Optimal Sequential Stochastic Shortest Path Interdiction

Juan S. Borrero

School of Industrial Engineering and Management, Oklahoma State University, juan.s.borrero@okstate.edu

Denis Sauré, Natalia Trigo

Department of Industrial Engineering, University of Chile, Santiago, Chile, dsaure@dii.uchile.cl, n.trigo@isci.cl

We consider the periodic interaction between a leader and a follower in the context of network interdiction where, in each period, the leader first blocks (momentarily) passage through a subset of arcs in a network, and then the follower traverses the shortest path in the interdicted network. We assume that arc costs are stochastic, and that while their underlying distribution is known to the follower, it is not known by the leader. We cast the problem of the leader, who aims at maximizing the cumulative cost incurred by the evader, using the multi-armed bandit framework. Such a setting differs from the traditional bandit in that the feedback elicited by playing an arm is the reaction of an adversarial agent. After developing a fundamental limit in the achievable performance by any admissible policy, we adapt traditional policies developed for linear bandits to our setting. We show that a critical step in such an adaptation is to ensure that the cost vectors imputed by these algorithms lie within a polyhedron characterizing information that can be collected without noise and in finite time. Within such a polyhedron, the problem can be mapped into a linear bandit. The polyhedron has exponentially many constraints in the worst case, which are indirectly tackled by solving several mathematical programs. We test the proposed policies and relevant benchmarks through a set of numerical experiments. Our results show that the adapted policies can significantly outperform the performance of the base policies at the price of increasing their computational complexity.

Key words: OR in defence, Combinatorial Optimization, Network interdiction, Multi-armed bandits

1. Introduction

Motivation and Objective. In shortest-path interdiction, a Leader (or interdictor) blocks passage through a set of arcs or vertices in a network to disrupt the operation of a Follower (or evader) who later traverses a path within the interdicted network. This interaction is adversarial: whereas the evader chooses the path of minimal cost, the interdictor aims at maximizing the evader's cost. Shortest-path interdiction is a particular case of network interdiction problems, which are problems that received significant attention in the literature, see Smith and Song (2020) for a survey on the subject.

Network interdiction problems originally conceived a single-period interaction between the leader and follower in which all parameters are deterministic and known to both agents (Fulkerson and Harding 1977, Malik et al. 1989). In the last decades, various extensions to the original setting have been presented, each aimed at introducing realistic features into the models (Cormican et al. 1998, Bayrak and Bailey 2008, Borrero et al. 2016, Kosmas et al. 2023). Motivated by their application in surveillance and homeland security domains, in this paper we consider settings where the follower has more knowledge about the cost structure of the network, relative to the leader, and the agents interact repeatedly through time. See, for example Steinrauf (1991), Gift (2010) that discuss such an application in the context of the US counter-narcotic efforts.

Starting with Borrero et al. (2016), there has been a handful of studies focusing on sequential interdiction problems where the leader, unlike the follower, does not know the underlying cost parameters (Borrero et al. 2019, Yang et al. 2021, Ketkov and Prokopyev 2020, Borrero et al. 2022*a*) and must infer them in an adaptive fashion from the observation of the follower's responses. Unlike traditional work in sequential decision-making under uncertainty, this stream of literature has considered so far (to the best of our knowledge) deterministic settings where the leader focuses on minimizing the number of periods it takes to reconstruct the *full-information* interdiction action implemented by an oracle interdictor who knows the underlying parameters. In our work, we extend upon this setting by considering the case of stochastic feedback, namely when the costs observed by the leader include random perturbations - which categorizes it naturally as a bandit.

Model. We consider a model of periodic interaction between a leader (or interdictor) and a non-strategic follower (or evader). Considering a fixed graph, in each period the interdictor acts first by blocking passage on a subset of arcs of the graph, and then the evader responds by traversing an unblocked path between a pair of fixed source and terminal nodes on the graph. The cost incurred by the evader is the sum of the costs of the arcs in the path which form an iid sequence through time.

We assume that the evader does not observe cost realizations but knows their probability distribution. Therefore, the evader uses the expected costs when choosing a path. In addition, we assume that the evader is not forward-looking and selects the path with the smallest expected cost (which we refer to as the shortest path). The first of these assumptions accommodates settings where cost variations are only observed when/after traversing a path (e.g. one can reasonably assume detecting drug movement through a pass is intrinsically binary and random, and cannot be predicted beyond its probability of occurrence). The second assumption avoids considering strategic inter-temporal interactions. However, when facing an oracle leader, a shortest-path response is part of any sub-game perfect strategy, thus this assumption can be thought of as an asymptotic approximation.

We assume that in each period the leader can block up to a finite and constant number of arcs in the graph, and only for the duration of the period. The leader does not know the cost distribution up front (except for its range), but instead, observes: (*i*) the path selected by the evader on each period, which the leader knows is a shortest path in the interdicted network; and (*ii*) the cost realizations associated with the arcs in the aforementioned path. Using this form of *semi-bandit* feedback (Audibert et al. 2013), the leader must select a sequence of interdiction actions to maximize the cumulative cost incurred by the follower. Following extant work, we frame the leader's problem as a multi-armed bandit with a combinatorial number of correlated arms (Cesa-Bianchi and Lugosi 2012, Gai et al. 2012, Modaresi et al. 2020). We then consider the problem of minimizing the leader's cumulative pseudo-regret relative to the policy adopted by an oracle leader with prior knowledge of the cost distribution.

Main contributions. The first contribution is establishing a fundamental limit on the performance of any admissible policy. Following the arguments in the seminal work of Lai and Robbins (1985), we restrict attention to policies that perform consistently well across instances and then use a change of measure argument to show that policies that perform well in alternative instances of the problem, must constantly elicit feedback from a special class of arcs. This result implies a minimal exploration frequency for a class of arcs in the network (see Theorem 1), which can be translated into a lower bound on performance that exhibits a logarithmic dependency on the length of the horizon, and whose definite form is the solution to a linear program of combinatorial size: see Corollary 1. The logarithmic dependence in the horizon confirms intuition coming from traditional bandits, speaks of the cost-efficient collection of information necessary to confirm the optimality of the full-information solution, and is aligned with similar results in combinatorial bandits (Modaresi et al. 2020).

A second contribution relates our problem to linear bandits (Auer 2002). As the evader's responses are the shortest paths (in expectation) in the interdicted network, it is possible to characterize a polyhedron $\mathcal{U}$ that contains the mean cost vector almost surely (see equation (4)). Within $\mathcal{U}$ the problem can be thought of as a linear bandit, and as such it could be tackled by state-of-the-art linear bandit policies (see Remark 3). However, such polyhedron has (a priori) an exponentially large number of constraints in the worst case. Thus, checking whether a cost vector belongs to $\mathcal{U}$ might be challenging, which implies that adapting policies for linear bandits for this setting might not be a straightforward task.

A third contribution is thus adapting standard policies for the linear bandit to the setting at hand, namely the Thompson Sampling (TS) (Thompson 1933) and UCB (Auer et al. 2002) policies. With a focus on asymptotic (in the horizon) optimality, the proposed policies begin with an initialization phase in which $\mathcal{U}$ is found by implicit enumeration. For the case of the TS policy, we first develop a short-term approximation for the posterior distribution over $\mathcal{U}$ based on a finite sample collected using a hit-and-run scheme (Berbee et al. 1987). Then we use such a sample to approximate the posterior sampling of cost vectors in $\mathcal{U}$. For the case of the UCB policy, we consider a Bayesian version of the policy, and a multi-variate normal approximation of the posterior distribution over $\mathcal{U}$, which allows us to use ellipsoidal uncertainty sets. We use mixed-integer programming to constrain the selection of the cost vector within the feasible polyhedron. The resulting policy is closely related to state-of-the-art implementations of UCB-type policies for the linear bandit.

Finally, we illustrate the practical implementation of the proposed policies, as well as their empirical performance, through a set of numerical experiments. In doing so, we also consider relevant benchmarks, namely more direct adaptations of the TS and UCB policies which do not consider the set $\mathcal{U}$. In this regard, our results show that carefully tailoring policies for the setting at hand results in significantly better empirical performance. A feature of the proposed policies is that their per-period practical complexity is closely related to that of solving a mixed-integer programming implementation of the shortest-path interdiction.

Organization of the manuscript. Section 2 provides a review of the relevant literature, and in Section 3 we present our model of stochastic sequential shortest-path interdiction. In Section 4 we develop a fundamental limit on the performance attainable by any admissible policy. Then, in Section 5, after presenting an initialization phase common to all policies (Section 5.1), we introduce our adaptations of the Thompson Sampling (Section 5.2) and Bayes-UCB (Section 5.3) policies. Then, Section 6 presents a series of numerical experiments in which we test the performance of the proposed policies and relevant benchmarks. Finally, Section 7 presents our final remarks. Proofs of all results are relegated to Appendix A.

Notation. Throughout the manuscript, we use the following notation. For $x \in \mathbb{R}$ we define $(x)^+ := \max\{0, x\}$ and $(x)^- := \max\{0, -x\}$. For $x \in \mathbb{R}^n$ and a positive-definite matrix $\Sigma \in \mathbb{R}^{n\times n}$, we define $\|x\|_\Sigma := (x^\top \Sigma^{-1} x) \in \mathbb{R}_+$. For $n \in \mathbb{N}$, we define $[n] \equiv \{1, \ldots, n\}$, and let $|B|$ denote the cardinality of a set B. Graphs are defined on a fixed set of nodes N, and by subsets of a set of arcs A: we denote by $G \equiv (N, A)$ the full underlying graph, and by $G(B)$ the graph defined excluding the arcs in B $(G(B) \equiv (N, A \setminus B))$, for all $B \subseteq A$. We let $\mathcal{P}(B)$ denote the set of paths in $G(B)$ between nodes 1 and $n \equiv |N|$, and denote $\mathcal{P} \equiv \mathcal{P}(\emptyset)$. Finally, we let 2^B denote the power set of a set B, and assume all random elements are defined within probability space $(\Omega, \mathbb{F}, P)$.

2. Literature Review

Our work borrows from and contributes to the network interdiction and multi-armed bandit literature. We position our work relative to both fields, highlighting the novelty of our model and results.

Network Interdiction. In shortest path interdiction (Fulkerson and Harding 1977) and max flow interdiction (McMasters and Mustin 1970, Ghare et al. 1971), a leader aims to maximally disrupt a follower's flow through a network. Whereas the original single-stage deterministic interdiction problems are NP-complete (Ball et al. 1989, Wood 1993), integer programming-based solution approaches are viable in practice (see e.g. Wood (1993), Israeli and Wood (2002)). See Smith and Song (2020) for a recent survey on network interdiction.

Many works relax the deterministic nature of the classical setup and have considered various forms of uncertainty. For example, Cormican et al. (1998), Morton (2010), Kang and Bansal (2023) consider settings in which the effectiveness of the interdiction effort is uncertain, and Hemmecke et al. (2003) consider the case where the network topology is not known upfront. Bayrak and Bailey (2008) consider deterministic settings where the leader and evader's costs do not coincide, thus departing from the max-min setup. Holzmann and Smith (2021), Sadana and Delage (2023) study settings where the leader is allowed to randomize his actions. Closer to our setting, Nguyen and Smith (2022*b*) and Nguyen and Smith (2022*a*) consider settings where costs are uncertain and the leader maximizes the expected or conditional value at risk of the evader's cost. These latter works can be thought of as also considering information asymmetries, as the evader does not know the effect of the interdiction actions with certainty. In a similar setup, Azizi and Seifi (2023) considers a robust approach to handling the evader's lack of knowledge.

In the last decade, multi-period network interdiction has been considered in deterministic settings where agents interact repeatedly over time and a budget is allocated through time (Ajay Malaviya and Sharkey 2012) or on each period (Soleimani-Alyar and Ghaffari-Hadigheh 2017), but not dynamically within a period (see Sefair and Smith (2016) for a dynamic setting). Closer to our work is the sequential shortest-path interdiction setting of Borrero et al. (2016) where the leader is initially unaware of the (deterministic) network costs and must learn them by observing the evader's responses through time. Yang et al. (2021) and Borrero et al. (2022*b*) extend such a setting by considering limited forms of feedback and by allowing the leader and evaders' costs to differ, respectively. Our work can be seen as extending that in Borrero et al. (2016) to settings where network costs are stochastic and form an iid sequence. In this regard, two observations are in order. First, by introducing cost uncertainty, the setting at hand can be envisioned as a multi-armed bandit with multiple simultaneous plays and correlated rewards; as such we consider the objective of minimizing the cumulative pseudo-regret (the standard criterion in the bandit literature) as opposed to the concept of time-stability. Second, the greedy and pessimistic policies in Borrero et al. (2016) can be interpreted as operating under the *optimism in the face of uncertainty* principle behind, for example, the celebrated UCB policy (Auer et al. 2002). This latter observation allows us to connect prior work in deterministic settings to the multi-armed bandit.

