*(B)$.*

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

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

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

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

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

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

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

We now show that for any B, $\bigcup_{S\in\mathcal{E}} S \subseteq C$ for all $(C,\mathcal{E}) \in \Gamma^*(B)$. For this, suppose that there exists a solution $(C,\mathcal{E}') \in \Gamma^*(B)$ such that there exists $S' \in \mathcal{E}'$ with $S' \setminus C \neq \emptyset$. We will show the result by noting that finding a minimum cost independent set or basis S such that $C \cap S' \subseteq S$ is achieved by greedily adding elements to $C' := C \cap S'$ (because $C' \subseteq S'$, it is an independent set), and by proving that such a procedure does not add elements in $A \setminus \bigcup_{S\in\mathcal{S}^*(B)} S$, thus contradicting the optimality of $(C,\mathcal{E}')$. Indeed, suppose that this is not the case and let a_k be the first one of these elements that is added (w.l.o.g. assume that $A = \left\{a_1,\ldots,a_{|A|}\right\}$ is such that $b_{a_1} \leq b_{a_2} \leq \ldots \leq b_{a_{|A|}}$). Define $D_j := \left\{a_i \in \bigcup_{S\in\mathcal{S}^*(B)} S : i \leq j\right\}$ for $j \leq k$. Because every solution in $\mathcal{S}^*(B)$ can be constructed through the greedy algorithm and $a_k \in A \setminus \bigcup_{S\in\mathcal{S}^*(B)} S$ we have that $D_k \cup \{a_k\}$ must have a circuit containing a_k. Hence, the only way the greedy augmentation of C' may add a_k is if it skipped some elements of D_k (which may happen as we were forced to start with elements of C'). Let $m \geq 1$ and $i_1,\ldots,i_m < k$ be such that $a_{i_j} \in D_k$ for all $j \leq m$ and such that the elements of D_k skipped by the greedy extension of C' are precisely $\{a_{i_1},\ldots,a_{i_m}\}$ (note that by the assumption on a_k, all elements in $\{a_1,\ldots,a_{k-1}\}$ picked by the greedy algorithm are the original elements in C' or elements in D_k). Hence for every $j \leq m$, $C' \cup D_{i_j-1} \setminus \left\{a_{i_1},\ldots,a_{i_{j-1}}\right\}$ has a circuit containing a_{i_j}. Then, by Lemma A.2 we have that $C' \cup D_k \setminus \{a_{i_1},\ldots,a_{i_m}\} \cup \{a_k\}$ has a circuit containing a_k, which contradicts the fact that the greedy augmentation of C' added a_k. This concludes the proof. □

A.2.2 A Limit on Achievable Performance

Proposition 4.9. *For any consistent policy π, regular F, and $D \in \mathcal{D}$ we have that*

$$\lim_{N\to\infty} \mathbb{P}_F\left\{\frac{\max\{T_{N+1}(\{a\}) : a \in D\}}{\ln N} \geq K_D\right\} = 1, \tag{7}$$

where K_D is a positive finite constant depending on F.

Proof. We begin by imposing some structure on F.

Preliminaries. We assume F_a, the *common* distribution of $b_{a,n}$, $n \in \mathbb{N}$, is uniformly continuous with respect to Lebesgue measure in $\mathbb{R}$ and let f_a denote its density function. To simplify the notation we assume that these functions accept parametric representations, and let θ_a denote the "true" parameter for $f_a, a \in A$. Finally, we let Θ_a denote the set of feasible parameters for $f_a, a \in A$. These *mild* conditions are fulfilled by most commonly used distribution functions.

For $\lambda_a \in \Theta_a$, let $I_a(\theta_a, \lambda_a)$ denote the Kullback-Leibler distance between $F_a(\cdot;\theta_a)$ and $F_a(\cdot;\lambda_a)$, i.e.,

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

Define $b_a(\lambda_a) := E_{F_a(\cdot:\lambda_a)}\{b_{a,n}\}$, $n \in \mathbb{N}$. In addition to the conditions above, we assume that $0 < I_a(\theta_a, \lambda_a) < \infty$ whenever $b_a(\theta_a) > b_a(\lambda_a)$. This *indistinguishability* condition implies that distributions with different mean costs are not distinguishable based on a finite sample. Finally, define $\theta = (\theta_a : a \in A)$ and let $\mathbb{E}_\lambda$ and P_λ denote the expectation and probability induced when F receives the parameter $\lambda := (\lambda_a : a \in A) \in \mathbb{R}^{|A|}$. Also, define $\mathcal{S}^*_\lambda := \mathcal{S}^*(\mathbb{E}_\lambda\{B_n\})$.

Proof of the result. Consider $D \in \mathcal{D}$ as defined in Section 4.4, and take $\lambda \in \mathbb{R}^{|A|}$ so that $\lambda_a = \theta_a$ for $a \notin D$, and that $D \subseteq S^*$ for all $S^* \in \mathcal{S}^*_\lambda$. By the consistency of π, one has that

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

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

$$\mathbb{E}_\lambda \{N - \max\{T_{N+1}(\{a\}) : a \in D\}\} \leq \mathbb{E}_\lambda \left\{ N - \sum_{S^* \in \mathcal{S}^*_\lambda} T_{N+1}(S) \right\} = o(N^\alpha). \tag{A-7}$$

Take $\epsilon > \alpha$. Define $I(D, \lambda) := |D| \max\{I_a(\theta_a, \lambda_a) : a \in D\}$, $D \in \mathcal{D}$. Applying Markov's inequality to $N - \max\{T_{N+1}(\{a\}) : a \in D\}$ and using (A-7), one has that

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

The above can be re-written as

$$\mathbb{P}_\lambda \left\{ \max \left\{ T_{N+1}(\{a\}) : a \in D \right\} < \frac{(1-\epsilon)\ln N}{I(D,\lambda)} \right\} = o(N^{\alpha-1}). \tag{A-8}$$

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

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

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

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

and note that

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

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

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

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

From above and (A-8) we have that

$$\lim_{N\to\infty} \mathbb{P}_\theta \left\{ \Xi(N) \right\} \leq \lim_{N\to\infty} N^{1-\alpha}\ \mathbb{P}_\lambda \left\{ \Xi(N) \right\} = 0. \tag{A-9}$$

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

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

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

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

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

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

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

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

which in turn implies

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

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

$$\lim_{N\to\infty} \mathbb{P}_\theta \left\{ L_{T_{N+1}(\{a\})}(a) > \frac{(1-\alpha)\ln N}{|D|} \text{ for some } a \in D\,, \max\{T_{N+1}(\{a\}) : a \in D\} < \frac{(1-\epsilon)\ln N}{I(D,\lambda)} \right\} = 0. \tag{A-10}$$

Thus, by (A-9), (A-10), and the definition of $\Xi(N)$ we have that

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

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

Proposition 4.10. *For any consistent policy π and any regular F we have that*

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

Proof. Define $\Upsilon_N := \bigcap_{D\in\mathcal{D}} \{\max\{T_{N+1}(\{a\}) : a \in D\} \geq K_D \ln N\}$ and note that $\zeta^\pi(F,N) \geq L(F)\ln N$ when Υ_N occurs, because $\left(x_a = \frac{T_{N+1}(\{a\})}{\ln N}, a \in A \right)$ and $\left(y_S = \frac{T_{N+1}(S)}{\ln N}, S \in \mathcal{S} \right)$ are feasible to LBP. Thus, one has that

$$\begin{aligned}\mathbb{P}_F\left\{\frac{\zeta^\pi(F,N)}{\ln N}<L(F)\right\} &= \mathbb{P}_F\left\{\frac{\zeta^\pi(F,N)}{\ln N}<L(F)\,,\,\Upsilon_N\right\}+\mathbb{P}_F\left\{\frac{\zeta^\pi(F,N)}{\ln N}<L(F)\,,\,\Upsilon_N^c\right\}\\ &\le \mathbb{P}_F\left\{\Upsilon_N^c\right\}. \qquad\qquad (A\text{-}11)\end{aligned}$$

From Proposition 4.9 and the union bound, we have that

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

because $|\mathcal{D}|<\infty$. Thus, taking the limit in (A-11) we have that

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

This concludes the proof. □

Lemma 4.12. *Let $K_D=K$ for some constant $K>0$, for all $D\in\mathcal{D}$ in formulation LBP. An optimal solution (x,y) to R-OCP$(\mathbb{E}_F\{B_n\})$ is also optimal to LBP.*

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

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

for $S^*\in\mathcal{S}^*((\mathbb{E}_F\{B_n\})_{A\setminus D})$, where (a) follows from the fact that $l_a=(l_a(1-x_a)+\mathbb{E}_F\{b_{a,n}\}x_a)$, for $a\in D$, and $\mathbb{E}_F\{b_{a,n}\}\ge(l_a(1-x_a)+\mathbb{E}_F\{b_{a,n}\}x_a)$, for $a\notin D$, and (b) follows from the fact that (x,y) satisfies constraints (6c) (because it is feasible to R-OCP). The last inequality above contradicts $z^*((\mathbb{E}_F\{B_n\})_{A\setminus D})<z^*(\mathbb{E}_F\{B_n\})$, thus we have that $\max\{x_a:a\in D\}=1$ for all $D\in\mathcal{D}$, therefore (x,y) is feasible to (8).

Now, let (x,y) be a feasible solution to (8) such that $x_a\in\{0,1\}$ for all $a\in A$, and that $x_a=1$ and $y_{S^*}=1$ for $a\in S^*$ and $S^*\in\mathcal{S}^*(\mathbb{E}_F\{B_n\})$ (note that one can restrict attention only to feasible solutions to (8) with x integral, and $\Delta^F_{S^*}=0$ for all $S^*\in\mathcal{S}^*(\mathbb{E}_F\{B_n\})$, thus this extra requirement does not affect the solution to (8)). Suppose (x, y) is not feasible to R-OCP, i.e., there exists some $S \in \mathcal{S}$ such that

$$\sum_{a \in S} (l_a(1 - x_a) + \mathbb{E}_F \{b_{a,n}\} x_a) < z^*(\mathbb{E}_F \{B_n\}). \tag{A-12}$$

Let S_0 be one such S that additionally minimizes the left-hand side in (A-12) (in case of ties we pick any minimizing solution S_0 with smallest value of $|\{a \in S_0 : x_a = 0\}|$). Then $D = \{a \in S_0 : x_a = 0\}$ belongs to $\mathcal{D}$ (or a subset of it does), This contradicts the feasibility of (x, y) to (8), because if (x, y) is feasible to (8), then we must have $\max \{x_a : a \in D\} \geq 1$ for all $D \in \mathcal{D}$. Thus, we conclude that (x, y) is feasible to R-OCP.

