Learning in Sequential Bilevel Linear Programming

Juan S. Borrero
School of Industrial Engineering & Management, Oklahoma State University, Stillwater, OK 74078,
juan.s.borrero@okstate.edu

Oleg A. Prokopyev
Department of Industrial Engineering, University of Pittsburgh, Pittsburgh, PA 15261, droleg@pitt.edu

Denis Sauré
Department of Industrial Engineering, University of Chile, Santiago, Chile dsaure@dii.uchile.cl

We consider a framework for sequential bilevel linear programming where a Leader and a Follower interact over multiple time periods. In each period, the Follower observes the actions taken by the Leader and reacts optimally, according to his own objective function, which is initially *unknown* to the Leader. By observing various forms of information feedback from the Follower's actions, the Leader is able to refine her knowledge about the Follower's objective function, and hence, adjust her actions at subsequent time periods, which ought to help in maximizing the Leader's cumulative benefit. We show that *Greedy and Robust* policies adapted from previous work in the max-min (symmetric) setting might fail to recover the optimal *full information* solution to the problem (i.e., a solution implemented by an oracle with complete prior knowledge of the Follower's objective function) in the asymmetric case. In contrast, we present a family of *Greedy and Best-Case* policies that are able to recover the full information optimal solution and also provide real-time certificates of optimality. In addition, we show that the proposed policies can be computed by solving a series of linear mixed-integer programs. We test policy performance through exhaustive numerical experiments in the context of asymmetric shortest path interdiction, considering various forms of feedback and several benchmark policies.

1. Introduction

Motivation. In this paper we study sequential bilevel programming, where a Leader and a non-strategic Follower interact over multiple time periods. In these problems, at each time period, the Leader acts first by selecting an action, which is observed by the Follower, who then responds in an optimal fashion. On the one hand, the Leader aims at maximizing her cumulative payoff, which depends on the Follower's response at each period. On the other hand, the Follower aims at minimizing his immediate cost on each period, which also depends on the Leader's actions.

We assume that the single-period interaction between these two agents is modeled using the bilevel linear programming framework: at each period the Follower responds to the Leader's actions by solving a lower-level linear program, whose feasible region depends on the Leader's action (note this rules out any strategic behavior on the Follower's behalf); the Leader in turn, knowing this, solves an upper-level mathematical program in which some variables (those corresponding to the Follower's response) are constrained to be solutions to the lower-level program. Because of its inherent ability to model hierarchical decision-making settings, bilevel programming has been extensively used in many application areas including defense (Brown et al. 2006), economics (Sherali et al. 1983), transportation (Lucotte and Nguyen 2013), revenue management (Côté et al. 2003), among many others; see the surveys by Colson et al. (2005, 2007) and the references therein.

In this paper we are motivated by bilevel linear programming applications in interdiction, particularly, sequential problems in military, surveillance and homeland security domains, where the objectives of the different agents are not necessarily aligned. For example, consider an interdiction setting where border patrol agents (the Leader) can block routes that drug-smugglers (the Follower) use to transport raw material (chemicals used for drug production) or finalized product (drug), by placing resources such as planes, ships, or military units to patrol those routes. This setting is in fact responsible for the renewed interest in network interdiction in the early nineties, in the context of the US counter-narcotic efforts to disrupt drug production and distribution in South America (Steinrauf 1991). A key feature of the early application of network interdiction to this context is that, while smugglers aim at maximizing drug-production/distribution, because of uncertainty about the precise location of production facilities, authorities' efforts aim at minimizing the flow of raw material/drug in/out of a region thought to include such facilities. Under the assumption of complete knowledge of the transportation network, its costs, the targeted region and facility locations, this problem can be framed as a network flow interdiction problem (Wood 1993, Chern and Lin 1995, Israeli and Wood 2002, Lim and Smith 2007, Bayrak and Bailey 2008), which is a particular type of bilevel linear optimization problem; see also a recent survey by Smith and Song (2020). However, in practice authorities do not have complete knowledge of the network or its costs, nor of the facility locations, thus it is not possible to directly instantiate the problem. Nonetheless, border patrol agents might observe (after-the-fact) the route used by smugglers, or infer transportation costs (even without knowing which route was used) by collecting information about drug activity from paid informants, or refine their assessment about the location of production facilities by collecting information about aircraft flights in and out covert airfields.

The setting also arises in other applications areas such as manufacturing (Cao and Chen 2006). Consider a firm (the Leader) that outsources production to various plants (the Follower); the Leader designs contracts to minimize acquisition costs; and once contracts are signed, the plants configure their operations so as to minimize their operational costs. In practice, firms often do not know the cost parameters associated with the plants' operations, and the plants might not have any incentive to reveal this information to the firm. However, the firm observes outsourced production, and might have access to some form of periodic financial statements by the plants, from which cost structures might be inferred.

The subject of learning through repeated sequential interaction between agents has been addressed in the Game Theory literature, mostly through an axiomatic approach (Fudenberg and Levine 1998). In this regard, we are motivated by Stackelberg-like interactions, focusing on the computation of optimal policies. For this, following the bulk of the literature on sequential bilevel programming, we assume that the Follower is non-strategic, i.e., a short-run player in a finite game (Fudenberg and Levine 1998), so as to simplify the analysis.

In this work, we aim at closing the gap between the theory and practice of bilevel programming by jointly incorporating two key aspects of the motivating examples above: the (initial) uncertainty surrounding certain parameters, and the fact that the Leader's profit does not necessarily coincide with the Follower's cost as in the max-min setting. In this regard, traditional bilevel optimization literature assumes that all parameters defining the upper/lower-level problems are known upfront by both agents, and when it does, it is usually in the max-min setting, which henceforth we refer to as the *symmetric* setting. While this might not fit most settings in practice, it is often the case that the Leader might have access to some form of feedback about the Follower's response to her decisions. Thus, while the Leader might have limited initial information about the Follower's operational parameters, such a knowledge might be refined periodically by using feedback from the Follower's reaction.

One possible approach to addressing the uncertainty surrounding the parameters guiding the Follower's response is via the use of probability distributions (Hemmecke et al. 2003), which can be constructed by using historical data, expert opinions, or well-understood physical processes. However, in many applications, such as the ones presented above, this type of information might either not be available or sufficient to reliably estimate such distributions.

A possible approach to addressing parametric uncertainty is the multi-armed bandit (Robbins 1952), which can be leveraged to tackle settings with combinatorially many arms (Cesa-Bianchi and Lugosi 2012, Modaresi et al. 2020). However, current methods do not extend to our setting. The reason for this is that, unlike in previous work in combinatorial bandits, which take advantage of the additive nature of the rewards to minimize the exploration of alternatives, in our problem rewards are the outcome of an optimization problem, and thus lack a structure that can be taken advantage of. Another 'distribution-free' alternative to deal with uncertainty is multi-stage robust optimization (Bertsimas and Georghiou 2015, Lorca et al. 2016). This approach, however, assumes a worst-case realization of the uncertainty and typically considers a one-level problem (rather than a bilevel) at each stage. Perhaps more importantly, this method deals with time-independent uncertainty, thus the learning process of the Leader cannot be addressed because modeling learning requires having *dependent* uncertainty sets.

A different approach that incorporates specific forms of feedback from the Follower's actions and deals with two-level optimization problems is given by sequential interdiction problems under incomplete information (Borrero et al. 2016, 2019). As mentioned earlier, these models assume a (symmetric) max-min relationship between a Leader's profit and a non-strategic Follower's cost. We are instead motivated by settings where the Leader and the Follower's objectives do not necessarily coincide, which we refer to as the *asymmetric* setting.

For example, in the context of drug smuggling, the border patrol might be interested in minimizing the evasion probability, while the evader might be maximizing the evasion probability, respectively. However, their costs coefficients (e.g., arc costs in the underlying transportation network) might be not necessarily aligned, i.e., "the players may not have the same perception of their problem data" (Smith and Song 2020). Alternatively, the evader might be optimizing some other measure, e.g., minimizing transportation costs (if either being unaware or simply ignoring possible border patrols), or maximizing the expected amount of the drugs smuggled. We refer the reader to the detailed survey Smith and Song (2020) and the references therein for other examples and the discussion of network interdiction problems with asymmetric settings.

Similarly, in the context of manufacturing, in general, the firm's procurement costs (e.g., regulated by a contract) do not necessarily match the plants' cash flows (which depends on how production is executed). Broadly speaking, many settings of interest are asymmetric in a sense that the Leader's profit does not necessarily coincide with the Follower's cost as in the max-min setting. Moreover, we will see that policies adapted from the existing work for the max-min case (Borrero et al. 2016, 2019) do not perform well in the asymmetric setting. In particular, such policies might stall and implement sub-optimal interdiction actions indefinitely.

Research goal. Considering the issues above, in this paper we analyze sequential bilevel linear programming where the Leader has incomplete information about the parameters defining the Follower's lower-level problem, but has access to feedback from the Follower's response in each period. Our research goal is two fold. First, we aim at studying the performance of policies adapted from extant research in terms of their convergence and the optimality guarantees they provide. Second, we aim at developing policies that converge to the optimal "full information" solution that a Leader with prior knowledge on the Follower's parameters would implement, and that are able to signal in real time when such a convergence has been achieved.

Specifically, we consider a class of online optimization problems, which we refer to as sequential bilevel linear problems with incomplete information (SBPI). We assume that the Leader and the Follower interact across a set of given time periods $\mathcal{T}$, and that the Leader knows all the parameters of the Follower's problem except for his cost vector, which she knows is time-invariant (see Section 6 for a discussion on time-variant settings) and belongs to a given *uncertainty set*. At each period

$t \in \mathcal{T}$, the Leader selects a feasible upper-level solution x^t, and then the Follower, who knows his cost vector with certainty, selects (rather non-strategically) an optimal response y^t to x^t. Such response in turns generates some information *feedback*, which is observed by the Leader. We consider three different types of feedback: *Standard*, where the Leader observes both the values of the upper- and lower-level objective functions; *Value-Perfect*, where in addition to Standard feedback, the Leader observes the lower-level objective coefficients associated with all activities performed by the Follower at t; and *Response-Perfect*, where in addition to Standard feedback, the Leader observes the Follower's response y^t. The Leader might use this feedback to refine her belief about the unknown cost coefficients, and thus improve her decision-making in subsequent time periods.

Following extant literature (Borrero et al. 2019), we assess policy performance in terms of their *time stability*, which is defined as the number of periods it takes a policy to converge to the solution implemented by an oracle Leader with complete prior knowledge of the Follower's objective function. Note that time stability is closely connected to the notion of regret in online optimization (Cesa-Bianchi and Lugosi 2006), as a finite upper bound on time stability implies a finite upper bound on regret. Our analysis follows closely that by Borrero et al. (2019), who studies sequential interdiction in the max-min setting, when the Leader is not only unaware of the Followers' objective function (and thus, her own), but also about other parameters defining the Followers' response. (While we restrict uncertainty to the Follower's objective, further uncertainty can be handled following the framework presented in Borrero et al. (2019); we do not consider such an extension here, so as to streamline the exposition.) Because our work can be seen as extending the model of Borrero et al. (2019) to asymmetric settings, we connect our results to those in the aforementioned work throughout the manuscript. The reader is directed to Appendix B for a summary of the setting and results in Borrero et al. (2019). There, we provide a comparative analysis of our setting and results.

Contribution. Our work contributes to the literature on online optimization in various fronts. First, we show that in the general asymmetric case, policies adapted from the max-min (symmetric) setting might fail to converge to the full-information solution, even when all information contained in the feedback is used, and thus it is not possible to bound their time stability. From the Leaders' perspective, such policies operate as if the Follower was also unaware of his objective function and adopt a robust approach to handling such uncertainty; assuming this, the Leader acts greedily, selecting the action that maximizes the immediate profit (Borrero et al. 2019).

Second, we reinterpret the ideas behind the aforementioned policies in the context of our setting, and propose the family of the *Greedy and Best-Case policies*, which, as we show, converge to the full-information solution, and provide certificates of optimality in real time, under mild conditions, even when feedback is Standard, see Theorem 2. From the Leaders' perspective, these policies operate as if the Leader was able to select the cost vector that the Follower would face, and acts greedily by selecting the action that maximizes the immediate profit. (This is only an artifact, as the Follower knows the actual cost vector and acts upon such a knowledge.) We show that convergence to the full-information solution can be checked on each period, simply by comparing the Leaders' expected benefit (if her choice of costs was correct) and the observed one, see Theorem 1. Furthermore, to alleviate potential scalability issues, we discuss a modified policy that ensures convergence to constant-factor approximate full-information solutions, see Corollary 1.

Third, a key distinctive feature of our work, relative to Borrero et al. (2019), is that our analysis assumes that the Leader have some discretion on the use of the information contained in Standard feedback beyond the value or response-perfect cases. This leads to a series of possible update mechanisms, which affect the practical implementation of the proposed policies. In particular, we show that the best "use" of the information leads to a *full update* of the uncertainty set. We show however, that such an update is rather intractable as it is non-convex and non-closed, and does not lend itself to a mixed-integer representation of the uncertainty set in each period, which prohibits the use of mixed-integer programming-based approaches to implement Greedy and Best-Case policies. For this reason, we consider additional *Cvx* and *NCvx* update mechanisms, which differ in the amount of information incorporated while trading tractability of their representation. These updates are amenable to implementation (in particular, are more easily incorporated into mixed-integer programming-based approaches for policy implementation), and thus are used in our computational experiments. In this regard, our numerical results suggest that update mechanisms that consider a better use of the information also provide better performance. Despite its good practical performance, we show that in general such updates do not guarantee convergence to the full-optimal solution under Greedy and Best-Case policies, as such policies might stall and implement suboptimal solutions indefinitely, even under Value-Perfect or Response-Perfect feedback.

Considering the above, our third contribution is showing that if the Follower's problem admits a linear-programming (LP) representation, then the uncertainty set is mixed-integer linear representable under both the Cvx and the NCvx update mechanisms. Moreover, we show that the proposed policies can be implemented by solving a series of mixed-integer linear programs. Regarding the case of the full update mechanism, we present an approximate mechanism that adds a series of "*non-repetitive*" linear constraints to the NCvx mechanism (and thus, it is mixed-integer representable), and show that such an approximation yields a bound on time stability, which coincides with that provided by the full update mechanism under certain conditions. Following our motivating example on drug smuggling, our experiments, as well as the illustrating examples presented throughout the paper, consider instances of asymmetric shortest path interdiction problems (see Example 1), understanding that the theoretical developments in the paper, as well as the proposed policies, apply to the more general bilevel linear setting.

To assess the performance of the various policies introduced, we perform an extensive series of computational experiments under various combinations of feedback and update mechanisms. In our results, the time stability of the proposed policies is considerably better than the worst-case theoretical bound, and quite close to the number of actions of the Follower. The latter observation suggests that under certain assumptions on the feedback and the update mechanisms, the Greedy and Best-Case policies might be worst-case optimal. That is, the proposed policies might the best possible, when evaluated against settings that while consistent with the prior information, are designed so as to result in the largest time-stability.

Structure of the paper. In the next section we formally introduce sequential bilevel linear problems with incomplete information, and present the different update mechanisms and feedback types. Section 3 analyzes the theoretical performance of the Greedy and Robust policies, while Section 4 presents such an analysis for the newly proposed Greedy and Best-Case policies. Details for policy implementation when the Follower's problem admits an LP representation are presented in Section 5, together with our numerical experiments. Finally, in Section 6 we provide conclusions and possible directions for future research. Proofs of all results are relegated to Appendix A.

2. Problem Formulation

Overview. Consider two decision–makers, the Leader and the Follower, that interact sequentially in each time period t in $\mathcal{T} = \{1, \dots, T\}$. At period t, the Leader acts first by selecting x_r^t, the usage level of each *resource* r in a set R; and after observing the Leader's decision, the Follower selects y_a^t, the usage level of each *activity* a in a set A.

We assume that the Leader's payoff is a time-invariant, known and linear function of x^t and y^t, and that the Follower's cost is the dot product between y^t and a time-invariant cost vector $\boldsymbol{c}$, known to the Follower but not to the Leader. Instead, we assume that the Leader maintains an *uncertainty set* $\mathcal{U}^t$, which is known to contain $\boldsymbol{c}$. (Thus, $\mathcal{U}^1$ represents the Leader's initial knowledge about $\boldsymbol{c}$.)

At each time $t \in \mathcal{T}$ the following sequence of events takes place:

1. Knowing that the cost vector $\boldsymbol{c}$ lies in an uncertainty set $\mathcal{U}^t$, the Leader chooses an (upper-level) resource-usage vector x^t within a feasible region X.
2. Observing the Leader's decision, the Follower chooses a (lower-level) activity-usage vector y^t from a region $Y(x^t)$, which depends on the Leader's decision.
3. The Leader collects the period's profit and observes a *feedback* $\mathcal{K}^t$, which is used, via an *update mechanism*, to update the uncertainty set $\mathcal{U}^{t+1}$.

Regarding this last step, and borrowing from extant literature, we consider various form of feedback, all of which include the Follower's and Leader's cost and profit, respectively.

Assuming that the Follower is non-strategic and greedy, and that the Leader maximizes her cumulative profit, a feasible policy (for the Leader) is a sequence of set functions that, on each period, map the history of the interaction to a feasible resource-usage vector. In presenting and analysing the proposed policies it is helpful to think about the uncertainty set $\mathcal{U}^t$ as mapping the history of the process to actionable information on $\boldsymbol{c}$, so that a policy is characterized by how it selects a usage vector as a function of $\mathcal{U}^t$; and how it uses the feedback generated to update $\mathcal{U}^{t+1}$. In this regard, our analysis shows that this two components are interconnected as, for example, the tractability of the update mechanism affects the complexity of choosing a resource-usage vector, and vice-versa.

Next, we formally introduce the components of interaction, as described above. For that purpose, we use bold symbols to denote the parameters of either the Leader's or the Follower's problems.

The Follower's response. Suppose the Leader selects $x^t := (x_r^t : r \in R)$ in period $t \in \mathcal{T}$. Following extant literature, we assume that the Follower then selects $y^t := (y_a^t : a \in A)$ from his rational reaction set

$$Z(x^t;\boldsymbol{c}) := \arg\min\left\{\boldsymbol{c}^\top y \colon y \in Y(x^t)\right\}, \tag{1}$$

with $Y(x) = \left\{y \in \mathbb{R}_+^{|A|} \colon \boldsymbol{F}y + \boldsymbol{L}x \leq \boldsymbol{f}\right\}$ for all $x \in \mathbb{R}^{|R|}$. Thus, the Follower minimizes a linear *cost* function, whose coefficients are given by the vector $\boldsymbol{c}$, subject to polyhedral constraints. Note that this prevents any strategic behavior on the Followers' behalf. The cost perceived by the Follower is given by

$$z(x^t;\boldsymbol{c}) := \min\left\{\boldsymbol{c}^\top y \colon y \in Y(x^t)\right\}. \tag{2}$$

Note that the Leader's decisions affect the feasible region in (1). We assume all parameters above are known to the Follower upfront, so that y^t can be computed upon observing the value of x^t.

The Leader's decision. The Leader collects a profit in each time period, and aims at maximizing the cumulative profit throughout the horizon. We assume that the Leader's profit in period t is a linear function of both the Follower's decision y^t, and the Leader's decision x^t, which is constrained to lie within a region $X := \left\{x \in \mathbb{R}_+^{|R|-I} \times \mathbb{Z}_+^I \colon \boldsymbol{H}x \leq \boldsymbol{h}\right\}$, where $I \leq |R|$. With this, the Leader's profit in period t is given by $w(x^t, y^t)$, where

$$w(x,y) = \boldsymbol{b}^\top x + \boldsymbol{d}^\top y, \quad y \in Y(x),\, x \in X.$$

Here, $\boldsymbol{b}$ relates to the profit generated by the Leader directly from her actions, and $\boldsymbol{d}$ relates to those generated by the Follower's reactions to said actions. Note that, given $x^t \in X$, the Follower's response lies within the set $Z(x^t;\boldsymbol{c})$, and assume that whenever such a set is not a singleton, the Follower's response y^t takes the value that is most beneficial to the Leader; this *optimistic approach* is standard in the bilevel literature (Dempe 2002). With this, the Leader's profit depends solely on x^t, and is given by

$$\tilde{w}(x^t;\boldsymbol{c}) := \boldsymbol{b}^\top x^t + v(x^t;\boldsymbol{c}),$$

where

$$v(x;\boldsymbol{c}) := \max\big\{\boldsymbol{d}^\top y\colon\, y \in Z(x;\boldsymbol{c})\big\}, \quad x \in X. \tag{3}$$

Note that we implicitly assume that the Follower knows $\boldsymbol{d}$ (i.e., knows how the lower-level actions impact the Leader, see below); and reacts in accordance with the optimistic approach to bilevel programming. For a given sequence of decisions $\{x^t : t \in \mathcal{T}\}$, the Leader's total cumulative profit is given by

$$\mathcal{P}(\big\{x^t : t \in \mathcal{T}\big\}; \boldsymbol{c}) := \sum_{t \in \mathcal{T}} \tilde{w}(x^t; \boldsymbol{c}).$$

Information setting. A key feature of this work is that we assume that, at time $t = 1$, the Leader does not know with certainty the parameters defining the Followers response, and thus can not optimize her profit $\mathcal{P}$ directly. In particular, we assume that while the Leader knows her profit function (i.e., knows $\boldsymbol{b}$ and $\boldsymbol{d}$), she *does not know* $\boldsymbol{c}$, the cost coefficients on the Follower's objective function. Note that, were the Leader certain about the value of $\boldsymbol{c}$, then in each period she would implement an optimal solution to the bilevel deterministic problem:

$$\tilde{w}^*(\boldsymbol{c}) := \max\{\boldsymbol{b}^\top x + v(x;\boldsymbol{c})\colon\, x \in X\}. \tag{4}$$

We call this the *full-information* solution to the Leader's problem. Instead, we assume that the Leader only knows that $\boldsymbol{c}$ lies within a known polyhedral uncertainty set $\mathcal{U}^1$ (Ben-Tal et al. 2009), given by

$$\mathcal{U}^1 = \{\hat{\boldsymbol{c}} \in \mathbb{R}^{|A|}\colon\, \boldsymbol{G}^1\hat{\boldsymbol{c}} \leq \boldsymbol{g}^1\}.$$

(Recall our assumption that, unlike the Leader, the Follower has all information needed to solve problem (1) and compute y^t.) Our modeling choice aims at representing that while the Leader might be aware of her own objective, resources and capabilities, she might not fully understand the Followers' rationale (a more general setting considering uncertainty on additional parameters defining the Followers response would admit a similar treatment (Borrero et al. 2019)). In this regard, we do not consider uncertainty surrounding the Leader's own parameters; however, we discuss the challenges associated with extending our approach to incorporate uncertainty on the Leader's parameters in our conclusions.

