SEQUENTIAL BILEVEL LINEAR PROGRAMMING WITH INCOMPLETE INFORMATION AND LEARNING

by

Juan Sebastián Borrero

B.Sc., Universidad de los Andes, 2008

M.Sc., Universidad de los Andes, 2010

Submitted to the Graduate Faculty of

the Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Juan Sebastián Borrero

It was defended on

June 15th 2017

and approved by

Oleg A. Prokopyev, Ph.D., Associate Professor, Department of Industrial Engineering

Denis Sauré, Ph.D., Assistant Professor, Department of Industrial Engineering,

Universidad de Chile

Jayant Rajgopal, Ph.D., Professor, Department of Industrial Engineering

Pavlo Krokhmal, Ph.D., Professor, Department of Systems & Industrial Engineering,

University of Arizona

Bo Zeng, Ph.D., Assistant Professor, Department of Industrial Engineering

Dissertation Director: Oleg A. Prokopyev, Ph.D., Associate Professor, Department of

Industrial Engineering

SEQUENTIAL BILEVEL LINEAR PROGRAMMING WITH INCOMPLETE INFORMATION AND LEARNING

Juan Sebastián Borrero, PhD

University of Pittsburgh, 2017

We present a framework for a class of sequential decision-making problems in the context of bilevel linear programming, where a leader and a follower repeatedly interact. At each period, the leader allocates resources that can modify the performance of the follower (e.g., as in interdiction or defender-attacker problems). The follower, in turn, optimizes 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, who decides as to optimize her objective function, has only partial information. As such, she 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 the performance of any given leader's decision-making policy in terms of its time-stability, defined as the number of periods it takes the policy to match the actions of an oracle decision-maker with complete information of the bilevel problem.

Three types of bilevel models are considered: Shortest path interdiction, max-min bilevel linear problems, and asymmetric bilevel linear problems. For shortest path interdiction we discuss greedy and pessimistic policies, and show that their time stability is upper-bounded by the number of arcs in the network; moreover, these policies are not dominated by any non-greedy or non-pessimistic policy. We refine these ideas into the more general max-min bilevel problems. Here we show that there is a class of greedy and robust policies that have the best possible worst-case performance, eventually match the oracle's actions, provide a real-time optimality certificate, and can be computed using mixed-integer linear programming. These policies, however, do not retain their features for asymmetric bilevel problems. For this set-

ting we study the performance of greedy and best-case policies and show that they keep many of the attractive properties that the greedy and robust policies have for the max-min case.

By performing computational experiments under different configurations, we show that the proposed policies compare favorably against different benchmark policies. Moreover, they perform reasonably close to the semi-oracle, that is a novel decision-maker we introduce that provides a lower bound on the time-stability of any policy.

Keywords: Bilevel optimization, Online Optimization, Robust Optimization, Interdiction, Learning.

TABLE OF CONTENTS

PREFACE x

1.0 INTRODUCTION 1

2.0 SEQUENTIAL SHORTEST PATH INTERDICTION WITH INCOMPLETE INFORMATION AND LEARNING 6

2.1 Introduction 6

2.2 Problem Formulation 10

2.3 Efficient Interdiction Policies 17

2.3.1 Efficient Policies When $\widehat{A}_0 = \emptyset$ 18

2.3.2 Efficient Policies When $\widehat{A}_0 \neq \emptyset$ 26

2.4 Lower Bounds for Policy Performance 30

2.4.1 Semi-oracle Policies 30

2.4.2 Lower Bound for Regret 32

2.4.3 Lower Bound for Time-Stability 37

2.5 Computational Study 38

2.5.1 Test Instances, Benchmark Policies and Implementation Details . . 38

2.5.2 Computation of the Oracle-based Policy 39

2.5.3 Comparison to Benchmark Policies 40

2.5.4 Policy Performance: Sensitivity with Respect to $|\widetilde{A}_0|$ 44

2.5.5 Policy Performance: Sensitivity with Respect to Quality of Bounds in $\widehat{A}_0$ 47

2.6 Concluding Remarks 50

3.0 SEQUENTIAL MAX-MIN BILEVEL LINEAR PROGRAMMING WITH INCOMPLETE INFORMATION AND LEARNING 53
3.1 Introduction 53
3.2 Basic Model: Cost Uncertainty 57
3.2.1 Feedback 63
3.2.2 Optimality Criteria 67
3.3 Greedy and Robust Policies 71
3.3.1 General Results for Standard Feedback 71
3.3.2 Policies in Λ Under Value–Perfect Feedback 73
3.3.3 Policies in Λ Under Response–Perfect Feedback 77
3.4 Model for Matrix Uncertainty 79
3.4.1 Assumptions and Feedback in the Matrix Model 79
3.4.2 Extended Greedy and Robust Policies 81
3.4.2.1 Policies in Λ_E under Standard and Value–Perfect Feedback 82
3.4.2.2 Policies in Λ_E under Response–Perfect Feedback 82
3.5 Semi-Oracle Lower Bounds 84
3.6 Computational Study 87
3.7 Concluding Remarks 93
4.0 SEQUENTIAL ASYMMETRIC BILEVEL LINEAR PROGRAMMING WITH INCOMPLETE INFORMATION AND LEARNING . . 95
4.1 Introduction 95
4.2 Problem Formulation 97
4.3 Greedy and Robust Policies 99
4.4 Greedy and Best–Case Policies 108
4.4.1 Definition and General Convergence Results 108
4.4.2 The Basic Uncertainty Set Update 112
4.4.3 The Convex Uncertainty Set Update 114
4.4.4 The Non-Convex Uncertainty Set Update 117
4.5 Computational Study 121
4.6 Concluding Remarks 127

5.0 CONCLUSIONS 129
APPENDIX A. SUPPLEMENT FOR CHAPTER 2 132
A.1 Basic properties of k-most vital arcs 132
A.2 Additional proofs 133
A.3 Additional graphs 136
APPENDIX B. SUPPLEMENT FOR CHAPTER 3 140
B.1 Proofs of the results for the basic cost model 140
B.2 Proofs of the results for the matrix model 144
B.3 Additional Results and Complementary Material 152
B.3.1 Semi-Oracle Algorithm 152
B.3.2 Numerical Computation of Policies in Λ 155
B.3.3 Sequential Assignment Interdiction 158
BIBLIOGRAPHY 161

LIST OF TABLES

1 Brief summary of the key notation used in Chapter 2. 16

2 Running times (in seconds) to solve LB. The entry "-" implies that an optimal solution was not found within one hour. 40

3 Average cumulative regret ($\times 10^2$) and MAD (in parenthesis) for $k = 6$. . . . 41

4 Average time-stability and MAD (in parenthesis) for $k = 6$. 41

5 Average running times (in seconds) per replication and MAD (in parenthesis) for computing π^{oracle} using Algorithm 2 (regret performance metric), which correspond to the results reported in Table 3. Average times for computing γ are below 5 seconds across all configurations. 42

6 Average running times (in seconds) per replication and MAD (in parenthesis) for computing π^{oracle} using Algorithm 3 (time-stability performance metric), which correspond to the results reported in Table 4. Values of time-stability for γ are computed instantly given the regret. 42

7 Average regret ($\times 10^2$) and MAD (in parenthesis) for $k = 6$. Among the entries denoting regret, the entry in bold is the best value, and the other entries indicate the difference with respect to the best value. 43

8 Average time-stability and MAD (in parenthesis) for $k = 6$. Among entries denoting time-stability, the entry in bold is the best value, and the other entries indicate the difference with respect to the best value. The entries in italic and "$-$" mean that the policy did not attain time-stability for some instances. 44

9 Average regret ($\times 10^3$) and time-stability, and MAD (in parenthesis) for $k = 15$. The entries in bold denote the best value. 48

10 Time-stability mean and MAD for the hypercube uncertainty model and Value Perfect feedback. 90
11 Time-stability mean and MAD for the hypercube uncertainty model and Response Perfect feedback. 91
12 Time-stability mean and MAD for the general uncertainty model and Value Perfect feedback. 91
13 Time-stability mean and MAD for the general uncertainty model and Response Perfect feedback. 92
14 Mean for the time-stability (τ^{ψ}) and regret (R_T^{ψ}) for a policy $\psi \in \Psi$ when $k = 1$.123
15 Mean for the time-stability (τ^{ψ}) and regret (R_T^{ψ}) for a policy $\psi \in \Psi$ when $k = 2$.124
16 Mean for the time-stability (τ^{ψ}) and regret (R_T^{ψ}) for a policy $\psi \in \Psi$ when $k = 3$.124
17 MAD for the time-stability (τ^{ψ}) and regret (R_T^{ψ}) for a policy $\psi \in \Psi$ when $k = 1$.125
18 MAD for the time-stability (τ^{ψ}) and regret (R_T^{ψ}) for a policy $\psi \in \Psi$ when $k = 2$.125
19 MAD for the time-stability (τ^{ψ}) and regret (R_T^{ψ}) for a policy $\psi \in \Psi$ when $k = 3$.126
20 Number of replications for which optimality is guaranteed, $k = 1$ 126
21 Number of replications for which optimality is guaranteed, $k = 2$ 126
22 Number of replications for which optimality is guaranteed, $k = 3$ 127

LIST OF FIGURES

1 Networks used in Remark 1. 15

2 Networks used in Remark 2. 19

3 Networks used in Remark 3. 20

4 Networks used in Remark 4. 21

5 Networks used in the proof of Lemma 3. 23

6 Networks used in Remark 5. 24

7 Network G used for the proof of Proposition 1, $q_{ku} = (u-1)(k+2)+1$. . . . 25

8 Network used in Remark 7. 29

9 Behavior of the average time-stability, average total regret, time-stability MAD and total regret MAD as p_c increases. The cost distribution is right-skewed and $p_a = 1/2$. 45

10 Behavior of the average time-stability, average total regret, time-stability MAD and total regret MAD as p_c increases. The cost distribution is symmetric and $p_a = 1/2$. 46

11 Behavior of the average time-stability, average total regret, time-stability MAD and total regret MAD as the cost intervals widen for the case of $p_a = 2/3$ and $p_c = 1/3$. Given the interval-width multiplier m, the lower and upper bounds of the arc costs in $\widehat{A}_0$ are $l_a = c_a - mx_a$ and $u_a = c_a + my_a$, respectively. . . . 49

12 Example of an instance when $w^{t,\delta} > w_R^{t,*}$. The labeling of the arcs is given by $[\ell_a, u_a], c_a, d_a$. 102

13 Example of an instance when $w^{t,\delta} < w_R^{t,*}$. The labeling of the arcs is given by $[\ell_a, u_a], c_a, d_a$. 103

14 Example of an instance when $w^* < w_R^{t,*}$. The labeling of the arcs is given by $[\ell_a, u_a], c_a, d_a$. 104
15 Example of an instance when $w^{t,\delta} = w_R^{t,*}$ does not imply that $w^{t,\delta} = w^*$, and where $w_R^{t,\delta} < w^*$. The labeling of the arcs is given by $[\ell_a, u_a], c_a, d_a$. 105
16 Example of an instance when $z^{t,\delta} = z_R^{t,*}$ does not imply that $w^{t,\delta} = w^*$. The labeling of the arcs is given by $[\ell_a, u_a], c_a, d_a$. 106
17 A layered network with two layers and four nodes per layer. It has $|N| = 10$ nodes and $|A| = 24$ directed arcs. 122
18 Behavior of the average time-stability, average total regret, time-stability MAD and total regret MAD as p_c increases. The cost distribution is left-skewed and $p_a = 1/2$. 137
19 Behavior of the average time-stability, average total regret, time-stability MAD and total regret MAD as the cost intervals widen for the case of $p_a = 2/3$ and $p_c = 0$. Given the interval-width multiplier m, the lower and upper bounds of the arc cost in $\widehat{A}_0$ are $l_a = c_a - mx_a$ and $u_a = c_a + my_a$, respectively. 138
20 Behavior of the average time-stability, average total regret, time-stability MAD and total regret MAD as the cost intervals widen for the case of $p_a = 2/3$ and $p_c = 2/3$. Given the interval-width multiplier m, the lower and upper bounds of the arc costs in $\widehat{A}_0$ are $l_a = c_a - mx_a$ and $u_a = c_a + my_a$, respectively. . . . 139

PREFACE

Thanks to God for always showing me the way.

I would like to thank my advisor Dr. Oleg Prokopyev. His motivation, energy, and knowledge are exceptional, and I deeply appreciate his advice, ideas, and support throughout all these years. I am also grateful for all the opportunities he provided me, for his patience, and for always being available and in a good mood.

Special thanks to Dr. Denis Sauré for his continuous support, motivation, and his invaluable insights. His contributions were fundamental for the successful completion of this dissertation.

I would also like to thank all the committee members Dr. Jayant Rajgopal, Dr. Pavlo Krokhmal, and Dr. Bo Zeng for their suggestions, their time, and their encouragement, as well as Dr. Bopaya Bidanda for believing in me and giving me the opportunity to pursue my graduate studies at the University of Pittsburgh.