Summarizing, feasible solutions to R-OCP are feasible to (8), and feasible solutions to (8) that cover all optimal elements in A are feasible to R-OCP. The result follows from noting that there is always an optimal solution to (8) such that x is integral, and $x_a = 1$ and $y_{S^*} = 1$ for $a \in S^*$ for all $S^* \in \mathcal{S}^*(\mathbb{E}_F \{B_n\})$. □

A.2.3 A R-OCP/cover-based hybrid policy

A feasible solution (x, y) to R-OCP corresponds to incidence vectors of a sufficient exploration set $C \subseteq A$, a solution cover $\mathcal{E}$ of such a set, and vector p of fractions associated with solutions in $\mathcal{E}$. That is, $C = \{a \in A : x_a = 1\}$, $\mathcal{E} = \{S \in \mathcal{S} : y_S > 0\}$, and $p = \{y_S : S \in \mathcal{E}\}$. In the following we refer to solutions to R-OCP as triplets $(C, \mathcal{E}, p)$.

Define $\tilde{\Gamma}^*(B)$ as the set of optimal solutions to R-OCP, and $\hat{\Gamma}^*(B)$ as the set of minimum regret covers of A, for B feasible. Algorithm 4 presents a hybrid policy that focuses feedback collection mainly on the solution to R-OCP.

A.3 Appendix for Section 5

The time-constrained Asynchronous Policy in Section 5.1 is depicted in Algorithm 5. A variant of Algorithm 5 based on *R-OCP* is simply obtained by modifying Algorithm 4 along the same lines.

A.3.1 Complexity of OCP and Basic MIP Formulation

To prove Theorem 5.1 and Proposition 5.3 we will use the following lemma.

Algorithm 4 hybrid policy $\pi_h(\gamma)$

```
Set $i = 0$, $C = A$, $\mathcal{E}$ a minimal cover of $A$, $\mathcal{G} = \mathcal{E}$ and $p_S = 1$ for all $S \in \mathcal{E}$
for $n = 1$ to $N$ do
  if $n \in \Phi$ then
    Set $i = i + 1$
    Set $S^* \in \mathcal{S}^*\left(\bar{B}_n\right)$,                                   [Update exploitation set]
    Set $(C, \mathcal{E}, p) \in \tilde{\Gamma}^*\left(\bar{B}_n\right)$ and $\mathcal{G} \in \hat{\Gamma}^*\left(\bar{B}_n\right)$          [Update exploration set]
  end if
  if $T_n(\{a\}) < \gamma\, i$ for some $a \in A$ then
    Set $S_n = S$ with $S \in \mathcal{G}$ such that $a \in S$                   [Cover-based exploration]
  else if $\mathcal{E} \neq \emptyset$ then
    Set $S_n = S \in \mathcal{E}$ with probability $(1-\gamma)p_S$, and $S^*$ otherwise    [R-OCP-based exploration]
    Set $\mathcal{E} = \mathcal{E} \setminus \{S_n\}$.
  else
    Implement $S_n = S^*$                                                   [Exploitation]
  end if
end for
```

Lemma A.3. *We may restrict OCP or R-OCP to have at most $|A|$ non-zero y_S variables without changing the problems.*

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

For R-OCP consider an optimal solution with more than $|A|$ non-zero y_S variables. Fixing the x_a variables in formulation (6) for R-OCP (e.g. (6) with $y_S \in \{0, 1\}$ changed to $0 \leq y_S \leq 1$) to the values of this solution and leaving the y_S variables free, we obtain a Linear Programming (LP) problem with at most $|A|$ constraints. Any basic optimal solution to this LP gives an alternate optimal solution to R-OCP with at most $|A|$ non-zero y_S variables. □

Theorem 5.1. *If $f(B)$ is in P, then OCP(B) and R-OCP(B) are in NP.*

Proof. By Lemma A.3, optimal solutions to OCP and R-OCP have sizes that are polynomial in $|A|$ and their objective function can be evaluated in polynomial time. Checking the feasibility of these solutions can be achieved in polynomial time, because checking (6c) can be achieved by solving $f(B_x)$ where $B_x := (b_{a,x} : a \in A)$ for $b_{a,x} := l_a(1 - x_a) + b_a x_a$. This problem is polynomially solvable by assumption. □

Algorithm 5 Basic Time-Constrained Asynchronous Policy

```
Set $i = 0$, $C = A$, and $\mathcal{E}$ a minimal cover of $A$
Let $S^* \in \mathcal{S}$ be an arbitrary solution and $B_f = B_{OCP}$ be an initial cost estimate
Asynchronously begin solving $f\left(B_f\right)$ and $OCP\left(B_{OCP}\right)$
for $n = 1$ to $N$ do
  if $n \in \Phi$ then
    $i \leftarrow i + 1$
    if Asynchronous solution to $f\left(B_f\right)$ has finished then
      Set $S^* \in \mathcal{S}^*\left(B_f\right)$ and $B_f = \bar{B}_n$          [Update exploitation set]
      Asynchronously begin solving $f\left(B_f\right)$
    end if
    if Asynchronous solution to $OCP\left(B_{OCP}\right)$ has finished then
      if $(C, \mathcal{E}) \notin \Gamma\left(B_{OCP}\right)$ then
        Set $(C, \mathcal{E}) \in \Gamma^*\left(B_{OCP}\right)$          [Update exploration set]
      end if
      Set $B_{OCP} = \bar{B}_n$
      Asynchronously begin solving $OCP\left(B_{OCP}\right)$
    end if
  end if
  if $T_n(\{a\}) < i$ for some $a \in \bigcup_{S \in \mathcal{E}} S$ then
    Try such an element, i.e., set $S_n = S$ with $S \in \mathcal{E}$ such that $a \in S$          [Exploration]
  else
    Implement $S_n = S^*$          [Exploitation]
  end if
end for
```

Proposition 5.3. *Let y^S be the incidence vector of $S \in \mathcal{S}$, $M \in \mathbb{R}^{m \times |A|}$, and $d \in \mathbb{R}^m$ be such that $\left\{y^S\right\}_{S \in \mathcal{S}} = \left\{y \in \{0,1\}^{|A|} : My \leq d\right\}$ and $\operatorname{conv}\left(\left\{y^S\right\}_{S \in \mathcal{S}}\right) = \left\{y \in [0,1]^{|A|} : My \leq d\right\}$. Then an MIP formulation of $OCP(B)$ is given by*

$$\min \quad \sum_{i=1}^{|A|} \left(\sum_{a \in A} b_a y_a^i - z^*(B) \right) \tag{10a}$$

$$s.t. \quad x_a \leq \sum_{i=1}^{|A|} y_a^i, \qquad a \in A \tag{10b}$$

$$My^i \leq d, \qquad i \in \{1, \ldots, |A|\} \tag{10c}$$

$$M^T w \leq \operatorname{diag}(l)(\mathbf{1} - x) + \operatorname{diag}(b) x \tag{10d}$$

$$d^T w \geq z^*(B) \tag{10e}$$

$$x_a,\, y_a^i \in \{0,1\}, w \in \mathbb{R}^m, \qquad a \in A, i \in \{1, \ldots, |A|\}, \tag{10f}$$

where for $v \in \mathbb{R}^r$, $\operatorname{diag}(v)$ is the $r \times r$ diagonal matrix with v as its diagonal. A formulation for R-$OCP(B)$ is obtained by replacing $y_a^i \in \{0,1\}$ with $0 \leq y_a^i \leq 1$.

Proof. We begin with the result for OCP. For any feasible solution (x, y) to (10) we have that x is the incidence vector of a critical subset. This, because (10d) enforces dual feasibility of w when elements with $x = 0$ are not covered, and (10e) forces the objective value of the dual of $f(B')$ to be greater than or equal to $z^*(B)$, where $B' = \text{diag}(l)(\mathbf{1} - x) + \text{diag}(b)x$. With this, the optimal objective value of $f(B')$ is greater than or equal to $z^*(B)$. On the other hand, any y_a^i is the incidence vector of some $S \in \mathcal{S}$ because of (10c) and the assumptions on M and d. Finally, (10b) ensures that $\mathcal{E} = \left\{\text{supp}\left(y_a^i\right)\right\}_{i,a\in A}$ covers the critical subset. Lemma A.3 ensures that the $|A|$ variables in y are sufficient for an optimal solution. If less than $|A|$ elements are needed for the cover, then the optimization can pick the additional y variables to be the incidence vector of an optimal solution to $f(B)$ so that they do not increase the objective function value. The proof for R-OCP is analogous, while noting that in this case, by assumption $\text{conv}\left(\left\{y^S\right\}_{S\in\mathcal{S}}\right) = \left\{y \in [0,1]^{|A|} : My \leq d\right\}$, y^i is a convex combination of incidence vectors, so it might correspond to fractions of more than one element in $\mathcal{S}$. □

A.3.2 IP formulation for OCP when $f(B)$ admits a compact IP formulation

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

$$\min \quad \sum_{i=1}^{|A|}\left(\sum_{a\in A} b_a y_a^i - z^*(B)\right) \tag{A-13a}$$

$$s.t. \qquad x_a \leq \sum_{i=1}^{|A|} y_a^i, \quad a \in A \tag{A-13b}$$

$$My^i \leq d, \qquad i \in \{1,\ldots,|A|\} \tag{A-13c}$$

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

$$x_a,\, y_a^i \in \{0,1\}, \quad a \in A, i \in \{1,\ldots,|A|\}. \tag{A-13e}$$

Like in formulation (10), a feasible solution (x, y) to (A-13) is such that x is the incidence vector of a critical subset (this is enforced by (A-13d)), and the y^i's are a cover of such set, due to (A-13c) and the assumptions on M and d. Note that an efficient cover includes at most $|A|$ elements (the optimization can pick the additional y^i to be the incidence vector of an optimal solution).

