Sequential Interdiction with Incomplete Information and Learning

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 present a framework for a class of sequential decision-making problems in the context of max-min bilevel programming, where a *leader* and a *follower* repeatedly interact. At each period, the leader allocates resources to disrupt the performance of the follower (e.g., as in defender-attacker or interdiction problems), who in turn minimizes some cost function over a set of activities that depends on the leader's decision. While the follower has complete knowledge of his problem, the leader has only partial information, and needs to learn about the cost parameters, available resources, and the follower's activities from the feedback generated by the follower's actions. We measure policy performance in terms of its *time-stability*, defined as the number of periods it takes for the leader to match the actions of an *oracle* with complete information. In particular, we propose a class of *greedy* and *robust* policies and show that these policies are *weakly optimal*, eventually match the oracle's actions, and provide a real-time certificate of optimality. We also study a lower bound on any policy performance based on the notion of a *semi-oracle*. Our numerical experiments demonstrate that the proposed policies consistently outperform reasonable benchmark, and perform fairly close to the semi-oracle.

Key words: bilevel programming, attacker-defender, interdiction, learning, incomplete information, online optimization, robust optimization

1. Introduction

Bilevel optimization deals with problems where a subset of the *lower-level* decisions are constrained to be a solution of another mathematical program that depends on the remaining *upper-level* decisions. This general structure makes bilevel programs useful for modeling hierarchical decision-making problems between multiple, typically, two actors, commonly referred to as the *leader* (an upper-level decision-maker) and the *follower* (a lower-level decision-maker), see Dempe (2002). In this perspective, the leader solves an optimization problem that depends on the optimal solution to the follower's problem, and this latter problem is, in turn, parameterized by the leader's decisions. Bilevel programs are used in several application areas such as law enforcement (Morton et al. 2007), defense (Brown et al. 2006), economics (Sherali et al. 1983), transportation (Lucotte

and Nguyen 2013), energy (Bard et al. 2000), revenue management (Côté et al. 2003), among others; see Dempe (2002), Colson et al. (2005, 2007) and the references therein.

An important class of bilevel programs, known as *max-min* problems, deals with settings where the leader and follower are adversaries. More precisely, in these problems the leader's objective is to maximally degrade the performance of the follower. As an example, consider network flow interdiction problems, which have applications in military and smuggling prevention settings. Here, the follower operates a network with the objective to move between two fixed vertices through a shortest path (Fulkerson and Harding 1977, Israeli and Wood 2002), to send the maximum flow possible between two vertices (Ratliff et al. 1975, Wood 1993), or more generally, to move flow in the network at minimum cost subject to some demand balance constraints (Chern and Lin 1995, Smith and Lim 2008). The leader, by using the resources at her disposition, can block (either totally or partially) a limited number of arcs and nodes in the network. Her objective is to allocate her resources so as to maximize the length of the follower's shortest path, minimize the maximum flow, or maximize the minimum cost incurred by the follower, respectively. These types of models are also used in surveillance settings, where the leader places resources (e.g., sensors) in a network so as to minimize the follower's probability of evasion, see Morton et al. (2007).

Network flow interdiction models belong to a larger class of Attacker-Defender (AD) or Defender-Attacker (DA) models (Brown et al. 2006, Wood 2011). In a typical AD setting, an attacker (the leader) and a defender (the follower) interact during a war-time confrontation: the attacker allocates her forces so as to disable assets of the defender's infrastructure; the defender decides how to operate his system at minimum cost given the restrictions set by the leader's attack. The leader decides her allocation with the objective to maximize the defender's operational costs. Conversely, in a DA model, a defender (the leader) allocates her limited defensive resources to protect her assets, and an attacker (the follower), for a given defensive configuration, seeks for the most effective attacks. Here, the defender's objective is to allocate her resources so as to minimize the effectiveness of the attacks. In general, AD and DA models can be casted as max-min bilevel programs to model decisions in a broad range of application areas: see, e.g., Salmeron et al. (2004), Brown et al. (2005).

Typical formulations of max-min bilevel problems in the literature assume a single interaction between the leader and the follower, and that either the leader knows all the parameters of the follower's problem (as in the references discussed above), or that she knows a probability distribution over the set of problem configurations and parameters (see, e.g., Held et al. (2005), Janjarassuk and Linderoth (2008)). Hence, these models solve a single (possibly stochastic) max-min bilevel problem, assuming that even if the leader and the follower interact across several periods, the leader would implement the resulting *full-information* solution at every time period. In contrast, many applications inherently involve multiple interactions between the leader and

the follower (e.g., as in smuggling interdiction and AD-DA problems). More importantly, in these problems the leader does not always know with certainty the system that the follower operates, and cannot estimate it (a priori) reliably due to the adversarial nature of their confrontation. Consequently, she has *incomplete information* of the problem solved by the follower at each time period, and has to learn about it through time by observing the follower's reactions to her actions.

Departing from the existing literature, this paper studies *sequential max-min problems with incomplete information* (**SMPI**). In these problems, the leader and follower interact *repeatedly*: at each stage the leader implements a set of actions and then observes the follower's reaction; from the information, or *feedback*, she gets from the follower's response, the leader (potentially) updates her knowledge of the follower's problem, and incorporates this information into her decision-making process. Observe that in **SMPI**, besides determining how to allocate her resources, the leader faces additional questions outside the scope of traditional bilevel models, as she needs to recognize whether a given upper-level solution is the best possible, she needs to force the follower to disclose as much information as possible, and needs to exploit this newly learned information to best re-allocate her resources in future periods. Therefore, given the leader's limited knowledge of the follower's problem, at each time period she faces a form of the *exploitation vs. exploration trade-off*: she must choose either to exploit the current information so as to maximize her immediate reward, or to explore solutions that albeit not being maximally rewarding, may reveal new information that can be used to implement better solutions in future periods.

For the reasons above, **SMPI** can be viewed as a class of *online optimization problems*. In said problems, at each period a decision-maker with incomplete knowledge of the problem's structure has to choose a solution from a fixed action set, incurring a cost. She learns about the problem's structure and data from the feedback she collects, which depends on the implemented solution (Cesa-Bianchi and Lugosi 2006). Performance is typically measured in terms of *regret*, i.e., the difference between the costs incurred by the decision-maker and those incurred by an *oracle*, i.e., a decision-maker with complete up-front knowledge of the problem. In this regard, **SMPI** can be framed as an adversarial multi-armed bandit problem (Auer et al. 2002, Audibert and Bubeck 2009). However, naively using bandit policies would result in regret bounds that are *exponential in the primitives of the **SMPI***. This follows as the number of solutions of a bilevel linear problem is typically exponentially large in the number of its variables and constraints, in the worst-case (Dempe 2002).

Multi-armed bandits do not yield polynomial regret bounds for **SMPI** as they make no specific assumptions about the relationship between the actions of the decision-maker and the costs associated with these actions. In this sense, online models with particular structures have been studied in the literature, see for instance, online convex (Zinkevich 2003, Hazan 2015), online combinatorial (Cesa-Bianchi and Lugosi 2012, Audibert et al. 2013), and online linear (Agrawal

et al. 2014) models. However, these models assume a *single-level relationship* between the decision-maker's actions and the costs she observes. As a consequence, **SMPI** does not fit these frameworks due to the *hierarchical* (and generally non-convex) relationship between the leader actions and the responses she observes.

Given the limitations of current online models, in this paper we develop a general framework for **SMPI**. We represent the leader's and follower's decisions in terms of *resources* and *activities*, respectively. Initially, the leader does not know all the follower's activities and constraints, and as such, she might not know all of her resources or constraints. The leader learns about an unknown follower's activity as soon as she observes him *performing* it, and at the same time learns about all the lower-level constraints that *restrict* this activity, all the leader's resources that *interfere* with that activity, and all upper-level constraints associated with the newly learned resources.

From a technical point of view, we first make the assumption that for every activity, resource, and constraint she knows, the leader also knows the corresponding entries in the upper and lower-level constraint matrices and the right-hand side vectors in a typical bilevel programming formulation of the full-information problem. However, we suppose that the leader does not know with certainty the components of the follower's cost vector for the activities she knows; she only knows that they belong to certain (polyhedral) *uncertainty set*. Furthermore, in Section 4 we analyze a more general uncertainty model, where the uncertainty extends beyond the follower's cost vector.

Besides learning new activities, resources, and constraints, the leader can also observe additional information of the follower's problem from his response. In this sense, we introduce the notions of *Standard feedback*, and its specializations, *Value-Perfect* and *Response-Perfect feedbacks*. In Standard feedback, the leader observes the total cost the follower incurs at each time period; in Value-Perfect feedback she also observes the cost coefficient associated with each activity used by the follower at that time, while in Response-Perfect feedback she also observes the value of the decision vector for the activities performed by the follower.

We measure the performance of the leader's decision-making policy in terms of its *time-stability*. This is defined as the first time period by which the cost the follower incurs coincides with the best possible cost an *oracle* leader with complete knowledge of the problem attains from there on. Time-stability is closely related to the notion of regret used in online optimization; in particular any upper bound on the time-stability of a policy implies an upper bound in the regret of that policy .

In this paper we analyze a set of *greedy* and *robust* policies, which we denote by Λ. The policies are greedy because at any time they exploit the leader's information of the follower's problem so as to maximize the follower's costs at the current time period, and they are robust because they assume that the follower's cost vector realizes its worst case for the follower. For these reasons,

implementing the policies in Λ involve s solving at each time a max-min bilevel problem with lower-level robustness constraints. Hence their computation requires both bilevel and robust optimization techniques: we develop a method that first replaces the lower-level robust optimization problem by its equivalent linear program counterpart (Ben-Tal et al. 2009), and then reformulates the resulting linear bilevel program as a one-level mixed integer program (Audet et al. 1997).

We demonstrate that the time-stability of policies in Λ under Value-Perfect and Response-Perfect feedback is upper bounded by the number of follower's activities. We show that these policies are *optimal* in the sense that they attain the best possible worst-case time-stability across all possible problem instances. Furthermore, they provide a *certificate of optimality* in real time. We also develop a method to provide a lower bound for the time-stability of any policy based on the concept of a *semi-oracle*. The semi-oracle has full information of the problem beforehand, but can only use the leader's resources whose existence has been revealed by the follower's actions. As such, the semi-oracle combines the knowledge of the standard oracle with the practical limitations of the leader and thus, provides an informative lower bound on the performance of any policy. Our numerical results show that the policies in Λ consistently outperform reasonable benchmark, and perform reasonably close to the semi-oracle.

The present work is connected to the shortest-path interdiction model with incomplete information discussed in Borrero et al. (2016). Particularly, the aforementioned model can be viewed as an **SMPI** where the bilevel problem is a shortest-path interdiction problem, the feedback is Value-Perfect, the uncertainty set is a hypercube, and there is incomplete information only about the follower's cost vector. In this sense, our work generalizes the results of Borrero et al. (2016) to account for general max-min bilevel linear models with Value-Perfect feedback, polyhedral uncertainty sets, and uncertainty in the follower's cost vector. Moreover, we extend many of these results to **SMPI** problems where feedback is Response-Perfect or Standard and where there is uncertainty in the follower's constraint matrix. In addition, in this work we measure the performance of policies in terms of worst-case time-stability rather that in terms of *efficiency* (see Borrero et al. (2016)); the former is a more transparent and informative measure of performance, see Section 2.1.

The remainder of the paper is organized as follows. In Section 2 we provide a mathematical formulation of the problem. Section 3 discusses greedy and robust policies, while Section 4 extends most of the results of greedy and robust policies for the case of uncertainty in the lower-level constraint matrix. Section 5 discusses the semi-oracle benchmark and Section 6 presents numerical experiments. The proofs of the main results are provided in the manuscript; supporting material and the remaining proofs are given in the online supplement.

2. Basic Model: Cost Uncertainty

We study an adversarial decision-making process where a *leader* and a *follower* sequentially interact, and where the leader has incomplete information of the follower's problem. Before describing this model, first consider the *single-stage model with full information*. Here, the leader can use any *resource* $i \in I$, $|I| < \infty$, and for each $i \in I$ she chooses a value $x_i \geq 0$ such that $x \coloneqq (x_i \colon i \in I) \in X$, where X denotes the set of feasible resource levels. We let C_L denote the set of constraints faced by the leader and assume that X is given by

$$X \coloneqq \{x \in \mathbb{Z}_+^k \times \mathbb{R}_+^{|I|-k} \colon \boldsymbol{H}x \leq \boldsymbol{h}\},$$

where $0 \leq k \leq |I|$, $\boldsymbol{H} \coloneqq (H_{di} \colon d \in C_L, i \in I) \in \mathbb{R}^{|C_L| \times |I|}$ and $\boldsymbol{h} \coloneqq (h_d, d \in C_L) \in \mathbb{R}^{|C_L|}$.

The follower, on the other hand, reacts after the leader chooses x. He can pick different levels among his *activities* in a finite set A: we let y_a denote the level by which activity a is performed, and define $y \coloneqq (y_a \colon a \in A)$. By performing activity a at level y_a the follower incurs a cost of $c_a \cdot y_a$, and hence he desires to select y so as to minimize his total costs. His choices for y are limited, however, as y should satisfy all the constraints in a set C_F and should also be feasible given the leader's decision x. Therefore, the follower selects vector $y(x)$, where for any $x \in X$

$$y(x) \in \arg\min\{\boldsymbol{c}^\top y \ : \ y \in Y(x)\}, \text{ with } Y(x) \coloneqq \left\{y' \in \mathbb{Z}_+^b \times \mathbb{R}_+^{|A|-b} \ : \ \boldsymbol{F}y' + \boldsymbol{L}x \leq \boldsymbol{f}\right\}. \tag{1}$$

For any $x \in \mathbb{Z}_+^k \times \mathbb{R}_+^{|I|-k}$, $Y(x)$ is the follower's set of feasible actions given the leader's decision x. In (1), $\boldsymbol{c} \coloneqq (c_a \colon a \in A) \in \mathbb{R}^{|A|}$, $\boldsymbol{F} \coloneqq (F_{da} \colon d \in C_F, a \in A)$ belongs to $\mathbb{R}^{|C_F| \times |A|}$, $\boldsymbol{L} \coloneqq (L_{di} \colon d \in C_F, i \in I)$ belongs to $\mathbb{R}^{|C_F| \times |I|}$ and $\boldsymbol{f} \coloneqq (f_d, d \in C_F) \in \mathbb{R}^{|C_F|}$. In addition, $0 \leq b \leq |A|$ indicates the number of discrete variables of the follower's problem.

The objective of the leader in the full-information bilevel model is to choose the $x \in X$ that maximizes the cost the follower faces. Therefore, she solves the following bilevel problem

$$z^* = \max_{x,y} \ \boldsymbol{c}^\top y \tag{2a}$$

$$\text{s.t. } \boldsymbol{H}x \leq \boldsymbol{h},\ x \in \mathbb{Z}_+^k \times \mathbb{R}_+^{|I|-k} \tag{2b}$$

$$y \in \arg\min\left\{\boldsymbol{c}^\top y' \colon \boldsymbol{F}y' + \boldsymbol{L}x \leq \boldsymbol{f},\ y' \in \mathbb{Z}_+^b \times \mathbb{R}_+^{|A|-b}\right\}. \tag{2c}$$

This paper departs from this usual single-stage max-min bilevel problem with complete information and considers a *sequential max-min bilevel problem with incomplete information* (**SMPI**). We assume that the leader and the follower interact sequentially, once per each period in $\mathcal{T} = \{0, 1, \ldots, T\}$, that at all times *the follower has the information needed* to compute $y(x)$, but that *this is not the case for the leader*: we assume that at time $t = 0$ the leader does not fully know the set of activities A, and hence potentially neither C_F, nor the value of all the data defining

region $Y(x)$. In addition, as some leader's resources might be only available if some of the follower's activities are known, she might have only partial information regarding I, C_L and the set X.

Specifically, at the beginning of each time $t \in \mathcal{T}$ the leader is aware of a subset of the follower's activities $A^t \subseteq A$, a subset of the leader's resources $I^t \subseteq I$, a subset of upper-level constraints $C_L^t \subseteq C_L$ and a subset of lower-level constraints $C_F^t \subseteq C_F$. The contents of A^t, I^t, C_L^t, and C_F^t depend on the set of activities, resources, and constraints the leader initially knows, denoted by A^0, I^0, C_L^0, and C_F^0, and on all the activities, resources, and constraints she has learned from the feedback generated by follower's responses until time $t-1$, see Figure 1. Furthermore, the leader's knowledge of the follower's lower-level problem data is limited, and in this direction we make the following assumptions:

A1: At any time $t \in \mathcal{T}$ the leader knows with certainty the values of $\boldsymbol{F}^t := (F_{da} \colon d \in C_F^t, a \in A^t)$ and $\boldsymbol{f}^t := (f_d \colon d \in C_F^t)$. In addition, the leader knows with certainty all her data (both *upper-level* and *lower-level*) with respect to the resources in I^t, that is, at time t she knows with certainty $\boldsymbol{H}^t := (H_{di} \colon d \in C_L^t, i \in I^t)$, $\boldsymbol{h}^t := (h_d \colon d \in C_L^t)$ and $\boldsymbol{L}^t := (L_{di} \colon d \in C_F^t, i \in I^t)$.

A2: The leader does not know with certainty all the entries of $\boldsymbol{c}$ but she knows that $\boldsymbol{c}^t := (c_a \colon a \in A^t) \in \mathcal{U}^t$, with

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

If C_U^t is the set of constraints of polyhedron $\mathcal{U}^t$, then $\boldsymbol{G}^t \in \mathbb{R}^{|C_U^t| \times |A^t|}$ and $\boldsymbol{g}^t \in \mathbb{R}^{|C_U^t|}$. We assume that both $\boldsymbol{G}^t$ and $\boldsymbol{g}^t$ are known with certainty to the leader at time t.

A3: The matrix $\boldsymbol{H}$ and vector $\boldsymbol{h}$ take non-negative values.

A4: For any $x \in X$, $\boldsymbol{L}x \leq \boldsymbol{f}$.

Assumption **A1** implies that, with the *exception of the cost vector*, the leader knows with certainty all the problem data in (2) that is associated with activities in A^t, resources in I^t, and constraints in C_F^t and C_L^t. Particularly, the latter part of this assumption stems from the idea that the leader is always certain about her operational capabilities (hence, she always knows $\boldsymbol{H}$ and $\boldsymbol{h}$ for all activities and constraints known to her), and about the effect that her actions have on the follower (hence, she always knows $\boldsymbol{L}$ for all activities and constraints known to her). We note that the assumption regarding the leader's certain knowledge of the values of $\boldsymbol{F}^t$ can be relaxed, and most of the results can be extended to this more general setting, see Section 4.

Assumption **A2** states that the leader has a polyhedral uncertainty set for $\boldsymbol{c}^t$. Polyhedral sets capture many important classes of uncertainty for the data in $\boldsymbol{c}^t$ such as lower and upper bounds, linear relationships between the entries, 1-norms, infinity norms, among others, see Ben-Tal et al. (2009). Assumption **A3** reflects the fact that the leader aims to optimally use her assets subject to budgetary constraints. (Note that this assumption holds for broad classes of standard max-min bilevel problems arising in interdiction, AD and DA models.) This follows due to our convention

1. The leader chooses $x^t \in X^t$, where

$$X^t := \{x \in \mathbb{Z}_+^{k^t} \times \mathbb{R}_+^{|I^t|-k^t} \ : \ \boldsymbol{H}^t x \leq \boldsymbol{h}^t\}, \tag{3}$$

and k^t, $0 \leq k^t \leq |I^t|$, is the number of resources in I^t whose levels are discrete.