I am grateful to all my friends at the Department of Industrial Engineering, in particular to Arnab Bhattacharya, David Abdul-Malak, Ruichen Sun, Jung Lim, Colin Gillen, and Hosein Zare who made these years much more interesting and easier to navigate.

Many thanks to my family, my parents Victor and Nubia, my brother Victor, and my sister Maria Antonieta. This dissertation would not have been possible without their love and constant support throughout all my life

1.0 INTRODUCTION

Bilevel programming deals with optimization problems where one part of the decisions to be made (referred as *lower-level* decisions) are constrained to be solutions of a mathematical program that depends on the remaining *upper-level* decisions. This general structure makes bilevel programs useful to model hierarchical decision-making problems between two actors, usually casted as a *leader*, or an upper-level decision-maker, and as a *follower*, or a lower-level decision-maker (Dempe 2002). In this perspective, the leader solves an optimization problem that depends on the optimal decision of the follower's problem, and this latter problem is, in turn, parameterized by the decisions of the leader. As such, bilevel programs have many applications in different fields such as defense (Brown et al. 2006), economics (Sherali et al. 1983), transportation (Lucotte and Nguyen 2013), among many others (see Dempe (2002), Colson et al. (2007*a*), Migdalas et al. (2013) and the references therein).

In the typical bilevel formulation it is assumed that the leader and the follower only interact once, and that the leader knows with certainty all the parameters of the problem solved by the follower. There are settings, however, where they interact sequentially during a given time horizon, and where the leader has incomplete information regarding the optimization problem the follower faces at each period. As an example, consider a law enforcement force that patrols smuggling activities over a national border (see e.g., Brown et al. (2006), Morton et al. (2007), Gift (2010)). In this problem, the smugglers continuously attempt to bring inside the country illegal immigrants or illegal goods at the highest possible efficiency using different routes and means of transportation. Law enforcement, on the other hand, have to periodically reallocate their resources to patrol and block the routes used by the smugglers. The effectiveness of their decisions is limited, however, since they do not have all the information regarding some of the routes the smugglers use. This setting can be framed as a network

interdiction bilevel problem, where the follower corresponds to the smugglers, who at each time solve a shortest-path problem on the network that remains after arcs are interdicted. On the other hand, law enforcement are the leader, who must decide what arcs to interdict at each time, albeit not having complete knowledge of the network used by the follower.

Alternatively, consider a sequential version of a typical bilevel pricing (or tariff) model (see Labbé et al. (1998), Van Hoesel (2008)). In here, at each period the leader determines the price (or tariff) for a set of follower's activities with the objective to maximize the revenue that results after the follower performs such activities. After the prices are set, the follower decides upon the level of tariffed and untariffed activities to perform, subject to certain operational constraints, with the objective to minimize the operational costs associated with both types of activities. In certain applications of such models, the leader might be forced to set prices even though she does not know all the information related to the activities she does not tax. For instance, in Bouhtou et al. (2007), the leader is a telecommunications network operator who sets tariffs for links under her ownership (competing companies control the remaining links of the network). The network users correspond to the follower, and they choose what links of the network to use in order to minimize their operational costs. In this setting, the leader might have incomplete information regarding the real costs (or operational constraints) the users incur by using links that belong to the competition, and hence the leader decides without knowing with certainty all the entries of the constraint matrix and the costs vector of the follower's problem.

In this dissertation we study such sequential bilevel programming problems with incomplete information (SBPI). Specifically, our objective is to formalize the problem, to establish useful optimality criteria, and to derive solution methodologies that are theoretically sound, and implementable in practice. Observe that SBPI fall under the umbrella of sequential decision-making under uncertainty; however, they cannot be tackled using standard methods, e.g., Markov decision processes (MDP), stochastic programming (SP) or ever adaptive robust optimization (ARO). For one, in SBPI the leader has a very poor characterization of uncertainty as most of the data related to the follower's activities cannot be directly collected or estimated. Moreover, the relationship between the leader's and followers' decision variables is generally non-convex (Dempe 2002).

These observations motivate us to pursue the objectives of this dissertation by framing SBPI within the methods of *online optimization*. In these models, at each period a decision-maker has to choose a solution from a fixed action set, incurring a cost that depends on the chosen solution. The objective of the decision-maker is to minimize the costs she faces across all time periods, however, she does not know the probability distributions from which the costs are drawn, or the rules, if any, they follow (Cesa-Bianchi and Lugosi 2006*a*, Bubeck and Cesa-Bianchi 2012). As such, online optimization is useful to tackle sequential decision-making problems where the decision-maker has very limited knowledge regarding the structure of uncertainty (Hazan 2015), and/or where there are no functional assumptions that relate the actions with the costs.

Different from classical methods such as MDP, SP and ARO, where a single optimization problem is solved by taking into consideration all possible future scenarios (see e.g., Puterman (2014), Birge and Louveaux (2011), Ben-Tal et al. (2009)), in online models the decision policies solve a new optimization problem at each period. Moreover, due to the lack of reasonable estimates for future uncertainty, these optimization problems incorporate only information from past observations. Under this perspective, the main goal is to find policies that provide good bounds on the *regret*: the difference between the performance faced by the decision-maker (the leader in our case) and the performance of an idealized *oracle* decision-maker that has all the problem's information beforehand (Cesa-Bianchi and Lugosi 2006*a*). A crucial question is thus how to use the information the decision-maker collects from her observations to attain low regret values (i.e., the exploration-exploitation dilemma see, e.g., Bubeck and Cesa-Bianchi (2012)), which in our context translates in determining the best way to incorporate the *feedback*, i.e, the information the leader gets from the follower's responses, into the bilevel framework.

There are several different type of online optimization models in the literature that consider diverse assumptions on the action set, on whether the rewards are stochastic or deterministic, and on the particular type of feedback, see, for instance, Auer et al. (1995), Freund and Schapire (1997), Cesa-Bianchi et al. (1997), Auer, Cesa-Bianchi, Freund and Schapire (2002), Auer, Cesa-Bianchi and Fischer (2002*a*), Kalai and Vempala (2005), Koolen et al. (2010), Neu and Bartók (2013), Cesa-Bianchi and Lugosi (2006*a*), Bubeck and Cesa-

Bianchi (2012) and the references given therein. There are, however, three major drawbacks of existing online optimization methods that make them unfit to solve the class of SBPI:

(*i*) Current models consider a *fixed amount of possible actions* (with the exception of Kleinberg et al. (2010), that does not adequately specialize to our class of problems). That is, in most existing models the set of actions available to the decision-maker at each time is always the same, independent of any new information that is discovered. Hence, such models cannot handle the fact that the leader might learn new means to affect the follower's actions (e.g., new routes in the smuggling interdiction setting) from observing the latter.

(*ii*) Naively using existing online policies, such as those applied to multi-armed bandit settings (see, e.g. Cesa-Bianchi et al. (1997), Auer, Cesa-Bianchi, Freund and Schapire (2002), Audibert and Bubeck (2009), Cesa-Bianchi and Lugosi (2006*a*)), would result in regret bounds that are *exponential in the primitives of the problem* (i.e., the number of variables and constraints). Specifically, general multi-armed bandit policies provide bounds that are proportional to the number of possible actions available to the decision-maker. In our context the number of possible actions corresponds to the number of feasible solutions in the upper-level (leader's) optimization problem, and it is well-known (see, e.g., Dempe (2002)) that said number is typically exponential in the number of the leader's variables and constraints.

(*iii*) Online methods that explicitly consider problems with combinatorial structure (e.g. Cesa-Bianchi and Lugosi (2012), Audibert et al. (2013), Kalai and Vempala (2005), Gyorgy et al. (2007)), or infinitely many solutions, as in online convex optimization (e.g. Zinkevich (2003), Kalai and Vempala (2005), Awerbuch and Kleinberg (2004*a*), Hazan et al. (2007), Hazan (2015)), assume a *single-level relationship* between the decision-maker's actions and the costs she observes. Consequently, these models cannot address the hierarchical relationship between the leader and follower that is present in bilevel optimization models.

We divide this dissertation into three models for SBPI. In the first chapter, we consider that the bilevel problem corresponds to a shortest-path network interdiction problem (Israeli and Wood 2002), where the leader initially has no information for some of the arcs of the network the evader uses, as well as the real costs of some of the arcs she observes (for which she only knows that they lie in some given intervals). For this setting we propose a set of policies based on blocking a k-most vital arcs solution of the network the leader observes at

each time (Ball et al. 1989). Under certain assumptions on the feedback, we prove that these policies bound the regret linearly in the number of arcs of the network, detect optimality in real time, and are easily computable using state-of-the-art MIP solvers.

In the second chapter we generalize the concepts and results of shortest path interdiction to general adversarial (max-min) bilevel problems. In this setting, the leader does not know some of the follower's variables and constraints, and might not know with certainty all the values of the data of the lower-level problem for those variables and constraint she observes. Nonetheless, the leader knows that this data belongs to a given polyhedral set. For this setting, we propose a set of policies that are *greedy* and *robust*, which depending on the type of feedback, can bound the regret linearly with the number of total variables and constraints the follower uses. Moreover, these policies are *weakly optimal*, in the sense that they have the best possible worst–case performance across all instances of the problem. In addition, as in the shortest path case, these policies detect optimality in real time; and by using robust and bilevel optimization techniques, we show how the policies can be computed using MIPs.

In the last chapter, we consider the asymmetric bilevel problem. Here, the objective of the leader is not necessarily to maximize the disruption of the follower's performance and she might optimize some other objective function. We show that greedy and robust policies are no longer able to guarantee finite time-stability bounds nor provide certificates of optimality. Subsequently, we study a class of alternative greedy and 'best'–case policies that can provide a finite upper-bound for the time-stability across all instances, and generate certificates of optimality in real time. In addition, under certain *updating mechanisms*, these policies have an MIP formulation which can be viewed as the extensive form of a two-stage stochastic mixed-integer problem. Hence, these policies can be computing either by using state-of-the-art MIP solvers or decomposition techniques from the stochastic programming literature.

2.0 SEQUENTIAL SHORTEST PATH INTERDICTION WITH INCOMPLETE INFORMATION AND LEARNING

2.1 INTRODUCTION

In network interdiction, an *interdictor* or leader selects a series of interdiction measures that change the structure of a network with the objective of disrupting or stopping an *evader*'s (follower's) movement through the network. The problem, initially studied in the context of military applications, has received considerable attention during the past decades, considering various objectives for the interdictor, from *maximizing the shortest path in the interdicted network*, see, for instance, Fulkerson and Harding (1977), Corley and Sha (1982), Malik et al. (1989), Israeli and Wood (2002), to minimizing the maximum flow between given nodes as in Wollmer (1964), McMasters and Mustin (1970), Ghare et al. (1971), Corley and Chang (1974), Ratliff et al. (1975), Wood (1993), to minimizing the maximum probability of successful evasion as in, e.g., Washburn and Wood (1995), Morton et al. (2007) and Pan and Morton (2008). Stochastic variations of these problems have also been considered by Cormican et al. (1998), Hemmecke et al. (2003), Janjarassuk and Linderoth (2008), as well as multicommodity versions, by Lim and Smith (2007). See also Smith and Lim (2008) for a survey on several types of interdiction and fortification models.

In addition to these classical settings, recent work in the area focuses on critical node/edge detection problems, where the objective is to remove a set of nodes/edges in order to maximally degrade some connectivity measure of the remaining network. Such measures include, for instance, the total number of pairwise connections, the size of the largest connected component, and the total number of connected components (Walteros and Pardalos 2012). In particular, the shortest path between two (or more) fixed nodes can be viewed as a special

case of such a connectivity measure, see Veremyev, Boginski and Pasiliao (2014). We also refer to Shen et al. (2012*a*), Granata et al. (2013), Veremyev, Prokopyev and Pasiliao (2014), Veremyev et al. (2015), Shen and Smith (2012), a survey in Walteros and Pardalos (2012) and the references given therein.

While most studies assume complete knowledge of the network structure and costs, in many applications areas, like in military settings, the interdictor operates (at least initially) with limited information about the conditions on the ground. The same holds true for applications in drug and nuclear material smuggling, which have been casted as interdiction problems, see, e.g., Wood (1993) and Morton et al. (2007). In this work, we envision shortest path interdiction as a sequential process, where the interdictor initially has *partial* knowledge about the structure and costs of the network, and may adapt the interdiction actions as new information is collected from observing the evader's reaction to previous actions. More specifically, we focus our attention on a specific setting, where the interdictor knows that (the evader's) arc costs are *deterministic* and belong to a given set. In each time period, the interdictor blocks at most k arcs from the network (only for the duration of the period), and the evader then travels along a shortest $1-n$ path of the interdicted network, where nodes 1 and n are assumed to be the same in all time periods, arbitrary and fixed. Subsequently, the interdictor observes each arc on said path and its cost, and, hence, learns about the structure and costs of the network, and adjusts its actions so as to maximize the cumulative cost incurred by the evader.