Remark 1. It is important to note that the setting described above does *not* generalize the max-min problem studied by Borrero et al. (2019), since that setting considers uncertainty surrounding the Leader's objective (e.g., considering uncertainty about the value of $\boldsymbol{d}$). ■

Remark 2. While the assumption that the uncertainty set is polyhedral is required to compute our proposed policies using mixed-integer linear programming, the theoretical results (which the exception of those involving polyhedral dimension) do not require this assumption. In fact, these results, which include convergence guarantees, hold for any non-empty compact uncertainty set, see Appendix A. ■

Feedback. Our model assumes that, on period t, after implementing both x^t and y^t, a feedback $\mathcal{K}^t := \mathcal{K}(x^t, y^t; \boldsymbol{c})$ is observed by the Leader. We use the three types of feedback functions $\mathcal{K}$ introduced by Borrero et al. (2019):

(i) **Standard feedback**: At each time $t \in \mathcal{T}$ the Leader learns the values of $z(x^t; \boldsymbol{c})$ and of $v(x^t; \boldsymbol{c}) := \tilde{w}(x^t; \boldsymbol{c}) - \boldsymbol{b}^\top x^t$ (because $\boldsymbol{b}$ and x^t are known by the Leader, this is equivalent to learning $\tilde{w}(x^t; \boldsymbol{c})$). Thus, under standard feedback, we have

$$\mathcal{K}(x^t, y^t; \boldsymbol{c}) = \left\{ z(x^t; \boldsymbol{c}), \tilde{w}(x^t; \boldsymbol{c}) \right\}.$$

(ii) **Response-Perfect feedback**: At any time $t \in \mathcal{T}$, in addition to Standard feedback, the Leader learns the value of y_a^t for all $a \in A$ such that $y_a^t > 0$. Thus, under standard Response-Perfect feedback, we have

$$\mathcal{K}(x^t, y^t; \boldsymbol{c}) = \left\{ z(x^t; \boldsymbol{c}), \tilde{w}(x^t; \boldsymbol{c}),\, y_a^t \text{ for } a \in A \text{ s.t. } y_a^t > 0 \right\}.$$

(iii) **Value-Perfect feedback**: At any time $t \in \mathcal{T}$, in addition to Standard feedback, the Leader learns the value of c_a for all $a \in A$ such that $y_a^t > 0$. Thus, under standard Value-Perfect feedback, we have

$$\mathcal{K}(x^t, y^t; \boldsymbol{c}) = \left\{ z(x^t; \boldsymbol{c}), \tilde{w}(x^t; \boldsymbol{c}),\, c_a^t \text{ for } a \in A \text{ s.t. } y_a^t > 0 \right\}.$$

For example, in the smuggling interdiction setting, the Follower's objective may correspond to either maximizing the expected amount of the drugs smuggled, maximizing the evasion probability, or simply minimizing the transportation costs. Then standard feedback corresponds to observing the Leader's and Follower's objective function values given both the Leader's and the Follower's decisions. From the application perspective such values can be inferred by exploring various types of available data (e.g., prices in illegal markets, enforcement and punishment records from the law-enforcement agencies); see Buehn and Eichler (2009), Gathmann (2008), Magliocca et al. (2019), Yürekli and Sayginsoy (2010), Yang et al. (2019) and the references therein. Similarly, Response-Perfect and Value-Perfect feedback might correspond to the law-enforcement observing a particular network route and its arc costs, respectively, from the available intelligence information, e.g., satellite images, communication interceptions.

2.1. Sequential bilevel problem with incomplete information (SBPI).

Let us revisit the definition of $\mathcal{P}$. Because the Leader's decisions might adapt to the feedback collected on each period, we have that, in full generality, a (Leader's) policy $\pi := (\pi^t : t \in \mathcal{T})$ is a sequence of set functions such that $x^t = \pi^t(\mathcal{H}^t(\pi; \boldsymbol{c})) \in X$, where $\mathcal{H}^t(\pi; \boldsymbol{c}) := (x^1, \mathcal{K}^1, \dots, x^{t-1}, \mathcal{K}^{t-1})$ denotes the history of upper-level decisions made, and feedback collected, up to time $t \geq 1$, where we define $\mathcal{H}^1 = \emptyset$. Note that such a history depends on the actual value of $\boldsymbol{c}$ and π (although we have suppressed the dependencies of actions and feedback on these values, to streamline the exposition). Accounting for these dependencies, and considering that $\boldsymbol{c}$ is initially unknown, it is possible to define the Leader's problem by assuming that, given $\mathcal{U}^1$, the Leader focuses on maximizing profits assuming a worst-case realization of $\boldsymbol{c}$, for a given policy π. That is, the Leader's problem becomes

$$\max_{\pi \in \Pi} \inf_{\boldsymbol{c} \in \mathcal{U}^1} \sum_{t \in \mathcal{T}} \tilde{w}(\pi^t(\mathcal{H}^t(\pi; \boldsymbol{c})); \boldsymbol{c}),$$

where Π denotes the set of feasible policies. *In the remainder of the paper whenever discussing a particular policy π, we use a superscript π to discuss vectors and quantities associated with it.*

Fixing all parameters known upfront by the Leader, we define the time stability $\tau^\pi(\boldsymbol{c})$ associated with policy a $\pi \in \Pi$ and a cost vector $\boldsymbol{c} \in \mathcal{U}^1$ as

$$\tau^\pi(\boldsymbol{c}) := \inf\{t \in \mathcal{T} : \tilde{w}(x^{t,\pi}; \boldsymbol{c}) = \tilde{w}^*(\boldsymbol{c}) \text{ for all } s \geq t\}.$$

In words, the time stability of a policy is the first time period by which the Leader implements the optimal full-information solution to (4), from there on. Following extant work, we focus our attention in minimizing the worst-case time stability (across all instances of **SBPI**).

2.2. Learning New Information: Uncertainty Set Updates

As mentioned earlier, we can think of the uncertainty set $\mathcal{U}^t$ as mapping the history of the interaction into actionable information on $\boldsymbol{c}$, so that policies might be characterized, partially, by the process by which they select the vector x^t, as a function of this information. In this regard, the next sections show that policies from extant work and those proposed in this work choose x^t by solving mathematical programs that take $\mathcal{U}^t$ as input. In this context, a policy is characterized also by the *update mechanism* it uses on each period to incrementally combine current information on $\boldsymbol{c}$ (i.e. the uncertainty set $\mathcal{U}^t$) with the most recent information on the interaction (i.e. the feedback $\mathcal{K}^t$) to produce new knowledge on $\boldsymbol{c}$ (i.e. the set $\mathcal{U}^{t+1}$).

Formally, we define an update mechanism U as a mapping from uncertainty sets and feedback into a new uncertainty sets, so that

$$\mathcal{U}^{t+1} = U(\mathcal{U}^t, \mathcal{K}^t).$$

Consider $x \in X$ and $y \in Z(x;\boldsymbol{c})$: we say an update mechanism is *valid* if it does not cut off the cost vector $\boldsymbol{c}$, i.e. $\boldsymbol{c} \in U(\mathcal{U}, \mathcal{K}(x,y;\boldsymbol{c}))$ whenever $\boldsymbol{c} \in \mathcal{U}$, and does not increase the uncertainty about $\boldsymbol{c}$, i.e. $U(\mathcal{U}, \mathcal{K}(x,y,\boldsymbol{c})) \subseteq \mathcal{U}$.

Intuitively speaking, smaller uncertainty sets should translate into better decisions, so that one is to prefer update mechanisms that produce smaller sets. At the same time, because of their potential role in the computation of x^t, one should also prefer updates that result in tractable representations of the uncertainty sets (as such representations might be incorporated, for example, in a mathematical formulation). Next, we present three such valid update mechanisms for the case of Standard feedback, i.e., where the Leader observes $\mathcal{K}^t = (z(x^t;\boldsymbol{c}), \tilde{w}(x^t;\boldsymbol{c}))$. The first mechanism uses the fact that cost $z(x^t;\boldsymbol{c})$ is associated with the Follower's rational response, and results in a polyhedral representation of the uncertainty sets; the second mechanism adds the fact that profit $\tilde{w}(x^t;\boldsymbol{c})$ is also associated with the Follower's rational response, and results in uncertainty sets that are not necessarily convex; and the third one incorporates the fact that the Follower's response follows the optimistic approach to bilevel programming, and results in uncertainty sets that are not necessarily closed.

Cvx update. In this update, at any time $t \in \mathcal{T}$, the Leader only uses the fact that $z(x^t;\boldsymbol{c})$ is the optimal value of the Follower's problem. In particular,

$$U(\mathcal{U}^t, \mathcal{K}^t) = \mathcal{U}^t \cap \mathcal{C}^t, \text{ where } \mathcal{C}^t := \{\hat{\boldsymbol{c}} \in \mathbb{R}^{|A|} \colon z(x^t;\hat{\boldsymbol{c}}) \geq z(x^t;\boldsymbol{c})\}. \tag{5}$$

One can check the update is trivially valid. Note that the uncertainty set resulting from this update is a semi-infinite linear set (i.e., a set with infinitely many constraints), and thus can be transformed into a lifted polyhedron by exploiting LP duality. That is,

$$\mathcal{C}^t = \left\{\hat{\boldsymbol{c}} \in \mathbb{R}^{|A|} \colon \exists q^t \in \mathbb{R}^m_+ \text{ s.t. } (\boldsymbol{L}x^t - \boldsymbol{f})^\top q^t \geq z(x^t;\boldsymbol{c}), -\boldsymbol{F}^\top q^t - \hat{\boldsymbol{c}} \leq 0\right\},$$

where m denotes the size of $\boldsymbol{f}$. Thus, $\mathcal{U}^t$ is a polyhedron for all $t \in \mathcal{T}$ under this update.

NCvx update. In this update, at time $t \in \mathcal{T}$ the Leader uses the fact that $z(x^t;\boldsymbol{c})$ is the optimal value of the Follower's problem and that the Leader's profit is $w(x^t;\boldsymbol{c})$. Formally, $U(\mathcal{U}^t, \mathcal{K}^t) = \mathcal{U}^t \cap \mathcal{N}^t$ where

$$\mathcal{N}^t := \left\{\hat{\boldsymbol{c}} \colon \exists y \in Z(x^t;\hat{\boldsymbol{c}}) \text{ s.t. } \hat{\boldsymbol{c}}^\top y = z(x^t;\boldsymbol{c}) \text{ and } \boldsymbol{b}^\top x^t + \boldsymbol{d}^\top y = \tilde{w}(x^t;\boldsymbol{c})\right\}.$$

This update is trivially valid, and *stronger* than the Cvx update in the sense that $\mathcal{N}^t \subseteq \mathcal{C}^t$. The name of the update follows as $\mathcal{N}^t$ is the set of solutions of an *inverse value linear optimization problem* (Ahmed and Guan 2005) and as such, it is a non-convex set in general. It is readily checked that $\mathcal{N}^t$ can be viewed as the following union of polyhedra

$$\mathcal{N}^t = \bigcup_{y \in E(x^t, \tilde{w}(x^t;\boldsymbol{c}))} \left\{\hat{\boldsymbol{c}} \colon \hat{\boldsymbol{c}}^\top y = z(x^t;\boldsymbol{c}),\, z(x^t;\hat{\boldsymbol{c}}) = z(x^t;\boldsymbol{c})\right\}, \tag{6}$$

where $E(x,w) := \{y \in Y(x)\colon \boldsymbol{b}^\top x + \boldsymbol{d}^\top y = w\}$. (To see this, note that $z(x^t;\hat{\boldsymbol{c}}) = z(x^t;\boldsymbol{c}) = \hat{\boldsymbol{c}}^\top y$ implies that $y \in Z(x^t;\hat{\boldsymbol{c}})$.) Therefore, under the assumption that the Follower always chooses an extreme point optimal solution (in case of multiple optimal solutions), $Y(x)$ can be replaced by its set of extreme points, thus $E(x^t,\tilde{w}(x^t;\boldsymbol{c}))$ is finite and $\mathcal{N}^t$ is a finite union of polyhedra.

Observe that this update uses *more* information in the feedback than the Cvx update (and thus, one would expect that $\mathcal{N}^t \subset \mathcal{C}^t$) and hence it should lead to better decisions. However, this comes at the price of more convoluted representation of $\mathcal{U}^t$.

Full update: In this update, in addition to the information used in the NCvx update, the Leader uses the fact that y^t must "favor" the Leader, in the sense of (3). Thus, a feasible $\hat{c}$ must belong to $\mathcal{N}^t$ and be such that $\boldsymbol{d}^\top y^t = v(x^t;\hat{\boldsymbol{c}})$. In other words, $U(\mathcal{U}^t,\mathcal{K}^t) = \mathcal{U}^t \cap \mathcal{F}^t$ where

$$\mathcal{F}^t := \left\{\hat{\boldsymbol{c}}\colon \exists y \in Z(x^t;\hat{\boldsymbol{c}}) \text{ s.t. } \hat{\boldsymbol{c}}^\top y = z(x^t;\boldsymbol{c}),\, \boldsymbol{b}^\top x^t + \boldsymbol{d}^\top y = \tilde{w}(x^t;\boldsymbol{c}) \text{ and } \boldsymbol{d}^\top y = v(x^t;\hat{\boldsymbol{c}})\right\}. \quad (7)$$

An alternative equation for $\mathcal{F}^t$ is given by

$$\mathcal{F}^t = \bigcup_{y \in E(x^t,\tilde{w}(x^t;\boldsymbol{c}))} \left\{\hat{\boldsymbol{c}}\colon \hat{\boldsymbol{c}}^\top y = z(x^t;\boldsymbol{c}),\, z(x^t;\hat{\boldsymbol{c}}) = z(x^t;\boldsymbol{c}) \text{ and } \boldsymbol{d}^\top y = v(x^t;\hat{\boldsymbol{c}})\right\}.$$

We call this update "full" as it incorporates all information from Standard feedback. The update is in general non-convex, and it might result in non-closed sets. The following example illustrates the update mechanisms described above.

Example 1 (**Asymmetric Shortest Path Interdiction.**). Consider the bilevel problem known as the asymmetric shortest path interdiction problem (ASPI) (Bayrak and Bailey 2008). Here, the Follower's objective is to move between two fixed nodes at a minimum cost and the objective of the Leader is to interdict k arcs to maximize the profit of the shortest path that the Follower uses (note that it is also assumed that $\boldsymbol{b} = \boldsymbol{0}$). If the Follower uses arc $a \in A$ then he incurs a cost of c_a and the Leader gets a profit of d_a; in general, $c_a \neq d_a$. The Leader does not know the cost vector $\boldsymbol{c}$, but she knows that the cost of arc a lies in the interval $c_a \in [\ell_a, u_a]$, $\ell_a \leq u_a$.

Figure 1 illustrates an instance of ASPI, where we assume that $k = 1$. Next, we illustrate the various update mechanisms under Standard feedback by means on an example. For simplicity, we include in $\mathcal{U}^1$ only those costs which the Leader does not know with certainty, thus $\mathcal{U}^1 = [0,10]^5$. Suppose that at time $t = 1$ the Leader interdicts arc $(1,2)$; thus, the Follower's solution is $y^1 = 1-3-7$, which gives $z(x^1;\boldsymbol{c}) = 3$ and $\tilde{w}(x^1;\boldsymbol{c}) = 10$ (recall that under Standard feedback, the Leader is only aware of the values of $z(x^1;\boldsymbol{c})$ and $\tilde{w}(x^1;\boldsymbol{c})$). In this case, the set $\mathcal{C}^1$ induced by the Cvx update in equation (5), corresponds to the cost vectors that make the shortest path in the network after (1,2) is interdicted to be greater than or equal to 3. In other words, $\mathcal{C}^1$ reduces to

$$\mathcal{C}^1 = \{\hat{\boldsymbol{c}}\colon \hat{c}_{1j} \geq 2,\ j = 3,\ldots,6\},$$

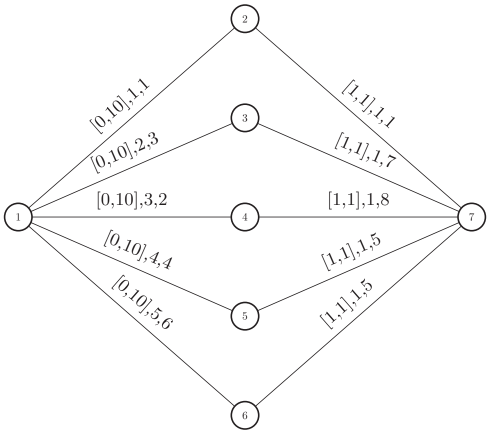


Figure 1 Instance that illustrates the different update mechanisms. We assume that $k=1$. The arcs' labels are given by $[\ell_a,u_a],c_a,d_a$.

or, equivalently $\mathcal{U}^2=[0,10]\times[2,10]^4$. On the other hand, to compute the NCvx update, consider first the set $E(x^1,10)$. This set consists of the paths that remain after (1,2) is interdicted that have an upper-level cost $\boldsymbol{b}^\top x^1+\boldsymbol{d}^\top y^1$ is equal to 10. Because $\boldsymbol{b}=0$, it is readily seen that these paths are 1-3-7 and 1-4-7, i.e., abusing the notation $E(x^1,10)=\{1-3-7,1-4-7\}$. Now, for path 1-3-7, $\{\hat{\boldsymbol{c}}\colon \hat{\boldsymbol{c}}^\top y=z(x^t;\boldsymbol{c}),\, z(x^t;\hat{\boldsymbol{c}})=z(x^t;\boldsymbol{c})\}$ is the set of cost vectors that make 1-3-7 a shortest path of length 3 in the network that remains after (1,2) is interdicted, i.e., this set is $\{\hat{\boldsymbol{c}}\colon \hat{c}_{13}=2, c_{1j}\geq 2, j=4,5,6\}$. Performing a similar analysis to 1-4-7, gives that the NCvx update (6) reduces to

$$\mathcal{N}^1=\{\hat{\boldsymbol{c}}\colon \hat{c}_{13}=2,c_{1j}\geq 2, j=4,5,6\}\cup\{\hat{\boldsymbol{c}}\colon \hat{c}_{14}=2,c_{1j}\geq 2,j=3,5,6\},$$

and thus, $\mathcal{U}^2=\big([0,10]\times\{2\}\times[2,10]^3\big)\cup\big([0,10]\times[2,10]\times\{2\}\times[2,10]^2\big)$.

Finally, the full update is very similar to the NCvx update, with the addition that for each path $y\in E(x^1,10)$ there should not be a $c\in\{\hat{\boldsymbol{c}}\colon \hat{\boldsymbol{c}}^\top y=z(x^t;\boldsymbol{c}),\, z(x^t;\hat{\boldsymbol{c}})=z(x^t;\boldsymbol{c})\}$ for which there exist a shortest path under c with an upper-level value greater than 10. For instance, for the case of 1-3-7, $\{\hat{\boldsymbol{c}}\colon \hat{c}_{13}=2,c_{1j}\geq 2,j=4,5,6\}$ should not include any vector with $c_{16}=2$, as any such vector would make 1-6-7 a shortest path with an upper-level value of 11. Repeating this analysis for 1-4-7 gives that

$$\mathcal{F}^1=\{\hat{\boldsymbol{c}}\colon \hat{c}_{13}=2,c_{14}\geq 2,c_{15}\geq 2,c_{16}>2\}\cup\{\hat{\boldsymbol{c}}\colon \hat{c}_{14}=2,c_{13}\geq 2,c_{15}\geq 2,c_{16}>2\}.$$

In this case, $\mathcal{U}^2$ becomes the non-closed and non-convex set $\mathcal{U}^2=\big([0,10]\times\{2\}\times[2,10]^2\times(2,10]\big)\cup\big([0,10]\times[2,10]\times\{2\}\times[2,10]\times(2,10]\big)$. ■

In addition to the update mechanisms described above for the case of Standard feedback, the Leader can also implement additional updates when she has access to Value-Perfect or Response-Perfect feedback.

Response-Perfect update. Suppose that the Leader has access to Response-Perfect feedback (i.e., in addition to standard feedback, the Leader observes the Follower's response), and consider the update in which the Leader, in addition to potentially using the information in one of the updates listed for the case of standard feedback (encoded in M^t) uses the fact that y^t is the Follower's rational response, i.e., $U(\mathcal{U}^t, \mathcal{K}^t) = \mathcal{U}^t \cap M^t \cap \mathcal{R}^t$, where

$$\mathcal{R}^t := \{\hat{\boldsymbol{c}} \in \mathbb{R}^n : \hat{\boldsymbol{c}}^\top y^t = z(x^t; \boldsymbol{c})\}$$

and $M^t \in \{\emptyset, \mathcal{C}^t, \mathcal{N}^t, \mathcal{F}^t\}$. Note that the set $\mathcal{R}^t$ is an hyper-plane.