2. The follower solves the lower-level linear program given that the leader implements x^t:

$$z^t := \min_y \Big\{ \boldsymbol{c}^\top y \colon \boldsymbol{F} y + \sum_{i \in I^t} \boldsymbol{L}_i x_i^t \leq \boldsymbol{f},\ y \in \mathbb{Z}_+^b \times \mathbb{R}_+^{|A|-b} \Big\},$$

where $\boldsymbol{L}_i$ is the i-th column of L. Denote by y^t the solution of this program that the follower implements.

3. The response of the follower generates *feedback* $\mathcal{F}^t$; see Section 2.2 for its definition. The leader observes the information in $\mathcal{F}^t$ and uses it to update her knowledge to I^{t+1}, C_L^{t+1}, A^{t+1}, C_F^{t+1} and $\mathcal{U}^{t+1}$ (thus, potentially, she also updates $\boldsymbol{H}^{t+1}$, $\boldsymbol{h}^{t+1}$, $\boldsymbol{F}^{t+1}$, $\boldsymbol{L}^{t+1}$, $\boldsymbol{f}^{t+1}$, and $\boldsymbol{c}^{t+1}$).

Figure 1 Interaction between the leader and the follower at period $t \in \mathcal{T}$. The matrices and vectors are defined as $\boldsymbol{H}^t := (H_{di} \colon d \in C_L^t, i \in I^t)$, $\boldsymbol{h}^t := (h_d \colon d \in C_L^t)$, $\boldsymbol{F}^t := (F_{da} \colon d \in C_F^t, a \in A^t)$, $\boldsymbol{L}^t := (L_{di} \colon d \in C_F^t, i \in I^t)$, $\boldsymbol{f}^t := (f_d \colon d \in C_F^t)$, and $\boldsymbol{c}^t := (c_a \colon a \in A^t)$. The set $\mathcal{U}^t$ is the uncertainty set for the cost vector.

that the upper-level vectors in X are non-negative. Thus, by using resource $i \in I$ at level x_i, the leader consumes $H_{di} x_i$ units of asset d, $d \in C_L$, and the total amount of such asset available to her at any given time is given by h_d. Finally, assumption **A4** is technical and is made to ensure that the follower's problem is not trivially infeasible.

Ideally, the leader would implement an optimal upper-level solution of the full-information problem (2) at any given time. However, given the lack of information, she might not be able to do so. For this reason, we assume that the objective of the leader is to select the value of x^t, for all $t \in \mathcal{T}$, in order to minimize the number of periods until she can implement an optimal upper-level solution of the full-information problem (2) from there on. More precisely, her objective is to find a *weakly optimal* decision-making policy with respect to *time-stability*. The formal definition of these concepts is given in Section 2.1 below. Before that, we illustrate the assumptions above and the flexibility of the framework by means of an example. (An additional example of **SMPI** in the context of network interdiction with incomplete information can be found in Borrero et al. (2016).)

Example 1. We consider a simple class of the attacker-defender linear models, which can be viewed as an adversarial knapsack problem (DeNegre 2011, Caprara et al. 2013). The defender has $n > 0$ assets; operating asset a during a time period costs him b_a and produces a profit of p_a. He has an operational budget of B per period, and has to decide a level $y_a \in [0, 1]$ at which the operation of asset a is performed for all $a = 1, \cdots, n$. Hence, at each period the follower would ideally solve the following knapsack problem absent the actions of the leader

$$y^* \in \arg\max_y \{\boldsymbol{p}^\top y \ : \ \boldsymbol{b}^\top y \leq B, 0 \leq y_a \leq 1 \ \forall a = 1, \cdots, n\},$$

where $\boldsymbol{p} := (p_a \colon a = 1, \cdots, n)$ and $\boldsymbol{b} := (b_a \colon a = 1, \cdots, n)$.

The attacker, on the other hand, can temporarily disable some of the defender's assets. Disabling asset a during any given period costs her r_a, and the attacker has a budget of R per period. Moreover, if an asset is disabled then the follower cannot operate it. In this setting, $A = I = \{1, \cdots, n\}$, C_F consist of $n+1$ constraints, and hence $\boldsymbol{F} = (\boldsymbol{b}^\top; \boldsymbol{I})$, where $\boldsymbol{I}$ is a $n \times n$ identity matrix. Here, the lower-level right-hand side vector is given by $\boldsymbol{f} = (B; \mathbf{1})$ ($\mathbf{1}$ is a vector of ones of size n) and the cost vector satisfies $\boldsymbol{c} = -\boldsymbol{p}$. On the other hand, C_L is a singleton that contains the leader budgetary constraint, so $\boldsymbol{H} = \boldsymbol{r}^\top$, with $\boldsymbol{r} = (r_a \colon a \in I)$, and $\boldsymbol{h} = (R)$. Observe that matrix $\boldsymbol{L}$ in this setting is given by $\boldsymbol{L} = (\mathbf{0}^\top; \boldsymbol{I})$ where $\mathbf{0}$ is a vector of zeros.

At time $t = 0$, we make the assumption that the attacker does not know all the assets operated by the defender, nor the corresponding profits. For those assets $A^0 \subseteq A$ she knows, she has interval estimates $\ell_a \leq c_a \leq m_a$ for the profits, which implies that $\mathcal{U}^0 = \{\hat{\boldsymbol{c}}^0 \in \mathbb{R}^{|A^0|} \colon \ell_a \leq \hat{c}^0_a \leq m_a \ \forall a \in A^0\}$. Thus, $\boldsymbol{G}^0 = [\boldsymbol{I}; -\boldsymbol{I}]$ and $\boldsymbol{g}^0 = (\boldsymbol{m}; \boldsymbol{\ell})$, with $\boldsymbol{m} = (m_a \colon a \in A^0)$ and $\boldsymbol{\ell} = (\ell_a \colon a \in A^0)$. ■

2.1. Optimality Criteria

We measure the performance of a leader's decision-making policy in terms of its *time-stability*, where the time-stability of a policy is the first time period by which the actions prescribed by the policy coincide with the actions of an oracle decision-maker from there on. Recall that the oracle has all the information about the problem and thus, implements an optimal full-information decision, which yields a cost of z^* to the follower, at all time periods $t \in \mathcal{T}$.

To formally introduce time-stability and the concept of optimality we use, we first define what we consider a problem's instance for the leader. The *initial information* of the problem is the collection $\mathcal{D}^0$, where

$$\mathcal{D}^0 := (A^0, I^0, C^0_F, C^0_L, \mathcal{U}^0, \boldsymbol{H}^0, \boldsymbol{h}^0, \boldsymbol{F}^0, \boldsymbol{L}^0, \boldsymbol{f}^0).$$

Note that given some initial information $\mathcal{D}^0$, there might be several different bilevel problems of the form (2) that agree with the information contained in $\mathcal{D}^0$. In view of this, we define $\mathbb{G}(\mathcal{D}^0)$ to be the collection that contains all possible bilevel problems given that the leader knows $\mathcal{D}^0$:

$$\mathbb{G}(\mathcal{D}^0) := \{(A, I, C_F, C_L, \boldsymbol{c}, \boldsymbol{H}, \boldsymbol{h}, \boldsymbol{F}, \boldsymbol{L}, \boldsymbol{f}) \colon \text{conditions } \mathbf{C1}\text{-}\mathbf{C5} \text{ below are satisfied}\}$$

C1: $A^0 \subseteq A$, $I^0 \subseteq I$, $C^0_F \subseteq C_F$, $C^0_L \subseteq C_L$.

C2: $I^0 = \cup_{a \in A^0} I(a)$, $C^0_L = \cup_{i \in I^0} C_L(i)$, $C^0_F = \cup_{a \in A^0} C_F(a)$.

C3: $\mathcal{U}^0$ has valid upper and lower bounds for all c_a, $a \in A^0$.

C4: $(c_a \colon a \in A^0) \in \mathcal{U}^0$.

C5: $\boldsymbol{H}^0$, $\boldsymbol{h}^0$, $\boldsymbol{F}^0$, $\boldsymbol{L}^0$, $\boldsymbol{f}^0$, are submatrices of $\boldsymbol{H}$, $\boldsymbol{h}$, $\boldsymbol{F}$, $\boldsymbol{L}$, $\boldsymbol{f}$.

In condition **C2**, the set $I(a)$ contains all the interdiction resources that interfere with activity $a \in A$. Likewise, $C_L(i)$ and $C_F(a)$ are, respectively, the sets of upper- and lower-level constraints that restrict resource $i \in I$ and activity $a \in A$; see Definition 1 in Section 2.2 below for further details on these notions in the context of feedback. Therefore, **C2** means that at time $t = 0$ the leader knows all interdiction resources and constraints associated with the follower's activities in A^0.

Using collection $\mathbb{G}(\mathcal{D}^0)$, we define an *instance* of the problem as a pair $(\mathcal{D}^0, \mathcal{D})$, where $\mathcal{D} \in \mathbb{G}(\mathcal{D}^0)$. We denote by $\mathbb{G}$ the set of all possible instances.

A decision-making *policy* π is a sequence of set functions $\pi = (\pi^1, \cdots, \pi^T)$, such that $x^t = \pi^t(\mathcal{H}^t(\mathcal{D}^0, \mathcal{D}))$, and $\mathcal{H}^t(\mathcal{D}^0, \mathcal{D})$ denotes the history of both the leader and follower decision-making process up to time t:

$$\mathcal{H}^t(\mathcal{D}^0, \mathcal{D}) \coloneqq (\mathcal{D}^0, x^0, \mathcal{F}^0, \cdots, x^{t-1}, \mathcal{F}^{t-1}), \qquad t \geq 1,$$

where we recall that $\mathcal{F}^t$ is the feedback the leader gets from the follower's response at time t, see Figure 1. The set of all policies is denoted by Π. When discussing a particular policy π, we include a superscript π on x^t and in all other quantities depending on it, and denote them by $x^{t,\pi}$, $y^{t,\pi}$, $z^{t,\pi}$, $I^{t,\pi}$, $A^{t,\pi}$, $\mathcal{U}^{t,\pi}$ and $\mathcal{F}^{t,\pi}$.

Let an instance $(\mathcal{D}^0, \mathcal{D})$ be given. We define the *time-stability* of a policy on $(\mathcal{D}^0, \mathcal{D})$, denoted by $\tau^\pi(\mathcal{D}^0, \mathcal{D})$, as the first time in $\mathcal{T}$ such that z^* is equal to $z^{t,\pi}$ from there on, i.e.,

$$\tau^\pi(\mathcal{D}^0, \mathcal{D}) \coloneqq \min\{t \in \mathcal{T} \,:\, z^{s,\pi} = z^* \text{ for all } s \geq t\}.$$

The leader would like to find an "optimal" time-stability policy, i.e., a policy that has a lower time-stability than any other policy across all instances. To this end, let us say that policy π is *absolutely better* than policy π' if and only if $\tau^\pi(\mathcal{D}^0, \mathcal{D}) \leq \tau^{\pi'}(\mathcal{D}^0, \mathcal{D})$ for any instance $(\mathcal{D}^0, \mathcal{D})$, and that π^* is *absolutely optimal* if it is absolutely better than any other policy. Unfortunately, absolute optimality is a very strong notion, and, in general, absolute optimal policies do not exist, see, e.g., Remark 1 in Borrero et al. (2016) for the sequential shortest-path interdiction problem with incomplete information, which can be viewed as a particular case in our general setting.

Henceforth, we study an alternative optimality notion referred to as *weak optimality*. For this, define the *size* of an instance $(\mathcal{D}^0, \mathcal{D})$ as the vector $(|A|, |A^0|)$, and define $\mathbb{G}_{\boldsymbol{s}}$ as the collection of instances of size $\boldsymbol{s} = (n, n^0)$ (with $n \geq n^0$):

$$\mathbb{G}_{\boldsymbol{s}} \coloneqq \{(\mathcal{D}^0, \mathcal{D}) \in \mathbb{G} \colon (|A|, |A^0|) = \boldsymbol{s}\}.$$

We say that π is weakly better than π' when for any given fixed instance size s, the worst-case time-stability of π across all possible instances of size s is at most the worst-case time-stability of π' across all possible instances of size s.

Observe that any direct information on $\mathcal{U}^0$ in the definition of $\boldsymbol{s}$ is not included. This follows as, from the worst-case analysis perspective, any reasonable notion of size of $\mathcal{U}^0$ is likely to be a function of n^0. Given the above considerations, we say that policy π is weakly better than π' if

$$\max_{(\mathcal{D}^0,\mathcal{D})\in\mathbb{G}_{\boldsymbol{s}}} \tau^{\pi}(\mathcal{D}^0,\mathcal{D}) \leq \max_{(\mathcal{D}^0,\mathcal{D})\in\mathbb{G}_{\boldsymbol{s}}} \tau^{\pi'}(\mathcal{D}^0,\mathcal{D}) \qquad \text{for all } \boldsymbol{s}\in S,$$

where $S := \{(n,n^0)\in\mathbb{Z}^2_+ \colon n\geq n^0\}$. We say that π^* is *weakly optimal* if it is weakly better than any other policy, that is, if

$$\pi^* \in \arg\min_{\pi\in\Pi} \max_{(\mathcal{D}^0,\mathcal{D})\in\mathbb{G}_{\boldsymbol{s}}} \tau^{\pi}(\mathcal{D}^0,\mathcal{D}) \qquad \text{for all } \boldsymbol{s}\in S. \tag{4}$$

Therefore, we define the objective of the leader as to find a weakly-optimal policy π^*, i.e., to find a solution to the optimization problem (4). It should be clear that the notion of weak optimality is an adaptation of the notion of min/max optimal policies used in the online optimization literature, specifically, in the multi-armed bandit settings, see Audibert and Bubeck (2009), Audibert et al. (2013).

Remark 1. Time-stability is closely connected to the concept of *regret*, the total loss of performance of the leader with respect to an oracle decision-maker with complete information of the problem at time $t=0$. Formally, the regret until time t is given by $R^{\pi}_t = \sum_{s\leq t}(z^* - z^{s,\pi})$. Importantly, a finite bound on the time-stability provides a finite upper-bound on the total regret, i.e., $R^{\pi}_t \leq U\tau^{\pi}$, where U is an upper-bound on the instantaneous regret at any given time, $z^* - z^{s,\pi}\leq U$. Most of the online optimization literature uses regret, or variations of it such as pseudo-regret, as a measure of performance and seeks to find decision-making policies that optimally upper-bound it, see, e.g., Cesa-Bianchi and Lugosi (2006), Bubeck and Cesa-Bianchi (2012).

We use time-stability instead of regret because of its practical implications: if the leader knows that time-stability is attained, then an optimal full-information solution has been found. Moreover, she can be assured that there will be no loss of performance after that time period. ■

2.2. Feedback

Recall that the feedback $\mathcal{F} := (\mathcal{F}^t, t\in\mathcal{T})$ is the information that the leader collects from the follower's response at each time period. Depending on the particular application, the feedback might include data from the follower's problem as well as from his response y^t, some information regarding the follower's activities and constraints that were unknown to the leader, as well as the leader's resources that were previously unavailable. In order to formalize these notions we introduce the following terminology:

Definition 1. Let time $t\in\mathcal{T}$ be given and consider the bilevel problem (2).

• We say that the follower *performs* activity $a \in A$ (leader *uses* resource $i \in I$) at time t if and only if $y_a^t > 0$ ($x_i^t > 0$).

• We say that a lower-level (upper-level) constraint $d \in C_F$ ($d \in C_L$) *restricts* follower's activity $a \in A$ (leader's resource $i \in I$) if and only if $F_{da} \neq 0$ ($H_{di} \neq 0$), and we denote by $C_F(a)$ ($C_L(i)$) the set of constraints that restrict $a \in A$ ($i \in I$).

• We say that a leader resource $i \in I$ *interferes* with follower activity $a \in A$ if and only if there exist s a lower-level constraint $d \in C_F$, such that $d \in C_F(a)$ and $L_{di} \neq 0$. We denote by $I(a)$ the set of all leader's activities that interfere with $a \in A$. ■

The first of the above definitions reflects the intuitive fact that if the follower's variable y_a takes the value 0 then it does not change the value of the follower's objective function nor the value of his constraints; hence this can be interpreted as if activity $a \in A$ is not performed. The second definition is a consequence of the fact that if $F_{da} = 0$ for a given $a \in A$, then y_a can take arbitrarily large values without compromising the satisfiability of constraint d; the remaining definitions are also inspired by the same observations.

Example 1 (continued). In the AD knapsack example, the follower performs activity $a \in A$ if he operates asset a. The leader uses resource $a \in A$ if she disables asset a (hence, $I = A$). For any $a \in A$, $C_F(a)$ consists of the defender's budget constraint and on the constraint $y_a \leq 1$. On the other hand, for any $a \in I$ it is clear that $C_L(a) = C_L$. Moreover, observe that in this setting, for any asset $a \in A$, we have that $I(a) = \{a\}$. ■

We are now in position to define a *standard feedback*:

Definition 2. We say that feedback $\mathcal{F}$ is *standard* if and only if for any $t \in \mathcal{T}$

S1: The leader observes the total cost z^t incurred by the follower.

S2: The leader observes the activities performed by the follower, that is, she can determine that the follower performed activity $a \in A$ at time t as long as $y_a^t > 0$. If $y_a^t > 0$ and $a \notin A^t$, the leader *learns* about the existence of $a \in A$, and of all the leader resources that can restrict $a \in A$. Therefore,

$$A^{t+1} = A^t \cup \bigcup_{a:\, y_a^t > 0} \{a\}, \qquad I^{t+1} = I^t \cup \bigcup_{a:\, y_a^t > 0} I(a).$$

S3: For every new follower's activity $a \in A$ learned by the leader, she learns all the lower-level constraints in $C_F(a)$, and all the upper-level constraints $C_L(i)$, for all $i \in I(a)$. Henceforth,

$$C_F^{t+1} = C_F^t \cup \bigcup_{a \in A^{t+1} \setminus A^t} C_F(a), \qquad C_L^{t+1} = C_L^t \cup \bigcup_{i \in I^{t+1} \setminus I^t} C_L(i).$$

S4: For any newly learned activity $a \in A$: the leader learns the value of F_{da} for all $d \in C_F(a) \cup C_F^t$; for any $i \in I(a) \cap I^t$ the leader learns the value of H_{di} for all $d \in C_L(i) \setminus C_L^t$ and the value of L_{di}

for all $d \in C_F(a) \setminus C_F^t$; for any $i \in I(a) \setminus I^t$ the leader learns the value of H_{di} for all $d \in C_L(i) \cup C_L^t$ and the value of L_{di} for all $d \in C_F(a) \cup C_F^t$. Finally, for any $d \in C_F(a) \setminus C_F^t$ the leader learns the value of f_d, and for any $i \in I(a)$ the leader learns the value of h_d for all $d \in C_L(i) \setminus C_L^t$. ■

Hereafter, we make the assumption that the feedback is always standard. Importantly, note that **S2** – **S4** imply that at any given time $t \in \mathcal{T}$ the leader knows all the resources and constraints associated with all the follower's activities she knows at time t. In other words, a condition analogous to **C2** in Section 2.1 for $t = 0$ holds at all periods $t \in \mathcal{T}$. These considerations, imply that at any given time $t \in \mathcal{T}$ the matrices $\boldsymbol{F}$, $\boldsymbol{L}$ and $\boldsymbol{H}$ can be partitioned in submatrices as follows:

$$\boldsymbol{F} = \begin{array}{c} \\ C_F^t \\ C_F \setminus C_F^t \end{array} \begin{array}{c} \begin{array}{cc} A^t & A \setminus A^t \end{array} \\ \begin{pmatrix} \boldsymbol{F}_1 & \boldsymbol{F}_2 \\ \boldsymbol{0} & \boldsymbol{F}_3 \end{pmatrix} \end{array}, \quad \boldsymbol{L} = \begin{array}{c} \\ C_F^t \\ C_F \setminus C_F^t \end{array} \begin{array}{c} \begin{array}{cc} I^t & I \setminus I^t \end{array} \\ \begin{pmatrix} \boldsymbol{L}_1 & \boldsymbol{0} \\ \boldsymbol{L}_2 & \boldsymbol{L}_3 \end{pmatrix} \end{array}, \quad \boldsymbol{H} = \begin{array}{c} \\ C_L^t \\ C_L \setminus C_L^t \end{array} \begin{array}{c} \begin{array}{cc} I^t & I \setminus I^t \end{array} \\ \begin{pmatrix} \boldsymbol{H}_1 & \boldsymbol{H}_2 \\ \boldsymbol{0} & \boldsymbol{H}_3 \end{pmatrix} \end{array}, \quad (5)$$

and it is clear that, in the notation of the above structure, the leader is only aware of $\boldsymbol{F}_1$, $\boldsymbol{L}_1$ and $\boldsymbol{H}_1$ at the beginning of time $t \in \mathcal{T}$. In particular, note that $\boldsymbol{F}^t = \boldsymbol{F}_1$, $\boldsymbol{L}^t = \boldsymbol{L}_1$, and $\boldsymbol{H}^t = \boldsymbol{H}_1$.