Multi-armed bandit. The multi-armed bandit (Thompson 1933, Robbins 1952) is a classical framework for studying dynamic decision-making under model uncertainty. In its traditional formulation, a decision-maker attempts to maximize his cumulative (stochastic) reward by pulling arms sequentially over time from an ex-ante identical set of arms. The setting features the classical exploration (learning reward distributions, including ex-post suboptimal ones) vs. exploitation (pulling the arm thought to deliver the largest reward) trade-off. Because in general optimal policies can not be computed in closed form (except for some settings, see Gittins (1979)), the literature focuses instead on achieving asymptotic optimality. In this sense, Lai and Robbins (1985) show that efficient policies must try every suboptimal arm $O(\log t)$ periods, where t denotes the length of the horizon.

Envisioning each interdiction action as an arm, our setting can be casted as a multi-armed bandit with combinatorially many arms and correlated rewards. Bandit settings with large sets of arms have been studied extensively in the last decades. Kleinberg et al. (2008) considers settings with a continuum of arms forming a metric space; see Bubeck et al. (2011) for a review of settings with continuum arms. Closer to our setting, (Cesa-Bianchi and Lugosi 2012, Gai et al. 2012, Modaresi et al. 2020) study adversarial and stochastic settings with combinatorially many arms. There, like in our work, reward correlation is a byproduct of the linearity of the cost function. Modaresi et al. (2020) adapt the arguments in Lai and Robbins (1985) and presents a $O(\log t)$ bound on performance that is related to the solution to a combinatorial problem. Our lower bound result can be seen as performing the same type of adaptation to the setting at hand.

A key result in our work is that after finite and deterministic information about the underlying parameters is collected, the structure of the setting is equivalent to that of a linear bandit (Auer 2002), which has been thoroughly studied in the past. A significant portion of the work in linear bandits presents different adaptations of the UCB policy of Auer et al. (2002): in such adaptations, the decision-maker selects an action while considering the best possible realization of the underlying mean cost vector within an uncertainty region, following the *optimism in the face of uncertainty* principle. The distinction between these policies comes from the uncertainty region used (see Section 6 for more details on this class of policies). In this regard, while some work adopts rectangular uncertainty sets (Chen and Zhang 2009, Gai et al. 2012), the policies with the best finite-time performance guarantees adopt ellipsoidal uncertainty sets (Dani et al. 2008, Rusmevichientong and Tsitsiklis 2010, Abbasi-Yadkori et al. 2011). Another important class of policies are those based on posterior sampling (Thompson 1933). Within a parametric Bayesian setup, in each period these policies sample from the underlying parameters' posterior distribution and select the best arm for such a sample. The finite-time performance guarantees of TS policies are closely related to that of UCB policies for the case of linear bandits (Russo and Van Roy 2014). Section 5 shows that adapting the UCB and TS policies to our setting requires non-trivial modifications to the uncertainty sets and priors used. Our work shows how to incorporate such modifications and maintain computational tractability in practice.

3. Model formulation

Consider sequential shortest path interdiction on graph G throughout a finite horizon of T periods. In each period $t \in [T]$, the interdictor first blocks a finite set of arcs B^t, and then the evader responds by traversing a path $S^t \in \mathcal{P}(B^t)$. We let $G^t \equiv G(B^t)$ denote the graph available to the evader at period $t \in [T]$.

The evader's response. We assume that in each period $t \in [T]$ the evader faces a linear cost function, modulated by a random cost vector $\boldsymbol{c}^t(\omega) := (c_a^t(\omega) : a \in A)$, so that the cost associated with traversing path S in period t, is given by $r(S, \boldsymbol{c}^t)$, where

$$r(S, \boldsymbol{c}) := \sum_{a \in S} c_a, \quad S \in \mathcal{P}, \, \boldsymbol{c} \in \mathbb{R}^{|A|}.$$

(We drop the dependence on $\omega \in \Omega$ of random elements when it is clear from the context.) We further assume that $(\boldsymbol{c}^t : t \in [T])$ form an i.i.d. sequence, and let F denote the (common) distribution of $\boldsymbol{c}^t$, $t \in [T]$. We let $\mathbb{C} := \prod_{a \in A} [l_a, u_a]$ denote the support of F, where l_a and u_a denote lower and upper bounds on the cost of arc $a \in A$, respectively. We make the following key assumption about the initial information the evader has.

Assumption 1 **(A1)**. *The evader knows F but does not observe $\boldsymbol{c}^t$, thus selects S^t minimizing its expected cost, $t \in [T]$.*

Considering **A1**, we assume that upon observing G^t, the evader chooses $S^t \in \mathcal{P}^t \equiv \mathcal{P}(B^t)$ to minimize its expected cost. Defining $\mu := \mathbb{E}_F[\boldsymbol{c}^t]$, in period $t \in [T]$, the evader traverses

$$S^t \in \arg\min \left\{ r(S, \mu) : \, S \in \mathcal{P}^t \right\}.$$

Note that, conditional on the interdictor action, the evader solves a deterministic shortest-path problem on each period. Despite this fact, our informational setup is such that in general the interdictor will not be able to anticipate the evader's response, as we detail next.

The interdictor's decision. We consider a constant interdiction budget $K < \infty$, so that on period $t \in [T]$ the interdictor chooses $B^t \in \mathcal{B} := \{B \subseteq A : \, |B| \leq K\}$, the set of feasible interdiction actions. We assume that the leader is interested in maximizing the cumulative cost incurred by the follower and make the following key assumption about the information available to the interdictor at the moment of selecting B^t.

Assumption 2 **(A2)**. *The interdictor does not know F, but knows its support $\mathbb{C}$. In addition, at period t he observes S^t and $\{c_a^t : a \in S^t\}$ immediately after the evader traverses S^t, for $t \in [T]$.*

Note that if the leader had access to F, then he would implement $B^t = B^*$ for all $t \in [T]$, where

$$B^* \in \arg\max \{\min \{r(S,\mu) \colon S \in \mathcal{P}(B)\} \colon B \in \mathcal{B}\}.$$

We call the formulation above the *full information problem* and refer to B^* as the *full information solution*. Similarly, we define the evader's response to the full information solution as

$$S^* := \arg\min \{r(S,\mu) : S \in \mathcal{P}(B^*)\},$$

which we assume is unique, to avoid unnecessarily cluttered notation (in the sequel, one can think that μ is chosen at random from an absolutely continuous distribution). The interdictor cannot implement the full information solution from the start as F is initially unknown (and so is μ). In this regard, the interdictor's actions must be adjusted to the information available at the beginning of period t. Define

$$\mathcal{F}^t := \sigma\left((S^s, \{c_a^s : a \in S^s\}) \colon s < t\right), \quad t \in [T],$$

and let $\mathcal{F} := \{\mathcal{F}^t : t \in [T]\}$ denote the filtration generated by the information revealed to the interdictor through his interaction with the follower. We say $\pi := \{\pi^t : t \in [T]\}$ is an admissible interdiction policy if it is a stochastic process adapted to $\mathcal{F}$ such that for all $t \in [T]$, we have that $B^{t,\pi} \equiv \pi^t \in \mathcal{B}$.

Remark 1. While a policy that implements B^* every period is admissible, it would perform poorly when applied to alternative cost distributions for which B^* is not a full-information solution. Thus, following the seminal work of Lai and Robbins (1985), we restrict our attention to policies that perform *consistently* well across all instances. ■

Following the bulk of the literature, we assume that the leader is interested in minimizing the expected regret associated with his actions. That is, for a distribution F and horizon T, we define the expected regret associated to an admissible policy π as

$$\mathcal{R}^\pi(T,F) := T \cdot r(S^*,\mu) - \sum_{t=1}^{T} r(S^t,\mu)$$

where the above depends on F through $\mu = \mathbb{E}_F\{\boldsymbol{c}^t\}$. (When it is clear from context, we will drop the dependence on the policy π when possible.)

4. Asymptotic Limit on Achievable Performance

In this section we establish an asymptotic limit on achievable performance for every admissible and *consistent* policy (we define the concept later in this section). Because of our interest in asymptotic performance, we first consider the information about μ that can be recovered by observing the evader's responses to the leader actions *in finite time*, and then the information that comes from observing the costs incurred by said responses. Later in the manuscript we borrow the ideas in this section to develop practical policies and focus on their finite-time performance.

Deterministic Finite-time Feedback. We make the following assumption on the follower's behavior, which states that the evader's decisions are consistent over time.

Assumption 3 **(A3)**. *If $S, S' \in \mathcal{P}^t$ for $t \in [T]$ and $S^t = S$, then $S^s \neq S'$ if $S \in \mathcal{P}^s$, for $s \in [T]$.*

Suppose that the leader interdicts the arcs in B^t in period $t \in [T]$: independent of $\boldsymbol{c}^t(\omega)$, the evader chooses a response, S^t, which (from **A3**) is the same for all periods $s \in [T]$ for which $B^s = B^t$. With this in mind, consider the information about μ that can be collected from observing the evader's response to every possible interdiction action, noting that this information can be obtained in finite time. For $\nu \in \mathbb{C}$, let $S(B,\nu)$ denote the evader's response when the leader blocks the arcs in $B \in \mathcal{B}$, and the mean cost vector is ν. After implementing all interdiction actions, the leader knows that

$$\mu \in \mathcal{U} := \{\nu \in \mathbb{C} : S(B,\nu) = S(B,\mu),\, B \in \mathcal{B}\}.$$

Note that $\mathcal{U}$ is the set of all (mean) cost vectors that can explain the feedback observed by the interdictor after implementing all possible interdiction solutions. We assume that the leader knows that $\mu \in \mathcal{U}$, as this information is deterministic (from **A3**) and can be obtained in finite time. Additional information on μ might be obtained by repeatedly observing cost realizations of chosen paths; from prior work on bandits, we anticipate that collecting such information will come at a cost (in terms of asymptotic regret), and will contribute to distinguishing settings where B^* is optimal, from others.

Remark 2. In order to compute $\mathcal{U}$, in principle it is necessary to implement a combinatorial number of interdiction actions. However, in Section 5 we show that in practice this task can be accomplished by implementing either a reduced set of actions (the proposed policies begin collecting such information in an initialization phase), or that the set $\mathcal{U}$ can be constructed on-the-go. ■

4.1. A lower bound on performance

We begin by narrowing down the set of policies of interest; following seminal work in multi-armed bandits (Lai and Robbins 1985), we restrict attention to policies that perform consistently well across all distributions F. Specifically, we consider policies such that for all distributions F are such that, for every $\alpha > 0$,

$$\mathcal{R}^{\pi}(T,F) = o(T^{\alpha}).$$

This set of *consistent* policies excludes those policies that perform well in a particular setting, at the expense of performing poorly in others. We *exploit* the consistency of the admissible policies to find a lower bound on the rate at which some suitably constructed sets of arcs must be traversed by the evader, asymptotically. In what follows, with some abuse notation, we let $B^*(\nu)$ and $S^*(\nu)$ denote the optimal interdiction action and evader's response to said action when the mean cost vector is $\nu \in \mathcal{U}$.

Suppose that $B^*(\nu) = B^*(\mu)$ for all $\nu \in \mathcal{U}$, then the interdictor can find $B^*(\mu)$ a.s. in finite time, and thus it is reasonable to expect that $\mathcal{R}^\pi(T,F) = O(1)$. Thus, assume that there exists $\nu \in \mathcal{U}$ such that $B^*(\nu) \neq B^*(\mu)$; moreover, observe that since $\nu \in \mathcal{U}$ then the feedback provided by implementing $B^*(\mu)$ under ν coincides with that obtained when the mean cost vector is given by μ (thus, by implementing $B^*(\mu)$ and observing the feedback one cannot differentiate whether the underlying mean cost vector is ν or μ). Finally, note that a consistent policy must necessarily deviate from implementing $B^*(\mu)$ (with a certain frequency, which we will characterize) so as to collect feedback that allows distinguishing ν from μ. More importantly, such feedback comes from observing the cost realizations for arcs $a \in A$ where $\mu_a \neq \nu_a$.

Considering the discussion above, we define the class $\mathcal{E} \subseteq 2^A$ of all sets $E \subseteq A \setminus S^*(\mu)$ such that if the mean cost of the arcs in E change (relative to μ), then it is possible that $B^*(\mu)$ is not longer optimal, i.e.

$$E \in \mathcal{E} \iff \exists\, \nu \in \mathcal{U} : \nu_a = \mu_a \text{ for } a \in A \setminus E \text{ and } B^*(\nu) \neq B^*(\mu).$$

As hinted above, there exist settings where $\mathcal{E} = \emptyset$, i.e. the feedback obtained in finite-time, in addition to that obtained from implementing $B^*(\mu)$, suffices to characterize $B^*(\mu)$, in which case we venture to say that a $O(1)$ regret might be attainable. Example 1 presents such an instance.