Our modeling approach is motivated by interdiction and evasion dynamics arising, for instance, when monitoring and patrolling illegal activities. In this context, Gift (2010) considers an interdictor (e.g., a U.S. law-enforcement or military task force) that periodically re-allocates resources such as ships, planes, helicopters and land units, over different geographical zones of known routes, so as to capture drug smugglers. The smugglers, on the other hand, learn by trial and error the route with highest probability of successful evasion for any given allocation; in particular, it is assumed that the smugglers solve the evasion learning problem via index-based (Gittin's) heuristics. Similarly, Morton et al. (2007) and Brown et al. (2006) consider the problem of detecting illegal material or immigrants entering through a border (their base models assume that the probabilities of successful evasion

are known upfront) when the interdictor allocates surveillance resources to modify detection probabilities throughout a network. The evader, who observes such an allocation, chooses a path of minimum detection probability. Assuming common (shared) information, Malaviya et al. (2012) study sequential allocation of police officers within an urban region (who monitor criminals and their trade links) so as to minimize the maximum flow of illegal drugs within a time horizon. In this work, learning is incorporated by means of side constraints (e.g. criminals are arrested only after being monitored for a number of time epochs, and/or if they have been denounced by lower ranked criminals). The studies above share distinctive features: the interdictor's problems are formulated using a bi-level framework; and are solved by using mixed-integer programming techniques. (Moreover, note that maximum probability path interdiction can be casted as a shortest path interdiction on the same graph with costs set to the logarithm of the reciprocal of the evasion probability).

Sequential attacker-defender and defender-attacker problems have been analyzed using game theory. Assuming perfect information, Hausken and Zhuang (2011) study how the government should balance defensive investments over time. Zhuang et al. (2010) study to what extent a defender with private information must be deceptive or secretive towards an attacker who updates its beliefs about the defender's "toughness" whenever a confrontation takes place. In a similar setting, albeit single-period, Xu and Zhuang (2014) study the attacker's trade-off between investing resources in either attacking or learning the defender's vulnerability. Non-sequential adversarial decision models where the attacker is uncertain about the defender's actions have been considered as well, see, e.g., McLay et al. (2012), and the references given therein.

The network interdiction and attacker-defender models above do not capture the particular model-learning component that arises in our setting. In this sense, sequential decision-making problems that involve both generic model uncertainty and learning are usually casted as multi-armed bandits (Robbins (1952)). However, the typical bandit formulation focuses on stochastic feedback, as in Lai and Robbins (1985), Auer, Cesa-Bianchi and Fischer (2002*b*), or models of adversarial nature, such as Auer et al. (2003), consider a trivial mapping between decisions and feedback. We also refer to the work by Modaresi et al. (2012) and Cesa-Bianchi and Lugosi (2012) for bandit settings with combinatorial structure. A salient

feature of our model, which distinguishes it from previous work, is that the feedback collected is not directly selected by the decision-maker, but can be used to infer the cost structure of the setting.

Our model makes several assumptions. Consistent with the literature discussed above (see, e.g., Israeli and Wood (2002)), we assume that the interdictor's objective is to maximize the cumulative cost incurred by the evader, thus implying that both agents perceive costs equally (a notable exception is Bayrak and Bailey (2008)). This might accommodate settings, for example, where costs represent travel times, and the interdictor adjusts time estimates after observing the evader go through a particular route. Note that we implicitly assume that the evader does not react strategically to the interdictor's actions, i.e., he/she always chooses a shortest $1-n$ path. We assume that all costs are deterministic and that arc costs are observed by the interdictor once these are used by the evader. These assumptions aim at isolating a first-order effect of model uncertainty: the cost of recovering the optimal interdiction action.[1] Note that absent uncertainty, when the evader and the interdictor only interact once, our problem reduces to the *k-most vital arcs problem*[2](Corley and Sha 1982, Malik et al. 1989), which is a special case of *shortest-path maximization*, see, e.g. Israeli and Wood (2002). In this latter problem, arcs might be interdicted "partially," which increases their traversing costs by some amount that depends on the interdiction effort.

In this chapter we analyze the performance of a simple policy in which the interdictor, at each period, removes a set of k-most vital arcs of the *observed* network, and separate the analysis for the cases when: (i) the costs of all initially observed arcs are known; and (ii) the costs of some initially observed arcs are unknown, but are in a known range. In this regard, the proposed policies are *greedy*, and *pessimistic* in that they assume a worst-case realization of the costs in setting (ii). The k-most vital arcs problem is *NP*-hard, as shown in Ball et al. (1989), but effective solution approaches are available in practice, see, e.g., Israeli and Wood (2002). Following similar work in sequential decision-making under uncertainty (see, e.g., Cesa-Bianchi and Lugosi (2006*b*)), we measure policy performance in terms of the *re-*

[1]Consider that under stochastic feedback, repeated implementation of an action should lead to reliable cost estimates, thus our model can be viewed as a certainty equivalent version of a model with stochastic feedback where actions are changed at a maximum frequency.

[2]A set of k-most vital arcs in graph G consists of (at most) k arcs whose removal from G results in the greatest increase in the length of the shortest path between two specified nodes.

gret, which is the cumulative loss in cost incurred by a policy relative to that achieved by an oracle interdictor with prior knowledge of the network's structure and arc costs, and on the *time stability* of a policy, which is the number of periods before the interdiction actions match those taken by said oracle.

The contribution of the chapter is two-fold. First, we show that the proposed class of policies is *efficient* (we define the concept of *efficiency* in the next section). In doing so, we identify attractive features of these policies: their regrets admit a finite horizon-independent upper bound; and they detect in real time when their actions match those taken by the oracle policy (thus, indicating that both regret and time stability not longer grow with time). In addition, we show that the pessimistic nature of the proposed policies in the case of uncertain arc costs is crucial for attaining a finite regret. Second, we propose a semi-oracle performance benchmark that contrasts cumulative cost against that induced by an oracle with advance knowledge of the cost vector, but that must not signal that such knowledge is available. We argue that this measure provides a better fit to the setting, relative to the cumulative regret, which is arguably impractical when feedback is deterministic. In addition, we perform numerical experiments to assess efficiency of the proposed policies.

The remainder of this chapter is organized as follows. Section 3.2 provides a detailed and more formal description of the problem as well as a definition of efficiency of an interdiction policy. In Section 2.3 we propose a simple class of interdiction policies and establish their efficiency. Section 2.4 develops fundamental lower bounds for policy performance, and Section 2.5 presents our numerical experiments. Finally, Section 2.6 presents our conclusions and highlights possible directions for future research.

2.2 PROBLEM FORMULATION

We begin this section by introducing some notation. Let $G := (N, A, C)$ be a directed network, where N and A denote nodes and arcs, respectively, and $C := (c_a)_{a \in A}$ is a nonnegative cost vector, where c_a is the cost or length of arc $a \in A$. Let $n = |N|$. For $A' \subseteq A$, we define the graph $G[A'] := (N, A', C)$, where it is understood that only the information in C about

arcs in A' is available. We denote by $\mathcal{S}(G)$ the set of all shortest $1-n$ paths in G, where nodes 1 and n are arbitrary, fixed and given. Observe that the set $\mathcal{S}(G)$ can be defined by:

$$\mathcal{S}(G) := \arg\min \left\{ \sum_{a \in P} c_a : P \text{ is an } 1-n \text{ path in } G \right\},$$

and let $z(G)$ denote the cost of a path in $\mathcal{S}(G)$. Finally, for any path P in G, let $\ell(P)$ denote the cost of the path, i.e., $\ell(P) := \sum_{a \in P} c_a$.

Consider an interdictor that initially observes a subnetwork $G[A_0]$ of $G = (N, A, C)$, and knows that the cost vector C lies in the set

$$\mathcal{C}_0 := \left\{ \left(c'_1, \cdots, c'_{|N \times N|}\right) \in \mathbb{R}_+^{|N \times N|} : c'_a = c_a \text{ for } a \in \widetilde{A}_0 \text{ and } \ell_a \leq c'_a \leq u_a \text{ for } a \in \widehat{A}_0 \right\},$$

for given sets $\widetilde{A}_0 \subseteq A_0$ and $\widehat{A}_0 \subseteq A_0$, where $0 \leq \ell_a < u_a < \infty$ for all $a \in \widehat{A}_0$ and $A_0 \subseteq A$ (with $\widetilde{A}_0 \cap \widehat{A}_0 = \emptyset$). That is, the interdictor is aware of arcs in A_0, she/he knows the costs of those in $\widetilde{A}_0$, and has some prior information about the costs (specifically, lower and upper bounds) of arcs in $\widehat{A}_0$. We refer to $\mathcal{C}_0$ as the initial information available to the interdictor, as it contains her/his initial knowledge about the structure and costs of the network (we assume that the set of nodes N is known to the interdictor upfront including the evader's source and destination nodes 1 and n, respectively).

In each time period $t \in \mathcal{T} := \{0, 1, \ldots, T\}$, the following sequence of events takes place:

(1) The interdictor blocks a set of arcs $I_t \subseteq A_t$ only for the duration of period t, with $|I_t| \leq k$, where A_t denotes the set of arcs the interdictor is aware of at the beginning of period t, and the constant k denotes the maximum number of arcs that can be removed in any time period.

(2) The evader traverses through path $P_t \in \mathcal{S}\left(G[A \setminus I_t]\right)$, incurring a cost of $z\left(G[A \setminus I_t]\right)$ and revealing the arcs in P_t as well as their costs to the interdictor, so that $A_{t+1} := A_t \cup P_t$ and

$$\mathcal{C}_{t+1} := \left\{ C' \in \mathcal{C}_t : c'_a = c_a \text{ for } a \in P_t \right\},$$

where $\mathcal{C}_t$ denotes the set of cost vectors that are consistent with the information available to the interdictor at the beginning of period $t \in \mathcal{T}$.

In the above, and throughout the chapter, we made the following assumptions:

A1. Each time period $t \in \mathcal{T}$ the interdictor observes path P_t and cost c_a of each arc $a \in P_t$ used by the evader.

A2. The evader acts myopically, always selecting a shortest $1-n$ path in the interdicted network. Also, the evader observes the interdictor's actions before choosing a path.

(A3). If there is more than one possible choice for P_t, then the evader chooses a path following a well-defined deterministic rule. Furthermore, this rule is *consistent*, in the sense that if P_t is chosen from a collection of paths $\mathcal{P}$, then it is also chosen from any collection $\widetilde{\mathcal{P}} \subseteq \mathcal{P}$ containing P_t.

(A1). $I_0 = \emptyset$ and $I_t \neq \emptyset$ for all $t \geq 1$; furthermore, $A_0 = \widetilde{A}_0 \cup \widehat{A}_0$, and G is not "trivially" k-separable.[3]

Assumption **A1** can be viewed as an instance of *perfect (or transparent) feedback*, which is common in the learning theory literature, see, e.g., Cesa-Bianchi and Lugosi (2006*b*) and requires some degree of monitoring of the evader's actions, thus its validity ultimately depends on the details of a particular application. In this regard, this assumption might accommodate situations when the interdictor observes the evader's actions, e.g., by using a satellite or a drone, but cannot immediately act upon those actions. Such situations occur, for example, when the interdiction actions require relocation of the interdictor's resources, e.g., ships, or land units, over different geographical zones. Furthermore, assumption **A1** naturally occurs when the information from the monitoring devices (e.g., satellite images) is of sufficiently good quality, but cannot be immediately used for interdiction, e.g., due to the need of its additional interpretation.

Furthermore, one can interpret this assumption in the context of a repeated interaction between the evader and the interdictor in a stochastic environment, as mentioned in the previous section. For example, consider the application to drug smuggling or illegal immigration detection, where evaders repeatedly choose a path of minimum detection probability (which can be formulated as a shortest path interdiction). In this setting, repeated interaction

[3]We refer to a directed network G as "trivially" k-separable if any set of k arcs in G forms an 1-n cut.

between the evader and the interdictor would account for successful as well as failed smuggling/trespassing attempts, thus providing the interdictor with some information regarding the success probability of the aforementioned path.

Admittedly, while this assumption simplifies our theoretical analysis, it is somewhat limiting as it also implies that our model cannot be applied directly to some practical settings where there are limited monitoring capabilities, e.g., when the interdictor observes only the total length of the path used by the evader, or a subset of the arcs used. Nonetheless, relaxing this assumption is an interesting topic of future research (see our additional discussion in Section 2.6). For example, one could potentially adapt the concept of the *barycentric spanners* used by Awerbuch and Kleinberg (2004*b*) for such generalizations (referred to as the *opaque feedback* case).

With regard to the first part of **A2**, this assumption imposes a rather simple behavior on the evader. However, one can show that the proposed policies are robust (with respect to their convergence) in settings with strategic evaders. Regarding the second part of **A2**, the assumption is that the evader has some degree of monitoring of the interdictor's actions. As outlined previously, it is possible to interpret this assumption in the context of repeated interactions in a stochastic setting, in which such monitoring might arise naturally from a process of learning by trial and error on the evader's side. Please see Section 6 for further discussion.