Formulation (A-13) has a polynomial number of variables, but the number of constraints described by (A-13d) is in general exponential. However, the computational burden of separating these constraints is the same as solving $f(B)$ (finding a violated inequality (A-13d) or showing that it satisfies all these inequalities can be done by solving $f(B')$ for $b'_a = l_a(1 - x_a) + b_a x_a$). Hence, if we can solve $f(B)$ sufficiently fast (e.g. when the problem is in P, or it is a practically solvable NP-hard problem) we should be able to effectively solve (A-13) with a Branch-and-Cut algorithm that dynamically adds constraints (A-13d) as needed. Finally, note that a similar formulation for R-OCP is obtained by introducing continuous variables $\{\delta_i : i = 1, \ldots, |A|\}$ with $\delta_i \in [0,1]$ and replacing y_a^i with $\delta_i y_a^i$ (which can be linearized with standard tricks) in the objective function and constraints (A-13b).

A.3.3 Linear-sized formulation for OCP for the shortest path problem

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

$$\begin{aligned}
\min \quad & \left(\sum_{a \in A} b_a p_a\right) - z^*(B)\ p_{(t,s)} && && \text{(A-14a)}\\
s.t. \quad & x_a \le p_a, && a \in A && \text{(A-14b)}\\
& \sum_{a \in \hat{\delta}_{out}(v)} p_a - \sum_{a \in \hat{\delta}_{in}(v)} p_a = 0, && v \in V && \text{(A-14c)}\\
& w_u - w_v \le l_{(u,v)}(1 - x_{(u,v)}) + b_{(u,v)} x_{(u,v)}, && (u,v) \in A && \text{(A-14d)}\\
& w_s - w_t \ge z^*(B) && && \text{(A-14e)}\\
& p_a \in \mathbb{Z}_+, && a \in \hat{A} && \text{(A-14f)}\\
& x_a \in \{0,1\},\ w_v \in \mathbb{R}, && a \in A,\ v \in V, && \text{(A-14g)}
\end{aligned}$$

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

The first difference between formulations (A-14) and (10) is the specialization of the LP duality constraints to the shortest path setting. The second one is the fact that the paths in cover $\mathcal{E}$ are aggregated into an integer circulation in augmented graph $\hat{G}$, which is encoded in variables p. Indeed, using known properties of circulations (Schrijver 2003, pp. 170-171), we have that $p = \sum_{S \in \mathcal{E}} y^{\hat{S}}$, where $y^{\hat{S}}$ is the incidence vector of the circulation obtained by adding (t,s) to path

S. Furthermore, given a feasible p we can recover the paths in $\mathcal{E}$ in time $O(|A||V|)$. To obtain a formulation for R-OCP we simply relax the integrality requirement for the p_a variables.

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

B Auxiliary Results and Omitted Proofs.

B.1 Auxiliary Result for the Proof of Theorem 3.4

Proposition A.1. *For any fixed $S \subseteq A$, $n \in \mathbb{N}$, $i \in \mathbb{N}$, and $\epsilon > 0$ we have that*

$$\mathbb{P}\left\{\left|\sum_{a\in S}\left(\bar{b}_{a,n} - \mathbb{E}\left\{b_{a,n}\right\}\right)\right| \geq \epsilon, T_n(\{a\}) \geq i\,,\ \forall a \in S\right\} \leq 2\ K(\epsilon)\,|S|\exp\left\{-\frac{2\epsilon^2 i}{|S|^2\,\mathcal{L}^2}\right\},$$

where $\mathcal{L} := \max\left\{u_a - l_a : a \in A\right\}$, and $K(\epsilon)$ is a positive finite constant that only depends on ϵ.

Proof. Consider $S \subseteq A$ and note that

$$\left\{\left|\sum_{a\in S}\left(\bar{b}_{a,n} - \mathbb{E}\left\{b_{a,n}\right\}\right)\right| \geq \epsilon\right\} \subseteq \bigcup_{a\in S}\left\{\left|\bar{b}_{a,n} - \mathbb{E}\left\{b_{a,n}\right\}\right| \geq \frac{\epsilon}{|S|}\right\},$$

hence using the union bound one has that

$$\mathbb{P}\left\{\left|\sum_{a\in S}\left(\bar{b}_{a,n} - \mathbb{E}\left\{b_{a,n}\right\}\right)\right| \geq \epsilon\,,\ T_n(\{a\}) \geq i\,,\ \forall a \in S\right\} \leq \sum_{a\in S}\mathbb{P}\left\{\left|\bar{b}_{a,n} - \mathbb{E}\left\{b_{a,n}\right\}\right| \geq \frac{\epsilon}{|S|}\,,\ T_n(\{a\}) \geq i\right\}. \tag{B-15}$$

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

$$\begin{aligned}
\mathbb{P}\left\{\left|\bar{b}_{a,n} - \mathbb{E}\left\{b_{a,n}\right\}\right| \geq \frac{\epsilon}{|S|}\,,\ T_n(\{a\}) \geq i\right\} &= \mathbb{P}\left\{\left|\sum_{m=1}^{T_n(\{a\})}\left(b_{a,t_m(a)} - \mathbb{E}\left\{b_{a,n}\right\}\right)\right| \geq \frac{\epsilon\ T_n(\{a\})}{|S|}\,,\ T_n(\{a\}) \geq i\right\} \\
&\leq \sum_{k=i}^{\infty}\mathbb{P}\left\{\left|\sum_{m=1}^{k}\left(b_{a,t_m(a)} - \mathbb{E}\left\{b_{a,n}\right\}\right)\right| \geq \frac{\epsilon\ k}{|S|}\right\} \\
&\overset{(a)}{\leq} 2\ \exp\left\{\frac{-2\ i\epsilon^2}{|S|^2\,\mathcal{L}^2}\right\}\sum_{k=0}^{\infty}\exp\left\{\frac{-2\ k\epsilon^2}{|S|^2\,\mathcal{L}^2}\right\} \\
&\leq 2\ K\exp\left\{\frac{-2\ i\epsilon^2}{|S|^2\,\mathcal{L}^2}\right\},
\end{aligned}$$

where (a) follows from the Hoeffding-Azuma Inequality (see, for example, Cesa-Bianchi and Lugosi (2006, Lemma A.7)), and K is a positive finite constant. Combining the above with (B-15) one has that

$$\mathbb{P}\left\{\left|\sum_{a\in S}\left(\bar{b}_{a,n} - \mathbb{E}\left\{b_{a,n}\right\}\right)\right| \geq \epsilon\,,\ T_n(\{a\}) \geq i\,,\ \forall a \in S\right\} \leq 2\ K\,|S|\exp\left\{\frac{-2\ i\epsilon^2}{|S|^2\,\mathcal{L}^2}\right\},$$

which is the desired result. □

B.2 Proof of Theorem 4.3

To prove Theorem 4.3 we need three probability bounding propositions, which in turn require two technical lemmas. In what follows, $(C_i, \mathcal{E}_i)$ denotes the solution to $OCP(\bar{B}_{n_i})$. Also, define the events $\tilde{U}_i := \{(C_i, \mathcal{E}_i) \in \Gamma(\mathbb{E}\{B_n\})\}$ and $U_i := \left\{\cup_{S \in \mathcal{E}_i} S \subseteq \cup_{S \in \mathcal{E}_{i-1}} S\right\}$. Also let $i'' := \max\{|A| + 2, i' + 1\}$ where i' is the first cycle in which one is sure to exploit. Finally, as in the proof of Theorem 3.4, we drop the dependence of $\mathbb{E}_F$ and $\mathbb{P}_F$ on F and π_a.

Lemma B.1. *Let $s := \max\{|S| : S \in \mathcal{S}\}$. Then, for any fixed $(C, \mathcal{E}) \in \Gamma(\mathbb{E}\{B_n\})$ we have that*

$$\left\{(C, \mathcal{E}) \notin \Gamma(\bar{B}_n)\right\} \subseteq \bigcup_{a \in \cup_{S \in \mathcal{E}} S} \left\{\left|\bar{b}_{a,n} - \mathbb{E}\{b_{a,n}\}\right| \geq \Delta_{\mathcal{C}}/(4\,|A|\,s)\right\}, \tag{B-16}$$

where $\Delta_{\mathcal{C}}$ is a positive finite constant.

Proof. We prove the lemma by establishing that the complement of the right-hand side of (B-16) implies the complement of its left-hand side.

Since $(C, \mathcal{E}) \in \Gamma(\mathbb{E}\{B_n\})$, by definition, $(C, \mathcal{E}) \in \Gamma^*(\tilde{B})$ where $\tilde{b}_a = \mathbb{E}\{b_{a,n}\}$ for all $a \in \cup_{S \in \mathcal{E}} S$ and $\tilde{b}_a = u_a$ for $a \in A \setminus \cup_{S \in \mathcal{E}} S$. Let $\Delta_{\min}^{(C,\mathcal{E})}$ denote the minimum optimality gap (i.e., the absolute difference between the objective value of the best and second best solution) in $OCP(\tilde{B})$, and define $\hat{\Delta}_{\mathcal{C}} := \inf\left\{\Delta_{\min}^{(C,\mathcal{E})} : (C, \mathcal{E}) \in \Gamma(\mathbb{E}\{B_n\})\right\}$. We assume that $\hat{\Delta}_{\mathcal{C}}$ is positive and bounded.

Let $\tilde{B}'$ be such that $\tilde{b}'_a = \bar{b}_{a,n}$ for all $a \in \cup_{S \in \mathcal{E}} S$ and $\tilde{b}'_a = \tilde{b}_a = u_a$ for $a \in A \setminus \cup_{S \in \mathcal{E}} S$. In what follows, we first prove that $\bigcap_{a \in \cup_{S \in \mathcal{E}} S} \left\{\left|\bar{b}_{a,n} - \mathbb{E}\{b_{a,n}\}\right| < \hat{\Delta}_{\mathcal{C}}/(4\,|A|\,s)\right\}$ implies that $(C, \mathcal{E}) \in \Gamma^*(\tilde{B}')$ which in turn implies $\left\{(C, \mathcal{E}) \in \Gamma(\bar{B}_n)\right\}$, provided that $(C, \mathcal{E})$ is feasible to $OCP(\bar{B}_n)$ (which we prove later in the proof).