Assumption **S1** on the standard feedback is typical in the online optimization literature (Cesa-Bianchi and Lugosi 2006) and can be seen as a minimum requirement to perform any optimization analysis. The role of the other assumptions, namely, **S2**-**S4** is to determine what information the leader gains when a new activity is learned; specifically, these assumptions ensure that at any time t the leader has the *structural* information of a version of problem (2). That is: (i) the leader always observes all the constraints associated with the resources/activities she knows, and hence, if she ignores the existence of a constraint (lower-level or upper-level) then she must ignore the existence of all the resources/activities associated with it; (ii) the leader is always aware of all the resources in I that can restrict the follower's activities she knows, and hence, if the leader ignores a resource at any given time, then it must be that said resource cannot interfere with the follower's activities that she already knows.

It is important to note that our assumptions on standard feedback do not rule out the possibility that there might exist resources that the leader knows at time t that might restrict the follower's activities she does not know at time t. In this sense, some of the leader's feasible vectors at time t might 'involuntarily' restrict the follower's activities.

Example 1 (continued). In the AD knapsack example, by assuming standard feedback, at each time t the leader observes the profit the follower receives from operating his assets. If the follower uses an asset unknown to the leader, then the leader learns about the existence of this asset, its cost b_a and the operating level upper bound. In addition, she discovers that she can disable the asset and that it costs her r_a to do so. ■

Observe that the assumptions on standard feedback impose no conditions on the *values* that are observed from the follower's response nor on the follower's cost vector. In this sense, stronger assumptions can be made in order to guarantee that the leader learns the follower's data in $\boldsymbol{c}$ or his response y^t with more accuracy. In this paper we consider the following two cases:

DEFINITION 3. Let $\mathcal{F}$ be standard. We say $\mathcal{F}$ is: (i) *Value-Perfect* if and only if at any time $t \in \mathcal{T}$ the leader learns the value of c_a for all $a \in A$ such that $y_a^t > 0$; or *(ii) Response-Perfect* if and only if at any time $t \in \mathcal{T}$ the leader learns the value of y_a^t for all $a \in A$ such that $y_a^t > 0$. ■

Standard feedback, as well as its Value-Perfect feedback version, can be viewed as adaptations of similar notions in the online optimization literature. For example, suppose that $A = A^0$ (hence, the leader knows all the follower's activities at time $t = 0$). In this case, standard feedback only requires the leader to observe the value of z^t at each $t \in \mathcal{T}$, and thus it parallels to the notion of *bandit feedback* that appears in the online convex and combinatorial optimization (see, e.g., Bubeck and Cesa-Bianchi (2012), Hazan (2015) and the references therein). Similarly, Value-Perfect feedback parallels the notion of *semi-bandit feedback* in the online combinatorial optimization (Audibert et al. 2013).

Example 1 (continued). In the AD knapsack setting, under Value-Perfect feedback, at each period the leader observes the follower's profit from the assets operated during the period. Under Response-Perfect feedback, she observes the corresponding values of y's. ■

3. Greedy and Robust Policies

In this section we introduce a set of leader's policies Λ that are greedy and robust. These policies are *greedy* in the sense that at each $t \in \mathcal{T}$ they aim to maximize the immediate cost that the follower faces at time t, and robust in the sense that they exploit the cost information in $\mathcal{U}^t$ in a worst-case scenario approach. Under the Value-Perfect and Response-Perfect conditions on the feedback $\mathcal{F}$, we show that these policies' time-stability are upper-bounded by $|A|$, and moreover, that they are weakly optimal. Also, we show that these policies also have additional features, such as that they can identify the value of time-stability in real time, yielding a *certificate of optimality*. Note that the proposed policies can be viewed, in a sense, as natural generalizations of known results for the shortest-path network interdiction problem, see Borrero et al. (2016). Throughout this section we omit any dependence on the instance $(\mathcal{D}^0, \mathcal{D})$ unless necessary to avoid confusion.

3.1. General Results for Standard Feedback

In order to define the set of greedy and robust policies, Λ, some additional concepts have to be introduced. For any $t \in \mathcal{T}$, and given any $x \in X^t$, define region $Y^t(x)$ as

$$Y^t(x) \coloneqq \left\{ y \in \mathbb{Z}_+^{b^t} \times \mathbb{R}_+^{|A^t|} \ : \ \boldsymbol{F}^t y + \boldsymbol{L}^t x \leq \boldsymbol{f}^t \right\},$$

with $0 \leq b^t \leq |A^t|$. Observe that $Y^t(x)$ is the leader's perception of the follower's feasible region given that she selects x, and that the leader completely knows $Y^t(x)$ at time t. For any $x \in X^t$, define $z_R^t(x)$ as the value of the robust linear program

$$z_R^t(x) \coloneqq \min\Big\{\max\big\{(\hat{\boldsymbol{c}})^\top y \colon \hat{\boldsymbol{c}}^t \in \mathcal{U}^t\big\} \colon \ y \in Y^t(x)\Big\}.$$

Note that $z_R^t(x)$ is the follower's (worst-case) objective function value given x if the leader's perception is correct. Let $z_R^{t,*}$ be the value that corresponds to the best possible decision the leader can take at time t if she estimates the follower's response using the robust approach above, that is,

$$z_R^{t,*} \coloneqq \max\{z_R^t(x) \ : \ x \in X^t\} \qquad \forall t \in \mathcal{T}.$$

Finally, for any policy π, define $\xi^\pi \coloneqq \xi^\pi(\mathcal{D}^0, \mathcal{D})$ as $\xi^\pi \coloneqq \min\{t \in \mathcal{T} \ : \ z_R^{t,*} = z^{t,\pi}\}$. We define policies in Λ as those policies that greedily optimize in a robust fashion from time $t = 0$ until time ξ^λ. From ξ^λ onwards, policies in Λ repeat the same solution used at time ξ^λ. Formally:

Definition 4. We say that $\lambda \in \Lambda \subseteq \Pi$ if and only if

$$x^{t,\lambda} \in \arg\max\{z_R^t(x) \ : \ x \in X^t\} \qquad \forall t \leq \xi^\lambda, \tag{6}$$

and $x^{t,\lambda} = x^{\xi^\lambda,\lambda}$ for all $\xi^\lambda < t \leq T$. ■

The greedy and robust policies Λ generalize the greedy and pessimistic policies given for the shortest-path interdiction problem in Borrero et al. (2016). In contrast to the policies of this earlier work, here Λ requires solving a general max-min bilevel linear problem, where the lower-level problem involves a robust optimization problem over a polyhedral uncertainty set. Despite this more general setting, it can be shown that the policies in Λ are computable by standard mixed integer programming (MIP) solvers as robust bilevel problem (6) can be reduced to a single-level MIP, see Appendix A.3 for further details.

The following result lists the main properties of the policies in Λ under the assumption of standard feedback. It establishes a simple relationship between the cost of the optimal oracle solution (z^*), the cost the follower faces at t ($z^{t,\lambda}$), and the cost the leader expects the follower to incur ($z_R^{t,*}$). In addition, it reveals the importance that time period ξ^λ has for time-stability. We note that this result generalizes Lemma 4 of Borrero et al. (2016), which is given for the shortest-path interdiction setting, to max-min bilevel problems.

Theorem 1. *Let $t \in \mathcal{T}$ be given and let $\lambda \in \Lambda$ be arbitrary. Then, $z^{t,\lambda} \leq z^* \leq z_R^{t,*}$ and $\tau^\lambda \leq \xi^\lambda$.*

Proof. Observe that for any $t \in \mathcal{T}$ and $x \in X^t$, $z_R^t(x) = \min\{(\boldsymbol{d}^t)^\top y' \ : \ y' \in Y_R^t(x)\}$, where $\boldsymbol{d}^t = (1, 0, \ldots, 0)^\top$ and $Y_R^t(x) := \left\{(y_0, y) \in \mathbb{R} \times \mathbb{R}_+^{|A^t|} \ : \ -y_0 + (\boldsymbol{c}^t)^\top y \leq 0 \ \forall \hat{\boldsymbol{c}}^t \in \mathcal{U}^t, y \in Y^t(x)\right\}$. Indeed, $z_R^t(x)$ can be equivalently written as

$$\min\Big\{y_0 \colon y_0 \geq \max_{\hat{\boldsymbol{c}}^t \in \mathcal{U}^t} (\hat{\boldsymbol{c}}^t)^\top y,\ \boldsymbol{F}^t y + \boldsymbol{L}^t x \leq \boldsymbol{f}^t,\ y \in \mathbb{R}_+^{|A^t|}, y_0 \in \mathbb{R}\Big\}.$$

The observation follows after noting that (y_0, y) satisfies the first constraint of the above problem if and only if $y_0 \geq (\boldsymbol{c}^t)^\top y$ for all $\hat{\boldsymbol{c}}^t \in \mathcal{U}^t$. Importantly, in the remaining proofs, we assume that $z_R^t(x)$ is given in terms of $Y_R^t(x)$. We proceed in two steps.

Step 1. We first prove that $z^{t,\lambda} \leq z^* \leq z_R^{t,*}$. Fix $x \in X$ and $x^t \in X^t$, and define $z(x)$ and $\bar{x}^t$ as

$$z(x) = \min\{\boldsymbol{c}^\top y \colon y \in Y(x)\}, \text{ and } \bar{x}_i^t = x_i^t \text{ if } i \in I^t;\ \bar{x}_i^t = 0 \text{ if } i \notin I^t. \tag{7}$$

For the leftmost inequality, the result follows from the definition of both z^* and $z^{t,\lambda}$ (see Equations (2) and Equation (4)). Indeed, observe that $z^* = \max\{z(x) \colon x \in X\}$, that $z^{t,\lambda} = z(\bar{x}^{t,\lambda})$, and that $\bar{x}^{t,\lambda} \in X$ because the feedback is standard and Assumption **A3** holds. For the rightmost inequality, let x^* be an element of X that attains z^*. Partition x^* as $x^* = (\hat{x}, \tilde{x})$, where $\hat{x} = (x_i^*)_{i \in I^t}$ and $\tilde{x} = (x_i^*)_{i \in I \setminus I^t}$. Recall the definition of the partition of matrices given by (5). Therefore, because $x^* \in X$ and **A3** holds, one has that $\hat{x} \in X^t$.

Now, suppose that $Y_R^t(\hat{x})$ is non-empty (if it is empty then it must be the case that $z_R^{t,*} = +\infty$ and the result holds) and let $(y_0, \hat{y})$ be such that $(y_0, \hat{y}) \in \arg\min\{(\boldsymbol{d}^t)^\top y' \ : \ y' \in Y_R^t(\hat{x})\}$ (hence $(\boldsymbol{c}^t)^\top \hat{y} = z_R^t(\hat{x})$). By the definition of $z_R^{t,*}$ we have that $(\boldsymbol{c}^t)^\top \hat{y} \leq z_R^{t,*}$. On the other hand, define $\bar{y}$ as $\bar{y}_a := \hat{y}_a$ if $a \in A^t$, and $\bar{y}_a := 0$ if $a \in A \setminus A^t$. Because $\mathcal{F}$ is standard, Assumption **A4** holds, and $\hat{y} \in Y^t(\hat{x})$, it follows that $\bar{y} \in Y(x^*)$; therefore, $z^* \leq \boldsymbol{c}^\top \bar{y}$. As $\boldsymbol{c}^\top \bar{y} = (\boldsymbol{c}^t)^\top \hat{y}$, and both $(\boldsymbol{c}^t)^\top \hat{y} \leq z_R^{t,*}$ and $z^* \leq \boldsymbol{c}^\top \bar{y}$ hold, then we have the desired result.

Step 2. Next, we show that $\tau^\lambda \leq \xi^\lambda$. For notational convenience, let $\xi = \xi^\lambda$ in the remainder of the proof. We claim that $\bar{x}^{\xi,\lambda} \in \arg\max\{z(x) \colon x \in X\}$. Indeed, the fact that the feedback is standard (recall equation (5)) implies that $\bar{x}^{\xi,\lambda} \in X$. Because by definition of ξ we have that $z^{\xi,\lambda} = z_R^{\xi,*}$, step 1 implies that $z(\bar{x}^{\xi,\lambda}) = z^*$ (recall that we have $z^{t,\lambda} = z(\bar{x}^{t,\lambda})$), and therefore the claim follows.

Now, by definition of λ, for all $s \geq t$ it must be the case that $x^{s,\lambda} = x^{\xi,\lambda}$. We claim that this implies that $z^{s,\lambda} = z^*$ for all $s \geq t$, and hence that $\tau^\lambda \leq \xi^\lambda$. In order to arrive at a contradiction, assume that $z^{s,\lambda} < z^*$ for $s > \xi^\lambda$. As $x^{s,\lambda} = x^{\xi,\lambda}$, one has that $y^{s,\lambda} \in Y(x^{\xi,\lambda})$, and by the definition of $y^{\xi,\lambda}$ it would follow that $z^{\xi,\lambda} \leq z^{s,\lambda} < z^*$, which contradicts the fact that $z^{\xi,\lambda} = z^*$. ∎

Theorem 1 has important practical implications. Note that the leader is always aware of the value of $z_R^{t,*}$, and (by standard feedback) always observes the value of $z^{t,\lambda}$. Therefore, she can determine whether a given period t is equal to ξ^λ. Let $t \in \mathcal{T}$ be given such that $t - 1 < \xi^\lambda$, then at time t exactly one of the following scenarios may occur:

(i) The follower faces the cost the leader expected ($z^{t,\lambda} = z_R^{t,*}$). In this case, $t = \xi^\lambda$, and Theorem 1 implies that the solution implemented by the leader at time t is an optimal solution of the full-information problem.

(ii) The follower faces a cost less than that the leader expects ($z^{t,\lambda} < z_R^{t,*}$). In this case, nothing can be said regarding the optimality of the solution that the leader implements at time t by only assuming standard feedback. However, if the stronger notions of either Value-Perfect or Response-Perfect feedback are assumed, then the leader must learn new information of the follower's problem, as we show in the following sections.

Particularly, observation (i) implies that policies in Λ provide certificates of optimality in real-time. That is, as soon as $t = \xi^\lambda$, the leader is sure that the best possible solution has been found. Given the importance of ξ^λ for greedy and robust policies, next we derive a sufficient condition in terms of the uncertainty set $\mathcal{U}^t$ that establishes whether a given time $t \in \mathcal{T}$ corresponds to ξ^λ.

PROPOSITION 1. *Let $t \in \mathcal{T}$ be given, suppose that $\mathcal{U}^t = \{\boldsymbol{c}^t\}$, and assume that $y_a^t = 0$ for all $a \notin A^t$. Then $\xi^\lambda \leq t$, and, in particular, $\tau^\lambda \leq t$.*

Proof. As $y^{t,\lambda} \in Y(x^{t,\lambda})$ and $y_a^t = 0$ for all $a \notin A^t$, it follows that $\sum_{a\in A^t} F_{da} y_a^{t,\lambda} + \sum_{i\in I^t} L_{di} x_i^{t,\lambda} \leq f_d$ for all $d \in C_F^t$, which implies that $(y_a^{t,\lambda})_{a\in A^t} \in Y^t(x^{t,\lambda})$. On the other hand, as $\mathcal{U}^t$ is a singleton, the set $Y_R^t(x^{t,\lambda})$ becomes $Y_R^t(x^{t,\lambda}) = \{(y_0, y) \in \mathbb{R}_+^{|A^t|} : \; -y_0 + (\boldsymbol{c}^t)^\top y \leq 0, y \in Y(x^{t,\lambda})\}$, and hence, $z_R^t(x^{t,\lambda}) \leq (\boldsymbol{c}^t)^\top (y^{t,\lambda})_{a\in A^t}$. Therefore, from the first set of inequalities of Theorem 1 and as $z_R^t(x^{t,\lambda}) = z_R^{t,*}$ by definition of $x^{t,\lambda}$, we have that $z^{t,\lambda} \leq z_R^{t,*} \leq (\boldsymbol{c}^t)^\top (y^{t,\lambda})_{a\in A^t}$. On the other hand, from the definition of $y^{t,\lambda}$, we have that $z^{t,\lambda} = (\boldsymbol{c}^t)^\top (y^{t,\lambda})_{a\in A^t}$. We can conclude that $z^{t,\lambda} = z_R^{t,*}$, and hence $\xi^\lambda \leq t$, as desired. The later part of the proposition is a consequence of the above result and the second set of inequalities of Theorem 1. ∎

In other words, whenever there is no uncertainty in $\mathcal{U}^t$, if the leader decides by using a policy in Λ, and the follower does not reveal any new activity, then the leader can be sure that the best solution has been found. Importantly, there is a connection between the fact that $\mathcal{U}^t$ is a singleton with the *polyhedral dimension* of $\mathcal{U}^t$, $\dim(\mathcal{U}^t)$, which is defined as the maximum number of affine independent points within $\mathcal{U}^t$ (see Wolsey and Nemhauser (2014)). Indeed, $\dim(\mathcal{U}^t) = 0$ if and only if $\mathcal{U}^t = \{\boldsymbol{c}^t\}$. As such, in the following sections we use the condition that $\dim(\mathcal{U}^t) = 0$ along with Proposition 1 to establish upper bounds on ξ^λ (and hence, on τ^λ) under Value-Perfect and Response-Perfect feedbacks.

3.2. Policies in Λ Under Value-Perfect Feedback

Recall that feedback $\mathcal{F}$ is Value-Perfect if the leader observes the value of c_a for all activities $a \in A$ such that $y_a^t > 0$. Under this feedback the leader should update the uncertainty set $\mathcal{U}^t$ to $\mathcal{U}^{t+1}$ as

$$\mathcal{U}^{t+1} = \{\hat{\boldsymbol{c}} \in \mathbb{R}^{|A^{t+1}|} : \; (\hat{c}_a)_{a\in A^t} \in \mathcal{U}^t, \; \hat{c}_a = c_a \text{ for all } a \text{ s.t. } y_a^t > 0\}.$$

For convenience we partition A^t as $A^t = \widetilde{A}^t \cup \bar{A}^t$, where for any follower action $a \in \widetilde{A}^t$ the leader knows with certainty the value of c_a, that is, $\widetilde{A}^t := \{a \in A^t \ : \ \hat{c}_a = c_a \ \forall \hat{\boldsymbol{c}} \in \mathcal{U}^t\}$, and $\bar{A}^t := A^t \setminus \widetilde{A}^t$. The next lemma establishes that if the cost the follower incurs is different from the one expected by the leader, then the leader must learn the real cost of a follower's activity.