Assumption **(A3)** ensures that the evader's decisions are consistent with his/her past decisions. Intuitively, one can think that the evader ranks all paths in the network based on their costs (resolving ties according to any criteria, or even randomly) *in advance.* In each time period the evader selects the highest-rank unblocked (shortest) path from such a list. Generally speaking, this assumption prevents the evader from using randomized algorithms *during* the evasion process.

Assumption **(A1)** is technical and made without loss of generality. Its first part implies that the evader acts first and always interdicts at least one arc. The second part simply states that the interdictor knows valid lower and upper bounds on the cost of any arc he/she is initially aware of (note that this assumption is not limiting as such lower and upper bounds can be set at zero and at an arbitrarily large value, respectively). Finally,

the non-k-separability condition implies that the problem does not admit a trivial solution. As a consequence, there are sets consisting of k arcs whose removal do not disconnect the network.

Considering **A1** and **(A1)**, for $t \in \mathcal{T} \setminus \{T\}$, we define recursively $\widetilde{A}_{t+1} := \widetilde{A}_t \cup P_t$, and $\widehat{A}_{t+1} := \widehat{A}_t \setminus P_t$, hence $A_t = \widetilde{A}_t \cup \widehat{A}_t$ for all $t \in \mathcal{T}$.

An *interdiction policy* is a deterministic sequence of set functions $\pi := (\pi_t,\ t \in \mathcal{T})$, such that for each $t \geq 1$, $I_t^\pi = \pi_t(\mathcal{F}_t^\pi)$ represents the set of arcs blocked in period t, and $I_t^\pi \subseteq A_t$, where $\mathcal{F}_t^\pi$ summarizes the initial information and history of the interdiction process up to time $t-1$. That is,

$$\mathcal{F}_t^\pi := \left(\mathcal{C}_0, I_0^\pi, P_0, I_1^\pi, P_1, \cdots, I_{t-1}^\pi, P_{t-1}\right),$$

where $I_0^\pi = \emptyset$ by Assumption **(A1)**. As P_t, $\widehat{A}_t$, $\widetilde{A}_t$, and A_t depend on I_s^π for all $s < t$, we add a π superscript to these sets to denote dependence in policy π, when necessary.

Let Π denote the set of all *feasible interdiction policies*. Given G and $\mathcal{C}_0$, from assumption **(A3)**, applying policy $\pi \in \Pi$ results in a *unique* sequence $\{(I_t^\pi, P_t^\pi) :\ t \in \mathcal{T}\}$ of *blocking* and *evasion* decisions. We define the *cumulative regret* incurred by policy π by time t as

$$R_t^\pi(G, \mathcal{C}_0) := \sum_{s \leq t} \left(z^*(G) - z\left(G[A \setminus I_s^\pi]\right)\right),$$

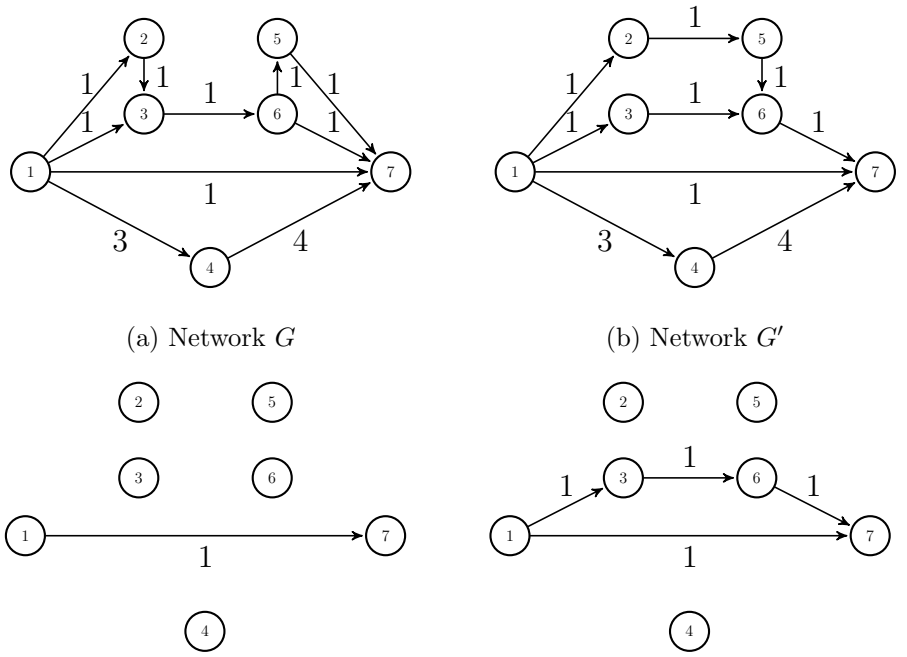
where $z^*(G)$ denotes the optimal cost in the k-most vital arcs problem on G, i.e.

$$z^*(G) := \max\left\{z\left(G[A \setminus I]\right) :\ I \subseteq A \text{ s.t. } |I| \leq k\right\}.$$

The regret represents the cumulative loss in cost incurred by a policy, relative to that of an oracle interdictor with prior knowledge of G. For a given graph G, regret minimization is equivalent to cumulative cost maximization. We say that $(z^*(G) - z\left(G[A \setminus I_t^\pi]\right))$ is the *instantaneous regret* incurred by policy π at time $t \in \mathcal{T}$. Note that when G is k-separable, then $z^*(G) = +\infty$, and $z\left(G[A \setminus I_t^\pi]\right) = +\infty$ when I_t^π is an 1-n cut. Thus, in such cases, we take the convention that $(z^*(G) - z\left(G[A \setminus I_t^\pi]\right)) = 0$.

Alternatively, one might instead focus on recovering the solution to the underlying k-most vital arcs problem as soon as possible, which is not necessarily aligned with the goal of regret minimization. Hence, we define the *time-stability* of policy $\pi \in \Pi$ as

$$\tau^\pi\left(G, \mathcal{C}_0\right) := \min\left\{t \in \mathcal{T} :\ z\left(G[A \setminus I_s^\pi]\right) = z^*(G) \text{ for all } s \geq t\right\}, \tag{2.1}$$



(a) Network G

(b) Network G'

(c) What π^1 and π^2 observe for both G and G' at time $t = 1$

(d) What π^1 and π^2 observe for both G and G' at time $t = 2$

Figure 1: Networks used in Remark 1.

where we assume the convention that $\min\{\emptyset\} = T + 1$. The time-stability of policy π corresponds to the first time period by which regret is made zero from there on, i.e., it is the earliest period by which for any $t \geq \tau^{\pi}(G, \mathcal{C}_0)$ the set I_t^{π} is a set of k-most vital arcs of G. Observe that minimizing time-stability, rather than regret, would be preferable in settings where the interdictor is willing to sacrifice the regret performance during the first $\tau^{\pi} - 1$ time periods in order to guarantee that the best solution is found as early as possible. Moreover, time-stability is still a useful measure of performance in cases where the last time period T is not known in advance (and the total regret would be ill-defined) or it is known to be large. Indeed, in such settings time-stability represents the earliest time period in which a k-most vital arc solution is implemented from there on, and this interpretation can be handled mathematically by setting $T = \infty$ in equation (2.1).

We summarize the notation used in the chapter in Table 1.

Table 1: Brief summary of the key notation used in Chapter 2.

G	Underlying directed graph	$\ell(P)$	Cost of $1-n$ path P
$G[A']$	Subgraph including only the arcs in A'	T	Time horizon
$\mathcal{S}(G)$	Set of all $1-n$ shortest paths in G	P_t	Path chosen by the evader in period t
$z(G)$	Cost of a shortest $1-n$ path in G	I_t	Set of arcs removed during period t
$z^*(G)$	Optimal cost of the k-most vital arcs problem on G	(l_a, u_a)	Lower and upper bounds (known to evader) on cost of arc $a \in A$
$\widetilde{A}_t$	Arcs with known cost in period t	Π_μ	Policies *efficient* with respect to μ
$\widehat{A}_t$	Arcs with known cost interval in period t	$\mathcal{F}_t^\pi$	History up to time t under policy π
$\mathcal{C}_t$	Cost vectors consistent with information in period t	k	Maximum number of arcs that can be interdicted in a time period
x^π	First period evader incurs a cost predicted by π	τ_t^π	Time-stability of policy π by time t
$\mathbb{G}(\mathcal{C}_0)$	Graphs compatible with initial information $\mathcal{C}_0$	R_t^π	Regret of policy π by time t

Ideally, we would like to find a policy $\pi' \in \Pi$ that performs better than any other policy for any graph G that is consistent with the initial information in $\mathcal{C}_0$. That is, given $\mathcal{C}_0$, we aim to find π' such that $R_T^{\pi'}(G, \mathcal{C}_0) \leq R_T^{\pi}(G, \mathcal{C}_0)$ and $\tau^{\pi'}(G, \mathcal{C}_0) \leq \tau^{\pi}(G, \mathcal{C}_0)$, for all $\pi \in \Pi$ and $G \in \mathbb{G}(\mathcal{C}_0)$, where

$$\mathbb{G}(\mathcal{C}_0) := \{G : \ G = (N, A, C),\ A \subseteq N \times N,\ C \in \mathcal{C}_0\}.$$

As shown in Remark 1 below, this is not always possible.

Remark 1. Consider networks $G = (N, A, C)$ and $G' = (N, A', C')$ depicted in Figures 1(a) and 1(b), respectively. Set $k = 2$, $T = 2$, and assume that $A_0 = A'_0 = \emptyset$ (thus, $\mathcal{C}_0 = \mathcal{C}'_0 = \mathbb{R}_+^{|N\times N|}$). Because $I_0 = \emptyset$ and $I_1 \neq \emptyset$ (by assumption **(A1)**), then $P_0 = 1{-}7$, $I_1 = \{(1,7)\}$ and $P_1 = 1{-}3{-}6{-}7$ for both networks under all policies. Define π^1 so that $I_2^{\pi^1} = \{(1,7),(3,6)\}$, and π^2 so that $I_2^{\pi^2} = \{(1,7),(6,7)\}$. Figures 1(c) and 1(d) depict the networks observed by the interdictor at times $t = 1$ and $t = 2$ for both G and G'. Observe that $z^*(G) = z^*(G') = 7$,

and thus for policy π^1 we have that on $(G, \mathcal{C}_0)$ the total regret is $R_T^{\pi^1}(G, \mathcal{C}_0)$ =(7-1)+(7-3)+(7-7)=10, while on $(G', \mathcal{C}_0')$ the total regret is $R_T^{\pi^1}(G', \mathcal{C}_0')$ =(7-1)+(7-3)+(7-4)=13. Similarly, for policy π^2, $R_T^{\pi^2}(G, \mathcal{C}_0) = 13$ and $R_T^{\pi^2}(G', \mathcal{C}_0') = 10$. Moreover, one can check that, for any policy $\pi \in \Pi$, $R_T^{\pi^1}(G, \mathcal{C}_0) = 10 \leq R_T^{\pi}(G, \mathcal{C}_0)$ and $R_T^{\pi^2}(G', \mathcal{C}_0') = 10 \leq R_T^{\pi}(G', \mathcal{C}_0')$. In particular, $R_T^{\pi^1}(G, \mathcal{C}_0) < R_T^{\pi^2}(G, \mathcal{C}_0)$ and $R_T^{\pi^2}(G', \mathcal{C}_0') < R_T^{\pi^1}(G', \mathcal{C}_0')$. Similar arguments can also be applied to time-stability. ■

In light of the discussion above, consider the properties that one would expect efficient policies to have. Generally speaking, for any policy π, let $\mu_T^{\pi}(G, \mathcal{C}_0)$ be a measure of performance (e.g., cumulative regret, time-stability) that depends on T, G and $\mathcal{C}_0$. We say that a subset of feasible policies $\Pi_\mu^* \subseteq \Pi$ is *efficient with respect to* μ if the following conditions hold:

C1: Any policy $\pi \in \Pi_\mu^*$ eventually finds and maintains a solution to the underlying k-most vital arcs problem for all T above some finite instance-dependent threshold.

C2: Π_μ^* is a *homogeneous* set in the sense that for any policy in Π_μ^* there is no other policy in Π_μ^* that is better, or worse, across all instances. Formally, for any policy $\pi \in \Pi_\mu^*$ there exist another policy $\pi' \in \Pi_\mu^*$, $\mathcal{C}_0$ and networks $G, G' \in \mathbb{G}(\mathcal{C}_0)$ such that $\mu_T^{\pi}(G, \mathcal{C}_0) < \mu_T^{\pi'}(G, \mathcal{C}_0)$ and $\mu_T^{\pi}(G', \mathcal{C}_0) > \mu_T^{\pi'}(G', \mathcal{C}_0)$.

C3: Π_μ^* is not *dominated* by another class of policies. That is, for any $\mathcal{C}_0$ and $\pi' \in \Pi \setminus \Pi_\mu^*$, there exist $\pi \in \Pi_\mu^*$, $G \in \mathbb{G}(\mathcal{C}_0)$ and T such that $\mu_T^{\pi}(G, \mathcal{C}_0) < \mu_T^{\pi'}(G, \mathcal{C}_0)$.