Value-Perfect update. Suppose that the Leader has access to Value-Perfect feedback (i.e. in addition to standard feedback, the Leader observes the some components of $\boldsymbol{c}$), and consider the update in which the Leader, in addition to potentially using the information in one of the updates listed for the case of standard feedback (encoded in M^t) uses the fact that some components of $\boldsymbol{c}$ are observed, $U(\mathcal{U}^t, \mathcal{K}^t) = \mathcal{U}^t \cap M^t \cap \mathcal{V}^t$, where

$$\mathcal{V}^t := \left\{\hat{\boldsymbol{c}} \in \mathbb{R}^n : \hat{c}_a = c_a \ \forall a \in A \text{ s.t. } y_a^t > 0\right\}$$

and $M^t \in \{\emptyset, \mathcal{C}^t, \mathcal{N}^t, \mathcal{F}^t\}$. Note that the set $\mathcal{V}^t$ is a polyhedron. Hereafter, we assume that the Leader implements a Value-Perfect/Response-Perfect update whenever the feedback is Value-Perfect/Response-Perfect. Regarding the standard-feedback component of an update mechanism (i.e., the choice of M^t above), we say the mechanism is *strong* whenever $M^t \neq \mathcal{C}^t$.

Remark 3. In the above, we include the possibility of selecting $M^t = \emptyset$, as in the analysis in Borrero et al. (2019).

3. Greedy and Robust Policies

The framework developed so far builds largely on that presented by Borrero et al. (2019), which studies sequential max-min bilevel interdiction problems. In particular, if one constrains uncertainty in such a framework to be restricted to objective function coefficients, one recovers a variation of the problem studied in our work, in the special case when $\boldsymbol{b} = 0$ and $\boldsymbol{d} = \boldsymbol{c}$ (the only difference would be that both $\boldsymbol{d}$ and $\boldsymbol{c}$ would be initially unknown by the Leader).

In the max-min context, Borrero et al. (2019) introduce a family of *greedy and robust* policies. One possible interpretation of the Leader's rationale behind such policies is the following (we will examine other in the following section): suppose that in period $t \in \mathcal{T}$, the Leader chooses x^t as her action; then she anticipates that the Follower will choose his response expecting that "nature" would ultimately choose a vector $\boldsymbol{c}$ from $\mathcal{U}^t$ so as to *damage him* as much as possible, hence the *robust* moniker. (Regarding the specification of the update mechanism, only the case of Value-Perfect or Response-Perfect feedback is considered, with the update being that of the previous section, using $M^t = \emptyset$.)

Remark 4. Note that this is only a construct that the Leader uses to rationalize the policy: recall that the Follower actually knows the value of $\boldsymbol{c}$, thus he *does not* actually use a *robust approach* to decision making; it is the Leader's uncertainty with regard to $\boldsymbol{c}$ that is somewhat projected into her perception of the Follower's decision-making process. ■

Imagining this interaction between the Follower and nature, the Leader then acts greedily, and chooses x^t so as to maximize period t's profit, hence the *greedy* moniker. Using this interpretation in the context of **SBPI**, one can adapt this set of policies (further denoted by Λ) for our asymmetric setting as follows. For $x \in X$ and a set $\mathcal{U}$, define

$$Z_{GR}(x;\mathcal{U}) := \arg\min\big\{\max\{\hat{\boldsymbol{c}}^\top y : \hat{\boldsymbol{c}} \in \mathcal{U}\} : y \in Y(x)\big\},$$

and $z_{GR}(x;\mathcal{U}) := \min\big\{\max\{\hat{\boldsymbol{c}}^\top y : \hat{\boldsymbol{c}} \in \mathcal{U}\} : y \in Y(x)\big\}$. We say $\lambda \in \Lambda$ if and only if

$$x^{t,\lambda} \in \arg\max\Big\{\boldsymbol{b}^\top x + \boldsymbol{d}^\top y : y \in Z_{GR}(x;\mathcal{U}^t),\, x \in X\Big\}. \tag{8}$$

Because the Follower's actions actually depend on the true value of $\boldsymbol{c}$, there will be in general a disconnect between the feedback that the Leader expects to observe (if the Follower used a robust approach) and what it is actually observed. Consider the case of Standard feedback: when using policy $\lambda \in \Lambda$, the Leader *expects* her profit at period t to be $\tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$, where

$$\tilde{w}_{GR}(x;\mathcal{U}) := \boldsymbol{b}^\top x + \max\left\{\boldsymbol{d}^\top y : y \in Z_{GR}(x;\mathcal{U})\right\},$$

and the Follower's cost to be $z_{GR}(x^{t,\lambda};\mathcal{U}^t)$.

The next lemma shows that, like in the max-min context, whenever the Leader's expectations about her own profit are different from what she observes (i.e . $\tilde{w}(x^{t,\lambda};\boldsymbol{c}) \neq \tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$, then the Follower must reveal new information to the Leader, under either Value-Perfect or Response-Perfect. In particular, the dimension of the uncertainty set containing $\boldsymbol{c}$ must decrease.

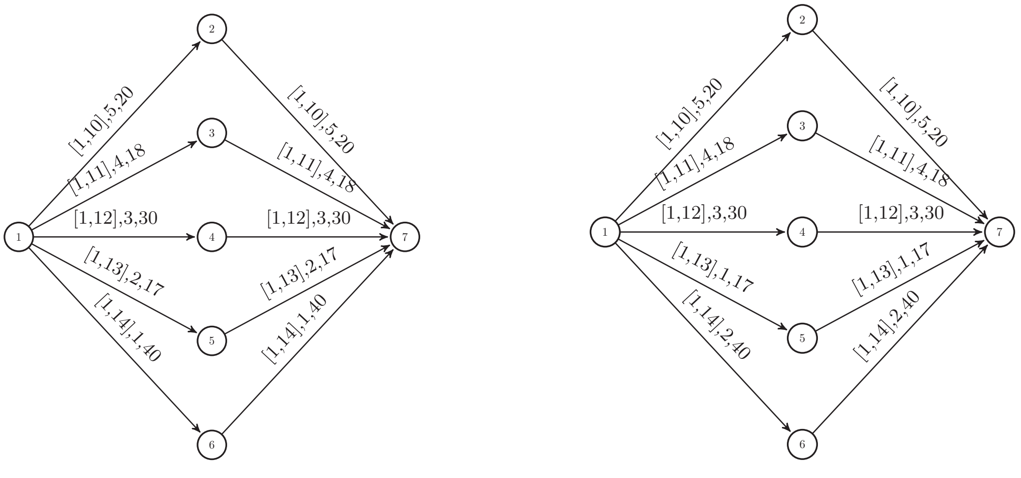
Lemma 1. *Suppose that $\lambda \in \Lambda$, that Standard Feedback is Value-Perfect or Response-Perfect, and that the Leader uses the Value-Perfect or Response-Perfect update mechanism with $M^t = \emptyset$, respectively. If $\tilde{w}(x^{t,\lambda};\boldsymbol{c}) \neq \tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$, then* $\dim(\mathcal{U}^{t+1}) < \dim(\mathcal{U}^t)$.

Remark 5. Note that the proof on Lemma 1 does not use the fact that the λ policies are greedy. Hence, the lemma holds for a broader class of policies. ■

Although policies in Λ assure new information is learned when expectations are not met, they do not assure that an optimal solution has been found when expectations are met. This behavior follows from the fact that, in contrast with the max-min setting (see Theorem 1 of Borrero et al. (2019)), in **SBPI** the expected profit $\tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$ is not a valid lower or upper bound for neither

$\tilde{w}(x^{t,\lambda};\boldsymbol{c})$ nor $\tilde{w}^*(\boldsymbol{c})$. Indeed, as we show in the next example, it is possible to come up with settings where $\tilde{w}(x^{t,\lambda};\boldsymbol{c})$, $\tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$ and $\tilde{w}^*(\boldsymbol{c})$ are arbitrarily ordered (with the exception that $\tilde{w}^*(\boldsymbol{c}) \geq \tilde{w}(x^{t,\lambda};\boldsymbol{c})$, which must always hold as per the optimality of the Follower's reaction).

Example 2. Next, we give various counterexamples that show that greedy and robust policies do not yield valid bounds for $\tilde{w}^*(\boldsymbol{c})$ in **SBPI**. First, in Figure 2a we show an example of an instance of the ASPI where $\tilde{w}(x^{t,\lambda};\boldsymbol{c}) > \tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$ and $\tilde{w}^*(\boldsymbol{c}) > \tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$ (recall that $\mathbf{b}=0$ in ASPI). Here (and in the remaining counterexamples), the objective of the Follower is to move between nodes 1 and 7, and we consider $k=2$. Observe that if the Leader is deciding robustly, then she assumes that the cost that the Follower incurs by traversing an arc is given by u_a. Hence the solution for any policy $\lambda \in \Lambda$ is to block arcs $(1,2)$ and $(1,3)$ (or more generally to block the two upper-most paths). Given this, the Leader expects that the Follower uses path 1–4–7, and hence she expects a profit of $\tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)=60$. However, observe that if the Leader blocks $(1,2)$ and $(1,3)$, then the path that the Follower uses is 1–6–7, which gives a profit of $\tilde{w}(x^{t,\lambda};\boldsymbol{c})=80$ to the Leader. Observe moreover, that $\tilde{w}^*(\boldsymbol{c})=\tilde{w}(x^{t,\lambda};\boldsymbol{c})$, so this example also shows that $\tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t) < \tilde{w}^*(\boldsymbol{c})$.



(a) (b)

Figure 2 Example of instances when (a) $\tilde{w}(x^{t,\lambda};\boldsymbol{c}) > \tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$ and (b) $\tilde{w}(x^{t,\lambda};\boldsymbol{c}) < \tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$. The arcs' labels are given by $[\ell_a, u_a], c_a, d_a$.

Conversely, Figure 2b shows an example of an instance of the ASPI where $\tilde{w}(x^{t,\lambda};\boldsymbol{c}) < \tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$. Here, the solution for any policy $\lambda \in \Lambda$ is to block again arcs $(1,2)$ and $(1,3)$, and as before, the Leader expects that the Follower uses path 1–4–7. This yields an expected profit of $\tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)=60$. However, observe that if the Leader blocks $(1,2)$ and $(1,3)$ then the path that the Follower uses is 1–5–7, which gives a profit of $\tilde{w}(x^{t,\lambda};\boldsymbol{c})=34$ to the Leader.

Finally, in Appendix C we provide additional examples (i) with $\tilde{w}^*(\boldsymbol{c}) < \tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$, and (ii) where the fact that $\tilde{w}(x^{t,\lambda};\boldsymbol{c}) = \tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$ does not imply that $\tilde{w}(x^{t,\lambda};\boldsymbol{c}) = \tilde{w}^*(\boldsymbol{c})$. ■

Following the discussion, one can also show that the fact that the expected cost for the Follower matches the observed cost (i.e., $z(x^{t,\lambda};\boldsymbol{c}) = z_{GR}(x^{t,\lambda};\mathcal{U}^t)$), does not imply that decision $x^{t,\lambda}$ is optimal. Moreover, when $z(x^{t,\lambda};\boldsymbol{c}) = z_{GR}(x^{t,\lambda};\mathcal{U}^t)$, the solution used by the Follower might not reveal any new information, and thus, the Leader might not learn anything; furthermore, at time $t+1$ the decision of time t is going to be repeated. In other words, policies in Λ might *stall*, i.e., might repeat a sub-optimal solution in all time periods indefinitely without forcing the Follower to reveal any new information. The next example illustrates these facts.

Example 3. Figure 3 shows an example in the ASPI where the fact that $z(x^{t,\lambda};\boldsymbol{c}) = z_{GR}(x^{t,\lambda};\mathcal{U}^t)$ does not imply that $\tilde{w}(x^{t,\lambda};\boldsymbol{c}) = \tilde{w}^*(\boldsymbol{c})$. Observe that any policy in λ blocks at least one arc among (1,3) and (3,7) but keeps path 1–4–7 unblocked, and hence, the Leader expects that the Follower uses path 1–4–7 at a cost of $z_{GR}(x^{t,\lambda};\mathcal{U}^t) = 14$. Given the real costs of arcs (1,4) and (4,7), then the Follower uses the same path 1–4–7 and incurs in the same cost of $z(x^{t,\lambda};\boldsymbol{c}) = 14$, which gives the Leader a profit of $\tilde{w}(x^{t,\lambda};\boldsymbol{c}) = 6$. However, it is seen that an optimal full-information solution is to remove (1,3) and (1,4), forcing the Follower to use path 1–5–7, and giving the Leader a profit of $\tilde{w}^*(\boldsymbol{c}) = 8$. ■

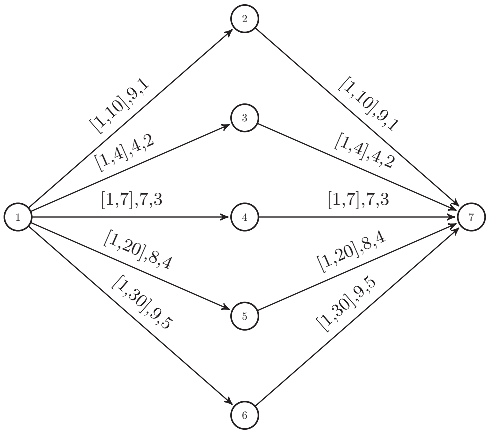


Figure 3 Example of an instance when $z(x^{t,\lambda};\boldsymbol{c}) = z_{GR}(x^{t,\lambda};\mathcal{U}^t)$ does not imply that $\tilde{w}(x^{t,\lambda};\boldsymbol{c}) = \tilde{w}^*(\boldsymbol{c})$. The arcs' labels are given by $[\ell_a, u_a], c_a, d_a$.

The above examples imply that nothing can be said in general about the optimality of $x^{t,\lambda}$ whenever the observed feedback matches that expected by the Leader (when she assumes the Follower uses a robust approach). This behavior is very troublesome as it implies that a policy $\lambda \in \Lambda$ might not find the optimal solution, and might not provide a certificate of optimality in real time. Thus, greedy and robust policies as in (8) might fail to converge to an optimal decision in **SBPI**.

Next, we consider an alternative interpretation of the rationale behind these policies in the **SBPI** setting, and show that such an interpretation leads to a new class of policies, that do converge to optimal decisions, and that provide optimality certificates in real time.

4. Greedy and Best–Case Policies

In the previous section, we show how Greedy and Robust policies arise in the max-min setting when the Leader *i*) acts greedily, and *ii*) assumes that the Follower is also unaware of the value of $\boldsymbol{c}$, and selects his response assuming a robust approach. In this section we consider an alternative interpretation, emanating from the following observation: in the max-min context, a worst-case cost realization for the Follower corresponds to a best-case realization for the Leader. Thus, we adapt the policies from the max-min context to the **SBPI** context by assuming that the Leader *i*) still acts greedily, but *ii*) assumes that the Follower is also unaware of the value of $\boldsymbol{c}$, and selects his response assuming that the actual realization will favor the Leader (i.e., it is the Leader who ultimately chooses $\boldsymbol{c}$, as opposed to nature).

The interpretation above gives rise to a class of *Greedy and Best-Case* policies, which we denote by Ψ. Thus, we say $\psi \in \Psi$ if before attaining time stability (we provide a formal definition below), one has that

$$(x^{t,\psi}, \cdot) \in \arg\max\Big\{\boldsymbol{b}^\top x + v(\hat{\boldsymbol{c}}, x),\, x \in X,\, \hat{\boldsymbol{c}} \in \mathcal{U}^t\Big\}. \tag{9}$$

(Note that nothing is said with regard to the update mechanism used.) In what follows we study the conditions under which policies in Ψ provide certificates of optimality in real time and accept finite upper bounds on time stability. Although the aforementioned upper-bounds are worst-case exponential, we demonstrate in the experiments of Section 5 that the time stability of these policies is typically much lower than the upper bounds. In addition, we show that these policies can handle a broader class of uncertainty sets, without compromising the tractability of the model. For this, we provide tractable mixed–integer programming formulations of (9), even when the initial uncertainty set is non-convex. This flexibility contrasts with robust (i.e., worst-case) approaches, where uncertainty sets are assumed to be convex (Ben-Tal et al. 2009).

4.1. Definitions

For a given set $\mathcal{U}$, define the following mathematical program,

$$\tilde{w}_E(\mathcal{U}) = \max_{x,y,\hat{\boldsymbol{c}}} \boldsymbol{b}^\top x + \boldsymbol{d}^\top y \tag{10a}$$

$$\text{s.t. } x \in X \tag{10b}$$

$$\hat{\boldsymbol{c}} \in \mathcal{U} \tag{10c}$$

$$y \in Z(x; \hat{\boldsymbol{c}}), \tag{10d}$$

and let $\mathcal{S}(\mathcal{U})$ be the set of optimal solutions, i.e., $\mathcal{S}(\mathcal{U}) := \arg\max\{\boldsymbol{b}^\top x + \boldsymbol{d}^\top y\colon$ (10b)–(10d) hold$\}$. For $t \in \mathcal{T}$ define $\bar{w}^t$ as the largest profit observed up to period t, i.e.

$$\bar{w}^t := \max\{\tilde{w}(x^s; \boldsymbol{c})\colon s \le t\}.$$

We define $\xi \in \mathcal{T}$ as the first time t in which the best observed profit so far $\bar{w}^t$ coincides with the Leader's "expectation", $\tilde{w}_E(\mathcal{U}^t)$. In other words,

$$\xi := \inf\{t \in \mathcal{T} \colon \bar{w}^t \geq \tilde{w}_E(\mathcal{U}^t)\}.$$

Also, let $s(\xi) \in \arg\max\{\tilde{w}(x^t; \boldsymbol{c}) : t \leq \xi\}$ denote a time period (prior to or equal to ξ) at which profit $\bar{w}^\xi$ is observed.

Definition 1. We say that $\psi \in \Psi$ if and only if there exist vectors $y^{t,E}$ and $\boldsymbol{c}^{t,E}$ such that $(x^{t,\psi}, y^{t,E}, \boldsymbol{c}^{t,E}) \in \mathcal{S}(\mathcal{U}^t)$ for all $t \leq \xi$, and $x^{t,\psi} = x^{s(\xi),\psi}$ for all $t > \xi$.

Note that $(x^{t,\psi}, y^{t,E}, \boldsymbol{c}^{t,E}) \in \mathcal{S}(\mathcal{U}^t)$ is such that $\tilde{w}_E(\mathcal{U}^t) = \boldsymbol{b}^\top x^{t,\psi} + \boldsymbol{d}^\top y^{t,E}$; we call $y^{t,E}$ the response that the Leader *expects* the Follower will use at time t, and $\boldsymbol{c}^{t,E}$ the *expected* cost vector at time t. Using this, we define $z^{t,E} := (\boldsymbol{c}^{t,E})^\top y^{t,E}$ as the cost the Leader expects the Follower to incur at time t.

Remark 6. Note that $\mathcal{S}(\mathcal{U}^t)$ is not necessarily a singleton; in such a case, a policy ψ is associated with a particular rule used to select a solution among those in $\mathcal{S}(\mathcal{U}^t)$, and this includes both the expected response and cost vector selected. Thus, any two policies ψ and ψ' selecting the same decision $x^{t,\psi} = x^{t,\psi'}$ in period t but expecting, for example, different Follower's responses are deemed as different policies. ■

The greedy and "best-case" nature of the policies in Ψ is in display in formulation (10), where we see that the Leader is *greedily* optimizing a single-period profit, but assumes a *best-case* realization for the cost vector. In this regard, the expected response by the Follower is only constrained to be a best-response to the Leader's decision, assuming that the cost vector is selected by the Leader. Note that the policies in Ψ follow such a pattern until period ξ, the point at which the Leader knows how to produce a profit not lower than that under her rather optimistic approach, thus she proceeds to implement the best decision tried so far until the end of the horizon.

The next result shows that, for policies in Ψ, the observed and expected profits are valid lower and upper bounds to the optimal single-period profit, respectively.

Theorem 1. *For any policy $\psi \in \Psi$ and under Standard Feedback, one has that $\tilde{w}(x^{t,\psi}; \boldsymbol{c}) \leq \tilde{w}^*(\boldsymbol{c}) \leq \tilde{w}_E(\mathcal{U}^t)$, for $t \in \mathcal{T}$. In particular, if for some period $t \in \mathcal{T}$ one has that $w_E(\mathcal{U}^t) \leq \bar{w}^t$, then $x^{s(\xi),\psi}$ is an optimal solution to the full-information problem, i.e. $\tilde{w}(x^{s(\xi),\psi}; \boldsymbol{c}) = \tilde{w}^*(\boldsymbol{c})$.*

Theorem 1 can be viewed as the analogous to Theorem 1 in the work by Borrero et al. (2019) in the max-min setting. Importantly, it implies that by using policies in Ψ the Leader can get a certificate of optimality in real time. To see this, note that under Standard Feedback, the Leader observes $\bar{w}^t$ for all $t \in \mathcal{T}$. Hence, she can verify in real time whether $\xi = t$, the moment at which Theorem 1 ensures $x^{s(\xi),\psi}$ is the full-information solution.