For any $(C', \mathcal{E}')$ feasible to $OCP(\tilde{B}')$ we have

$$\sum_{S \in \mathcal{E}} \Delta_S^{\tilde{B}'} \overset{(a)}{\leq} \sum_{S \in \mathcal{E}} \Delta_S^{\tilde{B}} + \hat{\Delta}_{\mathcal{C}}/2 \overset{(b)}{\leq} \sum_{S \in \mathcal{E}'} \Delta_S^{\tilde{B}} - \hat{\Delta}_{\mathcal{C}}/2 \overset{(c)}{\leq} \sum_{S \in \mathcal{E}'} \Delta_S^{\tilde{B}'},$$

where (a) comes from noting that for any $S \in \mathcal{E}$

$$\Delta_S^{\tilde{B}'} - \Delta_S^{\tilde{B}} = \sum_{a \in S} \left(\tilde{b}'_a - \tilde{b}_a\right) + \sum_{a \in S^*} \tilde{b}_a - \sum_{a \in \bar{S}^*} \tilde{b}'_a < \sum_{a \in S} \left(\bar{b}_{a,n} - \mathbb{E}\{b_{a,n}\}\right) + \hat{\Delta}_{\mathcal{C}}/(4\,|A|) < \hat{\Delta}_{\mathcal{C}}/(2\,|A|)$$

where $S^* \in \mathcal{S}^*(\mathbb{E}\{B_n\})$ and $\bar{S}^* \in \mathcal{S}^*(\bar{B}_n)$ (note that in what follows, we give conditions under which $\bar{S}^* \in \mathcal{S}^*(\mathbb{E}\{B_n\})$), and (b) and (c) follow from the definition of $\hat{\Delta}_{\mathcal{C}}$ and $\tilde{B}'$, respectively. Note that the feasibility of $(C, \mathcal{E})$ to $OCP(\bar{B}_n)$ (which will be verified next) also implies its feasibility to

$OCP(\tilde{B}')$ due to the definition of $\tilde{B}'$.

Now we prove that $(C, \mathcal{E})$ is feasible to $OCP(\bar{B}_n)$. Define

$$\tilde{\Delta}_{\mathcal{C}} := \inf \left\{ \sum_{a \in C \cap S} \mathbb{E}\left\{b_{a,n}\right\} + \sum_{a \in S \setminus C} l_a - \sum_{a \in S^*} \mathbb{E}\left\{b_{a,n}\right\} \,:\, S \in \mathcal{S} \setminus \mathcal{S}^*(\mathbb{E}\left\{B_n\right\}) \,,\, C \in \mathcal{C} \right\},$$

and consider the case of $\tilde{\Delta}_{\mathcal{C}} > 0$. Suppose the complement of the right-hand side of (B-16) holds true for the constant $\tilde{\Delta}_{\mathcal{C}}$. For any $S \in \mathcal{S} \setminus \mathcal{S}^*(\mathbb{E}\left\{B_n\right\})$ we have that

$$\sum_{a \in C \cap S} \bar{b}_{a,n} + \sum_{a \in S \setminus C} l_a > \sum_{a \in C \cap S} \mathbb{E}\left\{b_{a,n}\right\} + \sum_{a \in S \setminus C} l_a - \tilde{\Delta}_{\mathcal{C}}/(2\left|A\right|) \overset{(a)}{\geq} \sum_{a \in S^*} \mathbb{E}\left\{b_{a,n}\right\} + \tilde{\Delta}_{\mathcal{C}}/(2\left|A\right|) > \sum_{a \in S^*} \bar{b}_{a,n}$$

where (a) follows from the definition of $\tilde{\Delta}_{\mathcal{C}}$, and the fact that C includes the optimal solution S^*. Note that the argument above also implies that any $S^* \in \mathcal{S}^*(\mathbb{E}\left\{B_n\right\})$ is also optimal for $f\left(\bar{B}_n\right)$ and $f(\tilde{B}')$. Also note that S^* is also optimal for $f(\tilde{B})$ by definition of $\tilde{B}$. Setting $\Delta_{\mathcal{C}} = \min\left\{\hat{\Delta}_{\mathcal{C}}, \tilde{\Delta}_{\mathcal{C}}\right\}$ completes the proof. □

Lemma B.2. *Let $s := \max\left\{|S| : S \in \mathcal{S}\right\}$. Then, for any $(C, \mathcal{E})$ feasible to $OCP(\bar{B}_n)$, we have*

$$\left\{(C, \mathcal{E}) \notin \Gamma(\mathbb{E}\left\{B_n\right\})\right\} \cap \left\{\left|\bar{b}_{a,n} - \mathbb{E}\left\{b_{a,n}\right\}\right| < \Delta'_{\mathcal{C}}/(4\, s\left|A\right|) \,,\, a \in \cup_{S \in \mathcal{E}} S\right\} \subseteq \left\{(C, \mathcal{E}) \notin \Gamma(\bar{B}_n)\right\}, \tag{B-17}$$

where $\Delta'_{\mathcal{C}}$ is a positive finite constant.

Proof. Suppose the left-hand side of (B-17) holds true. Since $(C, \mathcal{E}) \notin \Gamma(\mathbb{E}\left\{B_n\right\})$, by definition, either $(C, \mathcal{E})$ is not feasible to $OCP(\mathbb{E}\left\{B_n\right\})$, and/or $(C, \mathcal{E}) \notin \Gamma^*(\tilde{B})$ where $\tilde{b}_a = \mathbb{E}\left\{b_{a,n}\right\}$ for all $a \in \cup_{S \in \mathcal{E}} S$ and $\tilde{b}_a = u_a$ for $a \in A \setminus \cup_{S \in \mathcal{E}} S$.

First, suppose that $(C, \mathcal{E})$ is not feasible to $OCP(\mathbb{E}\left\{B_n\right\})$. Define

$$\Delta''_{\mathcal{C}} := \min \left\{ \sum_{a \in S^*} \mathbb{E}\left\{b_{a,n}\right\} - \sum_{a \in C \cap S} \mathbb{E}\left\{b_{a,n}\right\} - \sum_{a \in S \setminus C} l_a > 0 \,:\, C \notin \mathcal{C} \right\},$$

where $S^* \in \mathcal{S}^*(\mathbb{E}\left\{B_n\right\})$. Note that $\Delta''_{\mathcal{C}} > 0$. The infeasibility of $(C, \mathcal{E})$ to $OCP(\mathbb{E}\left\{B_n\right\})$ implies that constraint (6c) must be violated for some $S \in \mathcal{S}$, for which

$$\sum_{a \in C \cap S} \mathbb{E}\left\{b_{a,n}\right\} + \sum_{a \in S \setminus C} l_a < \sum_{a \in S^*} \mathbb{E}\left\{b_{a,n}\right\}.$$

Let S' be one such S that additionally minimizes the left-hand side above. Using $\Delta''_{\mathcal{C}}$ in the left-hand side of (B-17), one has that for such S'

$$\sum_{a\in C\cap S'} \bar{b}_{a,n} + \sum_{a\in S'\setminus C} l_a < \sum_{a\in C\cap S'} \mathbb{E}\left\{b_{a,n}\right\} + \sum_{a\in S'\setminus C} l_a + \Delta''_{\mathcal{C}}/2 \overset{(a)}{\leq} \sum_{a\in \bar{S}^*} \mathbb{E}\left\{b_{a,n}\right\} - \Delta''_{\mathcal{C}}/2 \overset{(b)}{<} \sum_{a\in \bar{S}^*} \bar{b}_{a,n},$$

where $\bar{S}^* \in \mathcal{S}^*\left(\bar{B}_n\right)$, (a) follows from the definition of $\Delta''_{\mathcal{C}}$ and the fact that $\sum_{a\in S^*} \mathbb{E}\left\{b_{a,n}\right\} \leq \sum_{a\in \bar{S}^*} \mathbb{E}\left\{b_{a,n}\right\}$ for $S^* \in \mathcal{S}^*\left(E_F\left\{B_n\right\}\right)$, and (b) from the fact that $\bar{S}^* \subseteq C$ when $(C,\mathcal{E})$ is feasible to $OCP\left(\bar{B}_n\right)$. The last inequality above implies that constraint (6c) is not satisfied for $S' \in \mathcal{S}$, which contradicts the feasibility of $(C,\mathcal{E})$ to $OCP\left(\bar{B}_n\right)$. Thus $(C,\mathcal{E})$ must be feasible to $OCP(\mathbb{E}\left\{B_n\right\})$. Then, $(C,\mathcal{E}) \notin \Gamma(\mathbb{E}\left\{B_n\right\})$ implies that $(C,\mathcal{E}) \notin \Gamma^*(\tilde{B})$ where $\tilde{b}_a = \mathbb{E}\left\{b_{a,n}\right\}$ for all $a \in \cup_{S\in\mathcal{E}} S$ and $\tilde{b}_a = u_a$ for $a \in A \setminus \cup_{S\in\mathcal{E}} S$.