In the next section we show that such class of policies exists for the case when μ is either cumulative regret $R_T^{\pi}(G, \mathcal{C}_0)$ or time-stability $\tau^{\pi}(G, \mathcal{C}_0)$. Moreover, we show that $\Pi_R^* \cap \Pi_\tau^* \neq \emptyset$.

2.3 EFFICIENT INTERDICTION POLICIES

Guided by the discussion above, in this section we analyze a class of policies that are efficient with respect to regret and time-stability. First, in Section 2.3.1 we analyze the somewhat simpler case of $\widehat{A}_0 = \emptyset$, i.e., when there is no uncertainty with respect to the costs of arcs known to the interdictor at time $t = 0$. This setting reveals the *greedy* nature of the proposed

policies: in each period they remove a set of k-most vital arcs from the observed network. Later, in Section 2.3.2 we extend such policies for where $\widehat{A}_0 \neq \emptyset$, i.e., the case with possible cost uncertainty of the initially known arcs. This extension reveals the *pessimistic* nature of the proposed policies: when faced with uncertain costs on the observed arcs, they operate as in the case of $\widehat{A}_0 = \emptyset$, by using known upper bounds as proxies for unknown costs.

Later, we complement our theoretical analysis of these policies with numerical experiments, which demonstrate that their theoretical efficiency also translates into good regret and time-stability performance across different instances when compared to other benchmark policies.

2.3.1 Efficient Policies When $\widehat{A}_0 = \emptyset$

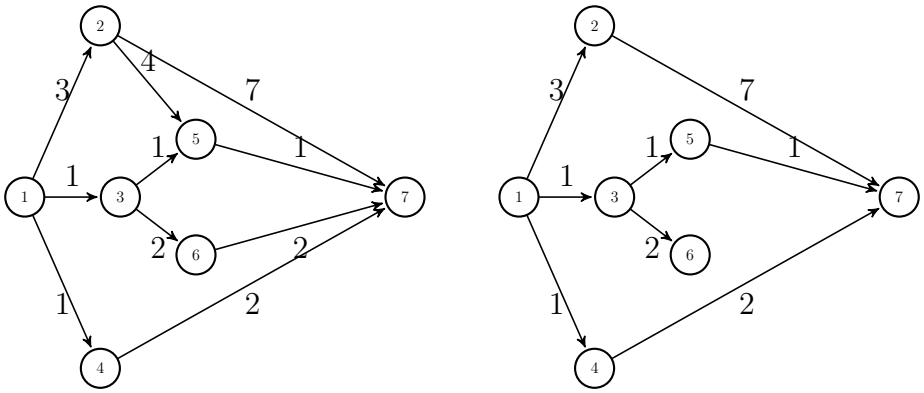
Assume that $\widehat{A}_0 = \emptyset$, which implies that $A_t = \widetilde{A}_t$ for all $t \in \mathcal{T}$. For any policy $\pi \in \Pi$, $\mathcal{C}_0$ and $G \in \mathbb{G}(\mathcal{C}_0)$, define

$$x^\pi(G, \mathcal{C}_0) := \min\{t \in \mathcal{T} : \ z(G[A_t^\pi \setminus I_t^\pi]) = z(G[A \setminus I_t^\pi])\}, \tag{2.2}$$

the first time period in which the evader uses a path whose length is expected by the interdictor (who follows policy π). Additionally, observe that, by the end of period t, the interdictor is aware of whether or not any time period t corresponds to x^π. Define Γ as the class of policies that at any time period t prior to x^π interdict a set of k-most vital arcs of $G[A_t^\pi]$, and then keep removing the same set of k-most vital arcs (used at time x^π) until T. That is, $\gamma \in \Gamma \subset \Pi$ if and only if

$$I_t^\gamma \in \arg\max\{z(G[A_t^\gamma \setminus I]) : \ I \subseteq A_t^\gamma, |I| \leq k\} \text{ for } t \leq x^\gamma, \qquad I_t^\gamma = I_{x^\gamma}^\gamma \text{ for } t > x^\gamma. \tag{2.3}$$

As we discuss later (see Lemma 1), regardless of any new information provided by path P_t^γ for any time period $t > x^\gamma$, $I_{x^\gamma}^\gamma$ remains a set of k-most vital arcs of $G[A_t^\gamma]$ for $t > x^\gamma$. Hence, the policies in Γ always interdict a set of k-most vital arcs of the observed network. Furthermore, by the definition of an interdiction policy given in Section 3.2, I_t^γ is a deterministic function. Thus, whenever a policy $\gamma \in \Gamma$ faces a tie at some time period (i.e., whenever there are multiple sets of k-most vital arcs in the observed network), the tie is broken in a deterministic



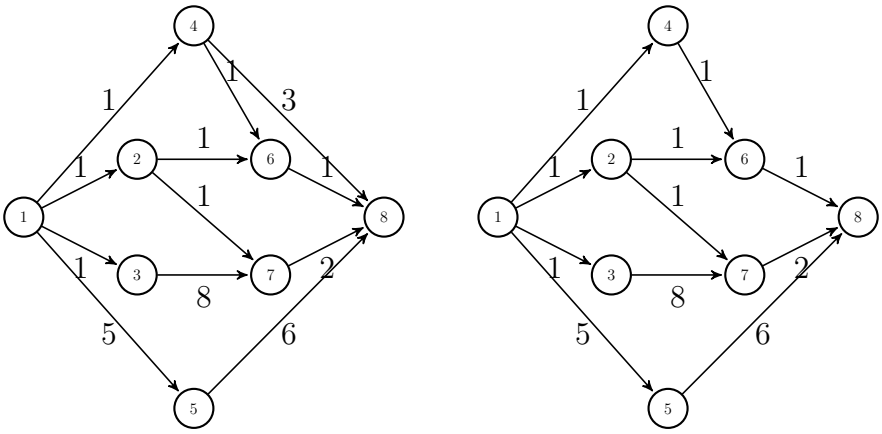
(a) Network G (b) Network $G[A_0^\gamma]$

Figure 2: Networks used in Remark 2.

fashion. This observation is similar in spirit to assumption **(A3)** describing the evader's behavior, in the sense that the leader has to be consistent with and cannot decide randomly.

Remark 2. One might expect that if the interdictor uses policies in Γ, then the lengths of the shortest paths used by the evader, i.e., $\{z(G[A \setminus I_t^\gamma]) : t \in \mathcal{T}\}$, define a non-decreasing sequence in t. However, it turns out not to be the case in general. For example, let $k = 2$, and assume that $G = (N, A, C)$ is as depicted in Figure 2(a), while $G[A_0^\gamma]$ is given in Figure 2(b). Observe that $P_0^\gamma = 1 - 3 - 5 - 7$ (this is a shortest path in $G = G[A \setminus I_0^\gamma]$, recall that $I_0^\gamma = \emptyset$), and that $I_1^\gamma = \{(1,3),(1,4)\}$ (this is a 2-most vital arc solution for $G[A_1^\gamma]$). Next, $P_1^\gamma = 1-2-5-7$ is a unique shortest path in $G[A\setminus I_1^\gamma]$. Suppose now that $I_2^\gamma = \{(1,4),(5,7)\}$ (this is a 2-most vital arc solution in $G[A_2^\gamma]$), which implies that $P_2^\gamma = 1-3-6-7$ (this is a unique shortest path in $G[A\setminus I_2^\gamma]$). Therefore, we have that $z(G[A\setminus I_0^\gamma]) = 3$, $z(G[A\setminus I_1^\gamma]) = 8$, and $z(G[A \setminus I_2^\gamma]) = 5$, yielding the desired counterexample. ■

In general, removing the same subset of arcs from a network with fewer arcs results in longer shortest paths, i.e. if $L \subseteq A' \subseteq A$, then $z(G[A' \setminus L]) \geq z(G[A \setminus L])$ (because $(A' \setminus L) \subseteq (A \setminus L)$). However, it is possible that such an action results in the same shortest path lengths in both networks. This observation motivates the following definition.



(a) Network G (b) Network $G[A']$

Figure 3: Networks used in Remark 3.

Definition 1. Given $G = (N, A, C)$, L and A' such that $L \subseteq A' \subseteq A$, the network $G[A']$ is called *L-spare* (with respect to G) if $z(G[A' \setminus L]) = z(G[A \setminus L])$. Moreover, if L is also a set of k-most vital arcs of $G[A']$, then the pair $(G[A'], L)$ is called *k-complete* (with respect to G), and L is referred to as a *k-set* of $G[A']$.

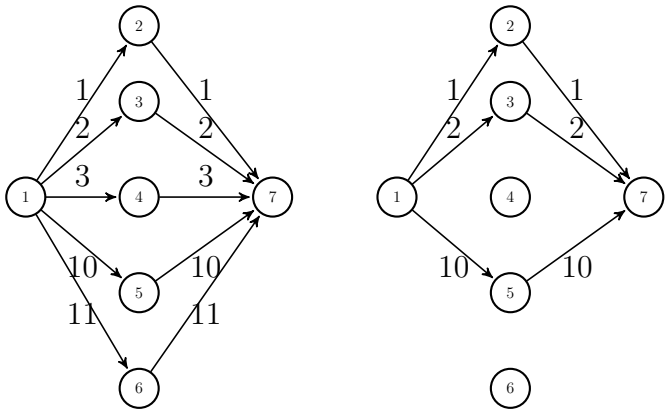
Observe that if, for a given policy $\pi \in \Pi$, time t is the first period in which $G[A_t^\pi]$ is I_t^π-spare, then $t = x^\pi$. The importance of the notion of k-completeness is illustrated by the following result.

Lemma 1. *Given $G = (N, A, C)$, let L and A' be such that $L \subseteq A' \subseteq A$ and $(G[A'], L)$ is k-complete. Then L is a set of k-most vital arcs of $G[U]$ for all U such that $A' \subseteq U \subseteq A$.*

Proof. See Appendix A.2. ∎

The practical importance of Lemma 1 lies in the fact that if one is to discover a k-complete solution of a partially observed graph, then one has indeed found a k-most vital arcs solution for the full network. This observation will play a role in showing the efficiency of the proposed policies.

Remark 3. Note that when $(G[A'], L)$ is k-complete it is not necessarily the case that every set of k-most vital arcs of $G[A']$ is a k-set of $G[A']$ (see Definition 1). That is, if $(G[A'], L)$ is k-



(a) Network G (b) Network $G[A_t^\tau]$

Figure 4: Networks used in Remark 4.

complete there might exist a set of k-most vital arcs $\widetilde{L}$ of $G[A']$ such that $G[A']$ is not $\widetilde{L}$-spare, i.e., such that $z(G[A' \setminus \widetilde{L}]) > z(G[A \setminus \widetilde{L}])$. For example, consider $k = 2$ and $G = (N, A, C)$ in Figure 3(a), and assume that $G[A']$ is as shown in Figure 3(b). Set $L = \{(1,2),(1,4)\}$ and observe that $(G[A'], L)$ is 2-complete. On the other hand, $\widetilde{L} = \{(6,8),(7,8)\}$ is a set of 2-most vital arcs of $G[A']$, but $\widetilde{L}$ is not a 2-set as $G[A']$ is not $\widetilde{L}$-spare. Indeed, $1-4-8$ is a shortest path in $G[A \setminus \widetilde{L}]$ and $z(G[A \setminus \widetilde{L}]) = 4$, while $1-5-8$ is a shortest path in $G[A' \setminus \widetilde{L}]$ and $z(G[A' \setminus \widetilde{L}]) = 11$. ■

The next two lemmas establish that the class Γ defined by (2.3) satisfies properties **C1** and **C2** (both with respect to cumulative regret and with respect to time-stability).

Lemma 2. *Let $\gamma \in \Gamma$. Then for any $\mathcal{C}_0$ and $G \in \mathbb{G}(\mathcal{C}_0)$:*

1. *$\tau^\gamma(G, \mathcal{C}_0) \leq x^\gamma(G, \mathcal{C}_0)$;*
2. *if $T > |A|$ then $\tau^\gamma(G, \mathcal{C}_0) \leq |A|$.*

Proof. See Appendix A.2. ■

Loosely speaking, the results above follow from noting that: (i) if k-completeness is satisfied, then one has found a k-most vital arcs solution, per Lemma 1; and (ii) while k-completeness is not met, new arcs are discovered in each period.