Example 1. Consider the shortest-path interdiction setting in Figure 1. We assume an interdiction budget of $K < k-1$, and consider the bounds $l_a = 0$ and $u_a = \infty$ for all $a \in A$. This setup is similar to that of traditional multi-armed bandit in the sense that there are $k-1$ non-intersecting paths. Note that, by implementing (a subset) of all possible interdiction actions, the leader is able to identify the $K+1$ paths with lowest mean cost, and their relative rank, and that such information suffices to solve the full information problem.

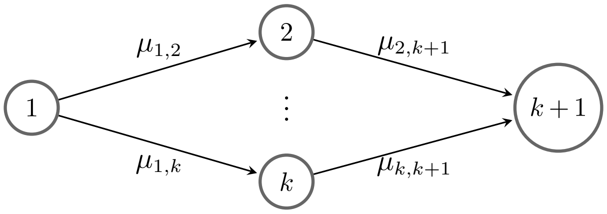

Figure 1 Shortest-path interdiction setting of Example 1. Each arc $a \in A$ is labeled with μ_a, its expected cost under F.

Specifically, suppose, without loss of generality, that

$$\mu_{1,j} + \mu_{j,k+1} \leq \mu_{1,j+1} + \mu_{j+1,k+1}, \quad j < k,$$

and define $B^j = \{(1,i)\colon i = 2,\ldots,j\}$, $j = 2,\ldots,K+1$. Note that

$$\mathcal{U} = \{\nu \in \mathbb{C}\colon S(B,\nu) = S(B,\mu),\, B = B^j, j = 2,\ldots,K+1\}$$

and note that when the mean cost vector is given by $\nu \in \mathcal{U}$, then any of the K paths of the form $1 - j - (k+1)$, $j = 2,\ldots,K+1$, are shorter than any path of the form $1 - j - (k+1)$, $j = K+2,\ldots,(k+1)$. Thus, for any $\nu \in \mathcal{U}^{K+1}$, the optimal interdiction solution is to interdict $B^*(\mu)$, i.e., to interdict the paths $1 - j - (k+1)$, $j = 2,\ldots,K+1$. This observation implies that $\mathcal{E} = \emptyset$, as desired. ■

Alternatively, there exist settings where $\mathcal{E} \neq \emptyset$ and thus the feedback obtained in finite time, and that coming from implementing $B^*(\mu)$ do not suffice to guarantee the optimality of $B^*(\mu)$. Example 2 presents such a setting.

Example 2. Consider the network interdiction setting depicted in Figure 2, and suppose that the interdiction budget is $K = 2$, and that $l_a = 0$ and $u_a = \infty$ for all $a \in A$. Define the sub-paths $l(j) \equiv (1, j+1) \to (j+1, 6)$ and $r(j) \equiv (6, j+6) \to (j+6, 11)$, for $j < 5$, and assume without lost of generality that

$$\sum_{a\in l(j)} \mu_a < \sum_{a\in l(j+1)} \mu_a, \quad \text{and} \quad \sum_{a\in r(j)} \mu_a < \sum_{a\in r(j+1)} \mu_a, \quad j < 4.$$

For any $j \leq 2$, define

$$\mathcal{B}(j) := \{B \in \mathcal{B} : B \cap l(i) \neq \emptyset,\, i \leq j \wedge B \cap r(i) \neq \emptyset,\, i \leq 2 - j\}.$$

Pick μ so that $B^*(\mu) \in \mathcal{B}(1)$, i.e. it is optimal to interdict the sub-paths $l(1)$ and $r(1)$, which implies that $S^*(\mu) = l(2) \to r(2)$. Observe that because $B^*(\mu) \in \mathcal{B}(1)$, the cost of $l(3) \to r(1)$ is smaller that $l(2) \to r(2)$; similarly, the cost of $l(1) \to r(3)$ is smaller than $l(2) \to r(2)$.

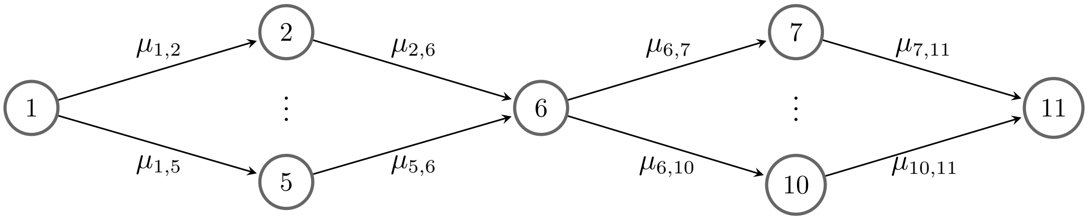

Figure 2 Interdiction setting of Example 2. Each arc $a \in A$ is labeled with μ_a, its expected cost under F.

Consider now the set $\mathcal{U}$, which contains all cost vectors consistent with the evader's answers to all interdiction actions. It is readily seen that the information in $\mathcal{U}$ implies that $l(j)$ is shorter than $l(j+1)$, $j=1,2,3$; likewise, it implies that $l(j)$ is shorter than $l(k)$, $k=4,5$, $j=1,2,3$. Note, however, that the information in $\mathcal{U}$ does not imply that $l(4)$ is shorter than $l(5)$ (because $l(4)$ and $l(5)$ are never used by the evader). Similar implications hold for the $r(j)$s. Importantly, the information in $\mathcal{U}$ does not imply that $l(3)\to r(1)$ is shorter than $l(2)\to r(2)$; neither that $l(1)\to r(3)$ is shorter than $l(2)\to r(2)$.

We claim that for any given μ, $E=\{(1,4),(1,5)\}$ is an element of $\mathcal{E}$. Indeed, let $\nu_a=\mu_a$ for all $a\in A$, $a\notin E$, and set $\nu_a=\mu_a+M$, $a\in E$, where M is an arbitrarily large constant. Then, it is readily seen that $\nu\in\mathcal{U}$. On the other hand, consider the action $\hat{B}$ that interdicts $(1,2)$ and $(1,3)$. Then $S(\hat{B},\mu)=l(3)\to r(1)$ is a shortest path and clearly, $r(S^*(\mu),\nu)<r(S(\hat{B},\mu),\nu)$. This argument also shows that $\{(1,4),(5,6)\}$, $\{(1,5),(4,6)\}$, and $\{(4,6),(5,6)\}$ are all elements of $\mathcal{E}$; following the same arguments, one can show that $\{(6,9),(6,10)\}$, $\{(6,9),(10,11)\}$, $\{(6,10),(9,11)\}$, and $\{(9,11),(10,11)\}$ are elements of $\mathcal{E}$. We note that these sets are not the only elements of $\mathcal{E}$. For instance, $E=\{(1,2)\}$ is also in $\mathcal{E}$. Indeed, let $\mu(s)$ denote the cost under μ of the path segment s and define $\nu_a=\mu_a$ for all $a\neq(1,2)$, and $\nu_{1,2}=\mu_{1,2}+\epsilon$, with $\epsilon\in(\mu(l(2))+\mu(r(2))-\mu(l(1))-\mu(r(3)),\mu_{1,3}-\mu_{1,2})$. Then $\nu\in\mathcal{U}$ and $r(S^*(\mu),\nu)<r(S(\hat{B},\mu),\nu)$. Finally, also note that not all subsets of arcs are elements of $\mathcal{E}$. For example, any subset of arcs in $l(4)\to r(4)$ is not an element of $\mathcal{E}$, as for any $\nu\in\mathcal{U}$ those arcs cannot be in any shortest path. ■

By construction, for each $E\in\mathcal{E}$ there exists (at least) one admissible alternative mean cost vector under which $B^*(\mu)$ is no longer optimal. The following lemma establishes that the optimal interdiction action under such an alternative cost configuration must necessarily interdict path $S^*(\mu)$, and vice-versa, the evader's optimal response under the new configuration is interdicted under $B^*(\mu)$. Moreover, the lemma shows that the shortest path after optimal interdiction under the alternative cost configuration must contain elements of E.

Lemma 1. *Assume that $\mathcal{E}\neq\emptyset$. For each $E\in\mathcal{E}$, there exist $\nu\in\mathcal{U}$ such that: $B^*(\nu)\neq B^*(\mu)$; $S^*(\nu)\cap E\neq\emptyset$; $S^*(\nu)\notin\mathcal{P}(B^*(\mu))$; and $S^*(\mu)\notin\mathcal{P}(B^*(\nu))$.*

By construction, for each $E\in\mathcal{E}$ there exists an alternative mean cost configuration ν under which: (i) the full-information solution $B^*(\mu)$ is suboptimal, and (ii) the feedback observed when implementing $B^*(\mu)$ coincides with that observed when the cost vector is μ. Because a consistent policy attains a sub-polynomial regret under both μ and ν, when the actual cost vector is μ, such a policy must necessarily collect feedback from arcs in E (because of *(ii)*) if it hopes to establish the sub-optimality of $B^*(\nu)$ (because of *(i)*). Because such a judgment call cannot be based on a finite sample if one is to attain asymptotic optimality, feedback from arcs in E must be collected at some minimal frequency. Theorem 1 below establishes a lower bound on such a frequency.

For $E \in \mathcal{E}$ and $t \in [T]$, define $\tau(E,t)$ as the number of periods up to (and including) t such that the follower's response included an arc in E. That is,

$$\tau(E,t) := \sum_{s=1}^{t} \mathbf{1}\left\{S^t \cap E \neq \emptyset\right\}.$$

THEOREM 1. *Assume that $\mathcal{E} \neq \emptyset$. For each $E \in \mathcal{E}$ there exists a constant $\kappa_E < \infty$ such that for any consistent policy π,*

$$\limsup_{t\to\infty} \mathbb{P}\left\{\frac{\tau^\pi(E,t)}{\ln t} \leq \kappa_E\right\} = 0.$$

Theorem 1 implies that consistency of a policy requires persistently observing cost realizations of arcs that are not observable when choosing the full information solution, thus implying a minimal regret growth with t. Theorem 1 can be leveraged to establish a fundamental bound on achievable performance. For that, define

$$\Delta_B := r(S^*(\mu),\mu) - r(S(B,\mu),\mu), \quad B \in \mathcal{B}.$$

Consider a collection of cyclic policies such that, in each cycle, each policy selects interdiction actions in $\mathcal{B}$ that assure that the follower's response contains an element of E, $E \in \mathcal{E}$, at least k_E times in each cycle. Consider the policy in this collection that attains the optimal (i.e., lowest possible) regret in each cycle. Observe that by Theorem 1, in the long run, the regret of any consistent policy is at least the regret of such an 'optimal' cyclic policy. The next corollary summarizes this discussion.

COROLLARY 1. *For any consistent policy, π one has that*

$$\liminf_{t\to\infty} \frac{\mathcal{R}^\pi(t,F)}{\ln t} \geq \kappa,$$

where κ is the objective value of the following Lower Bound Problem (LBP).

$$\begin{aligned}
\kappa := \min \; & \sum_{B\in\mathcal{B}} x_B \, \Delta_B \\
s.t. \; & y_a \leq \sum_{B\in\mathcal{B}:\, a\in S(B,\mu)} x_B, \quad a \in A \\
& \sum_{a\in E} y_a \geq k_E, \quad E \in \mathcal{E} \\
& y_a \in \mathbb{R}_+,\, a \in A,\, x_B \in \mathbb{R}_+,\, B \in \mathcal{B}
\end{aligned}$$

Limits on achievable performance in the bandit literature in settings with combinatorially many arms are typically put in terms of the cardinality of the sets involved, with the notable exception of Modaresi et al. (2020), where the bounds presented depend on formulations akin to the LBP above. In this regard, the logarithmic (in T) growth of the lower bound on the regret is now a commonplace in the bandit literature, so one can see Theorem 1 as proof that such dependence is also the best possible in the shortest-path interdiction setting. This is not surprising, as such a logarithmic dependence can be attained by traditional bandit policies that treat each of the leader's potential actions ($B \in \mathcal{B}$) as an arm. However, the constant accompanying the logarithmic term (i.e. playing the role of κ) in such policies turns out to be proportional to the size of $\mathcal{B}$, which is exponentially large in terms of A, in the worst-case. Policies designed for settings with combinatorially many arcs in network settings (which compute mean cost estimates at the arc level) fare better, achieving accompanying constants proportional to (a polynomial of) the size of A instead.

The dependence on both the size of $\mathcal{B}$ or A in finite-time performance upper-bounds for the case of, for example, UCB-type policies is driven by the need to estimate mean rewards associated with the elements of $\mathcal{B}$ or mean costs for the arcs in A, respectively. When seen through this lens, the results in Theorem 1 and Corollary 1 suggest that, in the setting of shortest-path interdiction, the regret is driven by the need to estimate mean costs of elements in each set $E \in \mathcal{E}$. However, such an estimation can be performed by implementing at most $\left(|\mathcal{E}| \wedge \left|\bigcup_{E\in\mathcal{E}} E\right|\right)$ solutions in $\mathcal{B}$. The bound in Theorem 1 does not explicitly consider the dependence on a set of solutions implemented, as it aims at achieving the lowest possible regret.