Remark 7. The multistage version of the greedy and best case policies would seek to solve

$$\max_{x^1,x^2,\dots,x^T\in X}\Big(\boldsymbol{b}^\top x^1+\max_{c^1\in\mathcal{U},y_1\in Z(c^1,x^1)}\Big\{\boldsymbol{d}^\top y^1+\boldsymbol{b}^\top x^2+\max_{c^2\in\mathcal{U}(x^1,y^1),y^2\in Z(c^2,x^2)}\Big\{\boldsymbol{d}^\top y^2+ \tag{11a}$$

$$\boldsymbol{b}^\top x^3+\max_{c^3\in\mathcal{U}(x^{[2]},y^{[2]}),y^3\in Z(c^3,x^3)}\Big\{\boldsymbol{d}^\top y^3+\dots\dots\Big\}\Big\}\Big\}\Big), \tag{11b}$$

where for any $t\in\mathcal{T}$ we define $x^{[t]}=(x^1,\dots,x^t)$; $y^{[t]}$ is defined in a similar way. Note that in order to include learning into the optimization, the uncertainty set at stage t has to depend on $x^{[t]}$ and $y^{[t]}$; this dependence implies the 'nested' structure in (11). Clearly, such an approach would generate upper bounds for $\tilde{w}^*(\boldsymbol{c})$ that are at most equal to $\tilde{w}_E(\mathcal{U}^t)$, and thus better, but this would come at the price of solving the intractable multistage bilevel problem in (11). An interesting relationship of (11) with adaptive multistage robust optimization (Bertsimas and Georghiou 2015, Bertsimas and Dunning 2016, Lorca et al. 2016) happens if the Leader takes a worst-case approach to uncertainty. Such worst-case policy would result in a problem of the form

$$\max_{x^1,x^2,\dots,x^T\in X}\Big(\boldsymbol{b}^\top x^1+\min_{c^1\in\mathcal{U}}\max_{y^1\in Z(c^1,x^1)}\Big\{\boldsymbol{d}^\top y^1+\boldsymbol{b}^\top x^2+\min_{c^2\in\mathcal{U}(x^1,y^1)}\max_{y^2\in Z(c^2,x^2)}\Big\{\boldsymbol{d}^\top y^2$$

$$\boldsymbol{b}^\top x^3+\min_{c^3\in\mathcal{U}(x^{[2]},y^{[2]})}\max_{y^3\in Z(c^3,x^3)}\Big\{\boldsymbol{d}^\top y^3+\dots\dots\Big\}\Big\}\Big\}\Big),$$

which would correspond to a general class of bilevel adaptive robust problems where the uncertainty set also depends on past decisions. The single-stage version of this policy, at any time $t\in\mathcal{T}$, would assume that $\hat{\boldsymbol{c}}$ realizes the value that results in the lowest possible value for $\boldsymbol{d}^\top y^t$. From the standpoint of learning, it can be shown that the adaptive robust policy does not provide valid upper bounds to $\tilde{w}^*(\boldsymbol{c})$ in general in the context of Theorem 1, which implies that it cannot provide certificates of optimality in real time. ■

4.2. Convergence of Policies in Ψ

Theorem 1 states a condition under which a policy $\psi\in\Psi$ provides an optimal solution, but does not ensure that such a condition would be met within the time horizon. In this regard, note that the result is independent of the update mechanism used to refine the Leader's belief about $\boldsymbol{c}$. Thus, for example, if no update is ever made ($\mathcal{U}^t=\mathcal{U}^1$ for all $t\in\mathcal{T}$), then (unless $\tilde{w}(x^1;\boldsymbol{c})=\tilde{w}^*(\boldsymbol{c})$) the optimal single-period profit $w^*(\boldsymbol{c})$ would never be achieved and time stability would be unbounded.

Convergence guarantees for policies in Ψ (i.e., finite upper bounds to their time stability) are likely to depend on the update mechanism used. In the max-min context, Borrero et al. (2019) establish finite upper bounds (at most linear in the dimension of the Follower's cost vector) for policies in Λ, *provided* that the feedback is either Value-Perfect or Response-Perfect. Such a result relies on the fact that whenever the Leader's expectations do not match what is observed, the Follower is forced to implement new solutions, which, in turn, reduces the polyhedral dimension of the uncertainty set (under Value-Perfect or Response-Perfect updates). This observation provides a bound to the number of periods until the Leader expectations are met.

In our setting, depending on the feedback available and the update mechanism used, policies in Ψ might not guarantee that the *size* of the uncertainty region decreases (which implies that the arguments in extant work do not apply). In particular, when using policies in Ψ under Standard Feedback, we have that the polyhedral dimension of the uncertainty set might not be reduced. Moreover, this fact, which is illustrated in the next example, continues to hold when, in addition, feedback is either Value-Perfect or Response-Perfect (and the update mechanism include Value-Perfect or Response-Perfect updates, respectively).

Example 4. In this example we show that even under the strongest possible update mechanisms (that is, full plus Value-Perfect or Response-Perfect updates), policies in Ψ might not force the Follower to explore a different solution each time. This observation implies that the dimension-reduction convergence arguments cannot be used to prove the convergence of policies in Ψ.

Consider the instance of ASPI in Figure 4 with $k = 2$, where the Leader does not know the costs of arcs (i, j), $i = 1$, $j = 2, \dots, 6$, with certainty, but knows the cost of all the other arcs with certainty. Observe that in this instance the optimal solution of the problem is $\tilde{w}^*(\boldsymbol{c}) = 8$, associated with blocking arcs $\{(1,3), (1,4)\}$.

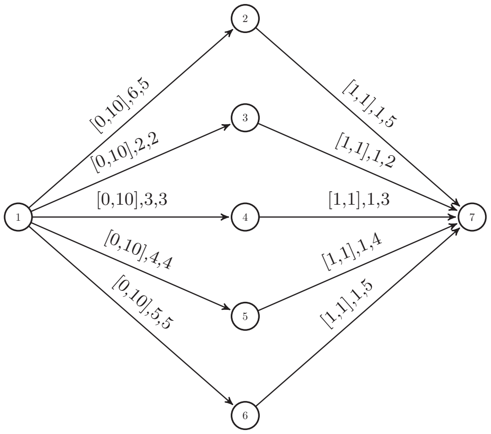


Figure 4 Instance for Example 4. We assume that $k = 2$. The arcs' labels are given by $[\ell_a, u_a], c_a, d_a$.

A possible action for policies in Ψ at period $t = 1$ is to interdict $x^1 = \{(1,2), (1,3)\}$, assuming that $c_{14}^{1,E} = c_{15}^{1,E} = 10$ and $c_{16}^{1,E} = 10$. This solution expects $y^{1,E}$ to be the path 1–6–7 and gives $\tilde{w}_E(\mathcal{U}^1) = 10$. The solution of the Follower is for y^1 to be 1–4–7, which gives $\tilde{w}(x^1; \boldsymbol{c}) = 6$. Assuming Value-Perfect feedback, the Leader updates the uncertainty set by jointly using the Value-Perfect and the full update, thus

$$\mathcal{U}^2 = \{\hat{c} \in [0,10]^5 : \hat{c}_{14} = 3, \hat{c}_{15} > 3, \hat{c}_{16} > 3\}.$$

At time $t=2$ an optimal solution for the Leader is to interdict $x^2=\{(1,3),(1,5)\}$, assuming that $c_{12}^{2,E}=0$, $c_{14}^{2,E}=3$, and $c_{16}^{2,E}=10$. This solution expects $y^{2,E}$ to be the path 1–2–7 and gives $\tilde{w}_E(\mathcal{U}^2)=10$. Note that the solution of the Follower is again to set y^2 equal to 1–4–7, which again gives $\tilde{w}(x^2;\boldsymbol{c})=6$. Hence, at time $t=2$ the Follower has not revealed any new solution to the Leader, and moreover, using the full update gives

$$\mathcal{U}^3=\{\hat{c}\in[0,10]^5\colon c_{14}=3, c_{12}>3, c_{15}>3, c_{16}>3\}.$$

Therefore, under the Value-Perfect and full update, the polyhedral dimension of $\mathcal{U}^3$ is the same as the dimension of $\mathcal{U}^2$. Finally, note that the same exact sequence of actions would happen if instead of Value-Perfect feedback the Leader has access to Response-Perfect feedback and uses the Response-Perfect update mechanism. ■

Despite the fact that the polyhedral dimension of the uncertainty set does not decrease in the previous example, the uncertainty itself shrinks, which might result in the Leader implementing a different solution at the next period. One might hope that such a variability on the Leader's actions might lead to convergence to an optimal solution (as the Leader's expectations are changing, so there is a chance that such expectations are met). Unfortunately, the next example shows that if the full update is not used, then decisions made by policies in Ψ might stall even when feedback is Value-Perfect and Response-Perfect.

Example 5. In this example we show that policies in Ψ might stall if the full update is not used. To this end, consider the same instance of Example 4 in Figure 4. Assume that the Leader has access to Value-Perfect feedback and that she only uses the Value-Perfect mechanism. Here, at time $t=2$ the uncertainty set becomes

$$\mathcal{U}^2=\{\hat{c}\in[0,10]^5\colon c_{14}=3\}.$$

In this case, at time $t=2$ an optimal solution for the Leader is $x^2=\{(1,2),(1,3)\}$, as she can assume that $c_{14}^{1,E}=3$, $c_{15}^{1,E}=10$ and $c_{16}^{1,E}=0$. Thus, $y^{2,E}$ is the path 1–6–7 which gives $\tilde{w}_E(\mathcal{U}^2)=10$. Clearly, at time $t=2$ the Follower will repeat the solution used at time $t=1$ (i.e., $w^2=6$) and therefore $\mathcal{U}^3=\mathcal{U}^2$. Moreover, in this case the policy stalls because $\mathcal{U}^3=\mathcal{U}^2$, and it will indefinitely repeat the same suboptimal solution across all time periods. Importantly, note that the same exact sequence of actions would happen above, if instead of Value-Perfect feedback the Leader has access to Response-Perfect feedback and uses the Response-Perfect update mechanism.

On the other hand, suppose that the Leader has only access to Standard feedback and that she uses the NCvx update. Under this assumption, $\mathcal{U}^2$ is as in the full update in Example 4, replacing $>$ with $\geq$. In this case, a solution for ψ can be given by $x^2=\{(1,2),(1,3)\}$ (i.e., $x^2=x^1$) by setting

$c_{14}^{2,E} = c_{16}^{2,E} = 3$ and $c_{15}^{2,E} = 10$, which expects $y^{2,E}$ =1–6–7 and $\tilde{w}_E(\mathcal{U}^2) = 10$. Clearly, as this solution repeats the actions implemented at $t = 1$, no new information will be learned by the Leader and the policy will stall. Moreover, this will be also be the case if the Leader uses jointly the NCvx and the Value-Perfect or Response-Perfect updates. Finally, in the Cvx update $\mathcal{U}^2 = \{\hat{c}\colon\, c_{1j} \geq 3,\ j = 4, 5, 6\}$, in which case the policy might stall at time $t = 2$ in a similar way as explained before. ■

The examples above show that convergence of policies in Ψ cannot be guaranteed in general when the full update mechanism is not used. Fortunately, it is possible to prove general convergence results for policies in Ψ when the full update mechanism is used. For this, consider the following equivalence relation between the elements of the upper-level solution set X.

We say $x \in X$ and $x' \in X$ are equivalent (written $x \sim x'$) if and only if $Y(x) = Y(x')$. It is readily seen that $\sim$ is an equivalence relation, and therefore it induces a partition of X into equivalence classes. For $x \in X$ we let $[x] := \{x' \in X : x \sim x'\}$ denote the equivalent class to which x belongs. We have the following result:

LEMMA 2. *Let $\psi \in \Psi$ and $t \in \mathcal{T}$ be given and assume that the Leader implements the full update mechanism. If $x^{t,\psi} \in \bigcup_{s<t} [x^{s,\psi}]$, then $\tilde{w}_E(\mathcal{U}^t) = \tilde{w}(x^{t,\psi}; \boldsymbol{c})$.*

REMARK 8. In the proof of Lemma 2, it is not sufficient to assume the NCvx update. To see this, note that under such an update, it is possible that $\boldsymbol{c}^{t,E} \in \mathcal{U}^t$ is such that there exist two vectors $y_1, y_2 \in \arg\min\{(\boldsymbol{c}^{t,E})^\top y' \colon\, y' \in Y(x^{s,\psi})\}$ such that $\boldsymbol{d}^\top y_1 = w^{s,\psi} - \boldsymbol{b}^\top x^{s,\psi}$ but with $\boldsymbol{d}^\top y_2 > w^{s,\psi} - \boldsymbol{b}^\top x^{s,\psi}$. In such a case, the optimality of $x^{t,\psi}$ would imply that $y^{t,E} = y_2$ and hence the conclusion of the Lemma would fail to hold. ■

When coupled with Theorem 1, Lemma 2 states that convergence to the full-information solution is guaranteed when implementing a decision belonging to the equivalence class of any solution implemented before (if $x^{t,\psi} \in \bigcup_{s<t} [x^{s,\psi}]$, then from Lemma 2 we have that $\tilde{w}_E(\mathcal{U}^t) = \tilde{w}(x^{t,\psi}; \boldsymbol{c})$ and thus, from Theorem 1, $\tilde{w}(x^{t,\psi}; \boldsymbol{c}) = \tilde{w}^*(\boldsymbol{c})$). The next result, which follows directly from this observation (and therefore, we state without proof), provides a bound for the time stability of the policies in Ψ.

THEOREM 2. *Let $\psi \in \Psi$ be given. If the Leader implements the full update mechanism, then $\tau^\psi \leq |\{[x] : x \in X\}|$.*

Theorem 2 states that the time stability is bounded by the size of the partition induced by $\sim$. Said upper bound can be exponential in the worst-case. Moreover, as per Example 4, the addition of Value-Perfect or Response-Perfect might not help. However, Theorem 2 states that in settings where the set of equivalent interdiction solutions is finite, then the proposed policies will not stall. In practice, as illustrated in the numerical experiments in Section 5, time stability under the full update mechanisms exhibits a linear behavior, which is far less than suggested by the upper-bound; furthermore, these results are significantly improved by the addition of Value-Perfect or Response-Perfect feedback.

4.3. α-optimal Greedy and Best-Case Policies

Implementing policies in Ψ requires solving formulation (10) with an increasingly more complex characterization of the set $\mathcal{U}^t$, for each update mechanism. For this reason, the amount of resources required to compute these policies might not scale well, as the size of the instances grow. In this section, we alleviate this scalability issue, by exploring the use of *approximate optimal solutions.*

Given a real number $\alpha \geq 1$, we say that $x \in X$ is an α-optimal solution to the bilevel problem (4) if it satisfies that $\alpha\ \tilde{w}(x;\boldsymbol{c}) \geq \tilde{w}^*(\boldsymbol{c})$; we assume both $\tilde{w}(x;\boldsymbol{c})$ and $w^*(\boldsymbol{c})$ are non-negative. We can extend the definition of the policies in Ψ to account for α-optimal solutions. To this end, let us denote by $\mathcal{S}_\alpha(\mathcal{U})$ the set of α-optimal solutions of problem (10), i.e.,

$$\mathcal{S}_\alpha(\mathcal{U}) := \left\{(x, y, \hat{\boldsymbol{c}}) : \alpha \max\left\{\boldsymbol{b}^\top x + \boldsymbol{d}^\top y,\, y \in Z(x;\hat{\boldsymbol{c}}),\, \hat{\boldsymbol{c}} \in \mathcal{U}\right\} \geq \tilde{w}_E(\mathcal{U}),\, x \in X\right\}.$$

Consider a policy that implements x^t such that exist $y^{t,\alpha}$ and $\boldsymbol{c}^{t,\alpha}$ so that $(x^t, y^{t,\alpha}, \boldsymbol{c}^{t,\alpha}) \in \mathcal{S}_\alpha(\mathcal{U}^t)$, for all $t \in \mathcal{T}$ and define

$$\xi_\alpha := \inf\{t \in \mathcal{T} : \bar{w}^t \geq \max\left\{\boldsymbol{b}^\top x^t + v(x^t;\hat{c}),\, \hat{c} \in \mathcal{U}^t\right\}\}.$$

Because $\tilde{w}_E(\mathcal{U}^t) \geq \tilde{w}^*(\boldsymbol{c})$, and at time ξ_α we have that $\alpha\,\bar{w}^t \geq \alpha\,\tilde{w}(x^t;\boldsymbol{c}) \geq w_E(\mathcal{U}^t)$, this is the first time we are sure that x^t is an α-optimal solution to (4). Let $s(\xi_\alpha) \leq \xi_\alpha$ denote the time period attaining the maximum $\bar{w}^{\xi_\alpha}$.

Definition 2. Let $\alpha \geq 1$ be given. We say that $\psi_\alpha \in \Psi_\alpha$ if and only if $x^{t,\psi_\alpha} \in \mathcal{S}_\alpha(\mathcal{U}^t)$ for all $t \leq \xi_\alpha$, and $x^{t,\psi_\alpha} = x^{s(\xi_\alpha),\psi_\alpha}$ for all $t > \xi_\alpha$.

The policies in Ψ_α operate as those in Ψ, but solving for α-optimal solutions to problem (10). Once an α-optimal solution has been found, it is repeated from there on. Note that, by construction, policies in Ψ_α provide certificates of α-optimality in real time ($t = \xi_\alpha$ can be checked in real time). For the same reason, it is not possible to prove finite bounds for the time stability of policies in Ψ_α, even under the full update mechanism. To see this, note that said policies can stall by implementing an α-optimal solution indefinitely. Nevertheless, it is possible to provide a finite upper bound on the number of periods until an α-optimal solution has been found.

For any policy π define its *α-time-stability*, $\tau_\alpha^\pi(\boldsymbol{c})$, as the first time period by which it can be assured that $\tilde{w}(x^t;\boldsymbol{c})$ is an α-optimal solution of (4) from there on, that is,

$$\tau_\alpha^\pi(\boldsymbol{c}) := \min\{t \in \mathcal{T} : \alpha\tilde{w}(x^s;\boldsymbol{c}) \geq \tilde{w}^*(\boldsymbol{c})\ \forall s \geq t\}.$$

Observe that for any $\psi_\alpha \in \Psi_\alpha$ it follows that $\tau_\alpha^{\psi_\alpha} \leq \xi_\alpha$. We have the following result.

Corollary 1. *Let $\alpha \geq 1$ and $\psi_\alpha \in \Psi_\alpha$ be given. If the Leader implements the full update mechanism, then $\tau_\alpha^{\psi_\alpha} \leq |\{[x] : x \in X\}|$.*

The proof of the result follows the same arguments in the proof of Theorem 2, and thus, it is omitted.

As a consequence of the above discussion, the Leader can use policies in Ψ_α as an alternative to the Greedy and Best-Case policies Ψ. These approximated policies inherit the most important properties of the policies in Ψ, namely, they ensure convergence to an α-optimal solution in finite time and provide a certificate of α-optimality in real time.

5. Computational Study

In this section we present a series of computational experiments designed to illustrate the performance of the Greedy and Best-Case policies in a set of instances of the Asymmetric Shortest Path Interdiction Bilevel Problem (ASPI), as described in Example 1. We begin by developing Mixed-Integer Programming (MIP) formulations to compute the policies in Ψ. We then show how said policies fare under the different update mechanisms, against the α-optimal policies, and against benchmark policies. Lastly, we perform a sensitivity analysis with respect to the quality of the initial information and the sizes of the instances.

5.1. MIP formulations of the Greedy and Best-Case Policies

This section presents MIP formulations for the policies in Ψ under all the update mechanisms. To make things concrete, we consider settings of ASPI, i.e., where the full-information problem (4) is an instance of ASPI. In this regard, we emphasize that the techniques used to derive these formulations (specifically, the use of the Follower's problem optimality conditions) can be adapted to broader classes of Follower's problems, such as the network flow problems or general linear programs; see the general approaches by Audet et al. (1997) and Zare et al. (2019).

The full-information formulation of ASPI (under the optimistic approach) is

$$\max \boldsymbol{d}^\top y \tag{13a}$$

$$\text{s.t. } \mathbf{1}^\top x = k \tag{13b}$$

$$y \in \arg\min\{\boldsymbol{c}^\top y' : \boldsymbol{M} y' = \boldsymbol{e}, y' \leq \mathbf{1} - x, y' \geq 0\} \tag{13c}$$

$$x \in \{0,1\}^{|A|}. \tag{13d}$$

In this formulation, matrix $\boldsymbol{M}$ is the node-arc adjacency matrix of a directed network $G = (N, A)$. We assume that node 1 is the source and node $m := |N|$ is the sink, thus $\boldsymbol{e} \in \mathbb{R}^m$ is such that $e_1 = -e_m = 1$, and $e_i = 0$ for $i \notin \{1, m\}$. Finally, $\mathbf{1}$ is a $|A| \times 1$ vector of ones. The Leader's decision variables x are binary, where $x_a = 1$ if and only if arc $a \in A$ is interdicted. In the sequel, following the work by Israeli and Wood (2002), we replace the optimality condition (13c) by an equivalent condition in the form:

$$y \in \arg\min\{(\boldsymbol{c} + Kx)^\top y' \colon \boldsymbol{M}y' = \boldsymbol{e}, y' \geq 0\}, \tag{14}$$

where K is a sufficiently large positive constant (note that in optimality $Kx^\top y = 0$). For a vector $c \in \mathbb{R}_+^{|A|}$, strong duality implies that y satisfies (14) if and only if there exists $\beta \in \mathbb{R}_+^m$ such that

$$\boldsymbol{M}y = \boldsymbol{e}, \ -\boldsymbol{M}^\top \beta \leq \boldsymbol{c} + Kx, \ \beta_m = \boldsymbol{c}^\top y, \ \beta_1 = 0. \tag{15}$$

This, because $\beta_m - \beta_1$ represents the dual objective function, which can be assumed non-negative as $c \in \mathbb{R}_+^{|A|}$; also, for any dual solution β, one can form an equivalent (non-negative) solution β' such that $\beta'_i = \beta_i - \beta_1$, thus one can set to $\beta_1 = 0$ without loss of generality. Therefore, a policy $\psi \in \Psi$ at time t can be computed by solving the equivalent of (10) for the ASPI setting, given by

$$w_E(\mathcal{U}) = \max_{x, \hat{\boldsymbol{c}}, y, \beta} \left\{ \boldsymbol{d}^\top y \colon \mathbf{1}^\top x = k, \ \hat{\boldsymbol{c}} \in \mathcal{U}, (y, \beta) \text{ satisfies } (15), x, y \in \{0,1\}^{|A|}, \beta \in \mathbb{R}_+^{|N|} \right\}. \tag{16}$$

Note that (16) is a nonlinear mixed-integer problem, because of the presence of the term $\beta_m = \boldsymbol{c}^\top y$ in (15) (even when $\mathcal{U}^t$ takes the form of a polyhedron). Because y is binary (this assumption can be made as the Follower solves the shortest path problem), the term $\boldsymbol{c}^\top y$ can be linearized using the McCormick envelopes (McCormick 1976). This gives us the following reformulation of (16):

$$\begin{aligned}
w_E(\mathcal{U}) = \max \ & \boldsymbol{d}^\top y && && \text{(17a)}\\
\text{s.t. } & \mathbf{1}^\top x = k && && \text{(17b)}\\
& \hat{\boldsymbol{c}} \in \mathcal{U} && && \text{(17c)}\\
& \boldsymbol{M}y = \boldsymbol{e} && && \text{(17d)}\\
& -\boldsymbol{M}^\top \beta - \hat{\boldsymbol{c}} - Kx \leq \mathbf{0} && && \text{(17e)}\\
& \beta_m - \mathbf{1}^\top q = 0 && && \text{(17f)}\\
& -\hat{c}_a + q_a \leq 0 && \forall a \in A && \text{(17g)}\\
& -u_a y_a + q_a \leq 0 && \forall a \in A && \text{(17h)}\\
& \hat{c}_a + u_a y_a - q_a \leq u_a && \forall a \in A && \text{(17i)}\\
& x, y \in \{0,1\}^{|A|}, \beta \in \mathbb{R}_+^{|N|}, q \in \mathbb{R}_+^{|A|}, && && \text{(17j)}
\end{aligned}$$

where u_a is a known upper bound on the value of c_a for any given $a \in A$. Thus, we have that if $\mathcal{U}^t$ is mixed-integer representable, then (17) is a linear mixed-integer problem. We show next that this is indeed the case when either the Cvx or NCvx updates are used. In the sequel, we assume that $\mathcal{U}^1 = \{\hat{c} : \ell_a \leq c_a \leq u_a, \, a \in A\}$.