Remark 4. In general, for $\pi \in \Pi$ the fact that $G[A_t^\pi]$ is I_t^π-spare does not necessarily guarantee that $z(G[A \setminus I_t^\pi]) = z^*(G)$. To see this, consider the following example. Set $k = 2$, and consider G and $G[A_t^\pi]$ in Figures 4(a) and 4(b), respectively. Suppose that $I_t^\pi = \{(1,2),(1,5)\}$ (so that $P_t^\pi = 1-3-7$). Note that $G[A_t^\pi]$ is I_t^π-spare as $z(G[A_t^\pi \setminus I_t^\pi]) = z(G[A \setminus I_t^\pi]) = 4$, but $z(G[A \setminus I_t^\pi]) < z^*(G) = 6$. The reason for this is that I_t^π is not a set of 2-most vital arcs of $G[A_t^\pi]$. This observation further highlights the necessity of interdicting a set of k-most vital arcs in order to achieve an instantaneous regret of zero. ∎

Lemma 3. *If $\widehat{A}_0 = \emptyset$, then Γ is a homogenous set both with respect to cumulative regret and with respect to time-stability.*

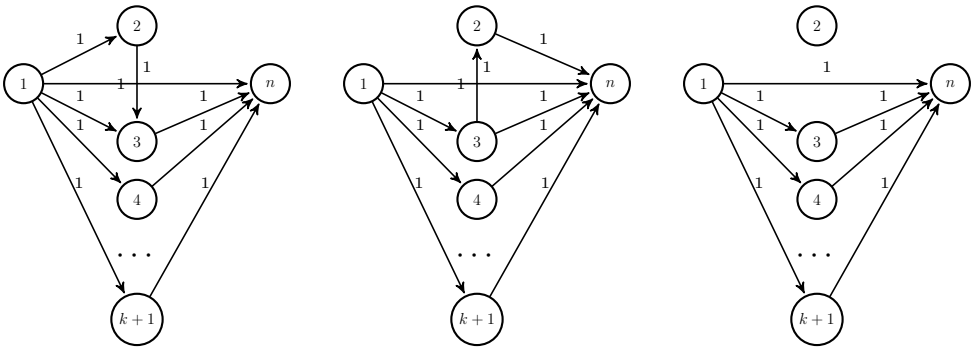
Proof. Let $k \geq 2$, $|N| \geq k+2$ and $\mathcal{C}_0$ be given by Figure 5(c),[4] where for simplicity we only show $k+2$ nodes. Consider networks G and G' depicted in Figures 5(a) and 5(b), respectively, and observe that $G, G' \in \mathbb{G}(\mathcal{C}_0)$. Clearly, $P_0^\pi = \{(1,n)\}$.

Let $\mathcal{F}_1^\pi = (\mathcal{C}_0, \emptyset, P_0^\pi)$. Observe that for the considered networks the set $\mathcal{F}_1^\pi$ is the same for all policies and the dependence on π can be dropped. Therefore, the set of policies Γ can be partitioned as $\Gamma = \Gamma^1 \cup \Gamma^2$, where $\Gamma^1 \cap \Gamma^2 = \emptyset$, $\Gamma^1 = \{\gamma\colon\ (3,n) \in I_1^\gamma = \pi_1(\mathcal{F}_1)\}$ and $\Gamma^2 = \{\gamma\colon\ (1,3) \in I_1^\gamma = \pi_1(\mathcal{F}_1)\}$. Note that for any $\gamma \in \Gamma^1$, $\tau^\gamma(G, \mathcal{C}_0) = 1$ and $\tau^\gamma(G', \mathcal{C}_0) = 2$, while for any $\gamma \in \Gamma_2$, $\tau^\gamma(G, \mathcal{C}_0) = 2$ and $\tau^\gamma(G', \mathcal{C}_0) = 1$. Likewise, if $\gamma \in \Gamma_1$ then $R_1^\gamma(G, \mathcal{C}_0) = 0$ and $R_1^\gamma(G', \mathcal{C}_0) = +\infty$, and if $\gamma \in \Gamma_2$, then $R_1^\gamma(G, \mathcal{C}_0) = +\infty$ and $R_1^\gamma(G', \mathcal{C}_0) = 0$. These observations provide the result. ∎

We have proven that there exist sets of initial information $\mathcal{C}_0$ for which there is no policy in Γ that is better (or worse) than all other policies in Γ across all $G \in \mathbb{G}(\mathcal{C}_0)$. A natural question at this point is if this result can be extended for any given $\mathcal{C}_0$. The answer is negative for both regret and time-stability as illustrated by Remark 5.

Remark 5. Consider $\mathcal{C}_0$ and $G \in \mathbb{G}(\mathcal{C}_0)$ as given in Figure 6. Let $T = 2$. At time $t = 1$ there are two sets of 2-most vital arcs: $I_1^\gamma = \{(3,4),(1,4)\}$ and $I_1^{\gamma'} = \{(1,3),(1,4)\}$. Observe that $P_1^\gamma = 1-3-2-4$ with cost $\ell(P_1^\gamma) = 6$, while $P_1^{\gamma'} = 1-2-3-4$ with cost $\ell(P_1^{\gamma'}) = 4$. Moreover, $x^\gamma(G, \mathcal{C}_0) = 1$ (thus, $\tau^\gamma(G, \mathcal{C}_0) = 1$) and $R_2^\gamma(G, \mathcal{C}_0) = 5$. On the other hand, $\tau^{\gamma'}(G, \mathcal{C}_0) = 2$, and $R_2^{\gamma'}(G, \mathcal{C}_0) = 7$.

[4] for $k = 1$ the same arguments apply after removing arc $(1,n)$ from $\mathcal{C}_0$, G and G'.



(a) Network G (b) Network G' (c) Initial information $\mathcal{C}_0$

Figure 5: Networks used in the proof of Lemma 3.

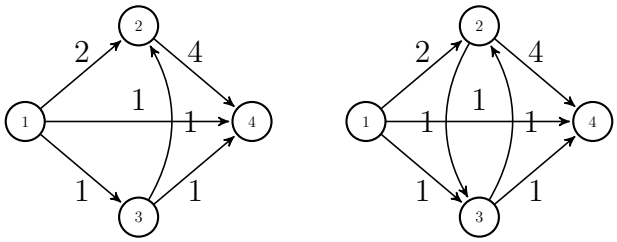
Consider any other $G' \in \mathbb{G}(\mathcal{C}_0)$ different from G. Note that adding arcs $(3,1)$, $(2,1)$, $(4,2)$ and/or $(4,3)$ to G does not affect P_1^{γ} and $P_1^{\gamma'}$. Therefore, the only possible modification is to change the cost of $(2,3)$. However, independent of this cost, $z(G'[A \setminus I_1^{\gamma'}]) \leq 6$. Thus, for any other $G' \in \mathbb{G}(\mathcal{C}_0)$ it follows that $\tau^{\gamma'}(G', \mathcal{C}_0) \geq \tau^{\gamma}(G', \mathcal{C}_0)$ and $R_2^{\gamma'}(G', \mathcal{C}_0) \geq R_2^{\gamma}(G', \mathcal{C}_0)$. Accordingly, we conclude that $\gamma \in \Gamma$ is better than (or at least as good as) any other policy in Γ for all $G \in \mathbb{G}(\mathcal{C}_0)$. ■

In view of the discussion above, it seems reasonable to define, for any given initial information $\mathcal{C}_0$, a subset of policies $\Gamma^* \subseteq \Gamma$ that contains all the policies that best resolve ties. Formally, for any set $\mathcal{C}_0$ we define

$$\Gamma^*(\mathcal{C}_0) = \left\{ \gamma \in \Gamma \colon R_T^{\gamma}(G, \mathcal{C}_0) \leq R_T^{\gamma'}(G, \mathcal{C}_0),\ \forall G \in \mathbb{G}(\mathcal{C}_0),\ \forall \gamma' \in \Gamma \right\}. \tag{2.4}$$

Observe that depending on $\mathcal{C}_0$ the set $\Gamma^*(\mathcal{C}_0)$ might be equal to Γ. Setting $\mathbb{C}^* = \{\mathcal{C}_0 \colon \Gamma^*(\mathcal{C}_0) \neq \Gamma\}$, define Γ^* as the set of policies that interdict using any k-most vital arc set of the observed network if $\mathcal{C}_0 \notin \mathbb{C}^*$, and that use the element of $\Gamma^*(\mathcal{C}_0)$ if $\mathcal{C}_0 \in \mathbb{C}^*$.

We note, however, that the set Γ^* is devoid of interest from a practical perspective. This follows as for any $t \in \mathcal{T}$ breaking a tie in advance requires the interdictor to consider, at least, all the potential replies P_t^{γ} over $(N, N \times N)$ that are consistent with $\mathcal{F}_t^{\gamma}$. Clearly, this is a task that is computationally prohibitive in general.



(a) Initial information $\mathcal{C}_0$ (b) Network G

Figure 6: Networks used in Remark 5.

Next, we establish the main result of this section:

Theorem 1. *If* $\widehat{A}_0 = \emptyset$, *then* $\Gamma \subseteq \Pi_\tau^* \cap \Pi_R^*$.

Proof. From Lemmas 2 and 3, Γ satisfies **C1** and **C2** (both with respect to cumulative regret and with respect to time-stability). We show next that Γ also satisfies **C3**.

Specifically, fix $\pi \in \Pi \setminus \Gamma$ and $\mathcal{C}_0$, and select T and $G \in \mathbb{G}(\mathcal{C}_0)$ such that at some $t_0 \in \mathcal{T}$ the set $I_{t_0}^\pi$ is *not* a set of k-most vital arcs of $G[A_{t_0}^\pi]$. Let t_0 denote the earliest among such periods. Define $\bar{G} := G[A_{t_0}^\pi]$, i.e., the arc set of $\bar{G}$ is given by $\bar{A} := A_{t_0}^\pi$, and note that $\bar{G} \in \mathbb{G}(\mathcal{C}_0)$. Also, let $(I_t^\pi, P_t^\pi)_{t\in\mathcal{T}}$ and $(\bar{I}_t^\pi, \bar{P}_t^\pi)_{t\in\mathcal{T}}$ be the unique sequences of blocking and evasion decisions generated by π for graphs G and $\bar{G}$, respectively. By the consistency assumption, namely, **(A3)**, it must hold that $P_t^\pi = \bar{P}_t^\pi$ and $I_t^\pi = \bar{I}_t^\pi$ for $0 \le t \le t_0 - 1$. Thus, $G[A_t^\pi] = \bar{G}[\bar{A}_t^\pi]$ for all $0 \le t \le t_0$. Moreover, as the interdictor acts first, then $I_{t0}^\pi = \bar{I}_{t0}^\pi$. Finally, set $\bar{T} = t_0$ and define $\bar{\mathcal{T}} = \{0, 1, \ldots, \bar{T}\}$.

By our construction there exists $\gamma \in \Gamma$ such that $\bar{I}_t^\gamma = I_t^\pi$ for $0 \le t \le t_0 - 1$, which also implies that $\bar{A}_t^\gamma = \bar{A}_t^\pi$ for $0 \le t \le t_0$. Also, π is such that set $I_{t_0}^\pi$ (which coincides with $\bar{I}_{t_0}^\pi$) is not a set of k-most vital arcs of $\bar{G}$. Therefore:

$$z(\bar{G}[\bar{A}_{t0}^\pi \setminus \bar{I}_{t_0}^\pi)) = z(\bar{G}[\bar{A} \setminus \bar{I}_{t_0}^\pi)) < z^*(\bar{G}). \tag{2.5}$$

Note that, because $\gamma \in \Gamma$, one has that $\bar{I}_{t_0}^\gamma$ is a set of k-most vital arcs of $\bar{G}[\bar{A}_{t_0}^\gamma]$. Therefore, one has that $z(\bar{G}[\bar{A}_{t_0}^\gamma \setminus \bar{I}_{t_0}^\gamma]) = z^*(\bar{G})$. Moreover, $x^\gamma(\bar{G}, \mathcal{C}_0) \le t_0$ and, by Lemma 2, $\tau^\gamma(\bar{G}, \mathcal{C}_0) \le$

t_0. In addition, from equation (2.5), we have that $\tau^{\pi}(\bar{G}, \mathcal{C}_0) > t_0$. Thus, $\tau^{\gamma}(\bar{G}, \mathcal{C}_0) < \tau^{\pi}(\bar{G}, \mathcal{C}_0)$. Therefore, **C3** holds for Γ with respect to time-stability. Finally, as the regret incurred for both π and γ from $t = 0$ to $t = t_0 - 1$ is the same, the previous observations also imply that $R^{\gamma}_{\bar{T}}(\bar{G}, \mathcal{C}_0) < R^{\pi}_{\bar{T}}(\bar{G}, \mathcal{C}_0)$ and Γ satisfies **C3** with respect to cumulative regret. ■

Remark 6. We note that in addition of having $\Gamma \subseteq \Pi^*_{\tau} \cap \Pi^*_{R}$, an implicit important feature of policies in Γ is that they provide the interdictor with a *certificate of optimality*, i.e., whenever $t = x^{\gamma}$ (which by Lemma 2 happens at a time period bounded from above by $|A|$) the interdictor has the certificate that I^{γ}_t is a set of k-most vital arcs of G. ■

Theorem 1 states that, despite its relative simplicity, the class of policies Γ defined by (2.3) is efficient with respect to regret and time-stability (i.e., it satisfies conditions **C1**-**C3**). In particular, such policies eventually attain an instantaneous regret of zero for sufficiently large values of T. However, as demonstrated next, the speed of convergence of policies in Γ may not be fast, and the bound implied in Lemma 3 may actually be tight.