We close this section by noting that if $\mathcal{E} = \emptyset$, then the optimal full-information solution might be found in finite time by employing the information in $\mathcal{U}$, see Example 3.

Example 3. Consider again Example 1 and consider a policy that initially implements $B^1 = \emptyset$, and $B^t = B^{t-1} \cup \{(1,t)\}$ for $t = 2, \ldots, K+1$. By period $t = K+2$, the interdictor can deduce that all paths of the form $1 - j - (k+1)$, $j \geq K+2$, are longer than the paths of the form $1 - j - (k+1)$, $j = 2, \ldots, K+1$, and thus the interdictor can conclude that $B^*(\mu) = B^{K+1}$. Note that this policy attains a finite regret (independent of T); in particular,

$$\liminf_{t\to\infty} \frac{\mathcal{R}(t,F)}{\ln t} = 0.$$

Compare the above to probable finite performance upper bounds for traditional benchmark policies (see next section) which ensure that, at most, $\liminf \frac{\mathcal{R}(t,F)}{\ln t} \geq o(|A|) \sim o(k)$. ■

5. Adapted Seminal Policies

In this section, we adapt classical multi-armed bandit policies to the sequential interdiction setting at hand. Because the implementation of the methods described in this section relies on mathematical programming, we benefit from describing the set of interdiction actions as elements of the set

$$\mathcal{X} := \left\{ (x_a : a \in A) : \sum_{a\in A} x_a \leq K,\, x_a \in \{0,1\}\, a \in A \right\},$$

where for $a \in A$, $x_a = 1$ encodes the action of blocking arc a, and $x_a = 0$ otherwise. Thus, in this section, we refer to $x \in \mathcal{X}$ and $B \in \mathcal{B}$ as interdiction actions, interchangeably. Similarly, for $t \in \mathbb{N}$, we let $\boldsymbol{z}^t := (z_a^t,\, a \in A)$ denote a vector representing the path observed in period t, thus $z_a^t := \mathbf{1}\{a \in S^t\}$ for $a \in A$; we refer to S^t and $\boldsymbol{z}^t$ interchangeably as well.

5.1. Initialization phase

A starting point, common to all the proposed policies, is an initialization phase that attempts to compute the set $\mathcal{U}$ without resorting to explicit enumeration. Such a procedure operates in a sequential fashion: starting from $\mathcal{B}^1 = \mathcal{B}$ at time $t = 1$, in period $t \geq 1$ the procedure implements an action B^t from the set $\mathcal{B}^t \subseteq \mathcal{B}$ of interdiction actions for which the evader's response cannot be anticipated based on the information available at the time. This is,

$$B^t \in \mathcal{B}^t := \{B \in \mathcal{B} : \exists\, \nu, \nu' \in \mathbb{C} \text{ s.t. } S^s = S(B^s,\nu) = S(B^s,\nu'),\, s < t \wedge S(B,\nu) \neq S(B,\nu')\}. \tag{1}$$

The procedure, which chooses $B^t \in \mathcal{B}^t$ arbitrarily until $\mathcal{B}^t = \emptyset$, is summarized in Algorithm 1. Note that, at the end of the procedure, we can (without loss of generality) eliminate the set of arcs in A that are not observed in any of the evader's responses. In addition, by construction, we have that

$$\mathcal{U} = \left\{ \nu \in \mathbb{C} : S(B^t,\nu) = S^t,\, t < t_0 \right\},$$

where t_0 denotes the first period after the initialization phase. Implementing the initialization procedure requires, for each period $t < t_0$, computing $\mathcal{B}^t$ and selecting B^t from such a set. In our experiments, we conduct such tasks jointly by solving the mixed integer program (2). The formulation, which is instantiated in each period t, finds two mean cost vectors, ν^1 and ν^2, that explain the evader responses observed prior to period t, and an interdiction action $x \in \mathcal{X}$ such that the evader responses $\bar{\boldsymbol{z}}^1$ and $\bar{\boldsymbol{z}}^2$, encoded through a linear programming (LP) formulation of the shortest path problem, under ν^1 and ν^2, respectively, differ.

$$\Gamma := \max\ (\nu^2)^\top \bar{\boldsymbol{z}}^1 - \bar{y}_n^2 + \bar{y}_1^2 \tag{2a}$$

Algorithm 1 Initialization phase

```
Set $t = 1$, $\mathcal{B}^t = \mathcal{B}$, $t_0 = 1$
while $\mathcal{B}^t \neq \emptyset$ do
    Choose $B^t \in \mathcal{B}^t$ arbitrarily, and observe $S^t$
    Set $t = t + 1$, $t_0 = t$, and update $\mathcal{B}^t$ according to (1)
end while
Set $A \equiv \bigcup_{t<t_0} S(B^t)$.
Set $\mathcal{U} \equiv \{\nu := (\nu_a,\, a \in A) \in \mathbb{C} \colon S(B^t, \nu) = S(B^t, \mu),\, t < t_0\}$.
```

$$
\begin{aligned}
\text{s.t.}\quad y_j^{k,s} - y_i^{k,s} &\leq \nu_{i,j}^k, \quad (i,j) \in A \setminus B^s, && s < t,\ k = 1,2 && \text{(2b)}\\
y_n^{k,s} - y_1^{k,s} &= \left(\nu^k\right)^\top \boldsymbol{z}^s, && s < t,\ k = 1,2 && \text{(2c)}\\
\mathcal{A}\,\bar{\boldsymbol{z}}^k &= b, && k = 1,2 && \text{(2d)}\\
\bar{z}_{i,j}^k + x_{i,j} &\leq 1, && (i,j) \in A,\ k = 1,2 && \text{(2e)}\\
\bar{y}_j^k - \bar{y}_i^k &\leq \nu_{i,j}^k + M\, x_{i,j}\,, && k = 1,2 && \text{(2f)}\\
\bar{y}_n^k - \bar{y}_1^k &= \left(\nu^k\right)^\top \bar{\boldsymbol{z}}^k, && k = 1,2 && \text{(2g)}\\
\boldsymbol{y}^{k,s}, \quad \nu^k, \quad & \bar{\boldsymbol{y}}^k \in \mathbb{R}^{|A|}, \bar{\boldsymbol{z}}^k \in \{0,1\}^{|A|},\ x \in \mathcal{X}. && && \text{(2h)}
\end{aligned}
$$

In this formulation, (2b) to (2c) enforce (through strong duality) that both cost vectors ν^1 and ν^2 are consistent with the feedback observed prior to period t. Constraints (2d) to (2g) ensure that $\bar{\boldsymbol{z}}^k$ is among the shortest paths in the interdicted network when the interdiction action $x \in \mathcal{X}$ is adopted and the mean cost vector is ν^k, for $k = 1,2$. (Here, $\mathcal{A}$ corresponds to the adjacency matrix of graph G, and $\boldsymbol{b} = (-1, 0, \ldots, 0, 1)$, thus (2d) amounts to primal feasibility of $\bar{\boldsymbol{z}}^k$).[1] The next result establishes that (2) accomplishes its purpose.

Lemma 2. *Let $\hat{x}$ be an optimal solution of* (2) *and let $\hat{B} = (a \in A \colon \hat{x}_a = 1)$ be the interdiction solution induced by $\hat{x}$. Then:* (i) *$\Gamma \geq 0$;* (ii) *If $\Gamma = 0$ then $\mathcal{B}^t = \emptyset$; and* (iii) *If $\Gamma > 0$ then $\hat{B}$ is an element of $\mathcal{B}^t$.*

Note that whereas formulation (2) includes nonlinear quadratic terms in its objective function and in constraints (2g), they correspond to sums of products of continuous and binary variables and as such can be linearized. Consequently, formulation (2) can be solved directly using out-of-the-shelf state-of-the-art MIP solvers.

Remark 3 (Connection to Linear Bandits.). Seminal linear bandit policies typically impute (in each period) a mean cost vector ν and select (either sequentially or jointly) an action that is optimal for such a mean cost vector. In our setting, when the selection of ν is restricted to lie in $\mathcal{U}$, the structure of the interdictor's problem is that of a linear bandit (Auer 2002). Specifically, if the interdictor knows that the mean cost vector lies in $\mathcal{U}$, then the interdictor can be seen as choosing $y^t \in \{y \in \mathbb{R}_+^{|A|} : y = y_B, B \in \mathcal{B}\}$ (y_B being the indicator vector of $S(B,\mu)$) after which the interdictor receives the linear profit of $(\boldsymbol{c}^t)^\top y^t$. In this regard, the challenges in the design of the policies outlined next (beyond dealing with a combinatorial set of actions, and a non-trivial map from an interdiction to a response) correspond to ensuring that the imputed mean cost vector remains within $\mathcal{U}$, so that the aforementioned equivalence, which ensures linearity in costs, holds true and thus can be exploited for finding a candidate action. ■

5.2. A Posterior (Thompson) sampling policy

Next, we propose an adaptation of the Thompson Sampling (TS) policy (Thompson 1933). For this, we adopt a parametric approach in which F is characterized by a vector of parameters - which for convenience we identify with μ - and consider an initial prior distribution over μ. The policy starts with the initialization phase described in Algorithm 1, which identifies $\mathcal{U}$ and redefines A. In each period after the initialization phase, the TS policy samples a vector from the posterior distribution of μ, and implements the full-information solution computed while assuming that the underlying mean cost vector equals the sample. We provide the details of our implementation next.

Consider a prior distribution $\lambda(\cdot)$ on μ such that $\lambda(\mathcal{U}) = 1$, and let $\lambda^t \equiv \lambda_{|\mathcal{F}^t}$ denote the posterior distribution conditional on the feedback observed by the beginning of period $t \in [T]$. For $t \geq t_0$, the TS policy draws a sample mean cost vector $\nu^t \sim \lambda^t$ and implements $B^t := B^*(\nu^t)$. The procedure is depicted in Algorithm 2.

Algorithm 2 TS phase

```
Set $t = t_0$ and $\lambda^t = \lambda_{|\mathcal{F}^{t_0}}$.
while $t \leq T$ do
    Sample $\nu^t$ from distribution $\lambda^t$.
    Implement $B^*(\nu^t)$ and observe $S^t$.
    Set $t = t + 1$ and update $\lambda^t := \lambda_{|\mathcal{F}^t}$.
end while
```

Implementing the TS policy requires sampling from the posterior distribution λ^t, and solving an instance of the full-information problem. In our numerical experiments, we conduct the latter task by solving the following MIP:

$$B^t \in \arg\max \big\{ y_n - y_1 : y_j - y_i \leq \nu^t_{i,j} + M\, x_{i,j} \;\; \forall (i,j) \in A, x \in \mathcal{X} \big\}. \tag{3}$$

Formulation (3) is well-known and follows directly from considering the dual of a linear program formulation of the shortest-path problem while envisioning the interdiction of an arc as increasing its cost by a sufficiently large (hence the big-M constant M) constant. Sampling from λ^t presents additional challenges, which we address next.

Sampling from the posterior distribution. Implementing the TS policy requires specifying a prior distribution λ such that $\lambda(\mathcal{U})=1$. However, in practice, $\mathcal{U}$ is not computed explicitly: one can see that $\mathcal{U}$ is the polyhedron

$$\mathcal{U}=\Big\{\nu\in\mathbb{C}:\sum_{a\in S^t}\nu_a\le\sum_{a\in S}\nu_a,\,S\in\mathcal{P}(B^t),\,t<t_0\Big\},\tag{4}$$

and thus is defined implicitly by the output of the initialization base. With no explicit representation of $\mathcal{U}$, defining a prior over it becomes challenging. In this regard, a method for defining the prior consists of first defining a prior distribution $\bar{\lambda}$ over $\mathbb{R}^{|A|}$ such that $\bar{\lambda}(\mathcal{U})>0$, and at time $t=t_0$ selecting λ as the restriction of $\bar{\lambda}$ to $\mathcal{U}$. That is,

$$\lambda(U):=\frac{\bar{\lambda}(U\cap\mathcal{U})}{\bar{\lambda}(\mathcal{U})},\quad U\in\mathbf{B}(\mathbb{R}^{|A|}).$$

(Here, $\mathbf{B}$ denotes the Borel sets). This method has the additional advantage that if $\bar{\lambda}$ and F are conjugate priors, then so are λ and F. Let $\bar{\lambda}^t(\cdot)$ denote the posterior distribution of μ given $\mathcal{F}^t$ relative to the prior $\bar{\lambda}(\cdot)$, i.e. $\bar{\lambda}^t:=\bar{\lambda}(\cdot)|_{\mathcal{F}^t}$; from above, one can check that

$$\lambda^t(U)=\frac{\bar{\lambda}^t(U\cap\mathcal{U})}{\bar{\lambda}^t(\mathcal{U})},\quad U\in\mathbf{B}(\mathbb{R}^{|A|}).$$

Thus, computing the posterior distribution λ^t becomes as *easy* as computing $\bar{\lambda}^t$ (which in the case of conjugate priors can be done analytically), provided that one can compute the constant $\bar{\lambda}^t(\mathcal{U})$ and carry the intersection operation with $\mathcal{U}$.