LEMMA 1. *Suppose $\lambda \in \Lambda$ and that feedback $\mathcal{F}$ is Value-Perfect. If $z^{t,\lambda} < z_R^{t,*}$ then $\widetilde{A}^{t+1} \setminus \widetilde{A}^t \neq \emptyset$. In particular, if $y_a^t = 0$ for all $a \notin A^t$, then $\dim(\mathcal{U}^{t+1}) < \dim(\mathcal{U}^t)$.*

Proof. First, note that if $y_a^{t,\lambda} > 0$ for some $a \notin A^t$, then the result follows from the assumption of Value-Perfect feedback. Therefore, suppose that $y_a^{t,\lambda} = 0$ for all $a \notin A^t$. We claim that there exists an activity $a \in A^t \setminus \widetilde{A}^t$ such that $y_a^{t,\lambda} > 0$; the existence of such an activity implies the desired result from the assumption of Value-Perfect feedback. Indeed, to proceed by contradiction, suppose that this is not the case, i.e., $y_a^{t,\lambda} = 0$ for all $a \in A^t \setminus \widetilde{A}^t$. As $y^{t,\lambda} \in Y(\bar{x}^{t,\lambda})$ (see Equation (7)) and $y_a^t = 0$ for all $a \notin A^t$, then it must be that $(y_a^{t,\lambda})_{a \in A^t} \in Y^t(x^{t,\lambda})$. Now, because $\hat{c}_a = c_a$ for all $a \in \widetilde{A}^t$, one has that for all $\hat{\boldsymbol{c}}^t \in \mathcal{U}^t$

$$(\hat{\boldsymbol{c}}^t)^\top (y_a^{t,\lambda})_{a \in A^t} = (\boldsymbol{c}^t)^\top (y_a^{t,\lambda})_{a \in A^t},$$

and therefore $((\boldsymbol{c}^t)^\top (y_a^{t,\lambda})_{a \in A^t}, (y_a^{t,\lambda})_{a \in A^t}) \in Y_R^t(x^{t,\lambda})$. Thus, by the definition of $x^{t,\lambda}$ we have that

$$z_R^{t,*} \leq (\boldsymbol{c}^t)^\top (y_a^{t,\lambda})_{a \in A^t}. \tag{8}$$

On the other hand, because $y_a^t = 0$ for all $a \notin A^t$, one has that $z^{t,\lambda} = (\boldsymbol{c}^t)^\top (y_a^{t,\lambda})_{a \in A^t}$, and hence, by Theorem 1 along with (8), we have that $z^{t,\lambda} = z_R^{t,*}$, yielding the desired contradiction. ∎

A direct consequence of the above result is that, in conjunction with Proposition 1, it provides an upper bound for the time-stability for any policy in Λ. We observe that this result generalizes Lemma 5 of Borrero et al. (2016) to max-min bilevel problems:

THEOREM 2. *Let $\lambda \in \Lambda$ and suppose that $\mathcal{F}$ is Value-Perfect. Then, $\tau^\lambda \leq \xi^\lambda \leq |A \setminus \widetilde{A}^0|$.*

Proof. Let $t \in \mathcal{T}$ be given such that $z^{t,\lambda} < z_R^{t,*}$. Lemma 1 implies that $\widetilde{A}^{t+1} \setminus \widetilde{A}^t \neq \emptyset$. Hence, $\widetilde{A}^t \neq A$ can happen at most for $|A \setminus \widetilde{A}^0|$ periods. Also, if $t \in \mathcal{T}$ satisfies $\widetilde{A}^t = A$, then $\dim(\mathcal{U}^t) = 0$ and Proposition 1 implies that $\xi^\lambda \leq t$. Therefore, $\xi^\lambda \leq |A \setminus \widetilde{A}^0|$ and the result follows. ∎

The previous results shed light into the importance of greedy and robust policies for solving the exploitation vs. exploration dilemma. Simply speaking, it states that as long as the leader is being robust with respect to uncertainty, then 'robust' exploitation (i.e., deciding greedily) always implies exploration (i.e., discovering new information). We emphasize that the key property behind the result is robustness: if the leader were to use another approach to deal with uncertainty, then she might not discover any new information; see Remark 7 in Borrero et al. (2016) for an example in the context of shortest path interdiction.

Next, we prove that the upper bound in Theorem 2 is tight across all instances and, more importantly, across all policies. In other words, we establish that policies in Λ are weakly optimal.

PROPOSITION 2. *Consider $\lambda \in \Lambda$ and suppose that $\mathcal{F}$ is Value-Perfect. Then, for any $\boldsymbol{s} = (n, n^0) \in S$*

$$\max_{(\mathcal{D}^0, \mathcal{D}) \in \mathbb{G}_{\boldsymbol{s}}} \tau^{\lambda}(\mathcal{D}^0, \mathcal{D}) \leq n. \tag{9}$$

Moreover, λ is weakly optimal.

Proof. First, observe that equation (9) is an immediate consequence of Theorem 2. In order to prove weak optimality, we show that for any given policy π and any $\boldsymbol{s} = (n, n^0) \in S$ there exists an instance $(\mathcal{D}^0, \mathcal{D})^{\pi}$ of size $\boldsymbol{s}$ such that $\tau^{\pi}((\mathcal{D}^0, \mathcal{D})^{\pi}) \geq n$.

Let $A = \{1, 2 \cdots, n^0, n^0+1, \cdots, n\}$, $A^0 = \{1, \cdots, n^0\}$ and $I = A$, $I^0 = A^0$. Let X (and hence $\boldsymbol{H}$ and $\boldsymbol{h}$) be given by

$$X = \Big\{ x \in \mathbb{Z}^n_+ \ : \ \sum_{j \in I^0} x_j = n^0 - 1, \ \sum_{j \in I} x_j \leq n-1, \ x_j \leq 1 \ \forall j = 1, \cdots, n \Big\},$$

and let X^0 (and hence, $\boldsymbol{H}^0$ and $\boldsymbol{h}^0$) be given by

$$X^0 = \Big\{ x \in \mathbb{Z}^n_+ \ : \ \sum_{j \in I^0} x_j = n^0 - 1, \ x_j \leq 1 \ \forall j = 1, \cdots, n^0 \Big\}.$$

On the other hand, for any $x \in X$ define $Y(x)$ as

$$Y(x) := \Big\{ y \in \mathbb{R}^n_+ \ : \ \sum_{j=1}^{n} y_j \leq 1, \quad y_j + x_j \leq 1 \ \forall j = 1, \cdots, n \Big\}.$$

That is, $\boldsymbol{F} = [\boldsymbol{1}^{\top}; \boldsymbol{I}]$ and $\boldsymbol{L} = [\boldsymbol{0}^{\top}; \boldsymbol{I}]$, where $\boldsymbol{I}$ is an identity matrix of size n, and $\boldsymbol{f}$ is a column vector of ones. Define $\boldsymbol{F}^0$, $\boldsymbol{L}^0$ and $\boldsymbol{f}^0$ as the corresponding submatrices of $\boldsymbol{F}$, $\boldsymbol{L}$ and $\boldsymbol{f}$ associated with $j = 1, \cdots, n^0$. Finally, consider $\boldsymbol{c}$ to be such that $c_{n^0+q} < c_{n^0+q+1}$, for $q = 1, \ldots n - n^0 - 1$, and for the cost coefficients of the first n^0 activities we assume that the leader knows that they belong to $\mathcal{U}^0$, where $\mathcal{U}^0 = \{\hat{\boldsymbol{c}}^0 \in \mathbb{R}^{n^0} : \ell \leq \hat{c}^0_j \leq u, \ j = 1, \ldots, n^0\}$, where in addition we assume that $c_n < \ell < u < 0$.

In order to adequately define the instance, a particular $\hat{\boldsymbol{c}}^0$ in $\mathcal{U}^0$ has to be fixed. However, independent of which specific $\hat{\boldsymbol{c}}^0$ is chosen (which will depend on the policy, see below), the above defined data constitutes an instance, i.e., $\mathcal{D}^{\pi} \in \mathbb{G}((\mathcal{D}^0)^{\pi})$, and its size is given by (n, n^0). Particularly, note that from the leader perspective, the problem consist of blocking those $n-1$ activities that are most profitable to the follower, constrained to the fact the she always need to block exactly $n^0 - 1$ out of the n^0 activities she knows at time $t = 0$. In addition, from the assumptions on $\boldsymbol{c}$, the follower's profit from any of the $n - n^0$ activities that the leader does not initially know is better than the profit generated by any activity that the leader initially knows.

From the definition of $(\mathcal{D}^0, \mathcal{D})$ it is clear that if x^* is an optimal oracle decision, then $x^*_j = 1$ for $j = n^0 + 1, \ldots, n$, which implies that the leader must learn all those activities before implementing a

solution where $z^{t,\pi} = z^*$. Hence, if t_0 denotes the first time after which the leader learns all activities from $A \setminus A^0$, it is clear from the structure of the instance that $t_0 \geq n - n^0$. In addition, note that until t_0 the follower has only used activities in $A \setminus A^0$, so by Value-Perfect feedback, he has not revealed to the leader any of the real costs of the activities in A^0.

In order to prove weak optimality we show that for any given policy π there is a cost vector $\boldsymbol{c}^0 \in \mathcal{U}^0$ such that it takes the leader at least another n^0 time periods to consistently implement x^* (this would imply that $\tau^\pi((\mathcal{D}^0, \mathcal{D})^\pi) \geq n$, yielding the desired result). First, assume that π does not repeat any solution from time t_0, until time $t_n = t_0 + n^0 - 1$. For any $t = t_0, \cdots, t_n$, let $j^{\pi,t}$ be the (unique) follower activity in A^0 that $x^{t,\pi}$ does not block at time t, and choose the values of $c_1, \ldots, c_{n^0}$ such that

$$\ell < c_{j^{\pi,t_0+1}} < c_{j^{\pi,t_0+2}} < \ldots < c_{j^{\pi,t_n}} < c_{j^{\pi,t_0}} < u,$$

and note that the above defined values are admitted by $\mathcal{U}^0$. Observe that fixing the costs of the actions in A^0 in this way, we have that x^* satisfies $x^*_j = 1$, for $j \neq j^{\pi,t_0}$ and $x^*_{j^{\pi,t_0}} = 0$, and that $z^* = c_{j^{\pi,t_0}}$. On the other hand, for $t = t_0 + 1, \cdots, t_n$,

$$z^{t,\pi} \leq c_{j^{\pi,t}} < z^* \tag{10}$$

(we note the first inequality above is, in general, not an equality, as it is not necessary for $x^{t,\pi}$ to block all the activities j with $j > n^0$). Henceforth, equation (10) implies that $\tau^\pi((\mathcal{D}^0, \mathcal{D})^\pi) > t_n$, and hence, as $t_0 \geq n - n^0$, $\tau^\pi((\mathcal{D}^0, \mathcal{D})^\pi) \geq n$, and the result follows.

Now, suppose that π repeats a solution once between t_0 and t_n, i.e., there exist $t_0 \leq u < v \leq t_n$ such that $x^{u,\pi} = x^{v,\pi}$. In this case $j^{\pi,u} = j^{\pi,v}$, and there exist $1 \leq b \leq n^0$ such that $b \neq j^{t,\pi}$ for all $t = 0, \cdots, n$. Let $\boldsymbol{c}^0$ satisfy

$$\ell < c_{j^{\pi,t}} < c_{j^{\pi,t+1}} \quad t = t_0, \cdots, v-2, \; c_{j^{\pi,t}} < c_{j^{\pi,t+1}} \quad t = v+1, \cdots, t_n,$$

and assume that $c_{j^{\pi,t_n}} < c_b < u$. Observe that the above defined $\boldsymbol{c}^0$ belongs to $\mathcal{U}^0$, and hence $(\mathcal{D}^0, \mathcal{D})^\pi$ is a valid instance, and moreover, x^* is given by $x^*_j = 1$ for all $j \neq b$, $x^*_b = 0$, with $z^* = c_b$. In addition, it is seen that for $t = t_0, \cdots, t_n$

$$z^{t,\pi} \leq c_{j^{\pi,t}} < z^*,$$

and hence $\tau^\pi((\mathcal{D}^0, \mathcal{D})^\pi) \geq n$, as desired. Also, note that if π repeats a solution between t_0 and t_n, then the same argument as above yields the result. ■

3.3. Policies in Λ Under Response-Perfect Feedback

Next, we establish convergence and weak optimality under Response-Perfect feedback. Recall that under this feedback the leader always observe the value of y_a^t for all $a \in A$ such that $y_a^t > 0$. In this setting, the leader should update the uncertainty set $\mathcal{U}^t$ to $\mathcal{U}^{t+1}$ by including the linear equality $\sum_{a\in A^{t+1}} y_a^{t,\lambda}\hat{c}_a = z^{t,\lambda}$. That is,

$$\mathcal{U}^{t+1} = \Big\{\hat{\boldsymbol{c}} \in \mathbb{R}^{|A^{t+1}|} \ : \ (\hat{c}_a)_{a\in A^t} \in \mathcal{U}^t, \ \sum_{a\in A^{t+1}} y_a^{t,\lambda}\hat{c}_a = z^{t,\lambda}\Big\}. \tag{11}$$

Observe that if $A^{t+1} = A^t$, i.e., if the leader does not learn any new activity at time t, then $\mathcal{U}^{t+1}$ has the same number of variables as $\mathcal{U}^t$, and moreover, equation (11) implies that $\mathcal{U}^{t+1} \subseteq \mathcal{U}^t$.

In Response-Perfect feedback, as in the Value-Perfect setting, by using a policy in Λ the follower must be forced to reveal new information whenever $z^{t,\lambda} < z_R^{t,*}$. Specifically, if $y_a^t = 0$ for all $a \notin A^t$, then it must be the case that $\dim(\mathcal{U}^{t+1}) < \dim(\mathcal{U}^t)$. This inequality follows because in this case $\dim(\mathcal{U}^t)$ cannot increase (since $\mathcal{U}^{t+1} \subseteq \mathcal{U}^t$), and, more importantly, from the fact that the linear equality $\sum_{a\in A^{t+1}} y_a^{t,\lambda}\hat{c}_a = z^{t,\lambda}$ is linearly independent from all the linear equalities in $\mathcal{U}^t$. These observations are formalized in the following result, which can be considered analogous to Lemma 1:

LEMMA 2. *Let $\lambda \in \Lambda$ and suppose feedback $\mathcal{F}$ is Response-Perfect. If $z^{t,\lambda} < z_R^{t,*}$ and $y_a^t = 0$ for all $a \notin A^t$ then* $\dim(\mathcal{U}^{t+1}) < \dim(\mathcal{U}^t)$.

Proof. As $z^{t,\lambda} < z_R^{t,*}$ there exists $\tilde{\boldsymbol{c}}^t \in \mathcal{U}^t$ such that $z^{t,\lambda} < (\tilde{\boldsymbol{c}}^t)^\top (y_a^{t,\lambda})_{a\in A^t}$. Because $A^{t+1} = A^t$, we have that $\mathcal{U}^{t+1} = \{\hat{\boldsymbol{c}}^t \in \mathbb{R}^{|A^t|} \ : \ (\hat{\boldsymbol{c}}^t)^\top (y_a^{t,\lambda})_{a\in A^t} = z^{t,\lambda}, \hat{\boldsymbol{c}}^t \in \mathcal{U}^t\}$, and therefore $\tilde{\boldsymbol{c}}^t \notin \mathcal{U}^{t+1}$.

Now, in view of the definition of $\mathcal{U}^{t+1}$ under Response-Perfect feedback, $\boldsymbol{G}^{t+1} = (\boldsymbol{G}^t; (y^{t,\lambda})^\top)$ and $\boldsymbol{g}^{t+1} = (\boldsymbol{g}^t; z^{t,\lambda})$. For any $t \in \mathcal{T}$ let us denote by $C_U^{t,=}$ those inequalities in the definition of $\mathcal{U}^t$ that must be satisfied as strict equalities, i.e.,

$$j \in C_U^{t,=} \Leftrightarrow \boldsymbol{G}_j^t \hat{\boldsymbol{c}}^t = g_j \ \forall \hat{\boldsymbol{c}}^t \in \mathcal{U}^t,$$

where $\boldsymbol{G}_j^t$ denotes j-th row of $\boldsymbol{G}^t$. Let us denote by $\boldsymbol{G}^{t,=}$ and $\boldsymbol{g}^{t,=}$ the corresponding submatrix and subvector of $\boldsymbol{G}^t$ and $\boldsymbol{g}^t$ associated with those elements in $C_U^{t,=}$. We have that (see, e.g., Wolsey and Nemhauser (2014))

$$\dim(\mathcal{U}^t) = |A^t| - \operatorname{rank}(\boldsymbol{G}^{t,=}, \boldsymbol{g}^{t,=}). \tag{12}$$

We claim that $\operatorname{rank}(\boldsymbol{G}^{t+1,=}, \boldsymbol{g}^{t+1,=}) \geq \operatorname{rank}(\boldsymbol{G}^{t,=}, \boldsymbol{g}^{t,=}) + 1$, and the desired result then follows from equation (12). Indeed, arguing by contradiction, suppose that $\operatorname{rank}(\boldsymbol{G}^{t+1,=}, \boldsymbol{g}^{t+1,=}) = \operatorname{rank}(\boldsymbol{G}^{t,=}, \boldsymbol{g}^{t,=})$. This implies that $((y^{t,\lambda})_{a\in A^t}; z^{t,\lambda})^\top$ can be written as a linear combination of the rows of $(\boldsymbol{G}^{t,=}, \boldsymbol{g}^{t,=})$, and thus it is readily seen that $\{\hat{\boldsymbol{c}}^t \ : \ \boldsymbol{G}^{t+1,=}\hat{\boldsymbol{c}}^t = \boldsymbol{g}^{t+1,=}\} = \{\hat{\boldsymbol{c}}^t \ : \ \boldsymbol{G}^{t,=}\hat{\boldsymbol{c}}^t = \boldsymbol{g}^{t,=}\}$.

Because $\tilde{\boldsymbol{c}}^t \in \mathcal{U}^t$, it belongs to $\{\hat{\boldsymbol{c}}^t : \boldsymbol{G}^{t,=}\hat{\boldsymbol{c}}^t = \boldsymbol{g}^{t,=}\}$, which by the above equation implies that it also belongs to $\{\hat{\boldsymbol{c}}^t : \boldsymbol{G}^{t+1,=}\hat{\boldsymbol{c}}^t = \boldsymbol{g}^{t+1,=}\}$ and thus to $\mathcal{U}^{t+1}$, which yields the desired contradiction. ■

Now, if the leader learns new activities at t, then $\mathcal{U}^{t+1}$ has $|A^{t+1} \setminus A^t|$ more variables than $\mathcal{U}^t$. The addition of the corresponding new variables potentially increases the dimension of $\mathcal{U}^{t+1}$ with respect to $\mathcal{U}^t$ by $|A^{t+1} \setminus A^t|$. However, it is readily seen that the linear equality $\sum_{a \in A^{t+1}} y_a^{t,\lambda}\hat{c}_a = z^{t,\lambda}$ is trivially linearly independent of previous inequalities in $\mathcal{U}^t$, and as such if the leader learns new activities at t it can be concluded that $\dim(\mathcal{U}^{t+1}) \leq \dim(\mathcal{U}^t) + |A^{t+1} \setminus A^t| - 1$. This observation, in conjunction with Lemma 2 immediately provides the following upper bound (whose proof we omit):

THEOREM 3. *Let $\lambda \in \Lambda$ be given. Then, under Response-Perfect feedback,*

$$\tau^\lambda \leq \xi^\lambda \leq \dim(\mathcal{U}^0) + |A \setminus A^0|.$$

The above results, as in the case of Value-Perfect feedback, have the same implications regarding the exploitation vs. exploration dilemma. That is, exploitation always implies exploration as long as the leader decides robustly. In addition, for Response-Perfect feedback weak optimality also holds. The proof of this fact applies the same arguments as in Proposition 2. Thus, its proof is omitted.