Proposition 1. *There exists $\mathcal{C}_0$, $G \in \mathbb{G}(\mathcal{C}_0)$ and $\zeta > 0$ such that if $T \geq |A|$, then $\tau^{\gamma}(G, \mathcal{C}_0) \geq \zeta|A|$. Moreover, the value of $R^{\gamma}_{T}(G, \mathcal{C}_0)$ can be made arbitrarily large.*

Proof. Consider G in Figure 7. There, we have that $|A| = 2(k+1)u$, for some positive integer u.

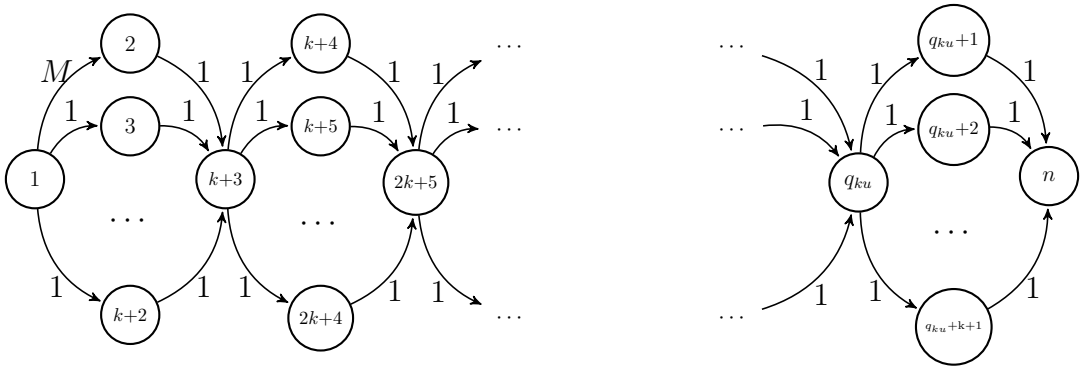


Figure 7: Network G used for the proof of Proposition 1, $q_{ku} = (u-1)(k+2)+1$.

Suppose that $M > 1$ and $A_0 = \emptyset$. Without loss of generality assume that $P^{\gamma}_0 = 1$-$(k+2)$-$(k+3)$-$(2k+4)$-$\cdots$ -$(q_{ku}+k+1)$-n (note that $(1,2) \notin P^{\gamma}_0$)), so that $A^{\gamma}_1 = P^{\gamma}_0$.

Suppose that we select $I_1^\gamma = \{(q_{ku}+k+1, n)\}$ (which is a set of k-most vital arcs) and that P_1^γ = 1-$(k+2)$-$(k+3)$-$(2k+4)$-$\cdots$-$(q_{ku}+k)$-n, so that $A_2^\gamma = P_0^\gamma \cup \{(q_{ku}, q_{ku}+k), (q_{ku}+k, n)\}$. Next, we select $I_2^\gamma = \{(q_{ku}+k+1, n), (q_{ku}+k, n)\}$, a set of k-most vital arcs, and P_3^γ is a new shortest path. Proceeding in this way, by the k^{th} period one has that $A_k^\gamma = P_0^\gamma \cup \{(q_{ku}, q_{ku}+k), (q_{ku}+k, n), \cdots, (q_{ku}, q_{ku}+2), (q_{ku}+2, n)\}$, so we select $I_k^\gamma = \{(q_{ku}+k+1, n), (q_{ku}+k, n), \cdots, (q_{ku}+2, n)\}$, which is a set of k-most vital because it is an 1-n cut with k arcs in A_k^γ.

Note that at this point, P_k^γ includes arc $(q_{ku}+1, n)$, and, thus, $A_{k+1}^\gamma = A_k^\gamma \cup \{(q_{ku}, q_{ku}+1), (q_{ku}+1, n)\}$. While $I_k^\gamma \cup \{(q_{ku}+1, n)\}$ is an 1-n cut, it is no longer a feasible interdiction as this set contains k+1 arcs. However, we can select $I_{k+1}^\gamma = I_k^\gamma \cup \{(q_{ku}-1, q_{ku})\} \setminus \{(q_{ku}+k+1, n)\}$, which is an 1-$n$ cut with k arcs, and thus, a set of k-most vital arcs. After removing those arcs, P_{k+1}^γ is of the form 1-$(k+2)$-$(k+3)$-$(2k+4)$-$\cdots$-$(q_{ku}-2)$-q_{ku}-$(q_{ku}+k+1)$-n, so that $A_{k+2}^\gamma = A_{k+1}^\gamma \cup \{(q_{k(u-1)}, q_{ku}-2), (q_{ku}-2, q_{ku})\}$. Then, we can select $I_{k+2}^\gamma = I_{k+1}^\gamma \cup \{(q_{ku}-2, q_{ku})\} \setminus \{(q_{ku}+k, n)\}$, which is a set k-most vital arcs. Proceeding in this fashion, each period (except the first one) we recover the costs of two arcs. The cost of the arc $(1, 2)$ is recovered at period ku. This implies that $\tau^\gamma(G, \mathcal{C}_0) = ku$ and hence, if $\zeta \leq k/(2k+2)$, then $\tau^\gamma(G, \mathcal{C}_0) \geq \zeta|A|$. While the latter fact depends on the selection of $\{(I_t, P_t) : t < uk\}$, one can see that, for any such selection, one can simply assign the cost M to the arc discovered last. Thus, the result holds true, independent of our selection. Finally, we observe that $R_T^\gamma(G, \mathcal{C}_0)$ can be made arbitrarily large by choosing the proper value of M. ■

Recall from Lemma 2 that $\tau^\gamma(G, \mathcal{C}_0) \leq |A|$ for any $\mathcal{C}_0$ and $G \in \mathbb{G}(\mathcal{C}_0)$. This result, in conjunction with Proposition 1, implies, loosely speaking, that $\tau^\gamma(G, \mathcal{C}_0)$ is a $\Theta(|A|)$ function.[5]

2.3.2 Efficient Policies When $\widehat{A}_0 \neq \emptyset$

In this section we assume that, in addition to $\widetilde{A}_0$, the interdictor is aware of another subset of arcs $\widehat{A}_0 \subseteq A_0$ ($\widetilde{A}_0 \cap \widehat{A}_0 = \emptyset$) for which only partial information is available. Specifically, the interdictor knows that $c_a \in [l_a, u_a]$ for some l_a and u_a, $l_a < u_a$, known upfront, for all $a \in \widehat{A}_0$.

[5]The definition of the Θ notation is given, for instance, in Ahuja et al. (1993).

Fix $\mathcal{C}_0$ and $\pi \in \Pi$, and let $(I_t^\pi, P_t^\pi)_{t\in\mathcal{T}}$ denote the unique sequence of blocking and evasion actions associated with $(G, \mathcal{C}_0)$ and π. Define a sequence of networks $\left\{G_t^\pi := (N, A_t^\pi, \widehat{C}_t^\pi) : t \in \mathcal{T}\right\}$, where for $t \in \mathcal{T}$, $\widehat{C}_t^\pi := \{\hat{c}_a,\ a \in A_t^\pi\}$ is given by

$$\hat{c}_a := \begin{cases} c_a & \text{if } a \in \widetilde{A}_t^\pi, \\ u_a & \text{if } a \in \widehat{A}_t^\pi. \end{cases} \tag{2.6}$$

In other words, for network G_t^π, the costs of arcs in $\widetilde{A}_t^\pi$ are at their known values, while the costs of arcs in $\widehat{A}_t^\pi$ are at their upper bounds (this information is part of $\mathcal{C}_0$). Note that, in general, $G_t^\pi \neq G[A_t^\pi]$. Similar to Section 2.3.1, for any policy $\pi \in \Pi$, $\mathcal{C}_0$ and $G \in \mathbb{G}(\mathcal{C}_0)$, define $\widehat{x}^\pi(G, \mathcal{C}_0)$ as

$$\widehat{x}^\pi(G, \mathcal{C}_0) := \min\{t \in \mathcal{T}:\ z(G_t^\pi[A_t^\pi \setminus I_t^\pi]) = z(G[A \setminus I_t^\pi])\}. \tag{2.7}$$

Also, as in the previous section, define Λ as the class of policies that at time t interdict a set of k-most vital arcs of G_t^π. That is, $\lambda \in \Lambda$ if and only if

$$I_t^\lambda \in \arg\max\left\{z(G_t^\lambda[A_t^\lambda \setminus I]) :\ I \subseteq A_t^\lambda,\ |I| \leq k\right\} \text{ for } t \leq \widehat{x}^\lambda, \qquad I_t^\lambda = I_{\widehat{x}^\lambda}^\lambda \text{ for } t > \widehat{x}^\lambda. \tag{2.8}$$

Note that $\widehat{x}^\pi(G, \mathcal{C}_0)$ and I_t^λ are obtained by replacing the terms $G[A_t^\pi \setminus I_t^\pi]$ and $G[A_t^\gamma \setminus I]$ by $G_t^\pi[A_t^\pi \setminus I_t^\pi]$ and $G_t^\lambda[A_t^\lambda \setminus I]$ in equations (2.2) and (2.3), respectively. Simply speaking, according to the policies in Λ the interdictor should act conservatively by assuming that the costs of arcs in $\widehat{A}_t^\lambda$ are at their upper bounds, and then apply the same approach as in Section 2.3.1. Next, we show that Λ preserves the attractive features (namely, properties **C1**-**C3**) of the policies described in Section 2.3.1. For this, we need the following technical lemma, whose proof is given in Appendix A.2.

Lemma 4. *Suppose that $t \in \mathcal{T}$ is such that $z(G_t^\lambda[A_t^\lambda \setminus I_t^\lambda]) = z(G[A \setminus I_t^\lambda])$, then $(G[A_t^\lambda], I_t^\lambda)$ is k-complete (with respect to G). Moreover, I_t^λ is a set of k-most vital arcs of $G[U]$ for all U such that $A_t^\lambda \subseteq U \subseteq A$.*

The next results, namely, Lemmas 5 and 6 and Theorem 2 generalize the results of Lemmas 2 and 3, and Theorem 1, respectively, for $\widehat{A} \neq \emptyset$ and the class of policies Λ. The proofs of the lemmas are similar to those of their counterparts in the previous section: see the details in Appendix A.2.

Lemma 5. *Let $\lambda \in \Lambda$. Then for any $\mathcal{C}_0$ and $G \in \mathbb{G}(\mathcal{C}_0)$:*

1. *$\tau^\lambda(G, \mathcal{C}_0) \leq \widehat{x}^\lambda(G, \mathcal{C}_0)$;*
2. *if $T \geq |A|$, then $\tau^\lambda(G, \mathcal{C}_0) \leq |A|$.*

Lemma 6. *Λ is a homogeneous set both with respect to cumulative regret and with respect to time-stability.*

Theorem 2. $\Lambda \subseteq \Pi^*_\tau \cap \Pi^*_R$.

Proof. From Lemmas 5 and 6, Λ satisfies conditions **C1** and **C2** (both with respect to cumulative regret and with respect to time-stability). To show that Λ also satisfies condition **C3**, we use the same construction as in the proof of Theorem 1. However, there is a subtle difference (explained in detail below, see equation (2.9) and the related discussion) due to existence of uncertain arc costs.

Fix $\pi \in \Pi \setminus \Lambda$ and $\mathcal{C}_0$. As in the proof of Theorem 1, select T such that at some $t_0 \in \mathcal{T}$ the set $I^\pi_{t_0}$ is *not* a set of k-most vital arcs of $G^\pi_{t0}[A^\pi_{t_0}]$, and let t_0 be the earliest time period among such periods. Define $\bar{G} := G^\pi_{t0}[A^\pi_{t_0}]$, i.e., the arc set of $\bar{G}$ is given by $\bar{A} := A^\pi_{t_0}$, and its costs by $\widehat{C}^\pi_{t_0}$. (Note that $\bar{G} \in \mathbb{G}(\mathcal{C}_0)$.) Also, let $(I^\pi_t, P^\pi_t)_{t\in\mathcal{T}}$ and $(\bar{I}^\pi_t, \bar{P}^\pi_t)_{t\in\mathcal{T}}$ be the unique sequences of blocking and evasions decisions generated by π for graphs G and $\bar{G}$, respectively. Next, we need to show that

$$I^\pi_t = \bar{I}^\pi_t \quad \text{and} \quad P^\pi_t = \bar{P}^\pi_t \quad \text{for } 0 \leq t \leq t_0 - 1. \tag{2.9}$$

However, unlike in the proof of Theorem 1, the costs of arcs in $\widehat{A}^\pi_{t_0}$ for G do not necessarily coincide with those for $\bar{G}$. We prove next (by contradiction) that (2.9) holds.