Let $(C',\mathcal{E}') \in \Gamma^*(\tilde{B})$, and $\tilde{B}'$ be such that $\tilde{b}'_a = \bar{b}_{a,n}$ for all $a \in \cup_{S\in\mathcal{E}} S$ and $\tilde{b}'_a = \tilde{b}_a = u_a$ for $a \in A \setminus \cup_{S\in\mathcal{E}} S$. We first establish some feasibility requirements which we need to complete the proof. Note that $(C',\mathcal{E}')$ is feasible to $OCP(\mathbb{E}\left\{B_n\right\})$ by definition of $\tilde{B}$. Using a similar argument as in the proof of Lemma B.1, we prove that $\bigcap_{a\in\cup_{S\in\mathcal{E}} S} \left\{\left|\bar{b}_{a,n} - \mathbb{E}\left\{b_{a,n}\right\}\right| < \tilde{\Delta}_{\mathcal{C}}/(4\left|A\right| s)\right\}$ implies $(C',\mathcal{E}')$ is feasible to $OCP(\tilde{B}')$, where $\tilde{\Delta}_{\mathcal{C}}$ is defined in that proof. To show this, note that for any $S \in \mathcal{S} \setminus \mathcal{S}^*(\mathbb{E}\left\{B_n\right\})$ we have

$$\sum_{a\in C'\cap S} \tilde{b}'_a + \sum_{a\in S\setminus C'} l_a \overset{(a)}{\geq} \sum_{a\in C'\cap S} \mathbb{E}\left\{b_{a,n}\right\} + \sum_{a\in S\setminus C'} l_a - \tilde{\Delta}_{\mathcal{C}}/(2\left|A\right|) \overset{(b)}{\geq} \sum_{a\in S^*} \mathbb{E}\left\{b_{a,n}\right\} + \tilde{\Delta}_{\mathcal{C}}/(2\left|A\right|) > \sum_{a\in S^*} \bar{b}_{a,n},$$

where $S^* \in \mathcal{S}^*\left(E_F\left\{B_n\right\}\right)$, and (a) and (b) follow from the definition of $\tilde{B}'$ and $\tilde{\Delta}_{\mathcal{C}}$, respectively. As in the proof of Lemma B.1, $\bigcap_{a\in\cup_{S\in\mathcal{E}} S} \left\{\left|\bar{b}_{a,n} - \mathbb{E}\left\{b_{a,n}\right\}\right| < \tilde{\Delta}_{\mathcal{C}}/(4\left|A\right| s)\right\}$ implies that S^* is optimal for $f\left(\bar{B}_n\right)$ and $f(\tilde{B}')$. Also note that by assumption $(C,\mathcal{E})$ is feasible to $OCP(\bar{B}_n)$ (and thus $OCP(\tilde{B}')$), and S^* is also optimal for $f(\tilde{B})$.

Let $\Delta^{(C,\mathcal{E})}$ denote the optimality gap of $(C,\mathcal{E})$ in $OCP(\tilde{B})$ and define

$$\hat{\Delta}'_{\mathcal{C}} := \inf\left\{\Delta^{(C,\mathcal{E})} : (C,\mathcal{E}) \text{ feasible to } OCP(\mathbb{E}\left\{B_n\right\}), (C,\mathcal{E}) \notin \Gamma^*(\tilde{B})\right\}.$$

Note that $\hat{\Delta}'_{\mathcal{C}} > 0$. Now, one has that $\bigcap_{a\in\cup_{S\in\mathcal{E}} S} \left\{\left|\bar{b}_{a,n} - \mathbb{E}\left\{b_{a,n}\right\}\right| < \hat{\Delta}'_{\mathcal{C}}/(4\left|A\right| s)\right\}$ implies that

$$\sum_{S\in\mathcal{E}} \Delta_S^{\tilde{B}'} > \sum_{S\in\mathcal{E}} \Delta_S^{\tilde{B}} - \hat{\Delta}'_{\mathcal{C}}/2 \geq \sum_{S\in\mathcal{E}'} \Delta_S^{\tilde{B}} + \hat{\Delta}'_{\mathcal{C}}/2 \geq \sum_{S\in\mathcal{E}'} \Delta_S^{\tilde{B}'},$$

via the arguments in the proof of Proposition B.1. Seeing that $(C,\mathcal{E})$ and $(C',\mathcal{E}')$ are both feasible to $OCP(\tilde{B}')$, above implies that $(C,\mathcal{E}) \notin \Gamma^*(\tilde{B}')$ and thus $(C,\mathcal{E}) \notin \Gamma(\bar{B}_n)$. Setting $\Delta'_{\mathcal{C}} = \min\left\{\Delta''_{\mathcal{C}}, \tilde{\Delta}_{\mathcal{C}}, \hat{\Delta}'_{\mathcal{C}}\right\}$ completes the proof. $\square$

Using the two lemmas above, we prove the following probability bounding propositions.

Proposition B.3. *If $i \geq i''$, then*

$$\mathbb{P}\left((C_i,\mathcal{E}_i) \notin \Gamma(\mathbb{E}\{B_n\})\right) \leq \tilde{K}\exp\left\{-\tilde{C}(i-|A|)\right\},$$

where $\tilde{C}$ and $\tilde{K}$ are some positive finite constants.

Proof. Remember that $(C_i,\mathcal{E}_i)$ denotes the solution to $OCP(\bar{B}_{n_i})$, $\tilde{U}_i = \{(C_i,\mathcal{E}_i) \in \Gamma(\mathbb{E}\{B_n\})\}$ and $U_i = \left\{\cup_{S\in\mathcal{E}_i} S \subseteq \cup_{S\in\mathcal{E}_{i-1}} S\right\}$. Note that

$$\mathbb{P}\left\{\tilde{U}_i^c\right\} \leq \sum_{j=i-|A|}^{i} \mathbb{P}\left\{\tilde{U}_i^c\,,\,U_j\right\} + \mathbb{P}\left\{\tilde{U}_i^c\,,\,U_j^c\,,\,\forall j \in \{i-|A|,\ldots,i\}\right\}. \tag{B-18}$$

We bound the first and second terms in the right-hand side of (B-18) separately.

Step 1. Here, we bound the first term in the right-hand side of (B-18). Consider $j \in \{i-|A|,\ldots,i\}$ and note that $T_{n_j}(\{a\}) \geq i-|A|$ for all $a \in \cup_{S\in\mathcal{E}_i} S$ under U_j. One has that

$$\mathbb{P}\left\{\tilde{U}_i^c\,,\,U_j\right\} = \mathbb{P}\left\{\tilde{U}_i^c\,,\tilde{U}_j\,,\,U_j\right\} + \mathbb{P}\left\{\tilde{U}_i^c\,,\,\tilde{U}_j^c\,,\,U_j\right\}. \tag{B-19}$$

We complete the Step 1 by bounding the two terms in the right-hand side of (B-19).

The first event in the right-hand side of (B-19) implies that $(C_j,\mathcal{E}_j) \notin \Gamma\left(\bar{B}_{n_{j'}}\right)$ for some $j' \in \{j,\ldots,i\}$. Then, from Lemma B.1 we have that

$$\begin{aligned}
\mathbb{P}\left\{\tilde{U}_i^c\,,\tilde{U}_j\,,\,U_j\right\} &\leq \sum_{j'=j}^{i} \mathbb{P}\left\{(C_j,\mathcal{E}_j) \notin \Gamma(\bar{B}_{n_{j'}})\,,\,T_{n_{j'}}(\{a\}) \geq i-|A|\;,\;a \in \cup_{S\in\mathcal{E}_j} S\,,\,\tilde{U}_j\right\} \\
&\leq \sum_{j'=j}^{i} \sum_{a\in\cup_{S\in\mathcal{E}_j} S} \mathbb{P}\left\{\left|\bar{b}_{a,n_{j'}} - \mathbb{E}\{b_{a,n}\}\right| \geq \Delta_{\mathcal{C}}/(4\,s\,|A|)\,,\,T_{n_{j'}}(\{a\}) \geq i-|A|\right\} \\
&\overset{(a)}{\leq} 2\;|A|^2\,K\exp\left\{-\tilde{C}_1(i-|A|)\right\},
\end{aligned}$$

where $\tilde{C}_1 := (\Delta_{\mathcal{C}})^2/(8\,s^2\,|A|^2\,\mathcal{L}^2)$, and (a) follows from Proposition A.1 and also noting that $i-|A| \leq j \leq j' \leq i$ and $\left|\cup_{S\in\mathcal{E}_j} S\right| \leq |A|$.

To bound the second term in the right-hand side of (B-19) we use Lemma B.2. We have that

$$\begin{aligned}
\mathbb{P}\left\{\tilde{U}_i^c\,,\,\tilde{U}_j^c\,,\,U_j\right\} &\leq \mathbb{P}\left\{(C_j,\mathcal{E}_j)\notin\Gamma(\mathbb{E}\left\{B_n\right\})\,,\,U_j\right\}\\
&\leq \mathbb{P}\left\{(C_j,\mathcal{E}_j)\notin\Gamma(\mathbb{E}\left\{B_n\right\})\,,\,(C_j,\mathcal{E}_j)\in\Gamma(\bar{B}_{n_j})\,,\,T_{n_j}(\{a\})\geq i-|A|\ ,\ a\in\cup_{S\in\mathcal{E}_j}S\right\}\\
&\leq \sum_{a\in\cup_{S\in\mathcal{E}_j}S}\mathbb{P}\left\{\left|\bar{b}_{a,n_j}-\mathbb{E}\left\{b_{a,n}\right\}\right|\geq\Delta'_{\mathcal{C}}/(4\,s\,|A|)\,,\,T_{n_j}(\{a\})\geq i-|A|\right\}\\
&\overset{(a)}{\leq} 2\ K\,|A|\exp\left\{-\tilde{C}_2(i-|A|)\right\},
\end{aligned}$$

where $\tilde{C}_2:=\left(\Delta'_{\mathcal{C}}\right)^2/(8\,s^2\,|A|^2\,\mathcal{L}^2)$, and (a) follows from Proposition A.1. Combining the results above we have

$$\sum_{j=i-|A|}^{i}\mathbb{P}\left\{\tilde{U}_i^c\,,\,U_j\right\}\leq 4\ K\,|A|^3\exp\left\{-\tilde{C}(i-|A|)\right\},$$

where $\tilde{C}:=\min\left\{\tilde{C}_1,\tilde{C}_2\right\}$.

Step 2. Here, we bound the second term in the right-hand side of (B-18). Note that the intersection of U_j^c for $j\in\{i-|A|\,,\ldots,i\}$ implies that there exists a $j''\in\{i-|A|\,,\ldots,i\}$ such that $T_{n_{j''}}(\{a\})\geq i-|A|$ for all $a\in\cup_{S\in\mathcal{E}_{j''}}S$. With this observation, one can apply the arguments in Step 1 to show that

$$\mathbb{P}\left\{\tilde{U}_i^c\,,\,U_j^c\,,\,\forall j\in\{i-|A|\,,\ldots,i\}\ ,\,\tilde{U}_{j''}\right\}\leq 2\,|A|^3\,K\exp\left\{-\tilde{C}_1(i-|A|)\right\},$$

and

$$\mathbb{P}\left\{\tilde{U}_i^c\,,\,U_j^c\,,\,\forall j\in\{i-|A|\,,\ldots,i\}\ ,\,\tilde{U}_{j''}^c\right\}\leq 2\,|A|^2\,K\exp\left\{-\tilde{C}_2(i-|A|)\right\},$$

where the extra $|A|$ in the right-hand side of the two inequalities above (compared to that in Step 1) comes from the fact that we do not know the exact value of j''.