PROPOSITION 3. *Let $\lambda \in \Lambda$ be given and suppose that $\mathcal{F}$ is Response-Perfect. Then, for any $\boldsymbol{s} \in S$*

$$\max_{(\mathcal{D}^0, \mathcal{D}) \in \mathbb{G}_{\boldsymbol{s}}} \tau^\lambda(\mathcal{D}^0, \mathcal{D}) \leq n.$$

Moreover, λ is weakly optimal.

4. Model for Matrix Uncertainty

In this section we consider a more general model referred to as the *matrix model* for the uncertainty of the leader regarding the data of the follower's problem. We assume that she knows with certainty the value of $\boldsymbol{c}^t$ at the beginning of time t, but that she does not know with certainty the values of matrix $\boldsymbol{F}^t$. We emphasize the generality of this model: if in a given problem $\boldsymbol{c}^t$ is uncertain as well, then it can be included in $\boldsymbol{F}^t$ w.l.o.g. by introducing a new variable y_0 that represents the cost function and adding the constraint $y_0 \geq (\boldsymbol{c}^t)^\top y$.

In this setup, and under the appropriate extensions of certain assumptions and feedback definitions, we show that the results for standard feedback for the basic model of Section 2 (which, in view of the current discussion, can be referred to as the *cost model*) are also valid. Moreover, we show that for the Value-Perfect feedback case, the time-stability upper bound of Theorem 2 also holds, while for Response-Perfect feedback, an extension of the upper bound in Theorem 3 holds under certain assumptions. The proofs of the results, except for that of the bound for Response-Perfect feedback (which can be found in the online supplement), follow from similar arguments as those for the cost model, and thus are omitted.

4.1. Assumptions and Feedback in the Matrix Model

In this model we assume that the leader knows $\boldsymbol{c}^t$ with certainty, but only knows that $\boldsymbol{F}^t$ belongs to an uncertainty set $\mathcal{U}^t$. For any $d \in C_F^t$ let us denote by n_d^t the number of the follower's activities in A^t that d restricts, that is, $n_d^t := |\{a \in A^t : d \in C_F(a)\}|$. We replace assumption **A2** from Section 2 with the following:

A2E: The leader does not know with certainty all entries of $\boldsymbol{F}$ but she knows that $\boldsymbol{F}^t \in \mathcal{U}^t$, with

$$\mathcal{U}^t = \{\hat{\boldsymbol{F}}^t \in \mathbb{R}^{\sum_{d \in C_F^t} n_d^t} \ : \ \boldsymbol{G}^t \hat{\boldsymbol{F}}^t \leq \boldsymbol{g}^t\},$$

where we make the convention that

$$\hat{\boldsymbol{F}}^t = (F_{11}, \ldots, F_{1n_1^t}, F_{21}, \ldots, F_{2n_2^t}, \ldots, F_{|C_F^t|1}, \ldots, F_{|C_F^t|n^t_{|C_F^t|}})^\top.$$

If C_U^t is the set of constraints of polyhedron $\mathcal{U}^t$, then $\boldsymbol{G}^t \in \mathbb{R}^{|C_U^t| \times \sum_{d \in C_F^t} n_d^t}$ and $\boldsymbol{g}^t \in \mathbb{R}^{|C_U^t|}$. We assume that both $\boldsymbol{G}^t$ and $\boldsymbol{g}^t$ are known by the leader at time t.

We also modify the definition of standard feedback; specifically we replace **S4** by **S4E**:

S4E: For any new learned activity $a \in A$, the leader learns the value of c_a (instead of learning the value of F_{da} for all $d \in C_F(a) \cup C_F^t$). The rest of the assumption is as S4.

Moreover, in this setting Value-Perfect feedback is extended to account for the values of the constraint matrix. That is, we refine the concept of Value-Perfect feedback as follows

Definition 5. In the matrix model, standard feedback $\mathcal{F}$ is called *Value-Perfect* if and only if at any time $t \in \mathcal{T}$ the leader learns the value of F_{da} for all a such that $y_a^t > 0$ and $d \in C_F^t \cup C_F(a)$. ■

Note that the definition of Value-Perfect feedback in the previous sections is a particular case of the above. On the other hand, we do not make additional assumptions on Response-Perfect feedback.

Finally, we modify the definition of an instance. The initial information in this setting consists of the vector $\mathcal{D}^0 := (A^0, I^0, C_F^0, C_L^0, \mathcal{U}^0, \boldsymbol{H}^0, \boldsymbol{h}^0, \boldsymbol{L}^0, \boldsymbol{f}^0, \boldsymbol{c}^0)$, and $\mathbb{G}(\mathcal{D}^0)$ becomes

$$\mathbb{G}(\mathcal{D}^0) := \{(A, I, C_F, C_L, \boldsymbol{F}, \boldsymbol{H}, \boldsymbol{h}, \boldsymbol{L}, \boldsymbol{f}, \boldsymbol{c}) : \text{conditions } \mathbf{C1,C2} \text{ and } \mathbf{C3E\text{-}C5E} \text{ below hold}\}$$

C3E: $\mathcal{U}^0$ has valid upper and lower bounds for all F_{da}, $d \in C_F^0$, $a \in A^0$.

C4E: $(F_{da} : d \in C_F^0, a \in A^0) \in \mathcal{U}^0$.

C5E: $\boldsymbol{H}^0$, $\boldsymbol{h}^0$, $\boldsymbol{L}^0$, $\boldsymbol{f}^0$, $\boldsymbol{c}^0$ are submatrices and subvectors of $\boldsymbol{H}$, $\boldsymbol{h}$, $\boldsymbol{L}$, $\boldsymbol{f}$, $\boldsymbol{c}$.

The above definitions are straightforward extensions of the assumptions and definitions of the basic cost model in Section 2. Using them, we extend most of the results in the next sections.

4.2. Extended Greedy and Robust Policies

In what follows we generalize the greedy and robust policies in Λ to the matrix model which we denote by Λ_E. Policies in Λ_E are greedy because they maximize the follower's costs at the next time period, and they are robust because they consider all possible realizations of $\hat{\boldsymbol{F}}^t$ over $\mathcal{U}^t$. As shown below, these policies share most properties of the policies in Λ under the different modes of feedback.

For any $t \in \mathcal{T}$, and given any $x \in X^t$ define the "robust" region $Y_E^t(x)$ as

$$Y_E^t(x) \coloneqq \Big\{ y \in \mathbb{R}_+^{|A^t|} \ : \ \hat{\boldsymbol{F}}^t y + \boldsymbol{L}^t x \leq \boldsymbol{f}^t \ \forall \hat{\boldsymbol{F}}^t \in \mathcal{U}^t \Big\}.$$

The robustness of $Y_E^t(x)$ follows from the fact that any element of this set must be feasible for any possible realization of the uncertain data in $\mathcal{U}^t$. Define

$$z_E^t(x) \coloneqq \min\big\{ (\boldsymbol{c}^t)^\top y \ : \ y \in Y_E^t(x) \big\}, \ x \in X^t \quad \text{and} \quad z_E^{t,*} \coloneqq \max\{ z_E^t(x) \ : \ x \in X^t \} \ t \in \mathcal{T}.$$

Additionally, for any policy π define $\xi_E^\pi \coloneqq \xi^\pi(\mathcal{D}^0, \mathcal{D})$ as $\xi_E^\pi \coloneqq \min\{ t \in \mathcal{T} \ : \ z_E^{t,*} = z^{t,\pi} \}$.

Definition 6. We say that $\lambda \in \Lambda_E \subseteq \Pi$ if and only if $x^{t,\lambda} \in \arg\max\{ z_E^t(x) \ : \ x \in X^t \}$ for all $t \leq \xi^\lambda$, and $x^{t,\lambda} = x^{\xi^\lambda,\lambda}$ for all $\xi_E^\lambda < t \leq T$. ■

As before, ξ_E^λ is the first time period when the follower uses a solution with the cost expected by the leader. Finally, from ξ_E^λ onwards, policies in Λ_E repeat the same solution used at time ξ_E^λ.

4.2.1. Policies in Λ_E under Standard and Value-Perfect Feedback The following proposition states that the standard feedback results that hold for Λ in Section 3.1, (i.e., Theorem 1 and Proposition 1) also hold for Λ_E.

Proposition 4. *Let $\lambda \in \Lambda_E$ be given and assume that $\mathcal{F}$ is standard. Then, (i) For any given $t \in \mathcal{T}$ it follows that $z^{t,\lambda} \leq z^* \leq z_E^{t,*}$; (ii) $\tau^\lambda \leq \xi_E^\lambda$; and (iii) Given $t \in \mathcal{T}$, if $\dim(\mathcal{U}^t) = 0$ and $y_a^t = 0$ for all $a \notin A^t$, then $\xi_E^\lambda \leq t$, and, in particular, $\tau^\lambda \leq t$.*

In addition, given the extended definition of Value-Perfect feedback, Lemma 1 and Theorem 2 can be generalized in a straightforward fashion for the policies in Λ_E. Indeed, define $\widetilde{A}_E^t$ as the set of the follower's activities for which the leader knows (with certainty) the values of the columns of A associated with them, that is, $\widetilde{A}_E^t \coloneqq \{ a \in A^t \ : \ \forall \hat{\boldsymbol{F}} \in \mathcal{U}^t \ \hat{F}_{da} = F_{da} \ \forall d \in C_F^t \}$.

Proposition 5. *Suppose $\lambda \in \Lambda_E$ and that feedback $\mathcal{F}$ is Value-Perfect. Then, (i) If $z^{t,\lambda} < z_E^{t,*}$ then $\widetilde{A}_E^{t+1} \setminus \widetilde{A}_E^t \neq \emptyset$; and (ii) $\tau^\lambda \leq \xi_E^\lambda \leq |A \setminus \widetilde{A}_E^0|$.*

4.2.2. Policies in Λ_E under Response-Perfect Feedback In this section we establish convergence under Response-Perfect feedback for policies in Λ_E. In contrast with the Value-Perfect case, the extended results are more involved. We begin with the following observation.

LEMMA 3. *Let $\lambda \in \Lambda_E$, and suppose that $z^{t,\lambda} < z_E^{t,*}$ and that $y_a^t = 0$ for all $a \notin A^t$. Then there exist a $\widetilde{\boldsymbol{F}}^t \in \mathcal{U}^t$ and a lower-level constraint $d \in C_F^t$ such that*

$$\left(\widetilde{\boldsymbol{F}}_d^t\right)^\top y^{t,\lambda} > f_d - \left(\boldsymbol{L}_d^t\right)^\top x^{t,\lambda}. \tag{13}$$

The above result implies that the leader can remove matrix $\widetilde{\boldsymbol{F}}^t$ from the uncertainty set at time t, as equation (13) means that $\widetilde{\boldsymbol{F}}^t \neq \boldsymbol{F}^t$. For any given $t \in \mathcal{T}$ and $\lambda \in \Lambda_E$, let us define $D^{t,\lambda}$ as the set of constraints for which equation (13) holds at time t, that is

$$D^{t,\lambda} := \left\{ d \in C_F^t \ : \ \exists \widetilde{\boldsymbol{F}}^t \in \mathcal{U}^t \text{ s.t. } \left(\widetilde{\boldsymbol{F}}_d^t\right)^\top y^{t,\lambda} > f_d - \left(\boldsymbol{L}_d^t\right)^\top x^{t,\lambda} \right\}.$$

Suppose that $z^{t,\lambda} < z_E^{t,*}$ and $y_a^t = 0$ for all $a \notin A^t$. Under the assumption of Response-Perfect feedback, one direct way to remove those elements of $\mathcal{U}^t$ that satisfy equation (13) is to define $\mathcal{U}^{t+1}$ as

$$\mathcal{U}^{t+1} = \{ \hat{\boldsymbol{F}}^t \in \mathcal{U}^t \ : \ \left(\hat{\boldsymbol{F}}_d^t\right)^\top y^{t,\lambda} \leq f_d - \left(\boldsymbol{L}_d^t\right)^\top x^{t,\lambda} \ \forall d \in D^{t,\lambda} \}, \tag{14}$$

where we note that $\mathcal{U}^{t+1} \subset \mathcal{U}^t$ by Lemma 3. On the other hand, if $y_a^t > 0$ for some $a \notin A^t$, then, in general, the existence of a $\widetilde{\boldsymbol{F}}^t$ such that (13) holds cannot be guaranteed, and hence the update in equation (14) can be vacuous (i.e., $\mathcal{U}^{t+1} = \mathcal{U}^t$).

From the above discussion it is clear that whenever the leader does not learn a new follower activity, then her uncertainty set reduces its size. However, the update defined by (14) does not necessarily reduce the dimension of $\mathcal{U}^t$, and hence an upper bound similar to that of Theorem 3 cannot be proved in this setting by using the polyhedral dimension arguments. However, if we make additional assumptions about the lower-level problem or about the leader's ability to observe the said problem, a finite upper bound can be established. These assumptions guarantee that the uncertainty update reduces the dimension of the uncertainty polyhedron at least by one.

PROPOSITION 6. *Let $\lambda \in \Lambda_E$ and suppose that $\mathcal{F}$ is Response-Perfect.*

(i) If all constraints of the lower-level problem are equalities, then $\tau^\lambda \leq \xi^\lambda \leq \dim(\mathcal{U}^0) + \sum_{a \in A \setminus A^0} |C_F(a)|$.

(ii) If for any period $t \in \mathcal{T}$ such that $y_a^t = 0$ for all $a \notin A^t$ the leader observes the slack associated with at least one of the constraints in $D^{t,\lambda}$, then $\tau^\lambda \leq \xi^\lambda \leq \dim(\mathcal{U}^0) + \sum_{a \in A \setminus A^0} \left(|C_F(a)| + 1\right)$.

Observe that all of the upper-bound results for policies in Λ (or Λ_E) proved so far rely on the fact that whenever the leader does not learn a new activity, then the dimension of $\mathcal{U}^{t+1}$ can be made strictly less than the dimension of $\mathcal{U}^t$. For the matrix model and under Response-Perfect feedback, if no additional assumptions are made, then this reduction in dimension cannot be guaranteed. In this general setting, however, we can prove that every time $\mathcal{U}^t$ is updated, the difference in 'size' between $\mathcal{U}^{t+1}$ and $\mathcal{U}^t$ is sufficiently large, see Borrero (2017).

5. Semi-Oracle Lower Bounds

In online optimization, the performance of a policy is compared against that of an *oracle*, who represents an ideal decision-maker who has all information of the problem beforehand, see Cesa-Bianchi and Lugosi (2006). Such an oracle faces no uncertainty and is able to make the best possible decision. In our problem setting, the oracle solves problem (2) at every period, and thus always attains a time-stability of zero. Unfortunately, such a lower bound is rather trivial and of not particular interest.

Consider instead a *weaker* oracle that, albeit knowing all the information of the problem in advance, has restrictions in the way she can use this information. Specifically, at any period such a weaker oracle can only use resources that she initially knows at time $t=0$, or that have been revealed to her by the follower in previous periods. Hence, this *semi-oracle*, see Borrero et al. (2016), represents a decision-maker that combines both the practical limitations of the leader, with all the knowledge of the traditional oracle. Specifically, the semi-oracle solves:

$$\min \sum_{t\in\mathcal{T}} \mathbb{1}_{\{\boldsymbol{c}^\top y^t < z^*\}} \tag{15a}$$

$$\text{s.t. } x^t \in X \qquad t\in\mathcal{T} \tag{15b}$$

$$y^t \in \arg\min\{\boldsymbol{c}^\top y\colon y\in Y(x^t)\} \qquad t\in\mathcal{T} \tag{15c}$$

$$x_i^t = 0 \qquad i\in I\setminus I^t, t\in\mathcal{T} \tag{15d}$$

$$I^{t+1} = I^t \cup \bigcup_{a\colon y_a^t>0} I(a) \qquad t\in\mathcal{T}\setminus\{T\}, \tag{15e}$$

where constraint (15d) prevents the semi-oracle from using activities which she does not know by time t. Observe that absent this constraint, the formulation corresponds to what the oracle (with full information) would solve. Constraints (15b) and (15c), on the other hand, imply that the semi-oracle has all the information of the problem. As a consequence, the leader cannot be expected to formulate nor optimally solve (at least, consistently) the problem given by (15) in practice.

There are two main advantages of using the notion of the semi-oracle, rather than the oracle, as a benchmark. First, it yields a more informative lower bound on the performance of any policy: the time-stability attained by the semi-oracle is not always zero; moreover, by using it we can evaluate the effect that the initial information has on the performance of any policy. Second, for any given instance, there is always a policy that attains the time-stability of the semi-oracle policy. Specifically, for any policy, any interaction between the leader and the follower can be mapped into a feasible solution of (15), and more importantly, given a *fixed* instance, there must exist a policy that yields the same values of x^t and y^t as an optimal solution of (15).

It is important to note, however, that the semi-oracle decision process does not constitute a feasible policy: given a same history $\mathcal{H}^t(\mathcal{D}^0,\mathcal{D})$, the semi-oracle might determine two different values

for x^t for different instances, see an example for the sequential shortest path interdiction in Borrero et al. (2016). This, because problem (15) is a function of the instance $(\mathcal{D}^0, \mathcal{D})$, rather than a function of the history (as it is the case with any admissible policies; recall their definition in Section 2.1).

It can be readily seen that the semi-oracle optimization problem (15) is NP-hard. Small and moderately sized instances of the problem, however, can be tackled by state-of-the-art MIP solvers. Indeed, a single-level MIP reformulation of (15) is given by:

$$\min_{u,v,w,y,x,\theta} \sum_{t\in\mathcal{T}} w^t \tag{16a}$$

$$\text{s.t. } \boldsymbol{H}x^t \le \boldsymbol{h} \qquad t\in\mathcal{T} \tag{16b}$$

$$\boldsymbol{F}y^t + \boldsymbol{L}x^t \le \boldsymbol{f},\ -\boldsymbol{F}^\top\theta^t \le \boldsymbol{c} \qquad t\in\mathcal{T} \tag{16c}$$

$$\theta^t \le \boldsymbol{M}^{\theta^t}u^t,\ y^t \le \boldsymbol{M}^{y^t}v^t \qquad t\in\mathcal{T} \tag{16d}$$

$$\boldsymbol{f} - \boldsymbol{F}y^t - \boldsymbol{L}x^t \le \boldsymbol{M}^{p^t}(\mathbf{1}-u^t) \qquad t\in\mathcal{T} \tag{16e}$$

$$\boldsymbol{c} + \boldsymbol{F}^\top\theta^t \le \boldsymbol{M}^{q^t}(\mathbf{1}-v^t) \qquad t\in\mathcal{T} \tag{16f}$$

$$x_i^t \le M^{x_i}\sum_{s=0}^{t-1}\sum_{a\in A(i)\setminus A^0} y_a^s \qquad t\in\mathcal{T},\ i\in I\setminus I^0 \tag{16g}$$

$$z^*(1-M^w w^t) \le \boldsymbol{c}^\top y^t \qquad t\in\mathcal{T} \tag{16h}$$

$$u^t\in\{0,1\}^{|C_F|}, v^t\in\{0,1\}^{|A|}, w^t\in\{0,1\} \qquad t\in\mathcal{T} \tag{16i}$$

$$y^t\in\mathbb{R}_+^{|A|}, x^t\in\mathbb{R}_+^{|I|-k}\times\mathbb{Z}_+^k, \theta^t\in\mathbb{R}_+^{|C_F|} \qquad t\in\mathcal{T}, \tag{16j}$$

where x^t is the solution of the semi-oracle at time t, and y^t is the solution of the follower at time $t\in\mathcal{T}$. The fact that $y^t\in\arg\min\{\boldsymbol{c}^\top y\colon y\in Y(x^t)\}$ is represented by its linear programming (LP) optimality conditions via constraints (16c) (primal and dual feasibility) and (16d), (16e), and (16f) (the linearized complementary slackness conditions). In these constraints, $\boldsymbol{M}^{\theta^t}$, $\boldsymbol{M}^{p^t}$, $\boldsymbol{M}^{y^t}$, and $\boldsymbol{M}^{q^t}$ are diagonal matrices that are upper bounds on θ^t, $\boldsymbol{f}-\boldsymbol{F}y^t-\boldsymbol{L}x^t$, y^t, and $\boldsymbol{c}+\boldsymbol{F}^\top\theta^t$, respectively. We refer the reader to Audet et al. (1997) for more details on single-level MIP reformulations of bilevel problems with the lower-level problem given by an LP.