In practical terms, sampling from λ^t can be done without computing the constant $\bar{\lambda}^t(\mathcal{U})$ - this is the essence of Monte Carlo Markov Chain methods (see, e.g. Geyer (1992)). In theory, a simple acceptance-rejection method suffices: letting $\frac{d\lambda^t(\omega)}{d\bar{\lambda}^t}$ denote the Radon-Nikodym derivative of λ^t with respect to $\bar{\lambda}^t$, we note that

$$\frac{d\lambda^t(\omega)}{d\bar{\lambda}^t}=\frac{\mathbf{1}\{\omega\in\mathcal{U}\}}{\bar{\lambda}^t(\mathcal{U})},\quad \omega\in\Omega.$$

From the above, an exact acceptance-rejection method for sampling from λ^t consists of drawing a sample from $\bar{\lambda}^t$, and then accepting (rejecting) the sample if it does (does not) belong to $\mathcal{U}$. The scheme is depicted in algorithmic form in Algorithm 3. Whereas Algorithm 3 does not require

Algorithm 3 AR sampling

```
For $t \in \mathbb{N}$, compute $\bar{\lambda}^t$ and sample $\nu$ from $\bar{\lambda}^t$.
while $\nu \notin \mathcal{U}$ do
  Sample $\nu$ from distribution $\bar{\lambda}^t$.
end while
set $\nu^t = \nu$.
```

computing the constant $\bar{\lambda}^t(\mathcal{U})$, it does require checking whether a sample ν lies in $\mathcal{U}$. In our numerical experiments, we perform such a check by solving the following LP:

$$\min \textstyle\sum_{s<t_0} w^s \tag{5a}$$

$$\text{s.t.} \quad y^s_j - y^s_i \leq \nu_{i,j}, \quad (i,j) \in A \setminus B^s, \quad s < t_0 \tag{5b}$$

$$y^s_n - y^s_1 = \nu^\top \boldsymbol{z}^s - w^s, \quad s < t_0 \tag{5c}$$

$$\boldsymbol{y}^s \in \mathbb{R}^{|A|}, \, w^s \in \mathbb{R}_+ \quad s < t_0. \tag{5d}$$

Constraints (5b) encodes (LP) dual-feasibility of the vector $\boldsymbol{y}^s$ at time s, and (5c) imposes strong duality: if $w^s = 0$, then $\boldsymbol{z}^s$ is indeed the shortest path on the interdicted network at time s; otherwise w^s compensates for the optimality gap associated with $\boldsymbol{z}^s$. Overall, if the objective function (5) is 0, we conclude that $\nu \in \mathcal{U}$. Otherwise, if the objective function is greater than zero, then there exists a time period $s < t$ for which $\boldsymbol{z}^s$ is not an optimal solution under ν.

Approximate Sampling from the posterior. The practical efficiency of acceptance-rejection schemes depends on the expected number of rejections made before accepting a sample. In our setting, one can anticipate that, depending on the *volume* of $\mathcal{U}$, as time progresses $\bar{\lambda}^t$ will concentrate around μ, which should result in small rejection rates. However, because $\bar{\lambda}$ is chosen irrespective of $\mathcal{U}$, it is possible that $\bar{\lambda}^{t_0}(\mathcal{U})$ is very small, which translates into very small initial acceptance rates, which in turn affects computational efficiency of implementing the TS policy (note that each accept/reject decision requires solving an instance of (5)).

In order to speed up the implementation of the TS policy, especially for early time periods, we consider the following approximate sampling scheme:

(*i*) Before sampling at time t_0, we construct a finite sample $\tilde{\mathcal{U}}$ of vectors in $\mathcal{U}$ by sampling at random from it using a *hit-and-run* scheme;

(*ii*) at time $t \geq t_0$, we sample ν from $\bar{\lambda}^t$; if the sample is rejected, then we sample from an approximate distribution $\tilde{\lambda}^t$ that has support on $\tilde{\mathcal{U}}$ (guarantying that the sample lies in $\mathcal{U}$).

We construct the finite set $\tilde{\mathcal{U}}$ following the hit-and-run scheme of Berbee et al. (1987) for (approximately) sampling uniformly from a polyhedron. The procedure works iteratively, starting from an incumbent mean cost vector $\nu^k \in \mathcal{U}$ at iteration k, and constructs a new cost vector ν^{k+1} by first sampling a random direction $\boldsymbol{d} \sim \mathcal{N}(0, I)$ (where $\mathcal{N}$ stands for the normal distribution and I is the identity matrix in $\mathbb{R}^{|A|\times|A|}$,[2] and then sampling ν^{k+1} at random (uniformly) from the (one dimensional) set $\{\nu^k + \boldsymbol{d}v : v \in \mathbb{R}\} \cap \mathcal{U}$.

In Berbee et al. (1987), the set $\{\nu^k + \boldsymbol{d}v : v \in \mathbb{R}\} \cap \mathcal{U}$ is computed by finding the values of v such that $\nu^k + \boldsymbol{d}v$ belongs to a facet of $\mathcal{U}$. Observe that, because $\mathcal{U}$ is bounded, there are exactly two such values, which we refer to as v_+ and v_-, associated with the directions $\boldsymbol{d}$ and $-\boldsymbol{d}$, respectively. A direct application of the method in Berbee et al. (1987) requires writing $\mathcal{U}$ as a polyhedron whose only variables are ν; in our case:

$$\mathcal{U} = \left\{\nu \in \mathbb{C} \colon (\boldsymbol{z}^s - \boldsymbol{z})^\top \nu \le 0 \;\forall \boldsymbol{z} \in Z^s, s < t_0\right\}, \tag{6}$$

where $Z^s = \{\boldsymbol{z} \in \{0,1\}^{|A|} \colon \mathcal{A}^s \boldsymbol{z} = b\} \setminus \{\boldsymbol{z}^s\}$ is a vector representation of $\mathcal{P}(B^t)$ (excluding S^t). It then computes

$$v_+ := \min\left\{\frac{(\boldsymbol{z}-\boldsymbol{z}^s)^\top \nu^k}{((\boldsymbol{z}^s-\boldsymbol{z})^\top \boldsymbol{d})^+} : \boldsymbol{z} \in Z^s,\, s < t_0;\, \frac{u_a - \nu_a^k}{d_a} : a \in A,\, d_a > 0;\, \frac{l_a - \nu_a^k}{d_a}, a \in A,\, d_a < 0\right\}, \tag{7a}$$

$$v_- := \min\left\{\frac{(\boldsymbol{z}-\boldsymbol{z}^s)^\top \nu^k}{((\boldsymbol{z}^s-\boldsymbol{z})^\top \boldsymbol{d})^-} : \boldsymbol{z} \in Z^s,\, s < t_0;\, \frac{\nu_a^k - u_a}{d_a} : a \in A,\, d_a < 0;\, \frac{\nu_a^k - l_a}{d_a}, a \in A,\, d_a > 0\right\}, \tag{7b}$$

samples $v \sim U[v_-, v_+]$ and sets $\nu^{k+1} = \nu^k + \boldsymbol{d}v$. Observe that this method cannot be applied directly to our case, as for $s < t_0$, Z^s can have exponentially many elements in the worst case. Nonetheless, the next result shows that v_+ and v_- can be found by solving a pair of LP formulations.

Lemma 3. *For $\boldsymbol{d} \in \mathbb{R}^{|A|}$, one has that $v_+ = w(\boldsymbol{d})$ and $v_- = w(-\boldsymbol{d})$, where for $\tilde{\boldsymbol{d}} \in \mathbb{R}^{|A|}$*

$$\begin{aligned} w(\tilde{\boldsymbol{d}}) :=\; & \max w && && \text{(8a)}\\ s.t. \quad & y_j^s - y_i^s \le \nu_{i,j}^k + \tilde{d}_{i,j}\, w, \quad (i,j) \in A \setminus B^s, && s < t_0 && \text{(8b)}\\ & y_n^s - y_1^s = (\nu^k + \tilde{\boldsymbol{d}}w)^\top \boldsymbol{z}^s, && s < t_0 && \text{(8c)}\\ & \nu^k + \tilde{\boldsymbol{d}}w \in \mathbb{C} && && \text{(8d)}\\ & \boldsymbol{y}^s \in \mathbb{R}^{|A|},\, w \in \mathbb{R}. && && \text{(8e)} \end{aligned}$$

The hit-and-run procedure is detailed in Algorithm 4. There, H denotes the size of the set $\tilde{\mathcal{U}}$.

Remark 4. The correctness of Algorithm 4 rests on the assumption that each vector ν^k lies inside of the interior of $\mathcal{U}$, which is guaranteed (with probability one) if $\nu^1 \in \operatorname{Int}(\mathcal{U}) \neq \emptyset$. Note that if μ is indeed distributed according to an absolutely continuous prior λ, then $\operatorname{Int}(\mathcal{U}) \neq \emptyset$ almost surely, in which case one can find $\mu^1 \in \operatorname{Int}(\mathcal{U})$ by solving the formulation

$$\max \;\epsilon \tag{9a}$$

Algorithm 4 Sampling from $\mathcal{U}$

```
Require: $\nu^1 \in \mathcal{U}$, $H \in \mathbb{N}$
  for $k = 1$ to $H$ do
    Sample $\boldsymbol{d} \sim \mathcal{N}(0, I)$
    Find $w_+$ and $w_-$ by solving (8) using directions $\boldsymbol{d}$ and $-\boldsymbol{d}$, respectively
    Sample $v \sim U[-w_-, w_+]$ and set $\nu^{k+1} = \nu^k + \boldsymbol{d}\, v$
  end for
```

$$\text{s.t.} \quad y_j^{s,k} - y_i^{s,k} \leq \nu_{i,j} + \epsilon\, \boldsymbol{e}^k, \quad (i,j) \in A \setminus B^s, \quad s < t_0,\ k = 0, \ldots, |A| \tag{9b}$$

$$y_n^{s,k} - y_1^{s,k} = (\nu + \epsilon\, \boldsymbol{e}^k)^\top \boldsymbol{z}^s, \qquad s < t_0,\ k = 0, \ldots, |A| \tag{9c}$$

$$\boldsymbol{y}^{s,k}, \nu \in \mathbb{R}^{|A|},\ \epsilon \in \mathbb{R}, \tag{9d}$$

where $\boldsymbol{e}^k \in \mathbb{R}^{|A|}$ is such that $e_i^k = \mathbf{1}\{k = i\}$, for $i \leq |A|$, $k \leq |A|$. Indeed, letting $(\{\boldsymbol{y}^{s,k}\}, \nu, \epsilon)$ denote a solution (9), then one has that if $\text{Int}(\mathcal{U}) \neq \emptyset$, then $\epsilon > 0$, and $\nu + \epsilon\, \boldsymbol{e}^k \in \mathcal{U}$ for all $k \in \{0, \ldots, |A|\}$, thus one can set $\nu^1 = \nu + \mathbf{1}\, \epsilon\, (2\,|A|)^{-1} \in \text{Int}(\mathcal{U})$. ■

Once $\tilde{\mathcal{U}}$ is constructed, our approximate sampling scheme samples ν at time $t \geq t_0$ from a distribution $\tilde{\lambda}^t$ with support on $\tilde{\mathcal{U}}$. We construct such a distribution so as to approximate λ^t. In particular, we consider

$$\tilde{\lambda}^t(\nu) \propto \partial \bar{\lambda}^t(\nu), \quad \nu \in \tilde{\mathcal{U}},$$

where $\partial \bar{\lambda}^t$ stands for the Radon-Nikodym derivative of $\bar{\lambda}^t$ with respect to the Lebesgue measure in $\mathbb{R}^{|A|}$. (Recall that $\bar{\lambda}^t$ is readily available for the case of conjugate distributions.) Algorithm (5) below summarizes our implementation of the TS policy.

5.3. A Bayesian Upper Confidence Bound policy

Next, we present an adaptation of a Bayesian upper confidence bound (Bayes-UCB) policy (Kaufmann et al. 2012). Our starting point is the parametric approach used in the previous section. In particular, we consider the posteriors λ^t over $\mathcal{U}$, and $\bar{\lambda}^t$ over $\mathbb{R}^{|A|}$.