Combining the above one has that

$$\mathbb{P}\left\{\tilde{U}_i^c\,,\,U_j^c\,,\,\forall j\in\{i-|A|\,,\ldots,i\}\right\}\leq 4\,|A|^3\,K\exp\left\{-\tilde{C}(i-|A|)\right\}.$$

Combining the results from Steps 1 and 2 one obtains

$$\mathbb{P}\left\{\tilde{U}_i^c\right\}\leq\tilde{K}\exp\left\{-\tilde{C}(i-|A|)\right\},$$

where $\tilde{K}$ is a finite positive constant. □

Proposition B.4. *If $i \geq i''$, then*

$$\mathbb{P}\left\{U_i^c\right\} \leq C' \exp\left\{-\tilde{C}(i - |A| - 1)\right\},$$

where C' is a positive finite constant and $\tilde{C}$ is as in Proposition B.3.

Proof. Remember that $U_i = \left\{\cup_{S\in\mathcal{E}_i} S \subseteq \cup_{S\in\mathcal{E}_{i-1}} S\right\}$ and $\tilde{U}_i = \{(C_i, \mathcal{E}_i) \in \Gamma(\mathbb{E}\{B_n\})\}$. Then

$$\mathbb{P}\left(U_i{}^c\right) = \mathbb{P}\left(U_i{}^c, \tilde{U}_{i-1}\right) + \mathbb{P}\left(U_i{}^c, \tilde{U}^c_{i-1}\right) \overset{(a)}{\leq} \mathbb{P}\left((C_{i-1}, \mathcal{E}_{i-1}) \notin \Gamma(\bar{B}_{n_i}), \tilde{U}_{i-1}\right) + \mathbb{P}\left(\tilde{U}^c_{i-1}\right), \tag{B-20}$$

where (a) follows from the fact that at each cycle, the OCP-based policy keeps the previous exploration set if $(C_{i-1}, \mathcal{E}_{i-1}) \in \Gamma(\bar{B}_{n_i})$. Also note that $T_{n_i}(\{a\}) \geq i-1$ for all $a \in \cup_{S\in\mathcal{E}_{i-1}} S$. By Lemma B.1, one has that

$$\begin{aligned}
\mathbb{P}\left((C_{i-1}, \mathcal{E}_{i-1}) \notin \Gamma(\bar{B}_{n_i}), \tilde{U}_{i-1}\right) &\leq \mathbb{P}\left\{\bigcup_{a\in\cup_{S\in\mathcal{E}_{i-1}} S} \left\{\left|\bar{b}_{a,n_i} - \mathbb{E}\{b_{a,n}\}\right| \geq \frac{\Delta_{\mathcal{C}}}{4\, s\, |A|}\right\} \cap \tilde{U}_{i-1}\right\} \\
&\leq \mathbb{P}\left\{\bigcup_{a\in\cup_{S\in\mathcal{E}_{i-1}} S} \left\{\left|\bar{b}_{a,n_i} - \mathbb{E}\{b_{a,n}\}\right| \geq \frac{\Delta_{\mathcal{C}}}{4\, s\, |A|}\right\},\ T_{n_i}(\{a\}) \geq i-1\right\} \\
&\overset{(b)}{\leq} 2\ |A|\, K \exp\left\{-\tilde{C}(i-1)\right\},
\end{aligned}$$

where $\tilde{C}$ is as in the proof of Proposition B.3, and (b) follows from applying the union bound and Proposition A.1. Using this and Proposition B.3 to bound the second term in the right-hand side of (B-20) gives

$$\mathbb{P}\left(U_i{}^c\right) \leq \left(\tilde{K} + 2K\,|A|\right) \exp\left\{-\tilde{C}(i - |A| - 1)\right\},$$

where $\tilde{K}$ is as in the proof of Proposition B.3. □

Proposition B.5. *If $i \geq i''$, then*

$$\mathbb{P}\left\{\bar{S}_{n_i} \notin \mathcal{S}^*(\mathbb{E}\{B_n\}), \tilde{U}_{i-1}\right\} \leq C'' \exp\left\{-\tilde{C}(i-1)\right\},$$

where $\bar{S}_{n_i} \in \mathcal{S}^*(\bar{B}_{n_i})$, C'' is a positive finite constant and $\tilde{C}$ is as in Proposition B.3.

Proof. When $\tilde{U}_{i-1}$ happens, then $(C_{i-1}, \mathcal{E}_{i-1}) \in \Gamma(\mathbb{E}\{B_n\})$. In particular, all $a \in S^*$ are included in C_{i-1}, for all $S^* \in \mathcal{S}^*(\mathbb{E}\{B_n\})$. Note that

$$\left\{\bar{S}_{n_i} \notin \mathcal{S}^*(\mathbb{E}\{B_n\})\right\} \subseteq \bigcup_{a \in \cup_{S \in \mathcal{E}} S} \left\{\left|\bar{b}_{a,n_i} - \mathbb{E}\{b_{a,n}\}\right| \geq \Delta_{\mathcal{C}}/(2\,s)\right\},$$

for any $(C, \mathcal{E}) \in \Gamma(\mathbb{E}\{B_n\})$, and $\Delta_{\mathcal{C}}$ is as in the proof of Lemma B.1. To prove the statement above, assume that the complement of the right-hand side holds for some fixed $(C, \mathcal{E}) \in \Gamma(\mathbb{E}\{B_n\})$. Then for any $S \in \mathcal{S} \setminus \mathcal{S}^*(\mathbb{E}\{B_n\})$

$$\sum_{a \in S} \bar{b}_{a,n_i} > \sum_{a \in C \cap S} \mathbb{E}\{b_{a,n}\} + \sum_{a \in S \setminus C} l_a - \Delta_{\mathcal{C}}/2 \overset{(a)}{\geq} \sum_{a \in S^*} \mathbb{E}\{b_{a,n}\} + \Delta_{\mathcal{C}}/2 > \sum_{a \in S^*} \bar{b}_{a,n_i},$$

where (a) follows from the definition of $\Delta_{\mathcal{C}}$. The last inequality above implies that $\bar{S}_{n_i} \in \mathcal{S}^*(\mathbb{E}\{B_n\})$. In addition, one has that $T_{n_i}(\{a\}) \geq i-1$ for all $a \in \cup_{S \in \mathcal{E}_{i-1}} S$. Therefore

$$\begin{aligned}
\mathbb{P}\left\{\bar{S}_{n_i} \notin \mathcal{S}^*(\mathbb{E}\{B_n\}),\, \tilde{U}_{i-1}\right\} &= \mathbb{P}\left\{\bar{S}_{n_i} \notin \mathcal{S}^*(\mathbb{E}\{B_n\}),\, T_{n_i}(\{a\}) \geq i-1,\, \forall\, a \in \cup_{S \in \mathcal{E}_{i-1}} S,\, \tilde{U}_{i-1}\right\} \\
&\leq \mathbb{P}\left\{\bigcup_{a \in \cup_{S \in \mathcal{E}_{i-1}} S} \left\{\left|\bar{b}_{a,n_i} - \mathbb{E}\{b_{a,n}\}\right| \geq \Delta_{\mathcal{C}}/(2\,s),\, T_{n_i}(\{a\}) \geq i-1\right\} \cap \tilde{U}_{i-1}\right\} \\
&\leq \mathbb{P}\left\{\bigcup_{a \in \cup_{S \in \mathcal{E}_{i-1}} S} \left\{\left|\bar{b}_{a,n_i} - \mathbb{E}\{b_{a,n}\}\right| \geq \Delta_{\mathcal{C}}/(2\,s\,|A|),\, T_{n_i}(\{a\}) \geq i-1\right\}\right\} \\
&\leq 2\,K\,|A| \exp\left\{-\tilde{C}(i-1)\right\},
\end{aligned}$$

where the last inequality above follows from the union bound and Proposition A.1, and $\tilde{C}$ is as in the proof of Proposition B.3. □

Theorem 4.3. *Let π_a denote the policy in Algorithm 1 and set $\varepsilon > 0$ arbitrary. If we choose $n_i := \max\left\{\lfloor e^{i^{1/(1+\varepsilon)}} \rfloor, n_{i-1}+1\right\}$, for all $i \geq 2$, then $(C_{n_i}, \mathcal{E}_{n_i})$ converges to $(C_\infty, \mathcal{E}_\infty) \in \Gamma(\mathbb{E}_F\{B_n\})$. Moreover,*

$$\frac{R^{\pi_a(\mathcal{E})}(F,N)}{(\ln N)^{1+\varepsilon}} \leq \mathbb{E}_F\left\{z^*_{OCP}\left(\bar{B}_\infty\right)\right\} + O\left(1/(\ln N)^{1+\varepsilon}\right) \leq \Delta^F_{\max}\, G + O\left(1/(\ln N)^{1+\varepsilon}\right),$$

where $G := \max\{|\mathcal{E}| : (C, \mathcal{E}) \in \Gamma(\mathbb{E}_F\{B_n\})\}$, and $\bar{B}_\infty$ is a random vector that coincides with $\mathbb{E}_F\{B_n\}$ for $a \in \bigcup_{S \in \mathcal{E}_\infty} S$.

Proof. Similar to the case of the simple policy, regret of the adaptive policy π_a stems from two sources: exploration and errors during exploitation. That is,

$$R^{\pi_a}(F,N) = \sum_{S\in\mathcal{S}} \Delta_S^F \; \mathbb{E}\left\{T_{N+1}(S)\right\} = R_1^{\pi_a}(F,N) + R_2^{\pi_a}(F,N), \tag{B-21}$$

where $R_1^{\pi_a}(F,N)$ is the exploration-based regret and $R_2^{\pi_a}(F,N)$ is the exploitation-based regret. We prove the result by bounding each term above separately. As in the proof of Theorem 3.4, we drop the dependence of $\mathbb{E}_F$ and $\mathbb{P}_F$ on F and π_a.