Variable w^t is binary and takes the value of zero if $\boldsymbol{c}^\top y^t = z^*$, i.e., if the optimal semi-oracle solution is used at time t, see constraint (16h). Here, $M^w = (z^*-\ell)/z^*$ and ℓ is a valid lower bound on the value of $\boldsymbol{c}^\top y$ for any feasible y. Finally, constraint (16g) implies that a resource cannot be used if it has not been revealed by the follower or if it is not in I^0. In this constraint, $A(i)$ is the set of follower activities that i interferes with, i.e., $A(i)=\{a\in A\colon i\in I(a)\}$, and $M^{x_i}=u^i/\ell_i$, where u^i is an upper bound on the value of the i-th entry of any $x\in X$, and ℓ_i is a strictly positive lower bound on the value that any y_a, $a\in A(i)$, can take whenever $y_a>0$. In general, the computation of these lower bounds can be highly involved, but for specific applications they can be computed rather efficiently from the problem's data, see Section 6 for an example.

We close this section by noting that although MIP problem (16) can be solved directly for moderately sized instances, it might require lengthy computational times due to the large number of variables and constraints, particularly if T is large. It turns out, however, that this problem can be made somewhat less "dependent" on the time horizon T by feeding to the solver an initial feasible solution. This approach can drastically reduce the size of the resulting MIP, and thus lead to shorter computational times; see the discussion on this approach in Appendix A.2.

6. Numerical Illustration

In this section we demonstrate the numerical performance of the policies in Λ. For this, we use a simple extension of the AD Knapsack problem of Example 1 where the follower has two budgetary constraints. We consider both Value-Perfect and Response-Perfect feedbacks, two different models for the initial uncertainty set along with the uncertainty either in the profits, or in the budgetary constraints, or in both the profits and the constraints. In order to provide a broader picture of the performance of the policies in Λ, we compare them against reasonable benchmark policies in the context of **SMPI**, and with respect to the semi-oracle lower-bounding procedure of the previous section. Our results show that the policies in Λ outperform the benchmark, and compare rather favorably with respect to the semi-oracle lower bound.

The decisions generated by the policies in Λ are computed by solving a one-level MIP reformulation of the bilevel problem (6), see Section A.3 of the Appendix for further details. Generally speaking, the transformation of optimization problem (6) into an MIP involves application of methods from bilevel optimization (to transform the hierarchical problem into a single-level problem) and robust optimization (to adequately optimize over the uncertainty set $\mathcal{U}^t$) areas. We note that, in general, problem (6) is NP-hard, as bilevel linear optimization is its special case.

Additional results regarding the performance of the policies in Λ can be found in Borrero et al. (2016). There, the numerical experiments are shown for when the max-min bilevel problem is a shortest path interdiction problem, the feedback is Value-Perfect, and there is interval uncertainty for the cost coefficients. We note that the results obtained for such configuration largely conform with the findings we obtain in this section.

Test Instances. We consider an extension of the AD knapsack problem from Example 1, where the defender has $n = 15$ assets (thus $|A| = |I|$=15). The upper-level information is given by $\boldsymbol{r} = \mathbf{1}$ and $R = 3$, thus the Attacker can disable at most three assets at any given time. In this extension the follower faces two budgetary constraints $\sum_{j=1}^{n} b_j^{(k)} y_j \leq B^{(k)}$, $k = 1, 2$, with $B^{(k)} = \lceil U(1, 10 \cdot n)/3 \rceil$, $U(a, b)$ denoting a number drawn from a uniform discrete distribution between a and b. We consider two models of initial uncertainty sets, namely, *hypercube* and *simplex*:

- In the hypercube model the defender's profits satisfy $p_a \in [\ell_a^p, u_a^p]$, $a \in A$, where for each $a \in A$ the values of ℓ_a^p, p_a, and u_a^p are drawn at random (and ordered) from a random variable V that is the sum of a $U(1,5)$ with a $U(1,20)$ distribution. Likewise, the budgets satisfy that $b_a^{(k)} \in [\ell_a^{(k)}, u_a^{(k)}]$, $a \in A$, where for each $k = 1,2$ and each $a \in A$ the values of $\ell_a^{(k)}$, $b_a^{(k)}$, and $u_a^{(k)}$ are drawn at random (and ordered) from a random variable W that is the sum of a $U(1,10)$ with a $U(1,20)$ distribution.
- For the simplex model we assume that (besides non-negativity constraints) $\boldsymbol{G}^0$ and $\boldsymbol{g}^0$ represent three inequalities, one for the profits, and one for each of the budgetary constraints:

$$\sum_{j=1}^{n} G_{1j}\hat{p}_j \leq g_1, \quad \sum_{j=1}^{n} G_{2j}\hat{b}_j^{(1)} \leq g_2, \quad \sum_{j=1}^{n} G_{3j}\hat{b}_j^{(2)} \leq g_3.$$

The coefficients G_{1j}, $j = 1,\ldots,n$, are drawn at random from a $U(1,5)$ distribution. Similarly for the budgets, G_{ij}, $i = 2,3$, $j = 1,\ldots,n$, are drawn at random from a $U(1,10)$ distribution. The right-hand sides satisfy that $g_i = 5n\sum_{j=1}^{n} G_{ij}$, $i = 1,2,3$. The real values for the profits are generated as $p_j = (g_1/G_{1j})Rw_j^p$, $j = 1,\ldots,n$, where R is a random number between 0 and 1, and the w_j^ps are weight values drawn at random from a continuous $U(0,1)$ distribution and normalized so that $\sum_{j=1}^{n} w_j^p \leq 1$. The real values for the budgets $b_j^{(k)}$, $j = 1,\ldots,n$, $k = 1,2$, are generated following the same logic.

Given the polyhedron $\mathcal{P}^0$ obtained by either of the above methods, we generate $\mathcal{U}^0$ by adding to $\mathcal{P}^0$ the constraints that specify the data that the leader knows with certainty. For instance, if the budgets are known with certainty, we add the constraints $\hat{b}_j^{(k)} \leq b_j^{(k)}$ and $-\hat{b}_j^{(k)} \leq -b_j^{(k)}$ for $j = 1,\ldots,n$, $k = 1,2$. If there is uncertainty for the profits and both constraints, then $\mathcal{U}^0 = \mathcal{P}^0$.

For each of the uncertainty models, we generated at random $N = 30$ instances, and assume the leader uses both Value-Perfect and Response-Perfect feedbacks. We consider three sets of initial information A^0: in the first, the leader knows five activities of the follower; in the second, she knows ten activities; and in the last, she knows all activities. Finally, we set $T = 20$ periods.

Benchmark Policies. In addition to policies in Λ, we consider the following benchmarks:

- The *analytical center policy* π_b: At each time $t \in \mathcal{T}$ the policy computes x^{t,π_b} by solving the deterministic bilevel problem

$$x^{t,\pi_b} \in \arg\max_{x \in X^t} \Big\{ (-\widetilde{\boldsymbol{p}}^t)^\top y \;:\; (\widetilde{\boldsymbol{b}}^{t,(k)})^\top y \leq B^{(k)} \quad k = 1,2,\; y + x \leq \mathbf{1}, y \in \mathbb{R}^{|A^t|} \Big\}, \tag{17}$$

where $(\widetilde{\boldsymbol{p}}^t, \widetilde{\boldsymbol{b}}^{t,(1)}, \widetilde{\boldsymbol{b}}^{t,(2)})$ is the analytical center of the polytope $\mathcal{U}^t$, see Bertsimas and Tsitsiklis (1997).

- The *random policy* π_r: At each time $t \in \mathcal{T}$ the policy computes x^{t,π_r} by solving problem (17) with $(\bar{\boldsymbol{p}}^t, \bar{\boldsymbol{b}}^{t,(1)}, \bar{\boldsymbol{b}}^{t,(2)})$ used instead of $(\widetilde{\boldsymbol{p}}^t, \widetilde{\boldsymbol{b}}^{t,(1)}, \widetilde{\boldsymbol{b}}^{t,(2)})$. Here we have that $(\bar{\boldsymbol{p}}^t, \bar{\boldsymbol{b}}^{t,(1)}, \bar{\boldsymbol{b}}^{t,(2)})$ is a randomly generated extreme point of $\mathcal{U}^t$ that is obtained by solving the linear program

$(\bar{\boldsymbol{p}}^t, \bar{\boldsymbol{b}}^{t,(1)}, \bar{\boldsymbol{b}}^{t,(2)}) \in \arg\max\{(\boldsymbol{\ell}^t)^\top v \colon v \in \mathcal{U}^t\}$. In this problem, at each time $t \in \mathcal{T}$ each entry of vector $\boldsymbol{\ell}^t$ is drawn at random from a Bernoulli distribution with parameter 1/2, i.e., each entry is zero or one with equal probability.

- The *"stopped" random policy* π_s: At each time $t \in \mathcal{T}$ the policy computes x^{t,π_s} in the same manner as policy π_r. However, whenever the follower's costs are as expected by the leader (i.e., z^{t,π_s} is the same as the value of (17) with $(\bar{\boldsymbol{p}}^t, \bar{\boldsymbol{b}}^{t,(1)}, \bar{\boldsymbol{b}}^{t,(2)})$ in place of $(\widetilde{\boldsymbol{p}}^t, \widetilde{\boldsymbol{b}}^{t,(1)}, \widetilde{\boldsymbol{b}}^{t,(2)})$), then the policy keeps using the same solution thereafter.
- We also consider the *lower bound* provided by the semi-oracle approach discussed in Section 5. While it is not an admissible policy, with a slight abuse of notation we denote it by π^* hereafter.

Results and Discussion. In Tables 1– 4 we show the mean time-stability and mean absolute deviation (MAD) of the time-stability across the $N = 30$ replications for each configuration. We make the convention that whenever a policy does not find the optimal full-information solution, then the time-stability is set to $\tau^\pi = 21$ (i.e., $\tau^\pi = T + 1$).

Table 1 Mean and MAD for time-stability for the hypercube uncertainty model and Value-Perfect feedback.

		λ		π_b		π_r		π_s		π^*	
Uncertainty	A^0	Mean	MAD	Mean	MAD	Mean	MAD	Mean	MAD	Mean	MAD
Profits	{1 . . . 5}	2.47	0.59	6.27	6.77	11.93	7.36	9.43	7.36	1.47	0.63
	{1 . . . 10}	2.53	0.55	7.93	8.05	15.50	5.58	11.93	6.16	1.03	0.53
	{1 . . . 15}	2.57	0.71	6.83	8.22	18.10	3.83	16.33	6.40	0.00	0.00
First Constraint	{1 . . . 5}	2.27	0.71	4.87	5.28	7.60	5.51	7.20	6.29	1.47	0.63
	{1 . . . 10}	1.83	0.80	5.93	5.60	9.17	7.56	7.83	7.70	1.03	0.53
	{1 . . . 15}	1.17	0.90	5.10	6.76	9.47	8.29	8.57	8.81	0.00	0.00
First and second constraints	{1 . . . 5}	2.47	0.67	5.67	5.88	13.33	7.50	9.97	8.12	1.47	0.63
	{1 . . . 10}	2.43	0.70	6.90	7.45	17.03	4.94	17.30	5.67	1.03	0.53
	{1 . . . 15}	2.17	0.63	9.73	9.67	19.60	2.24	16.80	6.11	0.00	0.00
Profits, first, and second constraints	{1 . . . 5}	2.77	0.62	5.80	5.29	18.37	2.95	15.60	6.86	1.47	0.63
	{1 . . . 10}	3.00	0.51	8.93	7.40	20.53	0.60	17.53	5.33	1.03	0.53
	{1 . . . 15}	3.13	0.45	11.50	6.42	20.80	0.32	21.00	0.00	0.00	0.00

Table 2 Mean and MAD for time-stability for the hypercube uncertainty model and Response-Perfect feedback.

		λ		π_b		π_r		π_s		π^*	
Uncertainty	A^0	Mean	MAD	Mean	MAD	Mean	MAD	Mean	MAD	Mean	MAD
Profits	{1 . . . 5}	3.37	0.67	8.60	7.41	14.00	6.67	10.97	8.10	1.47	0.63
	{1 . . . 10}	3.30	0.79	9.00	7.62	16.40	5.45	14.57	6.99	1.03	0.53
	{1 . . . 15}	3.23	0.75	8.77	7.29	18.20	3.68	18.10	4.50	0.00	0.00
First Constraint	{1 . . . 5}	2.80	0.97	7.80	7.53	12.03	7.03	7.23	6.81	1.47	0.63
	{1 . . . 10}	2.30	1.15	7.70	7.78	10.67	8.83	9.73	9.12	1.03	0.53
	{1 . . . 15}	1.60	1.21	7.63	8.86	12.27	8.85	9.30	9.49	0.00	0.00
First and second constraints	{1 . . . 5}	3.60	1.27	14.83	7.25	18.53	3.43	15.77	6.85	1.47	0.63
	{1 . . . 10}	3.57	1.25	16.23	6.25	20.20	1.35	17.50	4.93	1.03	0.53
	{1 . . . 15}	3.30	1.43	17.07	5.46	20.80	0.32	19.97	1.52	0.00	0.00
Profits, first, and second constraints	{1 . . . 5}	4.33	1.20	14.60	7.30	20.53	0.55	16.90	5.09	1.47	0.63
	{1 . . . 10}	4.47	1.21	15.87	6.41	20.67	0.48	20.77	0.41	1.03	0.53
	{1 . . . 15}	4.73	1.33	16.20	5.71	20.77	0.38	19.77	2.03	0.00	0.00

We observe that the proposed policies $\lambda \in \Lambda$ consistently outperform the benchmark by a large margin except for the semi-oracle lower bound π^*, which is to be expected. For virtually all configurations, policies π_r and π_s yield very poor time-stability results, while π_b is somewhat better.

Table 3 Mean and MAD for time-stability for the simplex uncertainty model and Value-Perfect feedback.

		λ		π_b		π_r		π_s		π^*	
Uncertainty	A^0	Mean	MAD	Mean	MAD	Mean	MAD	Mean	MAD	Mean	MAD
Profits	$\{1\dots5\}$	1.77	0.71	8.50	7.24	20.83	0.17	17.90	5.09	1.23	0.37
	$\{1\dots10\}$	1.77	0.71	4.10	3.38	21.00	0.00	19.73	2.28	1.00	0.20
	$\{1\dots15\}$	1.77	0.71	2.53	0.46	21.00	0.00	19.03	3.27	0.00	0.00
First Constraint	$\{1\dots5\}$	2.17	0.43	5.33	6.27	7.47	5.92	7.27	7.23	1.23	0.37
	$\{1\dots10\}$	2.27	0.49	11.73	9.25	12.90	7.61	12.23	8.59	1.00	0.20
	$\{1\dots15\}$	2.37	0.47	12.70	9.12	15.43	6.17	14.57	8.27	0.00	0.00
First and second constraints	$\{1\dots5\}$	2.20	0.37	8.10	7.99	11.53	7.87	11.17	8.50	1.23	0.37
	$\{1\dots10\}$	2.20	0.39	14.30	6.70	15.43	6.43	15.90	5.29	1.00	0.20
	$\{1\dots15\}$	2.37	0.43	14.83	8.35	17.67	4.68	18.83	2.87	0.00	0.00
Profits, first, and second constraints	$\{1\dots5\}$	1.77	0.71	10.73	9.51	19.10	1.79	13.93	8.87	1.23	0.37
	$\{1\dots10\}$	1.77	0.71	10.13	9.07	20.60	0.59	15.87	7.43	1.00	0.20
	$\{1\dots15\}$	1.77	0.71	5.63	5.89	20.63	0.47	16.07	7.15	0.00	0.00

Table 4 Mean and MAD for time-stability for the simplex uncertainty model and Response-Perfect feedback.

		λ		π_b		π_r		π_s		π^*	
Uncertainty	A^0	Mean	MAD	Mean	MAD	Mean	MAD	Mean	MAD	Mean	MAD
Profits	$\{1\dots5\}$	6.40	1.10	9.47	3.63	20.70	0.47	19.40	2.31	1.23	0.37
	$\{1\dots10\}$	6.40	1.10	9.70	3.89	21.00	0.00	20.50	0.80	1.00	0.20
	$\{1\dots15\}$	6.40	1.10	8.07	3.97	21.00	0.00	21.00	0.00	0.00	0.00
First Constraint	$\{1\dots5\}$	5.50	1.55	14.60	8.19	16.07	6.43	13.97	7.08	1.23	0.37
	$\{1\dots10\}$	5.23	1.72	13.73	8.60	16.43	6.11	14.00	8.23	1.00	0.20
	$\{1\dots15\}$	4.97	1.69	13.40	8.98	16.57	5.65	15.43	6.55	0.00	0.00
First and second constraints	$\{1\dots5\}$	5.43	1.56	18.27	4.16	18.63	3.15	18.93	2.70	1.23	0.37
	$\{1\dots10\}$	5.20	1.59	19.33	2.93	19.10	2.84	19.03	3.01	1.00	0.20
	$\{1\dots15\}$	5.03	1.53	19.13	3.03	19.27	2.78	18.83	3.33	0.00	0.00
Profits, first, and second constraints	$\{1\dots5\}$	7.80	1.43	17.63	4.51	20.77	0.37	19.80	1.83	1.23	0.37
	$\{1\dots10\}$	7.80	1.43	16.80	5.49	20.90	0.17	20.80	0.36	1.00	0.20
	$\{1\dots15\}$	7.80	1.43	16.23	6.22	20.97	0.06	20.37	1.14	0.00	0.00

Moreover, these policies, in contrast with λ, are not able to find an optimal solution for most cases within the time horizon. These results reflect one of the key advantages of the greedy and robust nature of the policies in Λ, namely, the fact that the leader is guaranteed to eventually find an optimal solution to the full information problem. In addition, it is remarkable that the time-stability of the policies in λ is considerable better than their theoretical worst-case of $|A|$.

Furthermore, we observe that the performance of the proposed policies is better for the case of Value-Perfect feedback when compared to Response-Perfect feedback, under both uncertainty models, and is particularly pronounced in the simplex model with Response-Perfect feedback. This is to be expected: under Value-Perfect feedback more linearly independent equations are added on average to $\mathcal{U}^t$ at each time period. It is also noticeable that the amount of initial information does not seem to have any significant impact on policy performance under both feedback types and uncertainty models. However, we note that in other bilevel settings the amount of initial information does have a very important effect (see, e.g., discussion in Borrero et al. (2016) for an example in the context of the shortest path interdiction).

An important feature of the policies in Λ is their extremely low variability, not much larger than the semi-oracle's. In contrast, the benchmark policies are orders of magnitude more variable. Importantly, while these policies have low MAD values some cases, this is due to the fact that for most instances their time-stability is infinity (recall our earlier remark that policies that do not find an optimal solution within the first 20 periods of an instance are assigned the value $\tau^\pi = T = 21$).