Cvx update. Using the linear programming formulation of the Follower's (shortest path) problem, under the Cvx update we have that for $t>1$

$$\mathcal{U}^t = \left\{\hat{\boldsymbol{c}} \in \mathcal{U}^0 : \exists\, \beta^s \in \mathbb{R}_+^{|N|} \text{ s.t.} -\boldsymbol{M}^\top \beta^s - \hat{\boldsymbol{c}} \le K x^{s,\psi},\ \beta^s_m \ge z(x^{s,\psi};\boldsymbol{c}),\ \forall\, s<t\right\},$$

which is mixed-integer representable (note that in the above, β^s represents a dual solution to the problem in period s, where consistent with (5) we do not impose strong duality).

NCvx update. Under the NCvx update, we can use strong duality optimality to obtain the following representation: for $t>1$

$$\mathcal{U}^t = \{\hat{\boldsymbol{c}} \in \mathcal{U}^0 : \exists\, y^s \in \{0,1\}^{|A|}, \beta^s \in \mathbb{R}_+^{|N|}, q^s \in \mathbb{R}_+^{|A|}, s<t \text{ s.t. (18) holds}\},$$

with

$$\begin{aligned}
&\boldsymbol{M} y^s = \boldsymbol{e} && \forall s = 1,\dots,t && \text{(18a)}\\
&-\boldsymbol{M}^\top \beta^s - \hat{\boldsymbol{c}} \le K x^{s,\psi} && \forall s<t && \text{(18b)}\\
&\beta^s_m - \mathbf{1}^\top q^s = 0 && \forall s<t && \text{(18c)}\\
&\beta^s_m - z(x^s;\boldsymbol{c}) = 0 && \forall s<t && \text{(18d)}\\
&\boldsymbol{d}^\top y^s = \tilde{w}(x^{s,\psi};\boldsymbol{c}) && \forall s<t && \text{(18e)}\\
&-\hat{c}_a + q^s_a \le 0 && \forall s<t,\ a\in A && \text{(18f)}\\
&-u_a y^s_a + q^s_a \le 0 && \forall s<t,\ a\in A && \text{(18g)}\\
&\hat{c}_a + u_a y^s_a - q^s_a \le u_a && \forall s<t,\ a\in A, && \text{(18h)}
\end{aligned}$$

which is mixed-integer representable. In the above, constraints (18a) enforce primal feasibility, (18b) enforce dual feasibility, (18c)-(18d) and (18f)-(18h) enforce strong duality, and (18e) look to match observed profits.

Full update and enhanced NCvx (Cvx) update. The full update requires characterizing the set of optimal solutions of (13), or more generally of (10); recall the general definition in (7). While there exist necessary conditions characterizing the optimal solutions of linear bilevel problems, they involve products of continuous variables with non-complementary type of constraints. Thus, an exact mixed-integer linear representation for these conditions is not readily available. Nonetheless, the full update can be approximated with the NCvx update by adding what we call the *"non-repetitive" constraints.* There are two types of such constraints: the first type forces $\tilde{w}_E(\mathcal{U}^t)$ to be equal to $\tilde{w}(x^{s,\psi};\boldsymbol{c})$ if $x = x^{s,\psi}$ for some $s<t$; the second type helps by reducing the size of $\mathcal{U}^t$.

For any given $t \in \mathcal{T}$ we can define the first type of constraint via the set N^t_R, where

$$N^t_R := \left\{(x,y) \in X \times \mathbb{R}^{|A|} :\ x = x^s \Rightarrow \boldsymbol{d}^\top y + \boldsymbol{b}^\top x \le \tilde{w}(x^s;\boldsymbol{c}),\ s<t\right\}.$$

Clearly, any $(x,y)\in N_R^t$ has an objective value in (10) at most equal to $\tilde{w}(x^s;\boldsymbol{c})$ as long as $x=x^s$ for some $s<t$. The second type of constraint is defined via the set $\mathcal{L}^t$, where

$$\mathcal{L}^t=\{\hat{c}\in\mathbb{R}^{|A|}:\,\hat{c}^\top y^{t,E}>z(x^t;\boldsymbol{c})\}.$$

While the benefit of adding N_R^t to (10) is evident, the usefulness of $\mathcal{L}^t$ is stated next.

LEMMA 3. *Let $t\in\mathcal{T}$, assume that the Leader implements a policy $\psi\in\Psi$, and that $\bar{w}^t<\tilde{w}_E(\mathcal{U}^t)$. Then: (*i*) $\boldsymbol{c}\in\mathcal{L}^t$; (*ii*) if $z^{t,E}\le z(x^t;\boldsymbol{c})$ then $c^{t,E}\notin\mathcal{L}^t$; and (*iii*) if $z^{t,E}>z(x^t;\boldsymbol{c})$ and the feedback is Value-Perfect (Response-Perfect), then* $\dim(\mathcal{U}^{t+1})<\dim(\mathcal{U}^t)$*, $c^{t,E}\notin\mathcal{V}^t$ ($c^{t,E}\notin\mathcal{R}^t$).*

Consider that any given update mechanism is modified by adding $\mathcal{L}^t$ to it. Lemma 3 then implies that this update is valid (as it does not remove $\boldsymbol{c}$). Moreover, it also assures that for a policy $\psi\in\Psi$ under Value-Perfect or Response-Perfect feedback, the previously assumed cost vector $c^{t,E}$ that led to the non-converging solution at time t is not considered again.

Besides the above properties, the next result shows that, independent of the update mechanism, the addition of the non-repetitive constraints potentially decreases the optimal value of (10) without compromising the real-time certificate of optimality of the policies in Ψ. Its proof follows from (i) of Lemma 3 and simple feasibility arguments. The strict inequality proof follows from Example 4.

PROPOSITION 1. *Let $t\in\mathcal{T}$ be given and assume that $\bar{w}^t<\tilde{w}_E(\mathcal{U}^t)$. Let $\widehat{\mathcal{U}}^{t+1}=\mathcal{U}^t\cap\mathcal{L}^t$ and define*

$$\hat{w}:=\max\big\{\boldsymbol{d}^\top y+\boldsymbol{b}^\top x:\,x\in X,\hat{c}\in\widehat{\mathcal{U}}^{t+1},y\in Z(\hat{c},x),(x,y)\in N_R^{t+1}\big\}.$$

Then, $\tilde{w}^*(\boldsymbol{c})\le\hat{w}\le\tilde{w}_E(\mathcal{U}^{t+1})$. Moreover, there exist instances, where the last inequality is strict.

For the case of ASPI, the non-repetitive constraints in N_R^t can be formulated as linear inequalities in terms of the variables of the problem. Indeed, let D be an upper bound on $\tilde{w}^*(\boldsymbol{c})$ (e.g., $D=\max\{\boldsymbol{d}^\top y:\,y\in\bigcup_{x\in X}Y(x)\}$) and, for $t\in\mathcal{T}$ given, consider the constraints

$$D(x^s)^\top x+\boldsymbol{d}^\top y\le Dk+\tilde{w}(x^s;\boldsymbol{c}),\quad s<t. \tag{19}$$

For $s<t$ the constraint states that if $x=x^s$, then $D(x^s)^\top x=Dk$ and therefore one must have that $\boldsymbol{d}^\top y\le\tilde{w}(x^s;\boldsymbol{c})$; otherwise, the constraint is trivially satisfied. In other words, constraints (19) imply that if solution x^s is repeated at any time after s, then its best possible objective value on problem (17) is $\tilde{w}(x^s;\boldsymbol{c})$. Consequently, Proposition 1 implies that if $G=(N,A)$ is a network where all interdiction solutions yield different networks for the Follower, then the full update can be computed by adding constraints (19) to (17).

We refer to the update mechanism that uses non-repetitive constraints jointly with the NCvx (Cvx) update as the *enhanced* NCvx (Cvx) update. While our aim is to improve practical performance, we note that the addition of the non-repetitive constraints for either the NCvx or Cvx updates, guarantees a (rather trivial) finite bound on the time stability. Indeed, by adding the non-repetitive constraints in N_R^t, policies in Ψ, will – in the worst case – exhaustively search over all $x \in X$, thus time stability is finite as long as $|X| < \infty$ (see also the work by Yang et al. (2019), where similar in spirits non-repetitive constraints are exploited for Greedy and Robust policies in the context of sequential shortest path interdiction with learning). In particular, this worst-case bound is equal to the worst-case bound of the full update in settings where $[x] = x$ for all $x \in X$. In this sense, for these particular settings, the enhanced updates and the full update can be considered equivalent.

5.2. Generation of the Instances

We assume that the Leader and the Follower are interacting in the context of smuggling. The Follower is a smuggler who uses a road network $G = (N, A)$, with cost vector $\boldsymbol{c}$, to smuggle goods from node $i = 1$ to node $i = m$ at minimum cost. The Leader estimates that the smuggler successfully traverses arc $a \in A$ with probability p_a. She must decide which arcs of the network to block (thus prohibiting its use by the smuggler), considering an interdiction budget of K blocked arcs, with the objective of minimizing the probability that a smuggler successfully reaches its destination.

Let $A' \subseteq A$ be a possible (unblocked) path to be used by the smuggler: the probability that he successfully traverses the path is given by $\prod_{a \in A'} p_a$. In order to minimize this probability, the Leader can equivalently maximize the negative of its logarithm. Hence, the full-information problem can be framed as an ASPI with $d_a = -\log(p_a)$, $a \in A$.

We assume that the road network G has a layered topology with n_ℓ layers and n_k nodes per layer. In this topology, each node has a directed arc only towards the nodes in the next layer and there are two additional source and sink nodes before the first and last layer, respectively. The real cost vector $\boldsymbol{c}$ is generated from the Euclidean distances between the locations of the nodes in the plane. The locations, in turn, are drawn according to the following procedure: the x-coordinates are drawn from an $U(0, B_x)$ distribution and ordered accordingly so that the smallest value is the x-coordinate of node 1, the following n_k smallest values are the x-coordinates of the first layer, and next smallest n_k values are the x-coordinates of the second layer, and so on. Once these values have been set, the y-coordinates of each node in each layer are drawn at random from a $U(0, B_y)$ distribution. Here, both B_x and B_y are tuning parameters.

The initial uncertainty set $\mathcal{U}^1 \subseteq \mathbb{R}_+^{|A|}$ is the hypercube $\prod_{a \in A} [\ell_a, u_a]$. The bounds ℓ_a and u_a are generated as follows: for a given tuning parameter $\Delta \in (0, 1]$, we set $\ell_a = c_a(1 - (1 - r_a)\Delta)$ and $u_a = c_a(1 + r_a\Delta)$, with $r_a \in [0, 1]$ drawn at random, for each $a \in A$. In our experiments we use three distributions for r_a: Uniform(0,1), Beta(5,2), and Beta(1,3), which imply that the location of c_a is uniform in the interval, or its located closer to ℓ_a (left skewed) or u_a (right skewed), respectively. Note that the width of the interval $[\ell_a, c_a]$ is given by $c_a\Delta$, thus the larger the value of Δ the greater the uncertainty faced by the Leader.

We generate the probabilities p_a by assuming that on each layer there is a 'sensor' whose coordinates are drawn from uniform distributions, that take values in the range of the x-coordinates and y-coordinates of the nodes of the layer. Specifically, if $s^\ell = (s^\ell_x, s^\ell_y)$ is the location of the sensor on layer ℓ and $a \in A$ is an arc whose initial node is in layer ℓ, then $p_a = 1 - \exp(-||s^\ell - s_a||^2/U)$, where s_a has the coordinates of the mean point between the nodes defining arc $a \in A$, $||\cdot||$ denotes the Euclidean distance, and U is a positive constant. Note that the farther away an arc is to the sensor, it is more probably that a smuggler traverses the arc undetected.

5.3. Results and discussion

We perform four sets of experiments. In the first set, we assess the performance of the policies in Ψ under the different update mechanisms, and compare them against their α-optimal counterparts. In the second set, we compare the policies in Ψ with respect to other benchmark policies. Later, we perform a sensitivity analysis with respect to the quality of the information known by the interdictor. Finally, we perform a sensitivity analysis with respect to the instance size. In assessing policy performance, in addition to time stability, we use the concept of regret, which measures the Leader's cumulative profit loss relative to an oracle Leader with full knowledge of the cost vector, i.e. for a policy π we define

$$\text{Total Regret}^\pi(\boldsymbol{c}) := \sum_t \big(\tilde{w}^*(\boldsymbol{c}) - \tilde{w}(x^{t,\pi};\boldsymbol{c})\big).$$

5.3.1. Performance of Ψ with respect to the update mechanism. For this set of experiments we set $n_\ell = n_k = 4$, $B_x = 50$, $B_y = 30$, $\Delta = 1/3$, $U = 12$, $k = 3$, and $T = 15$. We randomly generate ten instances and solve each instance using the four update mechanisms: Cvx (Cvx), enhanced Cvx (E-Cvx), NCvx (Ncvx), and enhanced NCvx (E-Ncvx). Note that $[x] = x$ for all $x \in X$ in this setup, so the enhanced updates are equivalent to the full update in terms of their worst-case time stability. In addition, we solve each instance using two α-optimal policies ($\alpha = 1.5$ and $\alpha = 2$) under the E-Ncvx update. We take the convention that if a policy stalls in a sub-optimal solution, then its time stability is ∞; likewise, if the policy has not attained the time stability before T (but has not stalled until T) we set its time stability to $T+1$. Tables 1 and 2 summarize our results.

From Table 1, we observe that the E-Ncvx update yields the best results for time stability and regret across all distributions when only Standard feedback is available. In most cases the

Feedback	Time-stability				Total Regret				Solution Time (secs)			
Uniform	Cvx	E-Cvx	Ncvx	E-Ncvx	Cvx	E-Cvx	Ncvx	E-Ncvx	Cvx	E-Cvx	Ncvx	E-Ncvx
Standard	∞ (8,2)	12 (0,6)	∞ (7,0)	**7.60**	40.81	32.25	32.71	**16.36**	**1.42**	4.26	6.14	12.28
VP	∞ (5,0)	6.8 (0,1)	∞ (6,0)	**6.70**	26.82	**15.06**	25.75	15.59	**1.56**	2.96	2.52	6.61
RP	∞ (8,0)	7.1 (0,1)	∞ (7,0)	**7.40**	33.21	15.51	31.88	**14.43**	**2.11**	3.01	3.79	8.00
Left Sk.	Cvx	E-Cvx	Ncvx	E-Ncvx	Cvx	E-Cvx	Ncvx	E-Ncvx	Cvx	E-Cvx	Ncvx	E-Ncvx
Standard	∞ (7,0)	13.3 (0,7)	∞ (1,0)	**7.80**	27.53	28.77	20.77	**17.25**	**1.66**	4.91	7.52	11.00
VP	∞ (1,0)	**5.20**	∞ (3,0)	**5.20**	10.02	**9.34**	12.59	11.26	2.46	**2.09**	5.50	4.57
RP	∞ (1,0)	8.00	∞ (5,1)	**7.50**	21.78	17.94	28.32	**15.91**	6.01	**3.28**	7.03	8.22
Right Sk.	Cvx	E-Cvx	Ncvx	E-Ncvx	Cvx	E-Cvx	Ncvx	E-Ncvx	Cvx	E-Cvx	Ncvx	E-Ncvx
Standard	∞ (9,0)	16 (0,10)	∞ (8,0)	**10 (0,2)**	40.36	45.27	39.64	**23.52**	**0.79**	6.45	4.45	15.95
VP	∞ (8,0)	**7.60**	∞ (8,0)	7.8 (0,1)	33.89	16.07	37.41	**16.02**	**1.51**	3.06	3.73	7.19
RP	∞ (9,0)	**10.30**	∞ (9,0)	9.7 (0,2)	38.39	24.63	35.90	**22.27**	**0.92**	4.54	2.34	10.81

Table 1 Average time stability, total regret, and solution time across ten instances for different update mechanisms: Cvx (Cvx), enhanced Cvx (Ecvx), NCvx (Ncvx), and enhanced NCvx (E-Ncvx), and different cost generation schemes for c_a in the test instances: Uniform in the interval, closer to either ℓ_a (Left skewed) or u_a (Right skewed). Whenever there is a parenthesis (a, b), a and b are the numbers of instances where the policy stalls and where the policy has a time stability greater than $T = 15$, respectively. Boldface indicates the best result.

time stability is found before $T = 15$. Note that in most instances, our policies either stall or do not guarantee an optimal solution within the first T periods, for all other update mechanisms. Regarding total regret, the results also favor the E-Ncvx update, getting roughly half the regret of the other update mechanisms for the uniform and right skewed distributions.

Whenever Value-Perfect or Response-Perfect feedback are available, the E-Cvx mechanism becomes a good alternative to the E-Ncvx update. The results show that these two mechanisms have a similar behavior for time stability and regret across arc cost distributions. For Response-Perfect feedback this similarity is expected, as from the definition of each mechanism, the NCvx (enhanced NCvx) update is equivalent to the Cvx (enhanced Cvx) update, see (6). In this sense, their differences are due to the way that the solver chooses alternative optimal solutions in the formulations. For Value-Perfect feedback the updates are not necessarily equivalent. However, Value-Perfect feedback plus the fact that the Follower's problem is binary, might help the MIP solver discover early in the optimization process what the actual value of y^t is. This knowledge reduces the number of potential feasible Follower's solutions.

Across the board it is seen that the Cvx and NCvx updates, without any enhancement, are not good update mechanisms. They stall in a significant number of instances and attain higher regret values relative to their enhanced counterparts. Their only advantage is that they are faster to solve for most of the cases. This stark difference between the Cvx and NCvx updates suggests that just the non-repetitive constraints are sufficient to attain a good performance (i.e., that updating the uncertainty set might not be necessary). However, this is not necessarily the case, as we show in the next section.

Feedback	Time-stability			$\tau_\alpha^{\psi_\alpha}$		Total Regret			Solution Time		
Uniform	ψ	$\psi_{1.5}$	ψ_2	$\psi_{1.5}$	ψ_2	ψ	$\psi_{1.5}$	ψ_2	ψ	$\psi_{1.5}$	ψ_2
Standard	**7.6**	∞ (2,0)	∞ (4,0)	4.40	**1.90**	16.36	**6.96**	7.23	12.28	4.85	**0.98**
VP	**6.70**	∞ (3,0)	∞ (4,0)	3.40	**2.50**	15.59	**8.57**	10.56	6.61	2.43	**1.36**
RP	**7.40**	∞ (1,0)	∞ (3,0)	3.80	**1.90**	14.43	**5.38**	5.83	8.00	3.26	**1.02**
Left Sk.	ψ	$\psi_{1.5}$	ψ_2	$\psi_{1.5}$	ψ_2	ψ	$\psi_{1.5}$	ψ_2	ψ	$\psi_{1.5}$	ψ_2
Standard	**7.8**	∞ (2,0)	∞ (3,0)	3.90	**2.50**	17.25	10.89	**6.37**	11.00	3.52	**1.64**
VP	**5.20**	∞ (3,0)	∞ (3,0)	3.80	**2.60**	11.26	**8.68**	8.87	4.57	2.85	**1.52**
RP	**7.50**	∞ (2,0)	∞ (4,0)	4.00	**1.90**	15.91	**7.31**	7.33	8.22	3.55	**1.02**
Right Sk.	ψ	$\psi_{1.5}$	ψ_2	$\psi_{1.5}$	ψ_2	ψ	$\psi_{1.5}$	ψ_2	ψ	$\psi_{1.5}$	ψ_2
Standard	**10 (0,2)**	∞ (2,0)	∞ (3,0)	5.60	**2.30**	23.52	12.13	**7.76**	15.95	6.43	**1.39**
VP	**7.8 (0,1)**	∞ (3,0)	∞ (2,0)	3.80	**2.70**	16.02	8.17	**4.16**	7.19	2.72	**1.47**
RP	**9.7 (0,2)**	∞ (2,0)	∞ (3,0)	5.40	**2.60**	22.27	9.74	**4.50**	10.81	5.24	**1.72**

Table 2 Average time stability, α-time-stability, total regret, and solution time across ten instances for the α-optimal policies, $\alpha \in \{1.5, 2\}$, with the E-Ncvx update. Whenever the is a parenthesis (a, b) in the time-stability column, a and b are the number of instances where the policy stalls and where the policy has a time stability greater than T, respectively. Boldface indicates the best result.