Step 1 (Exploration-based regret). We begin by setting up some notation. Let $(C_i, \mathcal{E}_i)$ denote the critical subset and exploration set used during cycle i, and for $S\in\mathcal{S}$ define $\Delta T_i(S) := T_{n_{i+1}}(S) - T_{n_i}(S)$. Also, define $i'' := \max\left\{|A|+2, i'+1\right\}$ where i' is the first cycle in which one is sure to exploit[19], and the events $U_i := \left\{\cup_{S\in\mathcal{E}_i} S \subseteq \cup_{S\in\mathcal{E}_{i-1}} S\right\}$ and $\tilde{U}_i := \left\{(C_i,\mathcal{E}_i)\in\Gamma(\mathbb{E}\left\{B_n\right\})\right\}$, for $i\geq i''$. Using these definitions, one has that

$$R_1^{\pi_a}(F,N) \quad \leq \Delta_{max}^F \; n_{i''} + \sum_{i=i''}^{\lceil(\ln N)^{1+\varepsilon}\rceil} \left(\mathbb{E}\left\{ \sum_{S\in\mathcal{E}_i} \Delta_S^F \; \Delta T_i(S) \mathbf{1}\left\{U_i\right\}\right\} + \Delta_{max}^F \; \mathbb{E}\left\{\sum_{S\in\mathcal{E}_i} \Delta T_i(S)\mathbf{1}\left\{U_i^c\right\}\right\}\right). \tag{B-22}$$

We bound the exploration-based regret in two steps.

Step 1-(a). First, we bound the second term in (B-22). We have that

$$\mathbb{E}\left\{\sum_{S\in\mathcal{E}_i} \Delta_S^F \; \Delta T_i(S)\mathbf{1}\left\{U_i\right\}\right\} \quad \leq \quad \mathbb{E}\left\{\sum_{S\in\mathcal{E}_i} \Delta_S^F \; \Delta T_i(S) \mathbf{1}\left\{U_i\cap\tilde{U}_i\right\}\right\} + \Delta_{max}^F \mathbb{E}\left\{\sum_{S\in\mathcal{E}_i} \Delta T_i(S)\mathbf{1}\left\{U_i\cap\tilde{U}_i^c\right\}\right\}.$$

Note that event U_i implies that $\Delta T_i(S)\leq 1$ for all $S\in\mathcal{E}_i$. Then, the event $\tilde{U}_i$ implies that

$$\mathbb{E}\left\{\sum_{S\in\mathcal{E}_i} \Delta_S^F \; \Delta T_i(S)\mathbf{1}\left\{U_i\cap\tilde{U}_i\right\}\right\} \leq \mathbb{E}\left\{\sum_{S\in\tilde{\mathcal{E}}_i} \Delta_S^F\right\} \leq \Delta_{max}^F \; \mathbb{E}\left\{\left|\tilde{\mathcal{E}}_i\right|\right\} \leq \Delta_{max}^F \; G,$$

for some random $(\tilde{C}_i, \tilde{\mathcal{E}}_i)\in\Gamma(\mathbb{E}\left\{B_n\right\})$, and where G is the constant in Theorem 4.3.

We can then use the bound on the probability of $\tilde{U}_i^c$ from Proposition B.3 to obtain

$$\mathbb{E}\left\{\sum_{S\in\mathcal{E}_i} \Delta T_i(S)\mathbf{1}\left\{U_i\cap\tilde{U}_i^c\right\}\right\} \overset{(a)}{\leq} |A| \; \mathbb{P}\left(\tilde{U}_i^c\right) \overset{(b)}{\leq} \tilde{K}\,|A| \exp\left\{-\tilde{C}(i-|A|)\right\},$$

[19] For instance, we can take $i' := \inf\left\{i\in\mathbb{N}, i\geq 2 : \lfloor e^{i^{1/(1+\varepsilon)}}\rfloor > n_{i-1} + i\,|A|\,, n_{i+1}-n_i > |A|\right\}$.

where (a) follows from noting that $\Delta T_i(S) \leq 1$ under U_i and $|\mathcal{E}_i| \leq |A|$, and (b) follows from Proposition B.3. From above, one can bound the second term in (B-22) as follows

$$\sum_{i=i''}^{\lceil (\ln N)^{1+\varepsilon} \rceil} \mathbb{E}\left\{ \sum_{S \in \mathcal{E}_i} \Delta_S^F \; \Delta T_i(S) \mathbf{1}\left\{U_i\right\} \right\} \leq \sum_{i=i''}^{\lceil (\ln N)^{1+\varepsilon} \rceil} \mathbb{E}\left\{ \sum_{S \in \tilde{\mathcal{E}}_i} \Delta_S^F \right\} + \Delta_{\max}^F \tilde{K} \, |A| \sum_{i=i''}^{\infty} \exp\left\{ -\tilde{C}(i - |A|) \right\}. \tag{B-23}$$

Note that the second term above is finite.

Step 1-(b). To bound the third term in (B-22), note that $\mathbb{E}\left\{\Delta T_i(S)\right\} \leq i$ for all $S \in \mathcal{E}_i$, and that $|\mathcal{E}_i| \leq |A|$, hence

$$\mathbb{E}\left\{ \sum_{S \in \mathcal{E}_i} \Delta T_i(S) \mathbf{1}\left\{U_i^c\right\} \right\} \leq i \, |A| \, \mathbb{P}\left(U_i^c\right).$$

Using the bounds on the probability of U_i^c from Proposition B.4 we have

$$\sum_{i=i''}^{\lceil (\ln N)^{1+\varepsilon} \rceil} \mathbb{E}\left\{ \sum_{S \in \mathcal{E}_i} \Delta T_i(S) \mathbf{1}\left\{U_i^c\right\} \right\} \quad \leq \quad \sum_{i=i''}^{\infty} i \, |A| \, C' \exp\left\{ -\tilde{C}(i - |A| - 1) \right\}. \tag{B-24}$$

Thus, combining (B-22), (B-23) and (B-24) we have that

$$\begin{aligned} R_1^{\pi_a}(F,N) \quad &\leq \quad n_{i''} \; \Delta_{max}^F + \sum_{i=i''}^{\lceil (\ln N)^{1+\varepsilon} \rceil} \mathbb{E}\left\{ \sum_{S \in \tilde{\mathcal{E}}_i} \Delta_S^F \right\} + \Delta_{max}^F \sum_{i=i''}^{\infty} |A| \left(C' i + \tilde{K} \right) \exp\left\{ -\tilde{C}(i - |A| - 1) \right\} \\ &= \quad \sum_{i=i''}^{\lceil H \ln N \rceil} \mathbb{E}\left\{ \sum_{S \in \tilde{\mathcal{E}}_i} \Delta_S^F \right\} + C_6, \end{aligned} \tag{B-25}$$

where C_6 is a positive finite constant.

Step 2 (Exploitation-based regret). From Proposition B.3 one has that

$$\begin{aligned} R_2^{\pi_a}(F,N) \quad &\leq \quad n_{i''} \; \Delta_{max}^F + \sum_{i=i''}^{\infty} \left(n_{i+1} - n_i\right) \mathbb{P}\left\{ \bar{S}_{n_i} \notin \mathcal{S}^*(\mathbb{E}\left\{B_n\right\}), \, \tilde{U}_{i-1} \right\} \Delta_{max}^F \\ &\qquad + \sum_{i=i''}^{\infty} \left(n_{i+1} - n_i\right) \tilde{K} \exp\left\{ -\tilde{C}(i - |A| - 1) \right\} \Delta_{max}^F, \end{aligned}$$

where $\bar{S}_{n_i} \in \mathcal{S}^*(\bar{B}_{n_i})$ is any solution with minimum average cost at time n_i. From proof of Theorem 3.4, we know that $n_{i+1} - n_i \leq e^{(i+1)^{1/(1+\varepsilon)}}$ for $i \geq i''$. Also, let $\tilde{i} \geq i''$ be such that $i \, \tilde{C}/2 > (i+1)^{1/1+\varepsilon}$

for $i \geq \tilde{i}$. Thus, using the results above and Proposition B.5 one has that

$$R_2^{\pi_a}(F,N) \leq n_{\tilde{i}} \Delta^F_{max} + \left(\tilde{K} + C''\right) \Delta^F_{max} \sum_{i=\tilde{i}}^{\infty} \exp\left\{-\tilde{C}(i - |A| - 1) + (i+1)^{1/(1+\varepsilon)}\right\}. \qquad \text{(B-26)}$$

From proof of Proposition B.3 one has that $\tilde{C} := \Delta_F^2/(8s^2 \, |A|^2 \, \mathcal{L}^2)$, where Δ_F is a distribution dependent constant. Therefore, one has that $R_2^{\pi_a}(F,N) \leq C_7$ for some positive finite constant C_7, independent of N.

Finally, combining (B-21), (B-25) and (B-26) results in the following bound

$$R^{\pi_a}(F,N) \leq \sum_{i=\tilde{i}}^{\lceil (\ln N)^{1+\varepsilon} \rceil} \mathbb{E}\left\{ \sum_{S \in \tilde{\mathcal{E}}_i} \Delta_S^F \right\} + C_8,$$

for a positive finite constant C_8.

Consider now the case when $n_i := \max\{\lfloor e^{i/H} \rfloor, n_{i-1}+1\}$ for $i \geq 2$, with $H := (1+\delta)/\tilde{C}$ and $\delta > 1$. Following the arguments in step 1 above, one has that

$$\begin{aligned} R_1^{\pi_a}(F,N) \quad &\leq \quad n_{i''} \, \Delta^F_{max} + \sum_{i=i''}^{\lceil H \ln N \rceil} \mathbb{E}\left\{ \sum_{S \in \tilde{\mathcal{E}}_i} \Delta_S^F \right\} + \Delta^F_{max} \sum_{i=i''}^{\infty} |A| \left(C' i + \tilde{K}\right) \exp\left\{-\tilde{C}(i - |A| - 1)\right\} \\ &= \quad \sum_{i=i''}^{\lceil H \ln N \rceil} \mathbb{E}\left\{ \sum_{S \in \tilde{\mathcal{E}}_i} \Delta_S^F \right\} + \tilde{C}_6, \end{aligned}$$

where $\tilde{C}_6 > 0$ is a finite constant. To bound $R_2^{\pi_a}(F,N)$ note that $n_{i+1} - n_i \leq C_2 \, e^{\frac{i}{H}}$ for $i \geq i''$, where $C_2 := e^{1/H} - e^{-1/H}$. The arguments in Step 2 above, and that $H > 1/\tilde{C}$ imply that

$$R_2^{\pi_a}(F,N) \quad \leq \quad n_{i''} \Delta^F_{max} + C_2 \left(\tilde{K} + C''\right) \Delta^F_{max} \sum_{i=i''}^{\infty} \exp\left\{-\tilde{C}(i - |A| - 1) + i/H\right\} = \tilde{C}_7,$$

for some finite positive constant $\tilde{C}_7$, independent of N. With this, one obtains that

$$R^{\pi_a}(F,N) \leq \sum_{i=i''}^{\lceil H \ln N \rceil} \mathbb{E}\left\{ \sum_{S \in \tilde{\mathcal{E}}_i} \Delta_S^F \right\} + \tilde{C}_8,$$

for a positive finite constant $\tilde{C}_8$.