We observe that our numerical experiments do not yield a clear conclusion regarding what type of uncertainty is more challenging for the leader. That is, note that for policy λ the results are fairly similar whether there is uncertainty only in the constraints or uncertainty only in the profits. In fact, for some configurations the time-stability is better for profit uncertainty while in other is better for constraint uncertainty. Note that the performance when there is uncertainty in both constraints is slightly worse than the case where only one constraint is uncertain. Interestingly, this is not always true, see, e.g., the simplex model for both Value-Perfect and Response-Perfect feedback.

For most configurations the worst performance is observed when there is uncertainty across both profits and budget constraints. The only exception being Value-Perfect feedback under the simplex model, where the results are the same as the case of profit uncertainty. Importantly, we observe that the policies in λ have a very good performance for Response-Perfect feedback whenever there is uncertainty in the constraints. Such behavior is remarkable, since as mentioned in Section 4.2.2, no theoretical results that upper bound the time-stability are yet available for these cases. This suggests that theoretical upper bounds for the matrix model under Response-Perfect feedback might be found, however their derivation might not depend on the notion of polyhedral dimension.

7. Conclusions

This paper presents a framework for addressing **SMPI** where at each period a leader allocates a series of resources so as to degrade the performance of a follower, who in turn aims at minimizing a cost function by performing a series of activities. The interaction at each time period is modeled as a bilevel program. We assume that, unlike the follower, the leader has incomplete information about the variables, constraints, and cost function of the follower's problem and has to learn them by observing the feedback generated by the follower's actions. Such feedback includes the total cost incurred by the follower, the activities performed, and any resource that might interfere with the said activities. Such settings naturally arise in military and law enforcement applications, e.g., attacker-defender and interdiction problems, which are often modeled as max-min bilevel problems.

We propose a class of policies Λ that are both greedy and robust, as they optimize the immediate performance considering worst-case realizations of the instance among those that are consistent with the information at hand. Under reasonable assumptions on the information that the leader collects from the follower's response, our theoretical results show that in **SMPI** exploitation always implies exploration as long as the leader is using policies in Λ, and moreover, their greediness and robustness are sufficient to guarantee weak optimality. Particularly, we show that the time-stability of policies in Λ is upper-bounded by the number of the follower's activities and the dimension of the cost's uncertainty polyhedron, which implies that they are guaranteed

to eventually match the actions of the oracle with prior knowledge of the instance. Moreover, we show that these policies provide the leader with a real-time certificate of optimality.

We also consider a more general setting where the leader has uncertainty regarding the follower's constraint matrix. We demonstrate that the extension of greedy and robust policies preserves most of the attractive features of their cost-model counterparts. Particularly, no extra assumptions are required to extend the time-stability upper bounds under Value-Perfect feedback, while only mild assumptions are required to preserve the upper bounds under Response-Perfect feedback.

Implementation of the proposed policies requires solving a linear MIP in each period: these problems can be solved by available commercial solvers. We also present a lower bound on the best possible achievable performance based on the actions of a *semi-oracle* that possess full information about the setting, but cannot signal it through her actions. We show that the said bound can also be computed via an MIP. Our theoretical results are supported by a series of numerical experiments that show that the proposed policies consistently outperform reasonable benchmark.

Several questions remain open at this point with regard to sequential bilevel problems with incomplete information. One of the most relevant is to study up to what point the results in this work can be extended to general (i.e., not necessarily max-min) bilevel programs. The key challenge for this broader class of problems is that, as it can be readily checked, Theorem 1 and Lemmas 1 and 2 do not hold for greedy and robust policies. This implies that these policies (*i*) no longer provide a certificate of optimality; (*ii*) do not imply that an optimal solution has been found whenever the expected cost of the leader is the same as that of the follower; and (*iii*) do not imply that the leader learns new information whenever the expectation and the value observed are different.

In addition, the study of models with more general assumptions on uncertainty, where, for instance, the leader is not certain about her upper-level data, provide an attractive avenue of future research. Regarding **SMPI**, the question of determining whether finite time-stability upper bounds can be proved for the matrix model under Response-Perfect feedback with no extra assumptions remains open, as well as to determine alternative feedback settings where finite bounds, and weak optimality, can be also be attained.

References

Agrawal, S., Wang, Z. and Ye, Y. (2014), 'A dynamic near-optimal algorithm for online linear programming', *Operations Research* **62**(4), 876–890.

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.

Audibert, J.-Y. and Bubeck, S. (2009), Minimax policies for adversarial and stochastic bandits, *in* S. Dasgupta and A. Klivans, eds, 'Proceedings of the 21st Annual Conference on Learning Theory (COLT)', Omnipress, pp. 217–226.

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

Auer, P., Cesa-Bianchi, N., Freund, Y. and Schapire, R. E. (2002), 'The nonstochastic multiarmed bandit problem', *SIAM Journal on Computing* **32**(1), 48–77.

Bard, J. F., Plummer, J. and Sourie, J. C. (2000), 'A bilevel programming approach to determining tax credits for biofuel production', *European Journal of Operational Research* **120**(1), 30–46.

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

Bertsimas, D. and Tsitsiklis, J. N. (1997), *Introduction to linear optimization*, Vol. 6, Athena Scientific Belmont, MA.

Borrero, J. S. (2017), Sequential Bilevel Linear Programming with Incomplete Information and Learning, PhD dissertation, University of Pittsburgh.

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

Brown, G., Carlyle, M., Diehl, D., Kline, J. and Wood, K. (2005), 'A two-sided optimization for theater ballistic missile defense', *Operations Research* **53**(5), 745–763.

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

Bubeck, S. and Cesa-Bianchi, N. (2012), 'Regret analysis of stochastic and nonstochastic multi-armed bandit problems', *CoRR* **abs/1204.5721**.
URL: *http://arxiv.org/abs/1204.5721*

Caprara, A., Carvalho, M., Lodi, A. and Woeginger, G. J. (2013), A complexity and approximability study of the bilevel knapsack problem, *in* 'Integer programming and combinatorial optimization', Springer, pp. 98–109.

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 programa 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*, Kluwer Academic Publishers.

DeNegre, S. (2011), Interdiction and discrete bilevel linear programming, PhD thesis, Lehigh University.

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

Hazan, E. (2015), 'Introduction to online convex optimization (Draft)', Foundations and Trends in Optimization.
URL: *http://ocobook.cs.princeton.edu/OCObook.pdf*

Held, H., Hemmecke, R. and Woodruff, D. (2005), 'A decomposition algorithm applied to planning the interdiction of stochastic networks', *Naval Research Logistics* **52**(4), 321–328.

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

Janjarassuk, U. and Linderoth, J. (2008), 'Reformulation and sampling to solve a stochastic network interdiction problem', *Networks* **52**(3), 120–132.

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

Morton, D., Pan, F. and Saeger, K. (2007), 'Models for nuclear smuggling interdiction', *IIE Transactions* **39**(1), 3–14.

Ratliff, H., Sicilia, G. and Lubore, S. (1975), 'Finding the n most vital links in flow networks', *Management Science* **21**(5), 531–539.

Salmeron, J., Wood, K. and Baldick, R. (2004), 'Analysis of electric grid security under terrorist threat', *IEEE Transactions on Power Systems* **19**(2), 905–912.

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 Lim, C. (2008), Algorithms for network interdiction and fortification games, *in* 'Pareto optimality, game theory and equilibria', Springer, pp. 609–644.

Wolsey, L. A. and Nemhauser, G. L. (2014), *Integer and combinatorial optimization*, John Wiley & Sons.

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

Wood, R. K. (2011), 'Bilevel network interdiction models: Formulations and solutions', *Wiley Encyclopedia of Operations Research and Management Science* .

Zare, M. H., Borrero, J. S., Prokopyev, O. A. and Zeng, B. (2017), 'A note on linearized reformulations for a class of bilevel linear integer problems', *To appear in Annals of Operations Research* .

Zinkevich, M. (2003), Online convex programming and generalized infinitesimal gradient ascent, *in* T. Fawcett and N. Mishra, eds, 'Proceedings of the Twentieth International Conference on Machine Learning', AAAI, pp. 928–936.

Appendix A: Additional Results and Complementary Material

A.1. Proof of Proposition 6

Before proceeding with the proof of Proposition 6, additional notation, concepts and results need to be introduced. In the discussion that follows, let us suppose that in Response-Perfect feedback, besides observing the values of y_a^t the leader is also able to observe the value of the left-hand side (or, equivalently, the slack q_d^t) for all constraints $d \in C_F^{t+1}$. For simplicity, let us denote $r_d^t := \sum_{a:\, y_a^t>0} F_{da} y_a^t = f_d - q_d^t - \boldsymbol{L}_d^\top x^t$. Then, by using the information from the feedback, the leader updates $\mathcal{U}^t$ by including the linear constraints

$$\sum_{a \in y_a^t > 0} y_a^t \hat{F}_{da} = r_d^t \qquad \text{for all } d \in C_F^{t+1}, \tag{A-1}$$

in the definition of polyhedron $\mathcal{U}^{t+1}$. Recall that for any $d \in C_F^t$, n_d^t denotes the number of the follower's activities in A^t that d restricts, that is $n_d^t := |\{a \in A^t : d \in C_F(a)\}|$. As such, for any given time $t \in \mathcal{T}$ we have that

$$\mathcal{U}^t \subseteq \mathbb{R}^{\sum_{d \in C_F^t} n_d^t}.$$

Denote $m^t = |C_F^t|$ and let us write $C_F^t = \{d_1, \cdots, d_{m^t}\}$. We organize the elements of $\mathcal{U}^t$ into blocks, so that $\hat{\boldsymbol{F}} \in \mathcal{U}^t$ is given by

$$\hat{\boldsymbol{F}} = [\hat{\boldsymbol{F}}^{d_1}; \hat{\boldsymbol{F}}^{d_2}; \cdots; \hat{\boldsymbol{F}}^{d_{m^t}}],$$

where $\hat{\boldsymbol{F}}^d \in \mathbb{R}^{n_d^t}$ for all $d \in C_F^t$. We also assume that the columns of matrix $\boldsymbol{G}^t$ are organized in this way. Using the conventions above, for any $d \in C_F^{t+1}$, constraint (A-1) can be rewritten as

$$\boldsymbol{v}_d^\top \hat{\boldsymbol{F}} = r_d^t, \tag{A-2}$$

where vector $\boldsymbol{v}_d$ is divided in subvectors as $\boldsymbol{v}_d := [\boldsymbol{v}_d^{d_1}; \boldsymbol{v}_d^{d_2}; \cdots; \boldsymbol{v}_d^{d_{m^{t+1}}}]$, and each subvector $\boldsymbol{v}_d^{d_j} \in \mathbb{R}^{n_{d_j}^{t+1}}$. If $d \neq d_j$, then $\boldsymbol{v}_d^{d_j}$ is a vector of zeros, i.e., $\boldsymbol{v}_d^{d_j} = \boldsymbol{0}_{n_{d_j}^{t+1}}^\top$. Otherwise, if $d = d_j$, then it has the information of $y^{t,\lambda}$ for those activities in A^{t+1} that are restricted by d, i.e, $(\boldsymbol{v}_d^d)_a = y_a^{t,\lambda}$ for all $a \in A^{t+1}$ such that $d \in C_F(a)$.

Let $\mathcal{D}^0$ and $\mathcal{D} \in \mathbb{G}(\mathcal{D}^0)$ be given, and suppose that T is sufficiently large. For any π, define $\mathcal{S}^\pi(\mathcal{D}^0, \mathcal{D}) := \{t \in \mathcal{T} : \exists a \notin A^t \text{ s.t. } y_a^t > 0\}$, that is, $\mathcal{S}^\pi(\mathcal{D}^0, \mathcal{D})$ is the set of time periods when at least a new activity is learned by the leader (who is using policy π). Suppose that $\mathcal{S}^\pi(\mathcal{D}^0, \mathcal{D}) = \{s_1, s_2, \cdots, s_p\}$, where w.l.o.g. we suppose that $s_k < s_{k+1}$ for all $k \leq p-1$ (observe p depends on π, we drop it for the notation for simplicity). In addition, for any $k = 1, \cdots, p$, define $N^k := \{a \in A \setminus A^{s_k} : y_a^{s_k} > 0\}$, i.e., N^k is the set of activities the leader learns by the end of time period s_k.

LEMMA 4. *Let $\lambda \in \Lambda$, suppose that feedback $\mathcal{F}$ is Response-Perfect and that the leader observes the values of all the slack variables of the follower problem at any time $t \in \mathcal{T}$. If $\xi^\lambda > s_p$ then,*

$$\dim(\mathcal{U}^{t+1}) - \dim(\mathcal{U}^t) \leq \begin{cases} \sum_{a \in N^k} |C_F(a)| - \left| \bigcup_{a \in N^k} C_F(a) \right|, & \text{if } t = s_k \text{ for some } k \leq p, \\ -1, & \text{otherwise.} \end{cases} \tag{A-3}$$

Proof. Let $k < p$ be given. Observe that at the end of period s_k the leader learns all the activities in N^k, and as such introduces a new variable $\hat{F}_{da}$ into $\mathcal{U}^{s_k+1}$ for all $d \in C_F(a)$ and $a \in N^k$. Hence, $\mathcal{U}^{s_k+1}$ has $\sum_{a \in N^k} |C_F(a)|$ more variables (columns) than $\mathcal{U}^{s_k}$ (observe that there is no new variable $\hat{F}_{da}$ for $a \in A^t$ from the standard feedback assumption). On the other hand, for every $d \in \bigcup_{a \in N^k} C_F(a)$ the leader includes the linear constraint (A-2) into $\mathcal{U}^{s_k+1}$ (in addition to the potentially new constraints associated with each $d \in C_F^t$).

From the definition of $\boldsymbol{v}_d$ in equation (A-2), it is readily seen that if $d \neq d'$, and both $d, d' \in \bigcup_{a \in N^k} C_F(a)$, then $(\boldsymbol{v}_d; r_d^{s_k})$ and $(\boldsymbol{v}_{d'}; r_{d'}^{s_k})$ are linearly independent. Moreover, it is also readily observed that these vectors are linearly independent of all the other (expanded) vectors that give equality constraints in $\mathcal{U}^{s_k}$.

The above analysis implies that, with respect to $\dim(\mathcal{U}^{s_k})$, $\dim(\mathcal{U}^{s_k+1})$ increases by $\sum_{a \in N^k} |C_F(a)|$ because of the new variables, but $\dim(\mathcal{U}^{s_k+1})$ decreases by (at least)

$$\left| \bigcup_{a \in N^k} C_F(a) \right|$$

because of the newly introduced linearly independent equality constraints. In other words,

$$\dim(\mathcal{U}^{s_k+1}) \leq \dim(\mathcal{U}^{s_k}) + \sum_{a \in N^k} |C_F(a)| - \left| \bigcup_{a \in N^k} C_F(a) \right|. \tag{A-4}$$

On the other hand, let $t < \xi^\lambda$ such that $t \notin \mathcal{S}^\lambda$; i.e., $y_a^t = 0$ for all $a \notin A^t$. Note that because $\xi^\lambda > t$ one has that $\boldsymbol{c}^\top y^{t,\lambda} < z_R^t$ (by part (i) of Proposition 4). We claim that (recall from the proof of Lemma 2 the definition of $\boldsymbol{G}^{t,=}$ and $\boldsymbol{g}^{t,=}$)

$$\operatorname{rank}([\boldsymbol{G}^{t+1,=}, \boldsymbol{g}^{t,=}]) > \operatorname{rank}([\boldsymbol{G}^{t,=}, \boldsymbol{g}^{t,=}]).$$

Indeed, because the assumptions of Lemma 3 hold, let $\widetilde{\boldsymbol{F}}^t$ such that

$$\left(\widetilde{\boldsymbol{F}}^t\right)_d^\top y^{t,\lambda} > f_d - (\boldsymbol{L}^t)_d^\top x^{t,\lambda}.$$

Now consider $\mathcal{U}^t$ after adding the equation $\boldsymbol{v}_d^\top \hat{\boldsymbol{F}} = r_d^t$. Because $q_d^t \geq 0$, one has that $\widetilde{\boldsymbol{F}}_d^T y^{t,\lambda} > f_d - (\boldsymbol{L}^t)_d^\top x^t - q_d^t$ and hence $\widetilde{\boldsymbol{F}} \notin \mathcal{U}^{t+1}$. Therefore, $\widetilde{\boldsymbol{F}}^t \in \mathcal{U}^t \setminus \mathcal{U}^{t+1}$ and, by the same arguments of Lemma 2, the vector $(\boldsymbol{v}_d; k_d)$ must be linearly independent from all the rows of $(\boldsymbol{G}^t, \boldsymbol{g}^t)$. Therefore, the desired claim follows and we can conclude that $\dim(\mathcal{U}^{t+1}) \leq \dim(\mathcal{U}^t) - 1$, as desired. ∎

LEMMA 5. *Let $\lambda \in \Lambda$ be given, suppose that the feedback $\mathcal{F}$ is Response-Perfect and that the leader observes the values of all the slack variables of the follower's problem at any time $t \in \mathcal{T}$. Then, $s_1 + \dim(\mathcal{U}^{s_1}) \leq \dim(\mathcal{U}^0)$, and*

$$s_{k+1} + \dim(\mathcal{U}^{s_{k+1}}) \leq s_k + \dim(\mathcal{U}^{s_k}) + 1 + \sum_{a \in N^k} |C_F(a)| - \Big| \bigcup_{a \in N^k} C_F(a) \Big| \qquad k = 1, \cdots, p-1.$$

Proof. By the definition of s_1, at periods $t = 0, 1, 2, \cdots, s_1 - 1$ we have that the leader does not learn any activity and hence, by Lemma 4, $\dim(\mathcal{U}^t) - \dim(\mathcal{U}^{t-1}) \leq -1$ for $t = 1, \cdots, s_1$. This implies that $\dim(\mathcal{U}^{s_1}) \leq \dim(\mathcal{U}^0) - s_1$ and the result follows. Suppose that $k = 1, \cdots, p-1$ is given. By definition of s_{k+1}, from $t = s_k + 1, \cdots, s_{k+1} - 1$ the leader does not learn any activity and Lemma 4 again implies that $\dim(\mathcal{U}^t) - \dim(\mathcal{U}^{t-1}) \leq -1$, $t = s_k + 2, \cdots, s_{k+1}$. This observation implies that

$$\dim(\mathcal{U}^{s_{k+1}}) \leq \dim(\mathcal{U}^{s_k+1}) - (s_{k+1} - s_k - 1).$$

Now, the above equation along with equation (A-4) imply that

$$\dim(\mathcal{U}^{s_{k+1}}) \leq \dim(\mathcal{U}^{s_k}) + \sum_{a \in N^k} |C_F(a)| - \Big| \bigcup_{a \in N^k} C_F(a) \Big| - s_{k+1} + s_k + 1,$$

which yields the desired result. ∎

Using the above Lemma 5 we have the following important result.