Table 2 shows the performance of our policies and their α-optimal counterparts under the E-Ncvx update. These results show that the α-optimal policies are a reasonable alternative to the policies in Ψ. Indeed, across all settings, they attain a lower regret and are solved faster. Moreover, as seen from their α-time-stability (the column of $\tau_\alpha^{\psi_\alpha}$), they can guarantee α-optimal solutions for all instances tested and these guarantees are found significantly earlier than the time it takes policies in Ψ to converge to the optimal solution. The only drawback of the α-optimal policies is that, as expected, they do not converge to the full-information optimal solution for all instances. Overall, they stall only for about 20%-30% of the instances. Nonetheless, as suggested from the regret values, it seems that even when they stall, the sub-optimal solution they find is closer in value to the full-information optimal than the α-guarantee suggests.

5.3.2. Performance of Ψ with respect to the benchmark policies. Next, we use the instances of the previous section to compare policies in ψ, under the E-Ncvx update, with a set of benchmark policies. The benchmark policies are all greedy in nature, and given by:

- Random policies (π_{R-Cvx} and π_{R-Ncvx}): At each $t \in \mathcal{T}$ these policies solve the bilevel problem (13), including the non-repetitive constraint N_R^t (19), by selecting $\boldsymbol{c}$ at random from $\mathcal{U}^t$. This random point is obtained by solving the MIP of the form: $\min\{\boldsymbol{r}^\top \hat{\boldsymbol{c}} \colon \hat{\boldsymbol{c}} \in \mathcal{U}^t\}$, where each entry of $\boldsymbol{r}$ is Bernoulli distributed with parameter $p = 1/2$. When $\mathcal{U}^t$ is updated using the Cvx (NCvx) update, we denote this policy by π_{R-Cvx} (π_{R-Ncvx}).
- Center policies (π_{C-Cvx} and π_{C-Ncvx}): These policies are computed as the random policies, with the difference that c is selected to be: (i) equal to the analytical center of $\mathcal{U}^t$ whenever using the Cvx update (π_{C-Cvx}); and (ii) the closest point in $\mathcal{U}^t$ (in the ℓ^1-norm) to the

analytical center of the relaxation of $\mathcal{U}^t$ whenever using the NCvx update (π_{C-Ncvx}). These policies can be rationalized by thinking that the Leader assumes that the real cost vector $\boldsymbol{c}$ is close to the 'center' point of the uncertainty set at each period $t \in \mathcal{T}$.

- Greedy and Best-case non-repetitive policy ($\psi_N \in \Psi$): these policies implement the solution to problem (10) at each time $t \in \mathcal{T}$ while including both the non-repetitive constraints N_R^t and $\mathcal{L}^t$; however, the Cvx or NCvx updates are not used, so that $\mathcal{U}^{t+1} = \mathcal{U}^t \cap \mathcal{L}^t$ under Standard feedback and $\mathcal{U}^{t+1} = \mathcal{U}^t \cap \mathcal{L}^t \cap \mathcal{V}^t$ ($\mathcal{U}^{t+1} = \mathcal{U}^t \cap \mathcal{L}^t \cap \mathcal{R}^t$) for Value-Perfect (Response-Perfect) feedback.
- Greedy and Robust policy ($\lambda \in \Lambda$): At each time $t \in \mathcal{T}$ these policies implement a solution to (8). We assume that $\mathcal{U}^t$ is updated only with Value-Perfect or Response-Perfect feedback, as it can be readily checked that the inclusion of the Cvx update does not change the optimal solution of (8); and that the NCvx update yields a formulation that is significantly more challenging to solve.

Tables 3 summarizes the average time stability for the experiments. From these results, we observe that ψ is the best policy overall, particularly under Standard feedback, while the random policies come in second. Remarkably, π_{R-Ncvx} has a time stability that is close to policies in Ψ; its main drawback being that for around 20% of the cases, across all feedback and distribution types, it cannot find a consistent optimal solution before T. On the other hand, the center and Λ policies always have instances where they stall. The rate of stalling for the center policies is uniform across feedback modes, update mechanisms, and distributions, while policies in Λ only stall once or twice under Value-Perfect and Response-Perfect feedback (an explanation for this behavior is that policies in Λ under these feedbacks always discover new information whenever the Follower uses a solutions different from what the Leader expects, see Lemma 1). Finally, the non-repetitive policy ψ_N has a worse performance than ψ across the board. The difference is more pronounced under Standard feedback, where ψ_N fails to get to the time stability before T for most instances. Consequently, we observe that both the Cvx and NCvx updates provide information that can substantially improve the performance of the greedy and best-case policies.

Table 4 gives the average regret for the benchmark policies. Observe that the random NCvx policy, the center policies, and policies in Λ, significantly outperform policies in Ψ across all feedback and distribution types. Therefore, although these policies do not guarantee finding an optimal solution nor provide certificates of optimality in real time, they can find an optimal or near-optimal solutions in most of the cases. On the other hand, observe that the non-repetitive policy fares worse than policies in Ψ across the board, which reinforces the observation that the Cvx and NCvx updates provide important information that should be exploited by the interdictor.

Feedback	Time-stability						
Uniform	ψ	π_{R-Cvx}	π_{R-Ncvx}	π_{C-cvx}	π_{C-Ncvx}	ψ_N	λ
Standard	**7.6**	14.1 (0,8)	7.9 (0,2)	∞ (3,2)	∞ (3,0)	13.8 (0,8)	∞ (5,0)
VP	**6.70**	7.8 (0,1)	7.6 (0,2)	∞ (2,0)	∞ (2,0)	7.80	∞ (1,0)
RP	**7.40**	9.2 (0,3)	7.5 (0,3)	∞ (1,0)	∞ (1,0)	9.9 (0,3)	∞ (2,0)
Left Sk.	ψ	π_{R-Cvx}	π_{R-Ncvx}	π_{C-cvx}	π_{C-Ncvx}	ψ_N	λ
Standard	**7.8**	14.8 (0,3)	9.2 (0,1)	∞ (3,0)	∞ (1,0)	14.7 (0,9)	∞ (4,0)
VP	**5.20**	8 (0,1)	7.6 (0,1)	∞ (2,0)	∞ (2,0)	6.50	∞ (1,0)
RP	**7.50**	9.5 (0,1)	11.4 (0,2)	∞ (2,0)	∞ (2,0)	9.7 (0,3)	∞ (1,0)
Right Sk.	ψ	π_{R-Cvx}	π_{R-Ncvx}	π_{C-cvx}	π_{C-Ncvx}	ψ_N	λ
Standard	10 (0,2)	14.6 (0,7)	8.4 (0,3)	∞ (2,0)	∞ (2,0)	16 (0,10)	∞ (4,0)
VP	**7.8 (0,1)**	8.2 (0,1)	11.1 (0,2)	∞ (2,0)	∞ (2,0)	10 (0,3)	∞ (1,0)
RP	9.7 (0,2)	9.6 (0,4)	**8.1 (0,2)**	∞ (2,0)	∞ (2,0)	11.7 (0,4)	∞ (1,0)

Table 3 Average time stability across ten instances for ψ with the E-Ncvx update and the benchmark policies. Whenever there is a parenthesis (a, b), a and b are the number of instances where the policy stalls and where the policy has a time stability greater than T, respectively. Boldface indicates the best result.

Interestingly, policies in Ψ have unsatisfactory regret performance. A possible explanation for this is that, during the initial time periods, policies in Ψ consider cost vectors $c^{t,E}$ that allow for shortest paths $y^{t,E}$ such that $\boldsymbol{d}^\top y^{t,E}$ is equal (or close to) $\max\{\boldsymbol{d}^\top y \colon y \in Y\}$. In practice, however, a Follower's feasible solution that has a high value of $\boldsymbol{d}^\top y$ is not a shortest path under the real cost vector $\boldsymbol{c}$, as both $\boldsymbol{c}$ and $\boldsymbol{d}$ are generally unrelated. Therefore, one would expect a better regret performance for policies in Ψ in instances where $\boldsymbol{d}$ and $\boldsymbol{c}$ are negatively correlated and a worse performance in instances where they are positively correlated.

Feedback	Total Regret						
Uniform	ψ	π_{R-Cvx}	π_{R-Ncvx}	π_{C-Cvx}	π_{C-Ncvx}	ψ_N	λ
Standard	16.36	21.63	8.92	11.28	**5.58**	33.89	22.74
VP	15.59	10.46	8.58	5.11	4.86	16.33	**4.67**
RP	14.43	9.85	7.48	**2.28**	3.18	22.11	5.68
Left Sk.	ψ	π_{R-Cvx}	π_{R-Ncvx}	π_{C-Cvx}	π_{C-Ncvx}	ψ_N	λ
Standard	17.25	21.56	8.53	5.50	**2.42**	33.62	7.64
VP	11.26	6.85	5.55	6.40	6.49	12.65	**5.26**
RP	15.91	7.95	7.68	6.46	**6.06**	21.00	5.91
Right Sk.	ψ	π_{R-Cvx}	π_{R-Ncvx}	π_{C-Cvx}	π_{C-Ncvx}	ψ_N	λ
Standard	23.52	18.01	10.47	7.74	**6.46**	44.17	13.47
VP	16.02	11.55	10.60	**6.46**	**6.46**	20.63	2.95
RP	22.27	6.15	10.36	6.32	6.23	29.29	**1.58**

Table 4 Average regret across ten instances for ψ with the E-Ncvx update and the benchmark policies. Boldface indicates the best result.

Finally, in Appendix D.1 we consider the average solution times across the benchmark policies. There it is shown that policies that use the NCvx update take more time to solve on average. This behavior is expected as optimizing over non-convex sets is harder than over convex sets.

5.3.3. Sensitivity to the Information Quality. In this section we assess policy performance when the structure of the bounds in $\mathcal{U}^1$ changes. For this, we assume that $u_a = \bar{u}$ and $\ell_a = 0$ for all $a \in A$, that is, all the lower and upper bounds are the same across all arcs, therefore $\mathcal{U}^1 = [0, \bar{u}]^{|A|}$. Observe that this type of uncertainty set gives far less information to the Leader because, at least initially, all arcs' costs have the same range and therefore all network paths can be shortest paths. In addition, in contrast to the previous instances, the initial bounds and the real cost vectors are uncorrelated.

We generate ten instances using the same parameters of Section 5.3.1, with the exception that here $T = 20$, and that we only use the uniform distribution to generate the true cost vector. For each instance we consider three types of bounds. In the *low variability* case, $\boldsymbol{c}$ is divided by two and $\bar{u} = \max\{c_a/2 \colon a \in A\} + 3$; in the *normal variability* case, $\boldsymbol{c}$ remains unchanged and $\bar{u} = \max\{c_a \colon a \in A\} + 5$; and in the *high variability* case, $\boldsymbol{c}$ is doubled and $\bar{u} = \max\{2\, c_a \colon a \in A\} + 10$.

In what follows, we compare the performance of policies in Ψ under the under the E-Ncvx update with that of: (i) their α-optimal counterparts, $\Psi_{1.5}$ and Ψ_2; (ii) theNCvx random policy π_R; and (iii) the greedy and robust policies in Λ. We use these benchmark policies as they performed reasonably well in the experiments of Sections 5.3.1 and 5.3.2. Table 5 summarizes the average time stability, α-optimal time stability, and total regret across all considered policies and instances.

Feedback	Time-stability					$\tau_\alpha^{\psi_\alpha}$		Total Regret				
Low	ψ	π_R	λ	$\psi_{1.5}$	ψ_2	$\psi_{1.5}$	ψ_2	ψ	π_R	λ	$\psi_{1.5}$	ψ_2
Standard	**11.9 (0,1)**	14.4 (0,3)	∞(9,0)	∞(3,0)	∞(6,0)	4.06	**2.24**	28.08	37.09	56.32	**21.58**	25.41
VP	**11.70**	14 (0,3)	∞(1,0)	∞(2,0)	∞(5,0)	4.80	**3.60**	26.94	30.24	15.96	**15.64**	23.89
RP	**11.00**	16.5 (0,3)	∞(2,0)	∞(5,0)	∞(5,0)	5.70	**4.10**	24.98	42.04	21.13	**19.75**	24.61
Normal	ψ	π_R	λ	$\psi_{1.5}$	ψ_2	$\psi_{1.5}$	ψ_2	ψ	π_R	λ	$\psi_{1.5}$	ψ_2
Standard	**12.3 (0,1)**	16.4 (0,6)	∞(9,0)	∞(2,0)	∞(5,0)	5.90	**3.80**	27.72	45.56	53.82	**17.16**	23.14
VP	**11.1**	13.6 (0,4)	∞(1,0)	∞(3,0)	∞(5,0)	4.60	**3.50**	25.69	38.04	**15.37**	15.97	19.26
RP	**11.70**	14.6 (0,2)	∞(2,0)	∞(3,0)	∞(4,0)	5.60	**4.70**	25.53	41.35	22.11	**17.25**	26.36
High	ψ	π_R	λ	$\psi_{1.5}$	ψ_2	$\psi_{1.5}$	ψ_2	ψ	π_R	λ	$\psi_{1.5}$	ψ_2
Standard	**12.3**	14.5 (0,3)	∞(9,0)	∞(4,0)	∞(6,0)	6.20	**3.60**	28.45	35.79	66.20	**22.04**	28.55
VP	**11.50**	15.4 (0,4)	∞(1,0)	∞(2,0)	∞(3,0)	4.80	**3.60**	27.44	34.21	15.49	**13.66**	15.46
RP	**11.10**	15 (0,1)	∞(4,0)	∞(4,0)	∞(5,0)	6.30	**4.00**	24.75	35.32	**21.38**	21.65	25.06

Table 5 Average time stability, α-time-stability, and regret across ten instances for different cost generation schemes for arc costs in the test instances: Low, Normal and High variability cases. Whenever there is a parenthesis (a, b), a and b are the number of instances where the policy stalls and where the policy has a time stability greater than $T = 20$, respectively. Boldface indicates the best result.

From Table 5 we observe that all policies are fairly insensitive to the interval size, indicating that their performance is robust to the scaling of the uncertain data. We also observe that the greedy and best-case policies (either exact or approximated) outperform other policies. Indeed, while for the experiments in Section 5.3.2 policies π_R and λ had close to the best (or the best) performances with respect to time stability and regret, here that is no longer the case. For instance, the average time stability of π_R is roughly 35% more than that of policies in Ψ, and in roughly 30% of the instances it fails to consistently implement the optimal solution within the first $T = 20$ periods. On the other hand, policy λ has a similar time-stability performance to that of the experiments in Section 5.3.2, but its average regret is far worse for this set of experiments.

The above considerations hint that the Greedy and Best-Case policies are more robust to the uncertainty, because they have a similar behavior independent of the uncertainty that the Leader faces. In contrast, the other benchmark policies are far more sensitive with respect to the uncertainty, and perform better whenever $\mathcal{U}^1$ provides more information about the real cost vector $\boldsymbol{c}$.

Finally, in Appendix D.2 we analyze the scalability of the policies, in terms of computation time, with respect to the instance size. We show that running time for policies in Λ scale well with size, but that this explained partly because they tend to stall earlier than the other policies.

5.3.4. Performance of Ψ in non-grid instances Next, we apply the proposed policies to an instance of the smuggling interception problem described in Section 1. In particular, we consider the "infiltration network" described by Unsal (2010). In said network, shown in Figure 5, the nodes are the locations used by the smuggler (the Follower) and an arc between two nodes means that the smuggler can move between the corresponding locations (the network has 38 nodes and 109 arcs). The interceptor (an US military task force, the Leader) has positioned inspectors of each arc, who have a positive probability of detecting the smuggler upon passage.

The objective of the interceptor is to allocate the patrol units so as to maximize the probability that the smuggler is intercepted by the inspectors (or, equivalently, to minimize the probability that the smuggler is successful). Assuming that detection (as a random variable) is independent across inspectors, we write the Leader's problem as that of maximizing the logarithm of the probability of detection, so that we can define the upper-level vector as $d_a = -\log(1-p'_a)$, $a \in A$, where p'_a is the probability that the smuggler is detected by an inspector while moving across arc $a \in A$; We consider three sets probabilities from Unsal (2010) (one for each "type" of inspector).

In our setting, we build upon this problem and assume that, in addition to the inspectors, the US task force has a limited number of k patrol units, which can be used to install roadblocks along some set of arcs. We assume that roadblocks are both effective and observable, thus the smuggler does not traverse blocked arcs. We also assume that the smuggler is oblivious to the presence of inspectors and focuses on minimizing transportation costs, which are proportional to the distances between the locations (computed using Google Earth (Gorelick et al. 2017)). Finally, we assume that the interdictor does not know the precise route used by the smuggler to travel

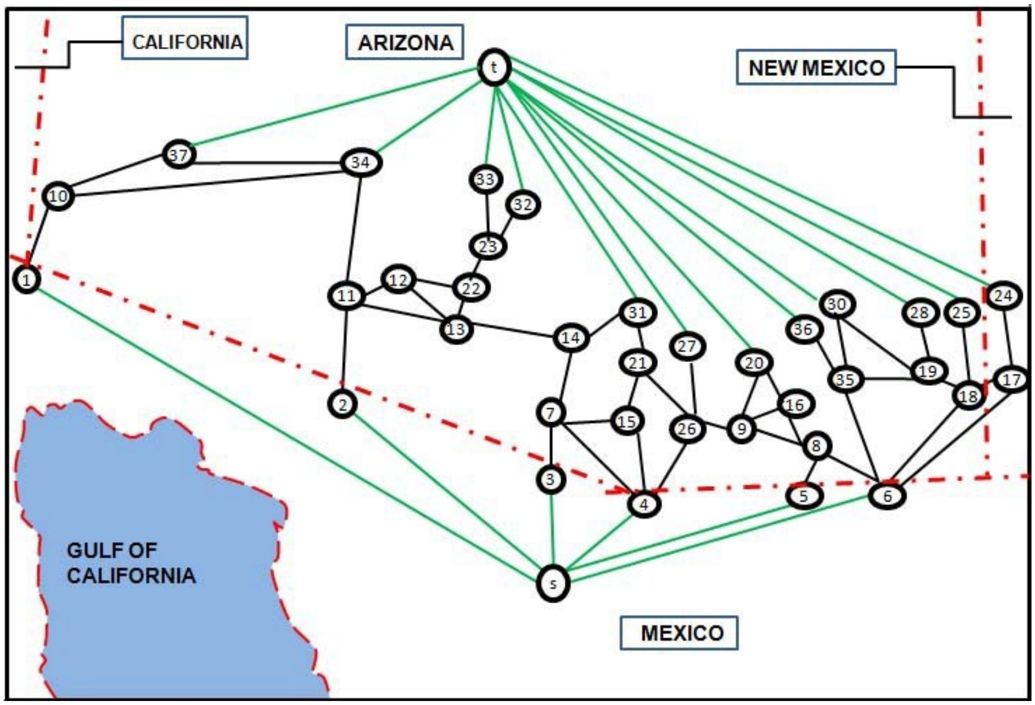


Figure 5 Infiltration network used by the smugglers in the Mexico-USA border, taken from the work by Unsal (2010). Node s is where the smugglers start, node t is their destination. The probability of detection in the green arcs is zero; the probability is non-zero for the black arcs.

between locations (and hence their costs), thus she estimates that the cost c_a of traversing arc $a \in A$ lies within a neighborhood $[\ell_a, u_a]$ of said cost. The values of ℓ_a and u_a are generated as in Section 5.2 using $\Delta = 1/3$ (so the width of each interval is $c_a/3$).

We run three instances of the problem (the difference between the instances being the value of $\boldsymbol{d}$), which we cast as ASPI, under various feedback modes, for various values of k, using policies ψ, π_R, and λ. Policy ψ is run with the enhanced NCvx update mechanism and π_R is run with the NCvx update mechanism. The results for one of the instances is shown in Table 6. The results for the other two instances can be found in Tables 9 and 10 in Appendix D.3. The results for these instances confirm our previous findings: policy ψ has the best average time-stability performance across feedback modes. Its regret is not the best all the time and it's second only to the randomized policy on average. The solution times of ψ, as observed in previous experiments, is longer than λ but comparable to that of the randomized policy. Policy π_R attains good regret performance but its time stability is consistently worse that policies in ψ. Policy λ is the fastest of the three, however, it has a bi-modal behavior. For some cases it attains very short time-stability values with regrets close to zero, while for some other cases it stalls and yields fairly large values of regret.

		Time-Stability			Regret			Solution time (secs.)		
	Feedback	π_R	λ	ψ	π_R	λ	ψ	π_R	λ	ψ
$k=2$	Standard	21	∞	7	1.27	12.82	4.54	28.89	0.14	25.56
	RP	11	2	7	3.07	0.64	3.93	28.41	0.30	17.02
	VP	15	3	5	2.28	0.67	2.34	26.29	0.58	11.46
$k=3$	Standard	10	∞	10	1.31	0.66	7.87	32.21	0.22	54.20
	RP	14	2	10	4.17	0.03	7.94	33.38	1.11	44.95
	VP	6	2	8	0.17	0.03	5.99	26.20	0.80	28.04
$k=4$	Standard	8	∞	7	4.94	11.20	5.84	51.80	0.34	141.90
	RP	18	∞	7	8.17	11.20	5.84	43.87	0.89	50.66
	VP	6	∞	4	5.03	12.09	2.81	39.74	1.83	22.59

Table 6 Time stability, regret, and solution time for one of the smuggling instances with various feedback modes and values of k. A time stability of ∞ means that the policy stalls.