Step 3. From Propositions B.3 and B.4,

$$\sum_{i=1}^{\infty} \mathbb{P}\left\{U_i^c \cup \tilde{U}_i^c\right\} \leq \hat{K} \exp\left\{-\tilde{C}\; i\right\} < \infty,$$

where $U_i = \left\{\cup_{S \in \mathcal{E}_i} S \subseteq \cup_{S \in \mathcal{E}_{i-1}} S\right\}$, $\tilde{U}_i = \{(C_i, \mathcal{E}_i) \in \Gamma(\mathbb{E}\{B_n\})\}$, and $\hat{K}$ is a positive finite constant. Thus, using Borel-Cantelli we have that $(C_i, \mathcal{E}_i)$ converges a.s. to some $(C_\infty, \mathcal{E}_\infty) \in \Gamma(\mathbb{E}\{B_n\})$. Let $i^*(\omega)$ denote the last cycle at which the exploration set changes on sample path ω. We have that

$$\mathbb{E}\left\{\sum_{i=1}^{i^*} \sum_{S \in \tilde{\mathcal{E}}_i} \Delta_S^F\right\} \leq G\Delta_{\max}^F \mathbb{E}\{i^*\} \leq G\Delta_{\max}^F \sum_{i=1}^{\infty} \mathbb{P}\{i^* \geq i\} \leq G\Delta_{\max}^F \hat{K} \sum_{i=2}^{\infty} \frac{\varrho^i}{1-\varrho} < \infty,$$

where $\varrho := \exp\{-\tilde{C}\} < 1$. Thus, we conclude that for both selections of $\{n_i : i \geq 1\}$

$$R^{\pi_a}(F,N) \leq (\ln N)^{1+\varepsilon}\, \mathbb{E}\left\{\sum_{S \in \mathcal{E}_\infty} \Delta_S^F\right\} + C_9, \text{ and } R^{\pi_a}(F,N) \leq \frac{(1+\delta)8s^2\,|A|^2\,\mathcal{L}^2}{\Delta_F^2} \ln N \mathbb{E}\left\{\sum_{S \in \mathcal{E}_\infty} \Delta_S^F\right\} + \tilde{C}_9$$

for positive finite constants C_9 and $\tilde{C}_9$. The result follows from noting that $\sum_{S \in \mathcal{E}_\infty} \Delta_S^F = z_{OCP}^*(B_\infty)$ (a.s.) for some B_∞ because $(C_\infty, \mathcal{E}_\infty) \in \Gamma(\mathbb{E}\{B_n\})$, by the definition of $\Gamma(\mathbb{E}\{B_n\})$. □

Theorem 5.2. *OCP and R-OCP are in P for weighted basis or independent set matroid minimization problems.*

Proof. From the proof of Lemma 4.8 we know that there exists a unique critical set. Moreover, such a set can be found in polynomial time (e.g. by solving $|A|$ instances of $f(\cdot)$). Let C^* denote the unique critical set and $R : 2^{\mathcal{N}} \to \mathbb{Z}_+$ be the rank function of the matroid. We claim that both OCP and R-OCP can be solved through the Linear Programming problems given by

$$\min \quad \sum_{l=1}^{r} \sum_{a \in A} b_a x_a^l \tag{B-27a}$$

$$s.t. \qquad 1 \leq \sum_{l=1}^{r} x_a^l, \quad a \in C^* \tag{B-27b}$$

$$\sum_{a \in S} x_a^l \leq R(S), \qquad S \subseteq A, l \in \{1,\ldots,r\} \tag{B-27c}$$

$$0 \leq x_a^l \leq 1, \qquad a \in A, l \in \{1,\ldots,r\}, \tag{B-27d}$$

for $r = 1$ to $r = |A|$ (for the basis problem we also add $\sum_{a \in S} x_a^l = R(S)$ for all l). Indeed, formulation (B-27) is the fractional covering of the critical set with at most r solutions of the matroid and if we change $0 \leq x_a^l \leq 1$ to $x_a^l \in \{0,1\}$ we have the standard covering with exactly r solutions of the matroid. By Lemma A.3, for both OCP and R-OCP it suffices to consider cases $r \in \{1, \ldots, |A|\}$; we just need to evaluate the regret for each case and pick the best. This proves the result for R-OCP. For OCP, we simply show that every extreme point of the feasible region of (B-27) is integral. Because the feasible region of the basis problem is a face of the independent set problem it suffices to prove this result for the latter one. For this, we need a few auxiliary results. We begin with a well known Lemma (Schrijver 2003).[20]

Lemma B.6 (Uncrossing Technique). *Let*

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

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

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

We use the following corollary of Lemma B.6.

Corollary B.7. *Let P be the independent set or base polytope of a matroid and let $x \in P$. If $x_a \in (0,1)$, then there exist $\varepsilon > 0$ and $a' \in A \setminus \{a\}$ such that $x_{a'} \in (0,1)$ and $x \in \operatorname{conv}\{\overline{x}, \underline{x}\}$, for $\overline{x}, \underline{x} \in P \setminus \{x\}$ defined by*

$$\overline{x}(\varepsilon, a, a') := x + \varepsilon \left(e^a - e^{a'}\right) \; ; \quad \underline{x}(\varepsilon, a, a') := x + \varepsilon \left(e^{a'} - e^a\right). \tag{B-28}$$

Proof of Corollary B.7. Let $W_0 \subset W_1 \subset W_2 \subset \ldots \subset W_k$ be the maximal chain from Lemma B.6 (with $W_0 = \emptyset$). If $k = 0$ then the result follows trivially ($x \in \operatorname{int}(P)$), so we will assume that $k \geq 1$. Let l_0 be the smallest $l \in \{1, \ldots, k\}$ such that $a \in W_l$. There exists $a' \in W_{l_0} \setminus \{a\}$ such that $x_{a'} \in (0,1)$: to see this, note that $R(W_{l_0-1}) \in \mathbb{Z}_+$, and that

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

thus one can always find such an a' in $W_{l_0} \setminus (W_{l_0-1} \cup \{a\})$. For any choice of a' we have that

[20] An explicit proof can be found in `http://theory.stanford.edu/~jvondrak/CS369P-files/lec10.pdf`

$y \in \{\overline{x}, \underline{x}\}$ defined in (B-28) satisfies $\sum_{h \in W_l} y_h = r(W_l)$ for all $l \leq k$, thus by Lemma B.6 $y \in P$ for $\varepsilon < \min\{x_a, x_{a'}\}$ (so that $y \geq 0$ and $\sum_{h \in S} y_h \leq R(S)$ for constraints not active at x). The result follows since $x \in \operatorname{conv}\{\overline{x}, \underline{x}\}$ by construction. □

Next, we use this corollary to show the integrality of (B-27).

Proposition B.8. *The feasible region of* (B-27) *has integral extreme points.*

Proof of Proposition B.8. Let x be a fractional extreme point of (B-27). Without loss of generality x^1 has a fractional component $x^1_{i_1} \in (0,1)$: we will reach a contradiction by constructing a set of solutions whose convex hull contains x. Corollary B.7 implies that there exist $\varepsilon_1 > 0$, j_1 such that $x^1 \in \operatorname{conv}\left\{\overline{x}^1(\varepsilon_1, i_1, j_1), \underline{x}^1(\varepsilon_1, i_1, j_1)\right\}$, with $\overline{x}^1(\varepsilon_1, i_1, j_1), \underline{x}^1(\varepsilon_1, i_1, j_1) \in P$ defined by (B-28).

Define $\tilde{C} := \left\{h \in C \,:\, 1 = \sum_{l=1}^r x^l_h\right\}$. If $\{i_1, j_1\} \cap \tilde{C} \neq \emptyset$, by symmetry we may assume w.l.o.g. that $j_1 \in \tilde{C}$ (if not rename i_1 and j_1). Because $x^1_{j_1} \in (0,1)$, $j_1 \in \tilde{C}$ and the definition of $\tilde{C}$, there exists $l \in \{2, \ldots, r\}$ such that $x^l_{j_1} \in (0,1)$. We assume w.l.o.g. that $l = 2$ and let $i_2 = j_1$. By Corollary B.7 applied to x^2 we have that there exist $\varepsilon_2 > 0$, j_2 and $\overline{x}^2(\varepsilon_2, i_2, j_2), \underline{x}^2(\varepsilon_2, i_2, j_2) \in P$ defined by (B-28) such that $x^2 \in \operatorname{conv}\left\{\overline{x}^2(\varepsilon_2, i_2, j_2), \underline{x}^2(\varepsilon_2, i_2, j_2)\right\}$. If $\{i_1, j_2\} \cap \tilde{C} \neq \emptyset$, again by symmetry we can assume $j_2 \in \tilde{C}$ and repeat this construction and continue until we obtain a sequence $i_1, j_1 = i_2, \ldots, j_{k-1} = i_k, j_k$ and $\varepsilon_1, \ldots, \varepsilon_k$ for $k \geq 1$ which satisfies one of the following conditions:

1. $\{i_1, j_k\} \cap \tilde{C} = \emptyset$.

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

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

The result follows by noting that (B-27) can be solved in polynomial time because (B-27c) can be separated in polynomial time (Schrijver 2003, Corollary 40.4c). □