LEMMA 6. *Let $\lambda \in \Lambda$ be given, suppose that the feedback $\mathcal{F}$ is Response-Perfect and that the leader observes the values of all the slack variables of the follower problem at any time $t \in \mathcal{T}$. Then,*

$$\tau^\lambda \leq \xi^\lambda \leq \dim(\mathcal{U}^0) + p + \sum_{k=1}^{p} \Big(\sum_{a \in N^k} |C_F(a)| - \Big| \bigcup_{a \in N^k} C_F(a) \Big| \Big). \tag{A-5}$$

Proof. By repeated application of Lemma 5 it is verified that

$$s_p + \dim(\mathcal{U}^{s_p}) \leq \dim(\mathcal{U}^0) + (p-1) + \sum_{k=1}^{p-1} \Big(\sum_{a \in N^k} |C_F(a)| - \Big| \bigcup_{a \in N^k} C_F(a) \Big| \Big). \tag{A-6}$$

Because by definition no new action is learned after s_p, $\dim(\mathcal{U}^t) - \dim(\mathcal{U}^{t-1}) \leq -1$ for $t \geq s_p + 2$. This implies that at most by time $s_p + \tilde{t}$, where $\tilde{t} := \sum_{a \in N^p} |C_F(a)| - \Big| \bigcup_{a \in N^p} C_F(a) \Big| + 1$, it must be the case that $\dim(\mathcal{U}^{s_p + \tilde{t}}) = 0$. Henceforth, part (iii) of Proposition 4 implies that $\xi^\lambda \leq s_p + \tilde{t}$, and hence equation (A-6) and the selection of t yield the desired result. ∎

Proof of Proposition 6. Suppose first that (i) holds, i.e., that all the constraints are equality constraints, thus the leader always knows that their slack is zero. Hence, a direct application of Lemma 6 implies that

$$\tau^\lambda \leq \xi^\lambda \leq \dim(\mathcal{U}^0) + p + \sum_{k=1}^{p} \Big(\sum_{a \in N^k} |C_F(a)| - \Big| \bigcup_{a \in N^k} C_F(a) \Big| \Big).$$

The desired result follows by noting that $\sum_{k=1}^{p}\sum_{a\in N^k}|C_F(a)| = \sum_{a\in A\setminus A^0}|C_F(a)|$ and that $\left|\bigcup_{a\in N^k} C_F(a)\right| \geq 1$. On the other hand, consider (ii), i.e., that the leader observes the slack of one of the constraints in $D^{t,\lambda}$ at every period $t\in\mathcal{T}$ such that $y_a^t = 0$ for all $a\notin A^t$. In this case, following the same arguments as in Lemma 4, equation (A-3) can be simplified to:

$$\dim(\mathcal{U}^{t+1}) - \dim(\mathcal{U}^t) \leq \begin{cases} \sum_{a\in N^k}|C_F(a)|, & \text{if } t = s_k, \text{ for some } k\leq p,\\ -1, & \text{otherwise.}\end{cases}$$

The result follows from Lemma 6, after mimicking the proofs of the previous results, as in this case equation (A-5) becomes

$$\tau^\lambda \leq \xi^\lambda \leq \dim(\mathcal{U}^0) + p + \sum_{k=1}^{p}\sum_{a\in N^k}|C_F(a)|.$$

■

A.2. Semi-Oracle Algorithm

In this section we discuss an algorithm that speeds-up the solution of problem (16) and that is particularly useful to determine the semi-oracle decisions for instances where T is large. The algorithm works by computing a time-stability upper bound, which is constructed by forcing the follower to reveal an 'optimal' set of resources I^* as soon as possible. Once this upper bound is computed, MIP (16) is solved by truncating the time to T^0, which, as it will be seen, can be bounded by the cardinality of I^*. Then, the optimal solution of the original MIP is obtained by extending the truncated solution until time T.

Before proceeding, we introduce some additional notation. Let x^* be an optimal solution of the full-information problem, and let $I^* \coloneqq \{i\in I\colon x_i^* > 0\}$ be the set of resources that x^* uses. For any $J\subseteq I^*$ define $x^{*,J}$ as $x_i^{*,J} \coloneqq x_i^*$ if $i\in J$ and zero otherwise, thus $x^{*,J}$ is the restriction of x^* to the resources in J. In addition, for any y define (with a slight abuse of notation)

$$I(y) \coloneqq \bigcup_{a\colon y_a>0} I(a)$$

i.e., $I(y)$ is the set of resources that interfere with the activities that y performs.

The computation of the upper bound T^0 is based on the two following observations: (i) as soon as the semi-oracle enforces the follower to reveal all the resources in I^*, then she can implement the optimal solution x^*; (ii) if for a given $J\subset I^*$ the semi-oracle implements $x^{*,J}$, then the response of the follower must reveal a new resource in $I^*\setminus J$, or else the response yields the optimal value z^*. While the proof of (i) is straightforward, the proof of (ii) is a consequence of the following:

LEMMA 7. *Let $J\subseteq I^*$ and $y^J\in\arg\min\{\boldsymbol{c}^\top y\colon y\in Y(x^{*,J})\}$. If $I(y^J)\cap I^*\subseteq J$, then $z^*\leq \boldsymbol{c}^\top y^J$.*

Proof. We proceed to prove that $y^J \in Y(x^*)$. Note that if this holds, then $z^* \leq \boldsymbol{c}^\top y^J$ by the definition of z^*. Indeed, let $d \in C_F$ and note that

$$\begin{aligned}\sum_{a\in A} F_{da} y_a^J + \sum_{i\in I^*} L_{di} x_i^* &= \sum_{a\in A} F_{da} y_a^J + \sum_{i\in J} L_{di} x_i^* + \sum_{i\in I^*\setminus J} L_{di} x_i^* \\ &= \sum_{a\in A} F_{da} y_a^J + \sum_{i\in J} L_{di} x_i^* + \sum_{i\in K_1} L_{di} x_i^* + \sum_{i\in K_2} L_{di} x_i^*, \end{aligned} \tag{A-1}$$

where in the last equation $K_1 = (I^* \setminus J) \cap I(y^J)$ and $K_2 = (I^* \setminus J) \setminus I(y^J)$. Our objective is to prove that the expression in (A-1) is at most f_d for all $d \in C_F$; from this the desired result follows.

First, suppose that $d \in C_F$ satisfies that $\sum_{a\in A} F_{da} y_a^J = 0$; then (A-1) is at most f_d by Assumption **A4**. Hence, suppose that $d \in C_F$ satisfies that $\sum_{a\in A} F_{da} y_a^J \neq 0$. Note that $K_1 = I^* \cap (I \setminus J) \cap I(y^J) = (I \setminus J) \cap (I(y^J) \cap I^*) = \emptyset$, because by hypothesis $I(y^J) \cap I^* \subseteq J$; therefore, $\sum_{i\in K_1} L_{di} x_i^* = 0$. On the other hand, suppose that $i \in K_2$. Then $i \notin I(y^J)$ and, since $\sum_{a\in A} F_{da} y_a^J \neq 0$, it must be the case that $L_{di} = 0$. As this holds for any $i \in K_2$, we have that $\sum_{i\in K_2} L_{di} x_i^* = 0$.

From the above observations, it follows that if $\sum_{a\in A} F_{da} y_a^J \neq 0$ then

$$\sum_{a\in A} F_{da} y_a^J + \sum_{i\in I^*} L_{di} x_i^* = \sum_{a\in A} F_{da} y_a^J + \sum_{i\in J} L_{di} x_i^* \leq f_d,$$

where the inequality in the above expression is a consequence of the assumption that $y^J \in Y(x^{*,J})$. Thus, (A-1) is at most f_d for any $d \in C_F$ and hence $y^J \in Y(x^*)$, as desired. ∎

Supported by the observations above, Algorithm 1 outputs an initial feasible solution. It starts by computing x^* and z^*. At any time t it implements the solution x^{*,J^t}, with $J^t = I^* \cap I^t$. If the follower's solution at t yields a value less than z^*, then, per observation (ii), the semi-oracle can use a new resource in I^* at the next time period; otherwise, the solution implemented at t is optimal. The value of T^0 is set to be the first time that z^* is equal to the follower's cost. We note that T^0 is upper-bounded by $|I^*|$ since in at most $|I^*|$ periods the semi-oracle discovers all the resources in I^*, and once these resources are available, the solution of the semi-oracle is optimal, per observation (i). The above considerations are formalized in Lemma 8.

LEMMA 8. *Let T^0 be as computed by Algorithm 1. Then, T^0 is an upper bound on the optimal value of problem* (16)*, and if $|I^* \setminus I^0| \leq T$, then $T^0 \leq |I^* \setminus I^0|$.*

Proof. First, if the algorithm outputs $T^0 = \infty$, the results holds trivially. Hence, suppose $T^0 < \infty$. In this case, it is readily checked that T^0 is an upper bound as the solution $\{(x^{*,J^t}, y^t) \colon t \in \mathcal{T}\}$ output by Algorithm 1 is feasible in (16) and yields an objective value of T^0.

On the other hand, suppose that $|I^* \setminus I^0| \leq T$ and let $s \in \mathcal{T} \setminus \{0\}$ be given such that $z^* > z^r$ for all $r \leq s$. Because $J^s \subseteq I^*$, $y^s \in \arg\min\{\boldsymbol{c}^\top y \colon y \in Y(x^{*,J^s})\}$, and $z^s = \boldsymbol{c}^\top y^s$, Lemma 7 implies that there exist $i \in I(y^s) \cap I^*$ such that $i \notin J^s$. Henceforth, $|J^{s+1} \setminus J^s| \geq 1$.

Algorithm 1 Finding an initial feasible solution to (16).

Require: $(\mathcal{D}^0, \mathcal{D})$, T

Compute x^* and z^*

$J^0 = I^0 \cap I^*$, $y^0 \in \arg\min\{\boldsymbol{c}^\top y \colon y \in Y(x^{*,J^0})\}$, $z^0 = \boldsymbol{c}^\top y^0$, $t = 0$

while $z^* > z^t$ and $t \leq T$ **do**

$J^{t+1} = J^t \cup (I(y^t) \cap I^*)$

$y^{t+1} \in \arg\min\{\boldsymbol{c}^\top y \colon y \in Y(x^{*,J^{t+1}})\}$, $z^{t+1} = \boldsymbol{c}^\top y^{t+1}$, $t = t+1$

end while

if $z^* = z^t$ **then**

$T^0 = t$, $z^s = z^*$, $x^{*,J^s} = x^*$, $y^s = y^t$ for $s = t+1, \cdots, T$

else

$T^0 = \infty$

end if

return T^0, z^*, $\{(x^{*,J^t}, y^t) \colon t \in \mathcal{T}\}$

In order to arrive at a contradiction, suppose that $T^0 > |I^* \setminus I^0|$. This implies that if we let $t = |I^* \setminus I^0|$, then $z^* > z^s$ for all $s \leq t$, and,

$$|J^t| = |J^0| + \sum_{s=1}^{|I^* \setminus I^0|} |J^s \setminus J^{s-1}| \geq |J^0| + |I^* \setminus I^0| = |I^* \cap I^0| + |I^* \setminus I^0| = |I^*|. \tag{A-2}$$

where the inequality follows as $|J^s \setminus J^{s-1}| \geq 1$ for all $s \leq t$. By construction, we have that $J^t \subseteq I^*$ for any t, thus inequality (A-2) implies that $J^t = I^*$, and hence, by observation (i) that $z^t = z^*$; which yields the desired contradiction. ■

By using Algorithm 1, an optimal solution of (16) can be readily computed via Algorithm 2. The correctness of Algorithm 2 follows from noting that T^0 is an upper bound for the time-stability. Hence, we have the following result, which we state without proof.

PROPOSITION 7. *Algorithm 2 correctly solves program* (16).

A.3. Numerical Computation of Policies in Λ

Next we establish that $x^{t,\lambda}$ and $z_R^{t,*}$ can be computed by solving a mixed-integer linear problem.

LEMMA 9. *Let $t \in \mathcal{T}$ be given and suppose that for all $x \in X^t$ the problem $z_R^t(x)$ has an optimal solution. Then,*

$$z_R^{t,*} = \max \ (\boldsymbol{g}^t)^\top p \tag{A-3a}$$

$$s.t. \ \boldsymbol{H}^t x \leq \boldsymbol{h}^t \tag{A-3b}$$

Algorithm 2 Finding an optimal solution to (16)

Require: $(\mathcal{D}^0, \mathcal{D})$, T

Compute $(T^0, z^*, \{(x^t, y^t) : t \in \mathcal{T}\})$ by calling **Algorithm 1** using $((\mathcal{D}^0, \mathcal{D}), T)$

if $T^0 \leq T$ **then**

Solve program (16) until time T^0 passing $\{(x^t, y^t) : t = 0, \cdots, T^0\}$ as an initial feasible solution, let τ^* be the objective value

else

Solve program (16) until time T passing $\{(x^t, y^t) : t = 0, \cdots, T\}$ as an initial feasible solution, let τ^* be the objective value

if $\tau^* = T + 1$ **then**

$\tau^* = \infty$

end if

end if

return τ^*

$$(\boldsymbol{G}^t)^\top p - y = \mathbf{0} \tag{A-3c}$$

$$\boldsymbol{F}^t y + \boldsymbol{L}^t x \leq \boldsymbol{f}^t \tag{A-3d}$$

$$\boldsymbol{G}^t \hat{c}^t \leq \boldsymbol{g}^t \tag{A-3e}$$

$$-(\boldsymbol{F}^t)^\top q - \hat{c}^t \leq \mathbf{0} \tag{A-3f}$$

$$q \leq \boldsymbol{M}^q v^1, \ \boldsymbol{f}^t - \boldsymbol{F}^t y - \boldsymbol{L}^t x \leq \boldsymbol{M}^q (1 - v^1) \tag{A-3g}$$

$$p \leq \boldsymbol{M}^p v^2, \ \boldsymbol{g}^t - \boldsymbol{G}^t \hat{c}^t \leq \boldsymbol{M}^p (1 - v^2) \tag{A-3h}$$

$$y \leq \boldsymbol{M}^y v^3, \ (\boldsymbol{F}^t)^\top q + \hat{c}^t \leq \boldsymbol{M}^y (1 - v^3) \tag{A-3i}$$

$$y \in \mathbb{R}_+^{|A^t|}, \hat{c}^t \in \mathbb{R}^{|A^t|}, q \in \mathbb{R}_+^{|C_F^t|} \tag{A-3j}$$

$$p \in \mathbb{R}_+^{|C_U^t|}, x \in \mathbb{R}_+^{|I^t| - k^t} \times \mathbb{Z}^{k^t} \tag{A-3k}$$

$$v^1 \in \{0,1\}^{|C_F^t|}, v^2 \in \{0,1\}^{|C_U^t|}, v^3 \in \{0,1\}^{|A^t|}. \tag{A-3l}$$

where in the above equations $\boldsymbol{M}^q$, $\boldsymbol{M}^p$, and $\boldsymbol{M}^y$ are diagonal matrices whose elements are large enough numbers. Specifically, if $(x, q, p, y, \hat{c}^t)$ satisfies equations (A-3b)–(A-3f)*, then $\boldsymbol{M}^q$ is such that $\max\{q_d, \boldsymbol{f}_d^t - \boldsymbol{F}_d^t y - \boldsymbol{L}_d^t x\} \leq \boldsymbol{M}_{dd}^q$ for any given $d \in C_F^t$ ($\boldsymbol{M}^p$ and $\boldsymbol{M}^y$ are defined analogously). Moreover, $x^{t,\lambda}$ can be computed as $x^{t,\lambda} = \tilde{x}$ where $(\tilde{x}, \tilde{q}, \tilde{p}, \tilde{y}, \tilde{c})$ is an optimal solution of the program* (A-3a)–(A-3l)*.*

Proof. The optimization problem $\max\{z_R^t(x) : x \in X^t\}$ can be written as

$$\max\Big\{\min\big\{y_0 : (\hat{\boldsymbol{c}}^t)^\top y \leq y_0 \ \forall \hat{\boldsymbol{c}}^t \in \mathcal{U}^t, -\boldsymbol{F}^t y \geq \boldsymbol{L}^t x - \boldsymbol{f}^t, y \in \mathbb{R}_+^{|A^t|}, y_0 \in \mathbb{R}\big\} : x \in X^t\Big\}. \tag{A-4}$$

Recall that $\mathcal{U}^t = \{\hat{\boldsymbol{c}}^t \colon \boldsymbol{G}^t\hat{\boldsymbol{c}}^t \leq \boldsymbol{g}^t\}$. The vector y satisfies the robust constraint $(\hat{\boldsymbol{c}}^t)^\top y \leq y_0 \ \forall \hat{\boldsymbol{c}}^t \in \mathcal{U}^t$ if and only if there exist $p \in \mathbb{R}_+^{|C_U^t|}$ such that

$$(\boldsymbol{g}^t)^\top p \leq y_0 \text{ and } (\boldsymbol{G}^t)^\top p = y$$

(see, e.g., Ben-Tal et al. (2009)). Moreover, due to the objective function and to the fact that there are no other constraints on y_0, it follows that problem (A-4) is equivalent to

$$\max_{x\in X^t}\Big\{\min\big\{(\boldsymbol{g}^t)^\top p\colon\ -y+(\boldsymbol{G}^t)^\top p=\boldsymbol{0},\ -\boldsymbol{F}^t y\geq \boldsymbol{L}^t x-\boldsymbol{f}^t,\ y\in\mathbb{R}_+^{|A^t|}, p\in\mathbb{R}_+^{|C_U^t|}\big\}\colon x\in X^t\Big\}. \tag{A-5}$$

Since for any $x \in X^t$ it is assumed that $z_R^t(x)$ has an optimal solution, any optimal solution y of the inner minimization problem satisfies its Karush-Kuhn-Tucker (KKT) optimality conditions (and vice-versa). Hence, replacing the minimization problem by the KKT conditions yields

$$\max_{x\in X^t}\ (\boldsymbol{g}^t)^\top p \tag{A-6a}$$
$$\text{s.t. } -y+(\boldsymbol{G}^t)^\top p=\boldsymbol{0} \tag{A-6b}$$
$$-\boldsymbol{F}^t y\geq \boldsymbol{L}^t x-\boldsymbol{f}^t \tag{A-6c}$$
$$-(\boldsymbol{F}^t)^\top q-\hat{c}^t\leq \boldsymbol{0} \tag{A-6d}$$
$$\boldsymbol{G}^t\hat{c}^t\leq \boldsymbol{g}^t \tag{A-6e}$$
$$(\boldsymbol{f}^t-\boldsymbol{F}^t-\boldsymbol{L}^t x)^\top q=0 \tag{A-6f}$$
$$(\boldsymbol{g}^t-\boldsymbol{G}^t\hat{c}^t)^\top p=0 \tag{A-6g}$$
$$((\boldsymbol{F}^t)^\top q+\hat{c}^t)^\top y=0 \tag{A-6h}$$
$$y\in\mathbb{R}_+^{|A^t|}, q\in\mathbb{R}_+^{|C_F^t|}, p\in\mathbb{R}_+^{|C_U^t|}, \hat{c}^t\in\mathbb{R}^{|A^t|}. \tag{A-6i}$$

Observe that problem (A-6) is a non-linear mixed-integer problem (due to the non-linear complementary slackness constraints). However, it can be linearized by introducing 0-1 variables. Indeed, q, y and x satisfy the constraint $(\boldsymbol{f}^t-\boldsymbol{F}^t-\boldsymbol{L}^t x)^\top q=0$ if and only if there exists $v^1\in\{0,1\}^{|C_F^t|}$ such that (see, e.g., Audet et al. (1997)) $q\leq \boldsymbol{M}^q v^1$ and $\boldsymbol{f}-\boldsymbol{F}^t y-\boldsymbol{L}^t x\leq \boldsymbol{M}^q(\boldsymbol{1}-v^1)$. A similar equivalence exists between the other two set of complementary slackness constraints in problem (A-6).

■

We note that whenever the leader's variables are all discrete, then $z_R^{t,*}$ and $x^{t,\lambda}$ can be computed using a different MILP. In this case, the transformation of problem (A-5) into a one-level problem involves using the strong duality optimality conditions, and then linearizing any resulting nonlinear product. We use this approach for the numerical experiments in Section 6 as it yields shorter solution times; details can be found in Zare et al. (2017).