The Bayes-UCB policy starts by implementing the initialization phase. Then, for period $t \geq t_0$, it considers an uncertainty region $U^t \subseteq \mathcal{U}$ such that $\lambda^t((U^t)^c) = O(t^{-1})$, and implements the solution to the following program

$$B^t \in \arg\max\left\{\max\left\{r(S(B), \nu) : \nu \in U^t\right\} : B \in \mathcal{B}\right\}. \tag{10}$$

For each interdiction action $B \in \mathcal{B}$, the interdictor adopts the *optimism in the face of uncertainty* principle and assumes a favorable realization of the underlying mean cost vector, but discarding

Algorithm 5 approximate TS phase

```
Require: $K \in \mathbb{N}$, $(\{(\boldsymbol{z}^s, B^s), s < t_0\})$
  Set $t = t_0$ and $\lambda^t = \lambda$.
  Compute $\nu^1$ by solving (9), and use it to compute $\tilde{\mathcal{U}}$ following Algorithm (4)
  while $t \leq T$ do
    Sample $\nu^t$ from distribution $\bar{\lambda}^t$, and find $w$ by solving (5)
    if $w > 0$ then
      Resample $\nu^t$ from $\tilde{\lambda}^t$
    end if
    Implement $B^*(\nu^t)$ and observe $S^t$.
    Set $t = t + 1$ and update $\bar{\lambda}^t := \bar{\lambda}_{|\mathcal{F}^t}$.
  end while
```

those values that have a probability lower than $O(t^{-1})$ of occurring (i.e. those outside U^t). When λ^t is absolutely continuous, one can choose $U^t := \{\nu \in \mathcal{U} : \partial\lambda^t(\nu) \leq \xi(t)\}$, where

$$\xi(t) := \sup\left\{\xi \in \mathbb{R}_+ : \lambda^t\left(\left\{\nu \in \mathcal{U} : \partial\lambda^t(\nu) \leq \xi\right\}\right) < t^{-1}\right\}. \tag{11}$$

The policy is depicted in Algorithm 6.

Algorithm 6 Bayes-UCB phase

```
Require: $(\{(\boldsymbol{z}^s, B^s), s < t_0\})$
  for $t = t_0$ to $T$ do
    Compute $U^t$, implement a solution to (10), and observe $S^t$
    Set $t = t + 1$ and update $\bar{\lambda}^t := \bar{\lambda}_{|\mathcal{F}^t}$.
  end for
```

Implementing the Bayes-UCB policy requires computing U^t and solving (10) for all $t \geq t_0$. Note that the former step requires access to λ^t, which is not readily available. Thus, in our numerical experiments, we consider a (truncated) multivariate normal approximation to λ^t when computing $\xi(t)$ according to (11). In particular, for $t \geq t_0$, we consider the following approximating scheme:

i) Let $\hat{\boldsymbol{c}}^t$ and $\hat{\Sigma}^t$ denote the mean and covariance matrix associated the posterior $\bar{\lambda}^t$, and consider the distribution $\bar{\lambda}^t_{\mathcal{N}} := \mathcal{N}(\hat{\boldsymbol{c}}^t, \hat{\Sigma}^t)$.

ii) Construct a large (but finite) sample of points $\mathbb{U}^t$ from $\bar{\lambda}^t_{\mathcal{N}}$, and use (5) to compute the $\rho^t := |\mathbb{U}^t \cap \mathcal{U}| \,/\, |\mathbb{U}^t|$. Note that ρ^t is a Monte Carlo approximation to $\bar{\lambda}^t_{\mathcal{N}}(\mathcal{U})$.

iii) Define the distribution $\lambda^t_{\mathcal{N}}$ so that

$$\partial\lambda^t_{\mathcal{N}}(\nu) = \frac{1}{\rho^t}\partial\bar{\lambda}^t_{\mathcal{N}}(\nu)\,\mathbf{1}\{\nu \in \mathcal{U}\}.$$

iv) We approximate the uncertainty region U^t with the $(1-t^{-1})$ quantile ellipsoidal confidence region of $\lambda^t_{\mathcal{N}}$. That is, we consider the approximation

$$\tilde{U}^t := \left\{\nu : \|\nu - \hat{\boldsymbol{c}}^t\|_{\hat{\Sigma}^t} \leq \chi^2_{|A|}((1-t^{-1})\rho^t)\right\},$$

where χ^2_k stands for the quantile function of the chi-squared distribution with k degrees of freedom, $k \in \mathbb{N}$

Because $\lambda^t(\mathcal{U}) \approx \rho^t$, the approximate uncertainty region might not be contained in $\mathcal{U}$ (especially in the short term). Moreover, it might be the case that $\tilde{U}^t \cap \mathcal{U} = \emptyset$. Thus, in our implementation, we first approximate the solution to (10) restricting attention to the set $\tilde{U}^t \cap \mathcal{U}$ by solving the formulation

$$\max\ y_n - y_1 \tag{12a}$$

$$\text{s.t.} \qquad y_j - y_i \leq \nu_{i,j} + x_{i,j}\,M, \quad (i,j) \in A \tag{12b}$$

$$\|\nu - \hat{\boldsymbol{c}}^t\|_{\hat{\Sigma}^t} \leq \chi^2_{|A|}((1-t^{-1})\,\rho^t) \tag{12c}$$

$$y^s_j - y^s_i \leq \nu_{i,j}, \quad (i,j) \in A \setminus B^s,\ s < t_0 \tag{12d}$$

$$y^s_n - y^s_1 = \nu^\top \boldsymbol{z}^s, \qquad s < t_0 \tag{12e}$$

$$x \in \mathcal{X} \qquad \boldsymbol{y}^s,\ \nu \in \mathbb{R}^{|A|} \tag{12f}$$

When this formulation is infeasible (i.e. when $\tilde{U}^t \cap \mathcal{U} = \emptyset$) we ignore constraints (12d)-(12e). When feasible, the formulation above finds an interdiction action x which results in a (dual) response $\boldsymbol{y}$ by the evader when ν is chosen so it is consistent with the information obtained up to time t and belongs to the approximate uncertainty region. Let $(x, \boldsymbol{y}, \boldsymbol{y}^s, \nu)$ denote the solution to (12) above in period t: we implement $x^t = x$ in period t. Because of the large computational cost of step ii) above, in our numerical experiments we construct $\mathbb{U}^t$ only at a certain frequency.

Remark 5 (Connection to Linear Bandits - Continued). UCB policies for linear bandits can be thought of as solving (12) while not having to consider (12d) and (12e) to select an action from a combinatorial set[3], with policies differing on the elliptical uncertainty set considered, see Dani et al. (2008), Rusmevichientong and Tsitsiklis (2010), Abbasi-Yadkori et al. (2011). In particular, implementing the OFUL policy Abbasi-Yadkori et al. (2011), a state-of-the-art variant of the UCB policy of Auer (2002) which exhibits superior practical performance, amounts to (under some mild assumptions) replacing (12c) with

$$\|\nu - \tilde{\boldsymbol{c}}^t\|_{\tilde{\Sigma}^t} \leq \sqrt{|A|\log\left(\frac{1 + t\,L\,\gamma^{-1}}{\delta}\right)} + \gamma^{1/2},$$

where L denotes the length (in number of arcs) of the largest path from 1 to n in G, γ and δ are tuning parameters, $\tilde{c}^t$ is the ℓ^2-regularized least-square estimate of μ, and

$$\tilde{\Sigma}^t := \gamma I + \sum_{s=1}^{t} \boldsymbol{z}^s (\boldsymbol{z}^s)^\top .$$

The set above is closely related to that used in our approximate Bayes-UCB policy: ℓ^2-regularized least-square estimates can be seen as incorporating information about a prior (encoded in the regularization term) into log-likelihood minimization under a multivariate normal assumption. ■

5.4. Oblivious adapted policies.

We close this section presenting direct adaptations of the seminal Thompson and UCB policies to this setting. These variations, which aim at reducing the burden of implementation, come from ignoring information about the mean cost vector deduced from observing the follower's response.

In the case of the UCB policy, ignoring information about the adversarial nature of the follower's response amounts to using a confidence region for the unrestricted posterior $\bar{\lambda}^t$ instead of λ^t, see (11). Depending on the choice of $\bar{\lambda}$ (in the case of a conjugate prior) such a region might have an analytical representation; otherwise, one may use the normal approximation for the posterior, i.e.,

$$U^t \approx \tilde{U}^t := \left\{ \nu \in \mathbb{R}^{|A|} : \|\nu - \hat{\boldsymbol{c}}^t\|_{\hat{\Sigma}^t} \le \chi^2_{|A|}(1 - t^{-1}) \right\}.$$

(we do the latter in our numerical experiments). For selecting B^t, we select

$$B^t \in \arg\max \left\{ \max \left\{ \min \left\{ r(S), \nu \right) : S \in \mathcal{P}(B) \right\} : \nu \in \tilde{U}^t \right\} : B \in \mathcal{B} \right\} \qquad \text{(OUCB-1)}.$$

That is, we select ν optimistically and implement the optimal shortest-path interdiction action for such a cost vector. Note that finding B^t for the oblivious policy amounts to solving (12) while relaxing constraints (12d) and (12e).

The case of Thompson sampling policy follows the same idea: in Algorithm 2 we sample directly from $\bar{\lambda}^t$ as opposed to approximate sampling from λ^t. This step, which is the most time-consuming in our implementation of the TS policy, is performed rather directly in the case of conjugate priors (we do the latter in our numerical experiments).

6. Numerical Experiments

In this section, we illustrate the practical performance of the proposed policies and algorithms using a set of numerical experiments. We first describe the test instances and then analyze the performance of the various policies.

6.1. Test instances and implementation details

Network structure. We test our policies using the class of layered graphs (Ryzhov and Powell 2011) in which nodes are located in a grid consisting of various layers, each having the same number of nodes; the source node is connected to all nodes in the first layer, all nodes in a layer are connected to all nodes in the next layer, and all nodes in the final layer are connected to the sink node. Our experiments use networks with three layers, and consider the cases of: two nodes per layer with an interdiction budget of $K=1$; four nodes per layer with a budget of $K=2$; and five nodes per layer with a budget of $K=3$. Additionally, we consider the case of a network with five layers, five nodes per layer, and a budget of $K=3$.

Cost and prior distribution. Given a network topology, an instance is determined by the distribution F of the costs. In our experiments, we consider Bernoulli distributed costs, independent across arcs, thus for $\mu \in [0,1]^{|A|}$ we have that

$$c_a^t = \begin{cases} 1 & \text{with probability } \mu_a \\ 0 & \sim . \end{cases},$$

and $c_a^t \perp c_{a'}^t$ for $a, a' \in A$, for all $t \in [T]$ (we consider $T = 1000$ in all our experiments). Thus, instances are defined by the value of μ. In our experiments, we sample μ_a independently for each arc $a \in A$ from a $U[0,1]$ distribution.

Our policies are embedded in a Bayesian framework. Practical implementation of the proposed policies requires specifying a prior $\bar{\lambda}$ over $\mathbb{R}^{|A|}$ for μ (as opposed to a prior λ over $\mathcal{U}$). Because costs are Bernoulli distributed, to have conjugate cost and prior distributions (which allow expedited computations), we consider a $U[0,1]$ prior for μ_a, independently for each $a \in A$. This implies that,

$$\bar{\lambda}_a^t \equiv \text{Beta}(1 + |\{s < t : a \in S^s \wedge c_a^s = 1\}|, 1 + |\{s < t : a \in S^s \wedge c_a^s = 0\}|), \quad a \in A, t \in [T].$$

The setup thus corresponds to Beta-Bernoulli distributed costs and mean costs. The experiments start with the initialization process described in (1).

Instance filtering and further details. In our experiments we focus on comparing the performance of various policies. With this in mind, for each sampled value of μ we conduct the initialization phase (common to all policies) and check whether $\mathcal{E} = \emptyset$. If this is the case, we discard the value of μ and resample it. For each network structure, we sample 30 "non-trivial" instances (i.e. settings in which $\mathcal{E} \neq \emptyset$), compute the decision of each policy at each $t \in [T]$, and report average performance across these instances.

In our implementation of the adapted TS policy, we used $|\tilde{\mathcal{U}}| = 1000$ points to conduct our approximate sampling scheme. In our implementation of the adapted UCB policy, we sample 1000 points from $\mathbb{U}^t$ to compute ρ^t, and we updated such a sample every 100 periods.

All algorithms were coded in Julia v1.6.7 (Bezanson et al. 2017). We use Gurobi 11.0.0 (Gurobi Optimization, LLC 2023) to solve all MIP and LP formulations. All experiments were performed on a Windows PC with AMD Ryzen 7 PRO 5850U processor with 8 cores and 16 threads, and 16 GB of RAM.

6.2. Policy performance

Consider the comparative performance of the policies in a small instance consisting of a graph with three layers, two nodes per layer, with an interdiction budget of $K = 1$. For such a setting, Figure 3 depicts the evolution of cumulative regret across time. There, each curve represents the evolution of the cumulative regret on average across 30 non-trivial instances. We observe that the adapted TS policy outperforms the adapted UCB policy (uniformly across the horizon), which is consistent with prior work on bandits. Note that both the adapted TS and UCB policies outperform their oblivious counterparts (as defined in Section 5.4), which is aligned with intuition and speaks of the benefits associated with exploiting the structural properties of the setting and restricting attention to mean estimates within $\mathcal{U}$. Notably, we observe that the oblivious UCB policy outperforms the oblivious TS policy, indicating that disregarding the structural properties of the setting affects the performance of the oblivious TS policy more than that of the oblivious UCB policy.