6. Concluding Remarks

This paper addresses sequential bilevel linear programming problems where a Leader and a Follower interact repeatedly. While the Follower always responds in an optimal fashion, the Leader is initially unaware of the Follower's (linear) objective function, except for the fact that its cost coefficients are within a given (initially known) uncertainty set. However, she might use feedback from the Followers' actions to refine her belief about the unknown objective, so as to maximize her cumulative benefit. Depending on the feedback available (Standard, Value-Perfect or Response-Perfect), we propose different updates mechanisms for the uncertainty set, differing on the amount of information that is incorporated into the updated set. We show that, in general, there is a trade-off between the reduction of the uncertainty set, and the tractability of its representation. In particular, we show that the strongest update leads to non-convex and non-closed regions, that are not amenable (at least, in a straightforward manner) to MIP-based solution techniques.

In the first approach discussed, we adapt the set of Greedy and Robust policies developed by Borrero et al. (2019) in the context of the max-min sequential interdiction: under these policies the Leader assumes that the Follower is also unaware of his own objective, and selects his response in a robust fashion. We show that the said adaptation fails to provide real-time optimality certificates, and might stall by implementing suboptimal solutions indefinitely. We then provide a second adaptation in the form of Greedy and Best-case policies, in which the Leader assumes that she is able to select (from the uncertainty set) the cost vector the Follower will use, and selects her action according to this optimistic approach. We show that these policies do provide real-time optimality guarantees, and converge to the full information solution, under the strongest update mechanism.

With regard to policy implementation, we first discuss our policy generalizations that ensure convergence to constant-factor α-approximate optimal solutions. Then we show that when the Follower's problem admits a linear programming formulation, one can compute the proposed policies by solving a series of MIPs, under all update mechanisms except for the full update. Nonetheless, we describe an approximation to the full update mechanism that adds a series of "non-repetitive" constraints to the MIP formulation arising from the use of the NCvx update.

We conduct a series of numerical experiments to test the performance of the proposed policies. Overall, we observe that the full update mechanism approximation dominates all others, at the cost of longer solution times. In this regard, the α-optimal approximate policies seem to achieve their objective of attaining good solutions in less time.

This work advances our understanding of sequential bilevel problems with a learning component, but there is still space to make progress. For example, the proposed policies are shown to converge to the full information solution, but from what we observe in our numerical experiments, the bound on time stability is quite loose in practice. In this regard, a question that remains unsolved is whether such a bound is tight (which seems to be the case from our counter-examples), and if not, how can it be improved. Moreover, it is not clear whether the proposed policies are the best, considering our performance measure. In addition, one might be tempted to claim that our setting generalizes the max-min setting of Borrero et al. (2019), who consider uncertainty beyond the Follower's objective. While uncertainty on the parameters defining the Follower's response can be handled through the framework presented in this paper, it is not clear how to model/rationalize uncertainty on the parameters defining the Leader's objective, feasible region or available resources.

Finally, our work assumes that the unknown parameters are time-invariant. Extant work in learning under parametric uncertainty has addressed the issue of time-changing environments, usually by applying policies designed for time-homogeneous settings in epochs, coupled with change-detection strategies. In this regard, the current work might serve as a building block for addressing time-changing environments. These are questions that demand a well-thought answer, driven by evidence from the interdiction applications that motivate our work. These all are exciting directions for future research.

7. Acknowledgments

The research of J.S. Borrero and O.A. Prokopyev was supported by the National Science Foundation grant CMMI-1634835. The research of J.S. Borrero was also supported in part by the grant of the Office of Naval Research (N00014-19-1-2329). The research of O.A. Prokopyev was also supported in part by the grant from the Office of Naval Research (N00014-19-1-2330). The research of D. Sauré was supported in part by the Complex Engineering Systems Institute, ISCI (CONICYT PIA/BASAL AFB180003). The authors would like to thank the Review Team (the Associate Editor and three anonymous referees) for their very constructive and helpful comments.

References

Ahmed, S. and Guan, Y. (2005), 'The inverse optimal value problem', *Mathematical Programming* **102**(1), 91–110.

Audet, C., Hansen, P., Jaumard, B. and Savard, G. (1997), 'Links between linear bilevel and mixed 0–1 programming problems', *Journal of Optimization Theory and Applications* **93**(2), 273–300.

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

Ben-Tal, A., El Ghaoui, L. and Nemirovski, A. (2009), *Robust optimization*, Princeton University Press.

Bertsimas, D. and Dunning, I. (2016), 'Multistage robust mixed-integer optimization with adaptive partitions', *Operations Research* **64**(4), 980–998.

Bertsimas, D. and Georghiou, A. (2015), 'Design of near optimal decision rules in multistage adaptive mixed-integer optimization', *Operations Research* **63**(3), 610–627.

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', *Operations Research* **67**(1), 72–89.

Brown, G., Carlyle, M., Salmerón, J. and Wood, K. (2006), 'Defending critical infrastructure', *Interfaces* **36**(6), 530–544.

Buehn, A. and Eichler, S. (2009), 'Smuggling illegal versus legal goods across the U.S.-mexico border: A structural equations model approach', *Southern Economic Journal* **76**(2), 328–350.

Cao, D. and Chen, M. (2006), 'Capacitated plant selection in a decentralized manufacturing environment: A bilevel optimization approach', *European Journal of Operational Research* **169**(1), 97–110.

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

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

Chern, M. and Lin, K. (1995), 'Interdicting the activities of a linear program: A parametric analysis', *European Journal of Operational Research* **86**(3), 580–591.

Colson, B., Marcotte, P. and Savard, G. (2005), 'Bilevel programming: A survey', *4OR* **3**(2), 87–107.

Colson, B., Marcotte, P. and Savard, G. (2007), 'An overview of bilevel optimization', *Annals of Operations Research* **153**(1), 235–256.

Côté, J.-P., Marcotte, P. and Savard, G. (2003), 'A bilevel modelling approach to pricing and fare optimisation in the airline industry', *Journal of Revenue and Pricing Management* **2**(1), 23–36.

Dempe, S. (2002), *Foundations of bilevel programming*, Springer Science & Business Media.

Fudenberg, D. and Levine, D. (1998), *The Theory of Learning in Games*, Vol. 1, 1 edn, The MIT Press.

Gathmann, C. (2008), 'Effects of enforcement on illegal markets: Evidence from migrant smuggling along the southwestern border', *Journal of Public Economics* **92**(10-11), 1926–1941.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. and Moore, R. (2017), 'Google earth engine: Planetary-scale geospatial analysis for everyone', *Remote Sensing of Environment* .
URL: *https://doi.org/10.1016/j.rse.2017.06.031*

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.

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

Lim, C. and Smith, J. C. (2007), 'Algorithms for discrete and continuous multicommodity flow network interdiction problems', *IIE Transactions* **39**(1), 15–26.

Lorca, Á., Sun, X. A., Litvinov, E. and Zheng, T. (2016), 'Multistage adaptive robust optimization for the unit commitment problem', *Operations Research* **64**(1), 32–51.

Lucotte, M. and Nguyen, S. (2013), *Equilibrium and advanced transportation modelling*, Springer Science & Business Media.

Magliocca, N. R., McSweeney, K., Sesnie, S. E., Tellman, E., Devine, J. A., Nielsen, E. A., Pearson, Z. and Wrathall, D. J. (2019), 'Modeling cocaine traffickers and counterdrug interdiction forces as a complex adaptive system', *Proceedings of the National Academy of Sciences* **116**(16), 7784–7792.

McCormick, G. P. (1976), 'Computability of global solutions to factorable nonconvex programs: Part i — convex underestimating problems', *Mathematical Programming* **10**(1), 147–175.

Modaresi, S., Sauré, D. and Vielma, J. (2020), 'Learning in combinatorial optimization: What and how to explore', *Operations Research* **68**(5), 1285–16240.

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

Sherali, H. D., Soyster, A. L. and Murphy, F. H. (1983), 'Stackelberg-Nash-Cournot equilibria: Characterizations and computations', *Operations Research* **31**(2), 253–276.

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

Steinrauf, R. (1991), Network interdiction models, PhD thesis, Naval Postgraduate School.

Unsal, O. (2010), Two-person zero-sum network-interdiction game with multiple inspector types, Technical report, Naval Postgraduate School.

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. (2019), 'Sequential shortest path interdiction with incomplete information and limited feedback', *Tehcnical report* .

Yürekli, A. and Sayginsoy, Ö. (2010), 'Worldwide organized cigarette smuggling: an empirical analysis', *Applied Economics* **42**(5), 545–561.

Zare, M. H., Borrero, J. S., Zeng, B. and Prokopyev, O. A. (2019), 'A note on linearized reformulations for a class of bilevel linear integer problems', *Annals of Operations Research* **272**(1-2), 99–117.

Appendix A: Proofs of the Formal Results

Lemma 1. *Suppose that $\lambda \in \Lambda$, that Standard Feedback is Value-Perfect or Response-Perfect, and that the Leader uses the Value-Perfect or Response-Perfect update mechanism with $M^t = \emptyset$, respectively. If $\tilde{w}(x^{t,\lambda};\boldsymbol{c}) \neq \tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$, then* $\dim(\mathcal{U}^{t+1}) < \dim(\mathcal{U}^t)$.

Proof of Lemma 1. Assume that $\tilde{w}(x^{t,\lambda};\boldsymbol{c}) \neq w_{GR}(x^{t,\lambda};\mathcal{U}^t)$. This implies that either (i) $y^{t,\lambda} \notin Z_{GR}(x^{t,\lambda};\mathcal{U}^t)$, or that (ii) $y^{t,\lambda} \in Z_{GR}(x^{t,\lambda};\mathcal{U}^t)$ and that there exists $y \in Z_{GR}(x^{t,\lambda};\mathcal{U}^t)$ such that $\boldsymbol{d}^\top y > \boldsymbol{d}^\top y^{t,\lambda}$. Let us first consider the case of Value-Perfect feedback. For that, define

$$A^{t,\lambda} := \left\{a \in A : \exists\, s < t \text{ s.t. } y_a^{s,\lambda} > 0\right\},$$

the set of activities for which c_a is known prior to time t (note that $A^1 = \emptyset$).

Consider first the case when (i) holds. As $y^{t,\lambda} \in Y(x^{t,\lambda})$, then it must be the case that there exists $\tilde{y} \in Y(x^{t,\lambda})$ such that $\max\{\hat{\boldsymbol{c}}^\top \tilde{y} \colon \hat{\boldsymbol{c}} \in \mathcal{U}^t\} < \max\{\hat{\boldsymbol{c}}^\top y^{t,\lambda} \colon \hat{\boldsymbol{c}} \in \mathcal{U}^t\}$. Suppose for a moment that $y_a^{t,\lambda} = 0$ for all $a \notin A^{t,\lambda}$ (we will arrive at a contradiction). Then $\max\{\hat{\boldsymbol{c}}^\top y^{t,\lambda} \colon \hat{\boldsymbol{c}} \in \mathcal{U}^t\} = \sum_{a \in A^t} c_a y_a^{t,\lambda} = \boldsymbol{c}^\top y^{t,\lambda}$, and hence

$$\boldsymbol{c}^\top \tilde{y} \leq \max\{\hat{\boldsymbol{c}}^\top \tilde{y} \colon \hat{\boldsymbol{c}} \in \mathcal{U}^t\} < \max\{\hat{\boldsymbol{c}}^\top y^{t,\lambda} \colon \hat{\boldsymbol{c}} \in \mathcal{U}^t\} = \boldsymbol{c}^\top y^{t,\lambda}.$$

This implies that $y^{t,\lambda} \notin Z(x^{t,\lambda};\boldsymbol{c})$, which contradicts the optimality of the Follower's response. Therefore, it follows that $y_a^{t,\lambda} > 0$ for some $a \notin A^{t,\lambda}$, and the result follows.

Consider now the case when (ii) holds. Then, we have that

$$\tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t) = \boldsymbol{b}^\top x^{t,\lambda} + \boldsymbol{d}^\top y > \boldsymbol{b}^\top x^{t,\lambda} + \boldsymbol{d}^\top y^{t,\lambda} = \tilde{w}(x^{t,\lambda};\boldsymbol{c}),$$

thus it is necessarily the case that $y \notin Z(x^{t,\lambda};\boldsymbol{c})$. Suppose for a moment that $y_a^{t,\lambda} = 0$ for all $a \notin A^{t,\lambda}$ (we will arrive at a contradiction). Then, we have that

$$\boldsymbol{c}^\top \tilde{y} \leq \max\{\hat{\boldsymbol{c}}^\top \tilde{y} \colon \hat{\boldsymbol{c}} \in \mathcal{U}^t\} = \max\{\hat{\boldsymbol{c}}^\top y^{t,\lambda} \colon \hat{\boldsymbol{c}} \in \mathcal{U}^t\} = \boldsymbol{c}^\top y^{t,\lambda},$$

contradicting the fact that $y^{t,\lambda} \notin Z(x^{t,\lambda};\boldsymbol{c})$. This proves the result for the case of Value-Perfect feedback.

Consider now Response-Perfect feedback, and note that $\mathcal{U}^{t+1} \subseteq \mathcal{U}^t \cap \left\{\hat{\boldsymbol{c}} : \hat{\boldsymbol{c}}^\top y^{t,\lambda} = z(x^{t,\lambda};\boldsymbol{c})\right\}$. Thus, if $\dim(\mathcal{U}^{t+1}) = \dim(\mathcal{U}^t)$ then it is necessarily the case that $\hat{\boldsymbol{c}}^\top y^{t,\lambda} = z(x^{t,\lambda};\boldsymbol{c})$ for all $\hat{\boldsymbol{c}}$ in $\mathcal{U}^t$.

Consider first the case when (i) holds. Fix $y \in Z_{GR}(x^{t,\lambda};\mathcal{U}^t)$ and note that

$$\boldsymbol{c}^\top y \leq \max\{\hat{\boldsymbol{c}}^\top y^{t,\lambda} \colon \hat{\boldsymbol{c}} \in \mathcal{U}^t\} < \max\{\hat{\boldsymbol{c}}^\top y^{t,\lambda} \colon \hat{\boldsymbol{c}} \in \mathcal{U}^t\} = \boldsymbol{c}^\top y^t,$$

thus contradicting the fact that $y^{t,\lambda} \in Z(x^{t,\lambda};\boldsymbol{c})$. We conclude that $i)$ can not hold.

Consider now the case when (ii) holds. As in the case of Value-Perfect feedback, we have that

$$\tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t) = \boldsymbol{b}^\top x^{t,\lambda} + \boldsymbol{d}^\top y > \boldsymbol{b}^\top x^{t,\lambda} + \boldsymbol{d}^\top y^{t,\lambda} = \tilde{w}(x^{t,\lambda};\boldsymbol{c}),$$

thus it is necessarily the case that $y \notin Z(x^{t,\lambda};\boldsymbol{c})$. However, if the dimension of the uncertainty set does not reduce, we have that

$$\boldsymbol{c}^\top \tilde{y} \le \max\{\hat{\boldsymbol{c}}^\top \tilde{y} \colon \hat{\boldsymbol{c}} \in \mathcal{U}^t\} = \max\{\hat{\boldsymbol{c}}^\top y^{t,\lambda} \colon \hat{\boldsymbol{c}} \in \mathcal{U}^t\} = \boldsymbol{c}^\top y^{t,\lambda},$$

contradicting the fact that $y^{t,\lambda} \notin Z(x^{t,\lambda};\boldsymbol{c})$. This proves the result for the case of Response-Perfect feedback. ■

Theorem 1. *For any policy $\psi \in \Psi$ and under Standard Feedback, one has that $\tilde{w}(x^{t,\psi};\boldsymbol{c}) \le \tilde{w}^*(\boldsymbol{c}) \le \tilde{w}_E(\mathcal{U}^t)$, for $t \in \mathcal{T}$. In particular, if for some period $t \in \mathcal{T}$ one has that $w_E(\mathcal{U}^t) \le \bar{w}^t$, then $x^{s(\xi),\psi}$ is an optimal solution to the full-information problem, i.e. $\tilde{w}(x^{s(\xi),\psi};\boldsymbol{c}) = \tilde{w}^*(\boldsymbol{c})$.*

Proof of Theorem 1. Let $x^* \in X$ and $y^* \in \arg\min\{\boldsymbol{c}^\top y \colon y \in Y(x^*)\}$ be such that $\boldsymbol{b}^\top x^* + \boldsymbol{d}^\top y^* = \tilde{w}^*(\boldsymbol{c})$. That is, (x^*, y^*) is an optimal solution of the full-information bilevel problem. Because $\boldsymbol{c} \in \mathcal{U}^t$ for all $t \in \mathcal{T}$, then $(x^*, y^*, \boldsymbol{c}) \in \mathcal{S}(\mathcal{U}^t)$, and thus, $\tilde{w}^*(\boldsymbol{c}) \le w_E(\mathcal{U}^t)$ for all $t \in \mathcal{T}$. In addition, because $x^{t',\psi} \in X$ for all $t' \in \mathcal{T}$, then, from the definition of x^*, we have that $\tilde{w}(x^{t',\psi};\boldsymbol{c}) \le \tilde{w}^*(\boldsymbol{c})$ for any given $t' \in \mathcal{T}$. These observations imply that for any $t, t' \in \mathcal{T}$ we have the following chain of inequalities

$$\tilde{w}(x^{t',\psi};\boldsymbol{c}) \le \tilde{w}^*(\boldsymbol{c}) \le \tilde{w}_E(\mathcal{U}^t). \tag{A-1}$$

Hence, since $\tilde{w}_E(\mathcal{U}^t) \le \bar{\tilde{}}w(x^{t,\psi};\boldsymbol{c})$, then $\tilde{w}_E(\mathcal{U}^t) \le \tilde{w}(x^{s,\psi};\boldsymbol{c})$, and Equation (A-1) implies that $\tilde{w}(x^{s,\psi};\boldsymbol{c}) = \tilde{w}^*(\boldsymbol{c})$, which gives the desired result. ■

Lemma 2. *Let $\psi \in \Psi$ and $t \in \mathcal{T}$ be given and assume that the Leader implements the full update mechanism. If $x^{t,\psi} \in \bigcup_{s<t}[x^{s,\psi}]$, then $\tilde{w}_E(\mathcal{U}^t) = \tilde{w}(x^{t,\psi};\boldsymbol{c})$.*

Proof of Lemma 2. First, observe that $x^{t,\psi} \sim x^{s,\psi}$ for some $s < t$ implies that $y^{t,\psi} = y^{s,\psi}$. Therefore,

$$\tilde{w}(x^{t,\psi};\boldsymbol{c}) = \boldsymbol{b}^\top x^{t,\psi} + \boldsymbol{d}^\top y^{t,\psi} = \boldsymbol{b}^\top x^{t,\psi} + \boldsymbol{d}^\top y^{s,\psi} = \boldsymbol{b}^\top x^{t,\psi} + \tilde{w}(x^{s,\psi};\boldsymbol{c}) - \boldsymbol{b}^\top x^{s,\psi}. \tag{A-2}$$

On the other hand, $\tilde{w}_E(\mathcal{U}^t) = \boldsymbol{b}^\top x^{t,\psi} + \boldsymbol{d}^\top y^{t,E}$, and by optimality in problem (10), $\tilde{w}_E(\mathcal{U}^t) \ge \tilde{w}(x^{t,\psi};\boldsymbol{c})$, which, in view of equation (A-2) is equivalent to say that

$$\boldsymbol{d}^\top y^{t,E} \ge \tilde{w}(x^{s,\psi};\boldsymbol{c}) - \boldsymbol{b}^\top x^{s,\psi}.$$

We prove that the above equation holds as an equality, i.e., $\boldsymbol{d}^\top y^{t,E} = \tilde{w}(x^{s,\psi};\boldsymbol{c}) - \boldsymbol{b}^\top x^{s,\psi}$, from which the result follows. Indeed, by the definition of the full update, for any $\hat{\boldsymbol{c}} \in \mathcal{U}^t$ there exists $y(\hat{\boldsymbol{c}}) \in \arg\min\{\hat{\boldsymbol{c}}^\top y' \colon y' \in Y(x^{s,\psi})\}$ such that

$$\tilde{w}(x^{s,\psi};\boldsymbol{c}) - \boldsymbol{b}^\top x^{s,\psi} = \boldsymbol{d}^\top y(\hat{\boldsymbol{c}}) \geq \boldsymbol{d}^\top y \quad \forall\, y \in \arg\min\{\hat{\boldsymbol{c}}^\top y' \colon y' \in Y(x^{s,\psi})\}. \tag{A-3}$$

Let $(x^{t,\psi}, y^{E,t}, \boldsymbol{c}^{E,t})$ be an optimal solution of problem (10). Then, because $Y(x^{t,\psi}) = Y(x^{s,\psi})$, it follows from (A-3) and the optimality of $(x^{t,\psi}, y^{E,t}, \boldsymbol{c}^{E,t})$ that $\boldsymbol{d}^\top y^{t,E} = \tilde{w}(x^{s,\psi};\boldsymbol{c}) - \boldsymbol{b}^\top x^{s,\psi}$, as desired. ■

Lemma 3. *Let $t \in \mathcal{T}$, assume that the Leader implements a policy $\psi \in \Psi$, and that $\bar{w}^t < \tilde{w}_E(\mathcal{U}^t)$. Then:* (i) *$\boldsymbol{c} \in \mathcal{L}^t$;* (ii) *if $z^{t,E} \leq z(x^t;\boldsymbol{c})$ then $c^{t,E} \notin \mathcal{L}^t$; and* (iii) *if $z^{t,E} > z(x^t;\boldsymbol{c})$ and the feedback is Value-Perfect (Response-Perfect), then* $\dim(\mathcal{U}^{t+1}) < \dim(\mathcal{U}^t)$*, $c^{t,E} \notin \mathcal{V}^t$ ($c^{t,E} \notin \mathcal{R}^t$).*

Proof of Lemma 3. For (i) observe that as $\tilde{w}(x^{t,\psi};\boldsymbol{c}) < \tilde{w}_E(\mathcal{U}^t)$, then $y^{t,E} \notin Z(x^{t,\psi};\boldsymbol{c})$. This follows by contradiction: If $y^{t,E} \in Z(x^{t,\psi};\boldsymbol{c})$, then by the optimistic assumption it would follow that $\tilde{w}_E(\mathcal{U}^t) = \tilde{w}(x^{t,\psi};\boldsymbol{c})$, yielding a contradiction. Hence, it can be concluded that $\boldsymbol{c}^\top y^{t,E} > z(x^{t,\psi};\boldsymbol{c})$ as desired. For (ii) if $z^{t,E} \leq z(x^{t,\psi};\boldsymbol{c})$, then by definition $(\boldsymbol{c}^{t,E})^\top y^{t,E} \leq z(x^{t,\psi};\boldsymbol{c})$ and it is clear that this implies that $\boldsymbol{c}^{t,E} \notin \mathcal{L}^t$.