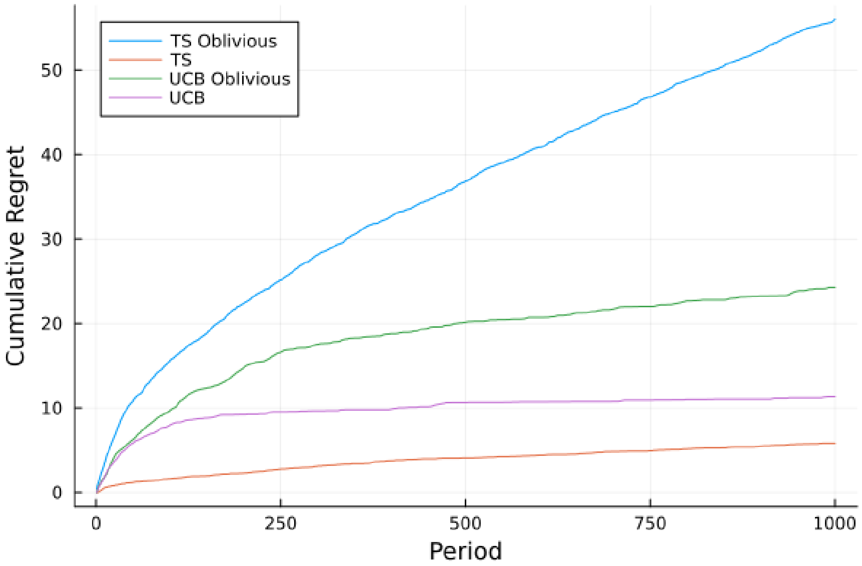

Figure 3 Performance of the first instance (3 layers, 2 nodes, budget of 1)

Table 1 compares the performance of the adapted and oblivious policies in the full set of test instances, both in terms of (average) total cumulative regret after 1000 periods and (average) running time. We set up an upper bound of 24 hours on the running time of our instances: we

Layers	Nodes	Budget(K)	Init. time (s)	Trivial instances		TS Oblivious	TS	UCB Oblivious	UCB
3	2	1	0.64	2	Policy Time	**6.83**	9.08	31.81	390.16
					Mean regret	35.32	**3.79**	18.27	9.80
3	4	2	0.86	1	Policy Time	**32.72**	81.44	145.33	177.40
					Mean regret	155.34	**2.80**	44.57	45.69
3	5	3	16.28	2	Policy Time	**58.05**	89.92	744.11	3018.62
					Mean regret	108.32	**24.06**	61.14	65.43
5	5	3	3,618.71	0	Policy Time	**215.32**	229.76	1886.88	-
					Mean regret	169.88	**80.19**	110.36	-

Table 1 Mean running times in seconds and mean regret (across 30 replications) for each policy. Best performance in boldface.

do not report the performance of the adapted UCB policy for networks with 5 layers, as running times in such instances consistently reached the upper bound. We only found 5 "trivial" instances using our procedures.

Overall, the adapted TS policy provides the best empirical performance without increasing running times significantly relative to that of linear bandit policies. Specifically, the TS policy outperforms the benchmark across all network structures in terms of total cumulative regret. In contrast, the performance of the UCB oblivious policy is comparable to that of its adapted counterpart, even outperforming it in some cases. It is also notable that the performance gap between the adapted and oblivious policies tends to decrease as the size of the instances increases. We suspect, however, that if the number of periods T increases with the instance size, then this gap should remain relatively constant.

From Table 1 we observe that running times for the initialization phase common to all policies are relatively small compared to the policy running times (which do not consider initialization time) for the smaller instances. However, this is not the case for the largest network structures in our experiments, where initialization time goes over one hour, significantly larger than the running time of the TS policies. These results suggest that the performance improvements resulting from using the information in $\mathcal{U}$ come at a computational price, which might be significant for larger instances.

As expected, running times for the TS policy are larger than those of its oblivious counterpart, as it requires additional computational time when a sampled mean cost is found to lie outside $\mathcal{U}$. However, such a computational cost remains below that associated with both UCB policies. In this regard, note that the UCB policy requires computing ρ^t every 100 periods, which is time-consuming as it requires checking whether each sampled point belongs to $\mathcal{U}$. This task, performed by solving a MIP formulation, explains why running times go over the upper limit for networks with 5 layers.

7. Conclusions

In this work, we study a model of periodic interaction on a network between a leader and a non-strategic follower where, in each period, the interdictor acts first by blocking passage on a subset of arcs of the graph, and then the evader responds by traversing the unblocked path with the smallest expected cost on the interdicted graph. We assume that initially, the interdictor does not know the cost distribution but observes cost realizations associated with the arcs traversed by the evader. By framing the problem as a multi-armed bandit, we established a fundamental limit on the asymptotic performance of any admissible policy. In particular, we showed that asymptotic optimality requires a minimal exploration frequency for a class of arcs in the network, which translates into a lower bound on performance by solving a linear program of combinatorial size. This program, broadly speaking, finds the most cost-efficient way of collection of information necessary to confirm the optimality of the full-information solution.

On a practical side, we showed that it is possible to adapt standard policies for the linear bandit to the setting at hand, namely the Thompson Sampling and Upper Confidence Bounds policies. A key step in this adaptation is to incorporate into posterior sampling and uncertainty region computation information about the setting that can be collected in finite time and without noise. While such a step presents various computational challenges, we proposed approximate methods based on the use of mathematical programming, which results in policies closely related to state-of-the-art implementations of UCB-type policies for the linear bandit. Our numerical experiments have shown that carefully tailoring policies for the setting at hand results in significantly better empirical performance, at a small computational cost. In this regard, while the benefits of the proposed adaptation are clear for the Thompson Sampling policy, that is not the case for UCB policies. Further research might focus on improving the construction of confidence regions, which is the most computationally expensive step in our adaptation of the UCB policy, and on finding faster ways to compute $\mathcal{U}$ in the initialization step.

Acknowledgments

The research of the second author is supported in part by the Complex Engineering Systems Institute, ISCI (ANID PIA AFB230002), and Fondecyt (project # 1211407). The research of the first author has been supported by the National Science Foundation (NSF) (Awards ENG/CMMI # 2145553) and by the Air Force Office of Scientific Research (AFOSR) (Award # FA9550-22-1-0236).

References

Abbasi-Yadkori, Y., Pál, D. and Szepesvári, C. (2011), Improved algorithms for linear stochastic bandits, *in* J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira and K. Weinberger, eds, 'Advances in Neural Information Processing Systems', Vol. 24, Curran Associates, Inc.

Ajay Malaviya, C. R. and Sharkey, T. (2012), 'Multi-period network interdiction problems with applications to city-level drug enforcement', *IIE Transactions* **44**(5), 368–380.

Audibert, J.-Y., Bubeck, S. and Lugosi, G. (2013), 'Regret in online combinatorial optimization', *Mathematics of Operations Research* **39**(1), 31–45.

Auer, P. (2002), 'Using confidence bounds for exploitation-exploration trade-offs', *Journal of Machine Learning Research* **3**(Nov), 397–422.

Auer, P., Cesa-Bianchi, N. and Fischer, P. (2002), 'Finite-time analysis of the multiarmed bandit problem', *Machine Learning* **47**(2-3), 235–256.

Azizi, E. and Seifi, A. (2023), 'Shortest path network interdiction with incomplete information: a robust optimization approach', *Annals of Operations Research* .

Ball, M., Golden, B. and Vohra, R. (1989), 'Finding the most vital arcs in a network', *Operations Research Letters* **8**(2), 73–76.

Bayrak, H. and Bailey, M. D. (2008), 'Shortest path network interdiction with asymmetric information', *Networks* **52**(3), 133–140.

Berbee, H., Boender, C., Ran, A., Scheffer, C., Smith, R. and Telgen, J. (1987), 'Hit-and-run algorithms for the identification of nonredundant linear inequalities', *Math Prog* **37**, 184–207.

Bezanson, J., Edelman, A., Karpinski, S. and Shah, V. B. (2017), 'Julia: A fresh approach to numerical computing', *SIAM review* **59**(1), 65–98.
URL: *https://doi.org/10.1137/141000671*

Borrero, J. S., Prokopyev, O. A. and Sauré, D. (2016), 'Sequential shortest path interdiction with incomplete information', *Decision Analysis* **13**(1), 68–98.

Borrero, J. S., Prokopyev, O. A. and Sauré, D. (2019), 'Sequential interdiction with incomplete information and learning', *Operations Research* **67**(1), 72–89.

Borrero, J. S., Prokopyev, O. A. and Sauré, D. (2022*a*), 'Learning in sequential bilevel linear programming', *INFORMS Journal on Optimization* **4**(2), 174–199.

Borrero, J. S., Prokopyev, O. A. and Sauré, D. (2022*b*), 'Learning in sequential bilevel linear programming', *INFORMS Journal on Optimization* **4**(2), 174–199.

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

Cesa-Bianchi, N. and Lugosi, G. (2012), 'Combinatorial bandits', *Journal of Computer and System Sciences* **78**(5), 1404–1422.

Chen, X. and Zhang, Y. (2009), 'Uncertain linear programs: Extended affinely adjustable robust counterparts', *Operations Research* **57**(6), 1469–1482.

Cormican, K., Morton, D. and Wood, R. (1998), 'Stochastic network interdiction', *Operations Research* **46**(2), 184–197.

Dani, V., 9, ., Hayes, T. and Kakade, S. M. (2008), 'Stochastic linear optimization under bandit feedback', *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland* pp. 355–366.

Fulkerson, D. and Harding, G. (1977), 'Maximizing the minimum source-sink path subject to a budget constraint', *Mathematical Programming* **13**(1), 116–118.

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.

Geyer, C. J. (1992), 'Practical markov chain monte carlo', *Statistical Science* **7**(4), 473–483.
URL: *http://www.jstor.org/stable/2246094*

Ghare, P., Montgomery, D. and Turner, W. (1971), 'Optimal interdiction policy for a flow network', *Naval Research Logistics Quarterly* **18**(1), 37–45.

Gift, P. D. (2010), Planning for an adaptive evader with application to drug interdiction operations, Master's thesis, Naval Postgraduate School, Monterey, California.

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

Gurobi Optimization, LLC (2023), 'Gurobi Optimizer Reference Manual'.
URL: *https://www.gurobi.com*

Hemmecke, R., Schultz, R. and Woodruff, D. L. (2003), Interdicting stochastic networks with binary interdiction effort, *in* 'Network Interdiction and Stochastic Integer Programming', Springer, pp. 69–84.

Holzmann, T. and Smith, J. C. (2021), 'The shortest path interdiction problem with randomized interdiction strategies: Complexity and algorithms', *Operations Research* **69**(1), 82–99.

Israeli, E. and Wood, R. (2002), 'Shortest-path network interdiction', *Networks* **40**(2), 97–111.

Kang, S. and Bansal, M. (2023), 'Distributionally risk-receptive and risk-averse network interdiction problems with general ambiguity set', *Networks* **81**(1), 3–22.

Kaufmann, E., Cappe, O. and Garivier, A. (2012), On bayesian upper confidence bounds for bandit problems, *in* N. D. Lawrence and M. Girolami, eds, 'Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics', Vol. 22 of *Proceedings of Machine Learning Research*, PMLR, La Palma, Canary Islands, pp. 592–600.

Ketkov, S. S. and Prokopyev, O. A. (2020), 'On greedy and strategic evaders in sequential interdiction settings with incomplete information', *Omega* **92**, 102161.

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

Kosmas, D., Sharkey, T. C., Mitchell, J. E., Maass, K. L. and Martin, L. (2023), 'Interdicting restructuring networks with applications in illicit trafficking', *European Journal of Operational Research* **308**(2), 832–851.

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

Malik, K., Mittal, A. and Gupta, S. (1989), 'The k-most vital arcs in the shortest path problem', *Operations Research Letters* **8**(4), 223–227.

McMasters, A. and Mustin, T. (1970), 'Optimal interdiction of a supply network', *Naval Research Logistics Quarterly* **17**(3), 261–268.

Modaresi, S., Sauré, D. and Vielma, J. P. (2020), 'Learning in combinatorial optimization: What and how to explore', *Operations Research* **68**(5), 1585–1604.

Morton, D. P. (2010), Stochastic network interdiction, *in* J. J. Cochran, L. A. Cox, P. Keskinocak, J. P. Kharoufeh and J. C. Smith, eds, 'Wiley Encyclopedia of Operations Research and Management Science', John Wiley & Sons, Inc.

Nguyen, D. H. and Smith, J. C. (2022*a*), 'Asymmetric stochastic shortest-path interdiction under conditional value-at-risk', *IISE Transactions* **0**(0), 1–13.

Nguyen, D. H. and Smith, J. C. (2022*b*), 'Network interdiction with asymmetric cost uncertainty', *European Journal of Operational Research* **297**(1), 239–251.

Robbins, H. (1952), 'Some aspects of the sequential design of experiments', *Bulletin of the American Mathematical Society* **58**(5), 527–535.

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

Russo, D. and Van Roy, B. (2014), 'Learning to optimize via posterior sampling', *Mathematics of Operations Research* **39**(4), 1221–1243.

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

Sadana, U. and Delage, E. (2023), 'The value of randomized strategies in distributionally robust risk-averse network interdiction problems', *INFORMS Journal on Computing* **35**(1), 216–232.

Sefair, J. A. and Smith, J. C. (2016), 'Dynamic shortest-path interdiction', *Networks* **68**(4), 315–330.

Smith, J. C. and Song, Y. (2020), 'A survey of network interdiction models and algorithms', *European Journal of Operational Research* **283**(3), 797–811.

Soleimani-Alyar, M. and Ghaffari-Hadigheh, A. (2017), 'Solving multi-period interdiction via generalized bender's decomposition', *Acta Math. Appl. Sin. Engl. Ser* (33), 633–644.

Steinrauf, R. (1991), Network interdiction models, Master's thesis, Naval Postgraduate School, Monterey, California.

Thompson, W. R. (1933), 'On the likelihood that one unknown probability exceeds another in view of the evidence of two samples', *Biometrika* **25**(3/4), 285–294.

Wood, R. K. (1993), ‘Deterministic network interdiction’, *Mathematical and Computer Modelling* **17**(2), 1–18.

Yang, J., Borrero, J. S., Prokopyev, O. A. and Sauré, D. (2021), ‘Sequential shortest path interdiction with incomplete information and limited feedback’, *Decision Analysis* **18**(3), 218–244.

Appendix A: Proofs of main results

Proof of Lemma 1. For (i), suppose by contradiction that $S(B) \in \mathcal{P}(B^*)$. Then, by definition of S^*, $r(S^*,\mu) \leq r(S(B),\mu)$. On the other hand, since $B \in \widetilde{\mathcal{B}}$, then $r(S(B),\mu) < r(S^*,\mu)$, and it can be concluded that $r(S(B),\mu) < r(S(B),\mu)$, which is a contradiction. Likewise, suppose by contradiction that $S^* \in \mathcal{P}(B)$ and let $\boldsymbol{c} \in \mathcal{U}(B)$. Then $r(S(B),\boldsymbol{c}) \leq r(S^*,\boldsymbol{c})$ because $\boldsymbol{c} \in \mathcal{U}$ and $r(S^*,\boldsymbol{c}) < r(S(B),\boldsymbol{c})$ as $\boldsymbol{c} \in \mathcal{U}(B)$. This implies that $r(S(B),\boldsymbol{c}) < r(S(B),\boldsymbol{c})$, which is a contradiction. For the second part, see the first part of the proof of Theorem 1 ■

Proof of Theorem 1. We assume that the cost distribution is parametrized on the mean cost vector, is absolutely continuous, and that its components are independent. Thus, we adopt the notation $f_a(\cdot|u_a)$ to represent the density of the cost associated with arc a, where u_a represents its mean value, for $a \in A$. Let $L_a(u_a|\nu_a)$ denote the Kullback-Leibler divergence between the cost vector distributions for arc a when the mean costs are given by u_a and ν_a respectively, i.e.

$$L_a(u_a|\nu_a) := \int_{\mathbb{R}} \ln(f_a(x|u_a)/f_a(x|\nu_a)) f_a(x|u_a)\, dx.$$

Consider $E \in \mathcal{E}$: by construction there exists $\nu \in \mathcal{U}$ with $\nu_a = \mu_a$ for all $a \in A \setminus E$, such that $\max\{r(S(B'),\nu) : B' \in \mathcal{B}\} > r(S^*,\nu)$. Set $B := \arg\max\{r(S(B'),\nu) : B' \in \mathcal{B}\}$, which we assume is unique (to avoid excessive notation). Note that $E \cap S^* = \emptyset$, so that

$$r(S(B),\mu) \leq r(S^*,\mu) = r(S^*,\nu) < r(S(B),\nu).$$

We conclude that $S(B) \cap E \neq \emptyset$.

For any consistent policy π we have that for any $\alpha > 0$

$$\begin{aligned}
\mathcal{R}^\pi(t, F(\cdot|\nu)) &\geq \Delta\, \mathbb{E}_\nu \left[t - \sum_{s \leq t} \mathbf{1}\{B^s = B\} \right] \\
&\geq \Delta\, \mathbb{E}_\nu \left[t - \sum_{s \leq t} \mathbf{1}\{S^s = S(B)\} \right] \\
&\geq \Delta\,(t - K \ln t)\, \mathbb{P}_\nu \left\{ \sum_{s \leq t} \mathbf{1}\{S^s = S(B)\} < K \ln t \right\} \\
&\geq \Delta\,(t - K \ln t)\, \mathbb{P}_\nu \{\tau(E,t) < K \ln t\} = o(t^\alpha).
\end{aligned}$$

for any positive constant $K \leq t/\ln t$, where Δ denotes the minimum optimality gap under ν, and $\mathbb{E}_u$ and $\mathbb{P}_u$ denote expectation and probability operators when the underlying mean cost vector is given by u. (While various random elements, e.g. S^t, depend on the policy π, we ignore such a dependence, to streamline the exposition.) From the consistency of π, we conclude that

$$\mathbb{P}_\nu \{\tau(E,t) < K \ln t\} = o(t^{\alpha-1}). \tag{13}$$

For $a \in E$ and $k \le \tau(\{a\},t)$, define random variable $\eta(a,k)$ as the period at which a cost realization for arc a is observed by the k-th time, i.e.

$$\eta(a,k)(\omega) := \inf\left\{ t \ge 1 : \sum_{s \le t} \mathbf{1}\{a \in S^s(\omega)\} = k \right\}, \quad k \le \tau(\{a\},t)(\omega),\, a \in E$$

and the partial log-likelihood random variable

$$\mathcal{L}_{a,k}(\omega) := \sum_{i=1}^{k} \ln\left(\frac{f_a(c_a^{\eta(a,i)(\omega)}(\omega)|\mu_a)}{f_a(c_a^{\eta(a,i)(\omega)}(\omega)|\nu_a)} \right).$$

Define the event

$$W := \left\{ \omega \in \Omega : \max_{a \in E}\left\{ \mathcal{L}_{a,\tau(\{a\},t)}(\omega) \right\} \le \frac{(1-\alpha)\ln t}{|E|},\, \tau(E,t)(\omega) < K \ln t \right\}.$$

Under our assumptions on F we have that

$$\begin{aligned}\mathbb{P}_\nu\{W\} = \int_W d\mathbb{P}_\nu &= \int_W \prod_{a \in E} \prod_{k=1}^{\tau(\{a\},t)} \frac{f_a(c_a^{\eta(a,k)}|\nu_a)}{f_a(c_a^{\eta(a,k)}|\mu_a)} d\mathbb{P}_\mu \\ &= \int_W \prod_{a \in E} \exp\left(-\mathcal{L}_{a,\tau(\{a\},t)}\right) d\mathbb{P}_\mu \ge \exp\left(-(1-\alpha)\ln t\right)\mathbb{P}_\mu\{W\} = \frac{\mathbb{P}_\mu\{W\}}{t^{1-\alpha}}.\end{aligned}$$

From (13), we conclude that

$$\lim_{t \to \infty} \mathbb{P}_\mu\{W\} = 0. \tag{14}$$

From the SLLN we have that $\lim_{k\to\infty} \frac{1}{k}\mathcal{L}_{a,k} = L_a(\mu_a|\nu_a)$ a.s. $(\mathbb{P}_\mu)$ for $a \in E$, thus

$$\lim_{k \to \infty} \frac{1}{k} \max\{\mathcal{L}_{a,l} : l \le k\} = L_a(\mu_a|\nu_a) \quad a.s.\,(\mathbb{P}_\mu),\, a \in E.$$

This implies that, for any $\delta > 1$,

$$\lim_{k \to \infty} \mathbb{P}_\mu\left\{ \frac{\mathcal{L}_{a,l}}{k} > \delta\, L_a(\mu_a|\nu_a) \text{ for some } l \le k \right\} = 0,$$

which in turn implies that, taking $k = \frac{(1-\alpha)\ln t}{\delta\,|E|\,L_a(\mu_a|\nu_a)}$,

$$\lim_{t \to \infty} \mathbb{P}_\mu\left\{ \mathcal{L}_{a,l} > \frac{(1-\alpha)\ln t}{|E|} \text{ for some } l \le \frac{(1-\alpha)\ln t}{\delta\,|E|\,L_a(\mu_a|\nu_a)} \right\} = 0.$$

The above equation implies that

$$\lim_{t \to \infty} \mathbb{P}_\mu\left\{ \mathcal{L}_{a,\tau(\{a\},t)} > \frac{(1-\alpha)\ln t}{|E|},\ \tau(\{a\},t) \le \frac{(1-\alpha)\ln t}{\delta\,|E|\,L_a(\mu_a|\nu_a)} \right\} = 0.$$

Define $\kappa_E := \frac{(1-\alpha)|E|}{\delta}\min\{L_a(\mu_a|\nu_a)^{-1} : a \in E\}$: noting that $\tau(\{a\},t) \le \tau(E,t)$ for $a \in E$, and using the union bound, we conclude from above that

$$\lim_{t \to \infty} \mathbb{P}_\mu\left\{ \mathcal{L}_{a,n(\{a\},t)} > \frac{(1-\alpha)\ln t}{|E|},\ \tau(E,t) \le \kappa_E \ln t\,, \text{for some } a \in E \right\} = 0,$$

Using $K = \kappa_E$, (14) and the above imply that

$$\lim_{t\to\infty} \mathbb{P}_\mu \left\{\tau(E,t) \leq \kappa_E \ln t\right\} = 0.$$

Because of the arbitrary nature of $\alpha > 0$ and $\delta < 1$, the result holds when redefining

$$\kappa_E := |E| \min\left\{L_a(\mu_a|\nu_a)^{-1} : a \in E\right\}.$$

Proof of Lemma 2. Observe that formulation (2) maximizes the gap between $\bar{\boldsymbol{z}}^1$ and $\bar{\boldsymbol{z}}^2$ when the mean cost vector is ν^2. Because $\bar{\boldsymbol{z}}^1$ is not necessarily optimal under ν^2, then $\Gamma \geq 0$. If $\Gamma = 0$, then we conclude that $\bar{\boldsymbol{z}}^1$ is optimal under ν^2, which implies, from the maximization sense of the formulation, that the conditions in (1) does not hold for any $B \in \mathcal{B}$. Alternatively, if $\Gamma > 0$ then $\hat{B}$ is an element of $\mathcal{B}$ for which the conditions in (1) hold. ■

Proof of Lemma 3. We focus on proving that $v^+ = w^+$, the other case is similar. Observe that given $\nu^k \in \text{int}(\mathcal{U})$ and the direction $\boldsymbol{d}$, v^+ can equivalently be found by finding the largest possible w such that $(\boldsymbol{z}^s - z)^\top(\nu^k + w\boldsymbol{d}) \leq 0 \ \forall z \in Z^s, s < t_0$ and $\nu^k + w\boldsymbol{d} \in \mathbb{C}$. In other words, we seek the w that solves the following problem:

$$\begin{aligned}
&\max\ w && && \text{(15a)}\\
&\text{s.t. } \min\left\{(\nu^k + w\boldsymbol{d})^\top z \colon \mathcal{A}^s z = b, z \in \{0,1\}^{|A\setminus B^s|}\right\} \geq (\nu^k + w\boldsymbol{d})^\top \boldsymbol{z}^s && s < t_0 && \text{(15b)}\\
&\qquad \nu^k + w\boldsymbol{d} \in \mathbb{C} && && \text{(15c)}\\
&\qquad w \geq 0. && && \text{(15d)}
\end{aligned}$$

Observe that the optimization problem in (15b) can be replaced by its linear relaxation because of the total unimodularity of $\mathcal{A}^s$. Additionally, one can replace such a relaxation by its dual, obtaining

$$\begin{aligned}
&\max\ w\\
&\text{s.t. } \max\left\{y_n^s - y_1^s \colon y_j^s - y_i^s \leq \nu_{i,j}^k + w d_{i,j} \ (i,j) \in A \setminus B^s, y \in \mathbb{R}^m\right\} \geq (\nu^k + w\boldsymbol{d})^\top \boldsymbol{z}^s \quad s < t_0\\
&\qquad \nu^k + w\boldsymbol{d} \in \mathbb{C}\\
&\qquad w \geq 0.
\end{aligned}$$

The result follows by noting that (8) is equivalent to the formulation above. ■