In order to prove (iii) we consider the Value-Perfect and Response-Perfect separately. For Value-Perfect, define $\widetilde{A}^t = \{a \in A \colon \hat{c}_a = c_a \text{ for all } \hat{\boldsymbol{c}} \in \mathcal{U}^t\}$. We show that if $z^{t,E} > z(x^{t,\psi};\boldsymbol{c})$ then there exists an $a \notin \widetilde{A}^t$ such that $y^t_a > 0$; such existence implies that $\dim(\mathcal{U}^{t+1}) < \dim(\mathcal{U}^t)$. The proof is by contradiction. Suppose that $y^t_a = 0$ for all $a \notin \widetilde{A}^t$. Then, for any $\hat{\boldsymbol{c}} \in \mathcal{U}^t$, $\hat{\boldsymbol{c}}^\top y^t = \boldsymbol{c}^\top y^t = z(x^{t,\psi};\boldsymbol{c})$. In particular $\boldsymbol{c}^{t,E} \in \mathcal{U}^t$, hence we would have that $(\boldsymbol{c}^{t,E})^\top y^t = z(x^{t,\psi};\boldsymbol{c}) < z^{t,E} = (\boldsymbol{c}^{t,E})^\top y^{t,E}$. This implies that, $y^{t,E} \notin Z(x^{t,\psi};\boldsymbol{c}^{t,E})$, which is a contradiction. Observe that the same argument shows that the fact that $\boldsymbol{c}^{t,E} \in \mathcal{V}^t$ yields a contradiction.

For Response-Perfect feedback, suppose that the result does not hold, thus $\dim(\mathcal{U}^{t+1}) = \dim(\mathcal{U}^t)$. This implies that y^t is linearly dependent of $y^1, \ldots, y^{t-1}$ and hence

$$\{\hat{\boldsymbol{c}} \in \mathbb{R}^{|A|} \colon (y^s)^\top \hat{\boldsymbol{c}} = z(x^{s,\psi};\boldsymbol{c}),\ s \leq t-1\} = \{\hat{\boldsymbol{c}} \in \mathbb{R}^{|A|} \colon (y^s)^\top \hat{\boldsymbol{c}} = z(x^{s,\psi};\boldsymbol{c}),\ s \leq t\}. \tag{A-4}$$

In particular, since $\boldsymbol{c}^{t,E} \in \bigcap_{s \leq t-1} \mathcal{R}^s$, then $\boldsymbol{c}^{t,E} \in \{\hat{\boldsymbol{c}} \in \mathbb{R}^{|A|} \colon (y^s)^\top \hat{\boldsymbol{c}} = z(x^{s,\psi};\boldsymbol{c}),\ s \leq t-1\}$, and by Equation (A-4) it follows that $(\boldsymbol{c}^{t,E})^\top y^t = z(x^{t,\psi};\boldsymbol{c})$. Now, as we are assuming that $z^{t,E} > z(x^{t,\psi};\boldsymbol{c})$, it follows that $(\boldsymbol{c}^{t,E})^\top y^{t,E} > (\boldsymbol{c}^{t,E})^\top y^t$, i.e., that $y^{t,E} \notin Z(x^{t,\psi};\boldsymbol{c}^{t,E})$, which is a contradiction. Finally, observe that the above arguments imply that $\boldsymbol{c}^{t,E} \notin \mathcal{R}^t$. ■

Appendix B: Summary and Comparative Analysis of Borrero et al. (2019) (BPS19)

We describe the setting in BPS19 adopting the notation in Section 2. BPS19 considers a Leader and Follower that interact sequentially during T periods: at period $t \in \mathcal{T} = \{1, \ldots, T\}$ the Leader acts first by selecting x_r^t, the usage level of each resource r in a set R; after observing the Leader's decision, the Follower selects y_a^t, the usage level of each activity a in a set A. Like in our work, the Follower's response $y^t := (y_a^t : a \in A)$ to the Leader decision $x^t := (x_r^t : r \in R)$ lies in the rational reaction set

$$Z(x^t; \boldsymbol{c}) := \arg\min\left\{\boldsymbol{c}^\top y \colon y \in Y(x^t)\right\}, \quad \text{with } Y(x) = \left\{y \in \mathbb{R}_+^{|A|} \colon \boldsymbol{F}y + \boldsymbol{L}x \leq \boldsymbol{f}\right\}, \tag{B-5}$$

where it is assumed that all parameters above are known to the Follower upfront, so that y^t can be computed upon observing the value of x^t. The Leader's decision x^t is constrained to lie within a region $X := \left\{x \in \mathbb{R}_+^{|R|-I} \times \mathbb{Z}_+^I \colon \boldsymbol{H}x \leq \boldsymbol{h}\right\}$, where $I \leq |R|$. Unlike in our work, BPS19 assumes that the Leader's profit in any period is the Follower's loss, i.e., the Leader's profit $w(x^t, y^t)$ in period t is given by

$$w(x^t, y^t) = \boldsymbol{c}^\top y^t, \quad y^t \in Y(x^t), \, x^t \in X.$$

Considering the above, BPS19 defines the full-information problem as follows

$$\tilde{w}^*(\boldsymbol{c}) = \max\left\{\boldsymbol{c}^\top y : y \in Z(x, \boldsymbol{c}), \, x \in X\right\}.$$

Note that BPS19 implicitly assumes the optimistic approach to bilevel programming. Defining $\tilde{w}(x; \boldsymbol{c}) := \max\left\{\boldsymbol{c}^\top y \colon y \in Z(x; \boldsymbol{c})\right\}$ for $x \in X$, we have that, for a given sequence of decisions $\{x^t : t \in \mathcal{T}\}$, the Leader's total cumulative profit is given by

$$\mathcal{P}(\left\{x^t : t \in \mathcal{T}\right\}; \boldsymbol{c}) := \sum_{t \in \mathcal{T}} \tilde{w}(x^t; \boldsymbol{c}).$$

BPS19 assumes that all parameters are known upfront by the Follower. On the Leader's side, a *cost uncertainty* model assumes that $\boldsymbol{c}$ is known to lie within an initial polyhedral uncertainty set $\mathcal{U}^1$, and that only a subset of rows and columns from F and L are known upfront, with unknown columns and rows revealed by the Follower as part of the feedback. A more general *Matrix uncertainty* model allows for uncertainty on the coefficients of matrices F and L; here, we discuss the cost uncertainty model, and consider the case where all columns and row of L and F are initially known, so as to facilitate the comparison between the settings and results. Note that, even under these conditions, the informational settings are not directly comparable, as even if one is to set $\boldsymbol{b} = \boldsymbol{c}$ and $\boldsymbol{d} = 0$, our work assumes that $\boldsymbol{b}$ is known upfront by the Leader.

BPS19 measures policy performance in terms of time stability, puts emphasis in finding weakly optimal policies, and defines the concept of standard, value-perfect and response-perfect feedback, which we adopt in our work. Unlike in our setting, where updates might result in non-convex uncertainty sets, BPS19 considers specific updates under value- and response-perfect feedback, which preserve the polyhedral nature of the uncertainty set.

The authors propose a family of greedy and robust policies, which on each period implement

$$x^t \in \arg\max\left\{\tilde{w}_R(x;\mathcal{U}^t) := \min\left\{\max\left\{\hat{c}^\top y : \hat{c} \in \mathcal{U}^t\right\} : y \in Y(x)\right\} : x \in X\right\},$$

until achieving time-stability (which can be checked in real-time). Letting z^t the profit observed at time t, Theorem 1 in BPS19 shows that $\tilde{w}(x^t;\boldsymbol{c}) \leq \tilde{w}^*(\boldsymbol{c}) \leq \tilde{w}_R(x^t;\mathcal{U}^t)$. This result is akin to Theorem 1, and follows from the robust nature of their policies, in the same manner our result follows from the *best-case* nature of our policies.

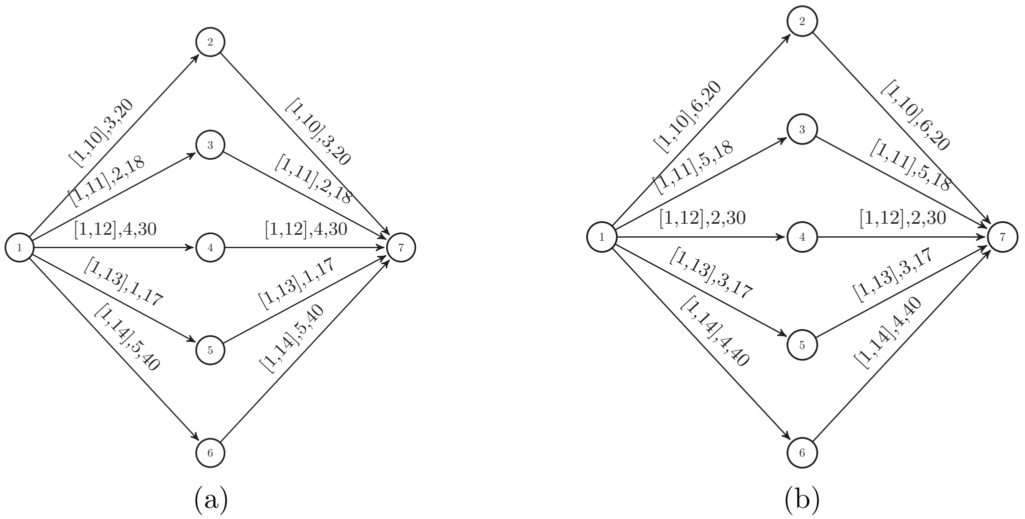
For the case of value-perfect feedback, BPS19 shows that the dimension of the uncertainty set is reduced every time the Leader's expectation is not met, which implies an immediate bound on the time-stability. In our work, refining the definition of the Leader's expectation, Lemma 1 establishes a similar result. However, this does not translate into an bound on time-stability, because no much can be inferred from the feedback when the Leader's expectation is met. This fact, illustrated in Example 2, is key to understanding the challenges in the asymmetric case, relative to BPS19.

Constructing a worst-case instance, BPS19 shows that time-stability of the proposed policies is of the order of the number of activities, under value-perfect feedback; in our work, no such a result is established, as it is shown that feasible upper-bounds on time-stability are exponential in the work case. An important distinctive feature of our work in this regard, is that partial convergence results need to be qualified by the type of update being use. In particular, our results show that partial convergence results (e.g., Theorem 2) do not necessarily hold when update mechanisms other than the full update are used.

On a more practical note, BPS19 provides MIP formulations for solving for x^t when their polyhedral-preserving updates mechanisms are used (provided that standard feedback is value- and/or response-perfect). In our work, we present such formulations for the more general case of standard feedback, and show how the polyhedral structure might be preserved under the different update mechanisms, in particular when an approximate full-date is used (under which, we show, the dimension-reduction results hold).

Appendix C: Additional examples for the discussion in Section 3

Next, we provide additional illustrations for the discussion in Example 2. Figure 6a shows an example of an instance of the ASPI where $\tilde{w}^*(\boldsymbol{c}) < \tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$. Here, the solution for any policy $\lambda \in \Lambda$ is to block again arcs $(1,2)$ and $(1,3)$, and as before, the Leader expects that the Follower uses path 1–4–7. This yields an expected profit of $\tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t) = 60$. The full-information optimal solution, however, is to remove $(1,3)$ and $(1,5)$. This makes the Follower use path 1–2–7, and gives a profit of $\tilde{w}^*(\boldsymbol{c}) = 40$, hence $\tilde{w}^*(\boldsymbol{c}) < \tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$.



(a) (b)

Figure 6 Example of instances when (a) $\tilde{w}^*(\boldsymbol{c}) < \tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$ and (b) when $\tilde{w}(x^{t,\lambda};\boldsymbol{c}) = \tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$ does not imply that $\tilde{w}(x^{t,\lambda};\boldsymbol{c}) = \tilde{w}^*(\boldsymbol{c})$, with $\tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t) < \tilde{w}^*(\boldsymbol{c})$. The arcs' labels are given by $[\ell_a, u_a], c_a, d_a$.

Finally, Figure 6b shows an example of an instance of the ASPI where the fact that $\tilde{w}(x^{t,\lambda};\boldsymbol{c}) = \tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$ does not imply that $\tilde{w}(x^{t,\lambda};\boldsymbol{c}) = \tilde{w}^*(\boldsymbol{c})$. In this case it is readily checked that the solution for any policy $\lambda \in \Lambda$ is to block arcs $(1,2)$ and $(1,3)$. The Leader expects that the Follower uses path 1–4–7, which yields an expected profit of $\tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t) = 60$. For this example the response of the Follower is the same the Leader expects, that is, to use 1–4–7, and hence $\tilde{w}(x^{t,\lambda};\boldsymbol{c}) = 60 = \tilde{w}_{GR}(x^{t,\lambda};\mathcal{U}^t)$. However, the optimal full-information solution for the Leader is to remove the arcs $(1,4)$ and $(1,5)$ which forces the Follower to use path 1–6–7 and gives an optimal profit of $\tilde{w}^*(\boldsymbol{c}) = 80$.

Appendix D: Additional tables and graphs

D.1. Average solution times for benchmark policies.

Feedback	Solution Time (seconds)						
Uniform	ψ	π_{R-Cvx}	π_{R-Ncvx}	π_{C-Cvx}	π_{C-Ncvx}	ψ_N	λ
Standard	12.28	3.81	9.80	2.89	5.12	1.47	**0.15**
VP	6.61	3.31	8.38	1.31	3.28	1.35	**0.50**
RP	8.00	3.06	9.16	2.46	12.87	1.43	**0.46**
Left Sk.	ψ	π_{R-Cvx}	π_{R-Ncvx}	π_{C-Cvx}	π_{C-Ncvx}	ψ_N	λ
Standard	11.00	4.07	9.57	0.79	2.69	1.80	**0.15**
VP	4.57	3.38	8.25	1.05	2.45	1.16	**0.53**
RP	8.22	3.25	9.54	1.18	7.67	1.79	**0.55**
Right Sk.	ψ	π_{R-Cvx}	π_{R-Ncvx}	π_{C-Cvx}	π_{C-Ncvx}	ψ_N	λ
Standard	15.95	4.51	10.60	1.05	2.14	1.92	**0.15**
VP	7.19	3.20	8.07	0.95	2.16	1.41	**0.42**
RP	10.81	3.13	9.36	1.19	7.57	1.62	**0.25**

Table 7 Average solution time across ten instances for ψ with the E-Ncvx update and the benchmark policies. Boldface indicates the best result.

Observe that the policies that use the NCvx update take on average more time to solve, and particularly, policies in Ψ have the longest solution times. By contrast, ψ_N achieves shorter solution times, which hints that exploiting the information of the NCvx update is the main driver of the longer solution times of policies in Ψ. Policies in Λ attain the shortest solution times by far, even faster than the Cvx counterparts of the random and center policies. It should be mentioned, however, that a plausible reason for this difference is that the policies in Λ tend to stall very early, while the random, center, or even policies in Ψ under the Cvx update (as in, e.g., Table 1) stall later on average or optimize across all the periods in $\mathcal{T}$.

D.2. Sensitivity to the Instance Size.

We use π_R, policies in Λ, and the α-optimal policies, $\Psi_{1.5}$ and Ψ_2 under the E-Ncvx update, to study the time it takes to solve increasingly larger instances. Networks with four different structures ($n_\ell \times n_k$) were considered: 5×4, 5×5, 6×5, and 6×6. Over each structure we ran five replications across all feedback types (the remaining parameters being as in Section 5.3.1). We set $T = 15$ for the instances with five layers and $T = 20$ for the instances with six layers, and run each instance for at most one hour. Table 8 summarizes the average running times; the remaining performance measures follow similar patterns as shown in the previous experiments.

From Table 8 we observe that only policies in Λ scale well with the size of the instance. As with previous experiments, this is partly explained because these policies stall fairly quickly, thus, the larger MIPs corresponding to larger values of t, do not need to be solved. The performance of the other policies clearly deteriorates, particularly when only Standard feedback is available,

Instance size	Solution Time (seconds)			
5×4 ($\lvert N\rvert$=22,$\lvert A\rvert$=72)	π_R	λ	$\psi_{1.5}$	ψ_2
Standard	37.72	0.27	520.61	494.24
VP	16.63	1.01	23.10	24.78
RP	18.04	0.87	22.17	12.59
5×5 ($\lvert N\rvert$=27,$\lvert A\rvert$=110)	π_R	λ	$\psi_{1.5}$	ψ_2
Standard	133.77	0.45	763.33	*722.13 (1,15.4)*
VP	49.27	1.10	124.00	10.17
RP	72.90	2.08	67.68	17.99
6×5 ($\lvert N\rvert$=32,$\lvert A\rvert$=135)	π_R	λ	$\psi_{1.5}$	ψ_2
Standard	*1267.67 (2,20)*	0.51	*2884.08 (4,13.4)*	*310.1 (1,19.4)*
VP	282.69	2.62	*2883.22 (4,14)*	747.86
RP	270.76	1.72	480.65	66.56
6×6 ($\lvert N\rvert$=38,$\lvert A\rvert$=192)	π_R	λ	$\psi_{1.5}$	ψ_2
Standard	*978.66 (1,19.8)*	0.72	*2618.45 (4,9)*	*2166.97 (3,14)*
VP	332.52	2.41	*2252.24 (3,14.2)*	433.09
RP	285.91	3.87	*1144.38 (1,20.2)*	188.00

Table 8 Average solution times for different instance sizes. Italics denote that not all instances were solved within the time limit. In these cases (a, b) denotes the number of instances not solved within the time limit (a), and the average period by which the time limit was reached (b).

while for Response-Perfect feedback, the decrease in performance is less noticeable (recall that in Response-Perfect feedback, the NCvx update is equivalent to the Cvx update and hence more tractable computationally). These results show that for policies in Ψ, state-of-the-art MIP solvers methods can provide solutions for small to medium scale instances. The solution of larger instances, particularly under Standard feedback, requires additional algorithms or pre-processing techniques.

D.3. Additional tables for the Mexico-USA border smuggling instances

We show the results for two of the three smuggling instances considered in Section 5.3.4 in Tables 9 and 10.

		Time-Stability			Regret			Solution time		
	Feedback	π_R	λ	ψ	π_R	λ	ψ	π_R	λ	ψ
$k=2$	S	13	∞	8	1.53	2.62	1.32	33.00	0.25	39.15
	RP	21	∞	8	1.67	2.62	1.32	29.95	0.29	26.42
	VP	10	∞	5	0.95	2.62	0.81	26.51	0.42	15.12
$k=3$	S	17	1	11	0.89	0.00	2.07	31.91	0.31	61.38
	RP	20	1	10	0.64	0.00	1.75	35.70	0.42	47.38
	VP	18	1	7	0.65	0.00	1.21	30.07	0.86	25.67
$k=4$	S	15	∞	5	3.34	4.00	1.53	57.42	0.50	72.25
	RP	17	4	5	3.31	0.88	1.53	48.62	2.20	57.19
	VP	9	2	4	1.81	0.20	1.24	46.32	1.48	30.36

Table 9 Time stability, regret, and solution time for one of the smuggling instances with various feedback modes and values of k. A time stability of ∞ means that the policy stalls.

		Time-Stability			Regret			Solution time		
	Feedback	π_R	λ	ψ	π_R	λ	ψ	π_R	λ	ψ
$k=2$	S	5	2	4	2.44	0.59	2.16	26.97	0.43	8.30
	RP	4	∞	4	1.76	12.30	2.16	29.08	0.69	9.86
	VP	7	∞	6	3.70	11.76	3.48	28.08	0.21	22.42
$k=3$	S	14	5	6	1.95	2.31	4.63	27.89	2.05	18.98
	RP	15	1	7	3.35	0.00	5.30	33.78	0.51	31.79
	VP	16	1	7	1.05	0.00	5.30	33.32	0.22	29.61
$k=4$	S	7	4	4	5.28	1.98	4.16	48.76	3.74	30.14
	RP	8	∞	4	4.91	9.31	4.16	54.45	1.49	36.32
	VP	2	∞	5	0.71	14.12	4.16	55.29	0.43	34.95

Table 10 Time stability, regret, and solution time for one of the smuggling instances with various feedback modes and values of k. A time stability of ∞ means that the policy stalls.