Let t' be the latest time period such that $0 \leq t' \leq t_0 - 1$ and (2.9) holds only for $0 \leq t \leq t' - 1$. Because the interdictor acts first, it follows that $\bar{I}^\pi_{t'} = I^\pi_{t'}$. Then it must be the case that $P^\pi_{t'} \neq \bar{P}^\pi_{t'}$. As $t' \leq t_0 - 1$, $P^\pi_{t'}$ does not contain arcs from $\widehat{A}^\pi_{t_0}$. Therefore, $P^\pi_{t'}$ is

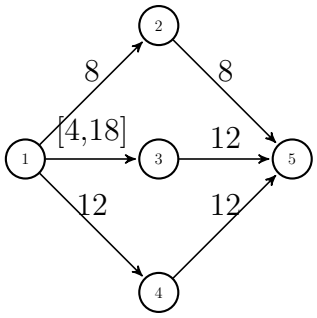


Figure 8: Network used in Remark 7.

also a shortest path in $\bar{G}[\bar{A} \setminus \bar{I}_{t'}^{\pi}]$ (recall that in $\bar{G}$ we increase only the costs of arcs in $\widehat{A}_{t_0}^{\pi}$). However, by Assumption **(A3)**, this implies that $P_{t'}^{\pi} = \bar{P}_{t'}^{\pi}$, and we arrive at a contradiction. Thus, (2.9) holds. The remainder of the proof is similar to the proof of Theorem 1. ∎

Remark 7. Suppose that instead of a conservative approach, we set the costs of arcs in $\widehat{A}_t$ to some other values (i.e., not upper bounds) when defining $\widehat{C}_t$. As we illustrate next, this leads to policies that do not necessarily converge, thus, violating **C1**. Consider the network depicted in Figure 8, and assume that the interdictor is aware of all the costs except for $c_{13} = 16$. However, its lower and upper bounds are known to the interdictor and given by $\ell_{13} = 4$ and $u_{13} = 18$, respectively. Set $k = 2$ and consider the policies that assign either (i) the lower bound, or (ii) the average of the upper and lower bounds to arcs with unknown costs. Observe that such policies assign the cost of 16 and 23 to path $1-3-5$, respectively. Thus, the interdictor would always remove one arc of this path and another one of path $1-2-5$. Note that the instantaneous regret associated with such interdiction is 4 and that if the interdiction decision does not change, the real cost of arc $(1,3)$ would never be revealed, implying that condition **C1** is not satisfied when the actual cost of $(1,3)$ is $c_{13} > 12$. Furthermore, one can see that the example extends to any value used by the interdictor, other than the upper bound. ∎

One can show that Remarks 2, 4, and 5 also hold in this setting (via similar counterexamples); likewise policies in Λ also provide a certificate of optimality for the interdictor, as in Remark 6. In addition, Proposition 1 can also be extended to the case of $\widehat{A}_0 \neq \emptyset$, as we show next.

Proposition 2. *There exists* $\mathcal{C}_0$, $G \in \mathbb{G}(\mathcal{C}_0)$ *and* $\zeta > 0$ *such that if* $T \geq |A|$, *then* $\tau^{\lambda}(G, \mathcal{C}_0) \geq \zeta|A|$. *Moreover, the value of* $R_T^{\lambda}(G, \mathcal{C}_0)$ *can be made arbitrarily large.*

Proof. See Appendix A.2. ∎

2.4 LOWER BOUNDS FOR POLICY PERFORMANCE

In the previous section we established the existence of efficient policies. A natural question arising at this point is that of how close is the total regret and time-stability performance of these policies to the performance of a best "practical" and "implementable" policy. This section is devoted to answering such question. First, in Section 2.4.1 we argue that additional restrictions must be imposed on how the decisions are made by these policies in the complete information setting, which leads to the definition of semi-oracle policies. Later, in Sections 2.4.2 and 2.4.3, we detail how semi-oracle policies can be computed for the case of regret and time-stability, respectively, and propose algorithms to improve the performance of their computation.

2.4.1 Semi-oracle Policies

To measure policy performance, the classical theory in sequential prediction often postulates a probabilistic model of uncertainty and computes, for example, the expected cost incurred by a policy, thus allowing one to search for an efficient policy based on this well-defined criterion. Assuming that such probabilistic model is initially unknown, the learning literature focuses instead, for the most part, on the concept of regret, namely, the cumulative loss in cost relative to that of an oracle (i.e., an oracle interdictor in our setting) with advanced knowledge of the underlying probabilistic model, see Cesa-Bianchi and Lugosi (2006*b*). In our problem such information consists of the structure and arc costs of the network.

In most situations the assumption of existence of such an oracle is impractical, as feasible policies do not posses such advanced information. Nonetheless, an oracle-based benchmark can be used for normalization (i.e., preventing optimal performance to grow with the time

horizon), and also for bounding the opportunity cost of missing information. While the first use also applies to our problem, the second one can be improved upon. Indeed, the oracle policy in our setting would block a set of k-most vital arcs of G in each period, i.e.,

$$I_t^{\text{oracle}} \in \arg\max \{z(G[A \setminus I]) : \ I \subseteq A \text{ s.t. } |I| \leq k\}, \quad t \in \mathcal{T}.$$

However, this might imply using information that is not available to *any* feasible policy because, in general, $A_0 \neq A$. Thus, in our setting the bound on the opportunity cost of information is trivial (note, for example, that $\tau^{\text{oracle}}(G, \mathcal{C}_0) = 0$) and not particularly meaningful. Fortunately, it is possible to tighten such a bound by asking the oracle not to *signal* the availability of advanced information through its actions. That is, we impose that $I_t^{\text{oracle}} \subseteq A_t$ for all $t \in \mathcal{T}$. We refer to such interdictor as a semi-oracle interdictor.

Simply speaking, the semi-oracle is an interdictor, who, while having complete knowledge of all the arcs and costs in the network, at any given time period can only remove either arcs that have been observed so far, i.e., those used in earlier time periods by the evader, or that lie within A_0. Note that because the semi-oracle interdictor knows G, it is capable of anticipating the evader's actions and feedback for any sequence of blocking decisions. Therefore, the semi-oracle is also capable of evaluating with certainty the cost incurred by the evader across all time periods for any such sequence, and as a result his/her actions are not necessarily adapted to the history of the process (see Remark 8). Moreover, one can show that the semi-oracle in fact uses the best non-adapted policy. Note that while the existence of such semi-oracle is still impractical, its actions for a given instance can be matched by some feasible adapted policies, and thus it serves as a more reasonable benchmark, relative to a traditional oracle. (Please see some additional discussion in Section 2.4.2, in particular, Remark 8.)

Finally, although the semi-oracle's actions are the result of computations far more complex than those of an oracle, it is still possible to formulate mathematically the decision problem faced by the semi-oracle interdictor, and to reconstruct its actions. In the following sections, we formulate such a problem as a mixed integer program (MIP), first, for the case of cumulative regret (Section 2.4.2), and then for time-stability (Section 2.4.3). In the remainder of the section we assume that $\widehat{A}_0 = \emptyset$. Also, to simplify the exposition and shorten the

notation we refer to the semi-oracle interdictor as an oracle interdictor, or simply, an oracle. Similarly, the resulting policies are referred to as oracle-based instead of semi-oracle-based.

2.4.2 Lower Bound for Regret

Given $\mathcal{C}_0$, $G \in \mathbb{G}(\mathcal{C}_0)$ and T, suppose the oracle interdictor knows G at time $t = 0$, aims at maximizing the evader's cumulative cost over $\mathcal{T}$, but is restricted to selecting $I_t \subseteq A_t$ for $0 \leq t \leq T$. That is, the oracle interdictor solves the following bilevel (max-min) optimization problem:

$$LB(G, \mathcal{C}_0, T): \ \max \sum_{t \in \mathcal{T}} \ell(P_t) \tag{2.10a}$$

$$\text{s.t. } P_t \in \arg\min \left\{ \sum_{a \in P} c_a : P \text{ is a path in } G[A \setminus I_t] \right\} \quad \forall t \in \mathcal{T}, \tag{2.10b}$$

$$I_0 = \emptyset, \ I_t \subseteq A_t, \ |I_t| \leq k, \ A_t = A_{t-1} \cup P_t \quad \forall t \in \mathcal{T} \setminus \{0\}. \tag{2.10c}$$

Note that in order to produce a valid lower bound, the bilevel problem (2.10) is *optimistic* in the sense that the interdictor has some degree of control over the decisions made by the evader. Specifically, if (2.10b) has multiple optimal solutions (i.e., multiple shortest paths), then the evader delegates the decision to the oracle interdictor (otherwise, it is potentially possible to improve upon the interdictor's actions). While this modeling assumption is common in the bilevel optimization literature (see, e.g., Beheshti et al. (2015), Colson et al. (2007*b*) and references therein), in our setting it is necessary to obtain valid lower bounds for the performance (with respect to regret) of *any* policy in Π.

In order to solve (2.10) we observe that the evader's (lower-level) problem (2.10b) is the shortest path problem, which admits a compact linear programming (LP) formulation, see Ahuja et al. (1993). Consequently, the initial bilevel problem (2.10) can be reformulated as a single-level mixed integer program by exploiting the LP duality. We should note that this is a standard approach in the bilevel optimization literature (Colson et al. 2007*b*), which can be applied as long as the lower-level optimization problem can be replaced by its necessary and sufficient optimality conditions.

For any arc $(i,j) \in A$ and period $t \in \mathcal{T}$, define $r_{ij}^t = 1$ if arc (i,j) is blocked at time t, and $r_{ij}^t = 0$ otherwise. Similarly, define $p_{ij}^t = 1$ if the evader travels along arc (i,j) at time t, and $p_{ij}^t = 0$, otherwise. For $t \in \mathcal{T}$, define $\{y_i^t\}_{i \in N}$ as the variables in the dual of the LP formulation of (2.10b). Then we have the following alternative formulation of (2.10):

$$
\begin{aligned}
LB(G, \mathcal{C}_0, T): \ \max & \sum_{t=0}^{T} (y_1^t - y_n^t) \\
\text{s.t. } & y_i^t - y_j^t \leq c_{ij} + M\, r_{ij}^t && \forall t \in \mathcal{T},\ \forall (i,j) \in A, && \text{(2.11a)} \\
& \sum_{j:(i,j)\in A} p_{ij}^t - \sum_{j:(j,i)\in A} p_{ji}^t = \begin{cases} -1 & i = 1 \\ 1 & i = n \\ 0 & \text{otherwise} \end{cases} && \forall t \in \mathcal{T}, && \text{(2.11b)} \\
& \sum_{(i,j)\in A} c_{ij}\, p_{ij}^t = y_1^t - y_n^t && \forall t \in \mathcal{T}, && \text{(2.11c)} \\
& r_{ij}^t \leq \sum_{s=0}^{t-1} p_{ij}^s && \forall t \in \mathcal{T} \setminus \{0\},\ \forall (i,j) \in A \setminus A_0, && \text{(2.11d)} \\
& p_{ij}^t \leq 1 - r_{ij}^t && \forall (i,j) \in A,\ \forall t \in \mathcal{T}, && \text{(2.11e)} \\
& \sum_{(i,j)\in A} r_{ij}^t \leq k && \forall t \in \mathcal{T}, && \text{(2.11f)} \\
& r_{ij}^0 = 0 && \forall (i,j) \in A, && \text{(2.11g)} \\
& r_{ij}^t, p_{ij}^t \in \{0,1\} && \forall (i,j) \in A,\ \forall t \in \mathcal{T}, && \text{(2.11h)}
\end{aligned}
$$

where M is a sufficiently large constant parameter. Constraints (2.11a) and (2.11b) correspond to the dual and primal constraints of the LP formulation of the shortest path problem in the interdicted network, respectively. Note that the right-hand side of (2.11c) is the length of the shortest path in $G[A \setminus I_t]$. Thus, if suffices to consider $M = (n-1) \cdot \max\{c_a \mid a \in A\}$ (we do so in our numerical experiments). Constraints (2.11c) enforce strong duality at all times, so that $\left\{p_{ij}^t : (i,j) \in A\right\}$ corresponds to a shortest path in the interdicted network at time t. Constraints (2.11d) ensure that the blocking decision at time t includes only the arcs that have been observed prior to time t, and constraints (2.11e) prevent the evader from using blocked arcs. Finally, constraints (2.11f) impose that at most k arcs are interdicted in any period.

Let (r^*, p^*, y^*) denote an optimal solution to (2.11). We have that

$$I_t^*(G, \mathcal{C}_0, T) := \left\{(i,j) \in A : \ r_{ij}^{*t} = 1\right\} \text{ and } P_t^*(G, \mathcal{C}_0, T) := \left\{(i,j) \in A : \ p_{ij}^{*t} = 1\right\}, \ t \in \mathcal{T}$$

is a feasible sequence of blocking and evasion decisions, which we refer to as the *oracle-based* policy.

Note that the oracle-based policy does not belong to Π, as it is not *adapted* to the history of the process, see Remark 8. However, its performance serves as a lower-bound for regret of any policy because each policy in Π can be mapped to a feasible solution to (2.11). Therefore:

$$\sum_{t \leq T} (z^*(G) - z(G[A \setminus I_t^*(G, \mathcal{C}_0, T)])) \leq R_T^\pi(G, \mathcal{C}_0), \quad \pi \in \Pi, \tag{2.12}$$

for any $\mathcal{C}_0$, $G \in \mathbb{G}(\mathcal{C}_0)$ and T.

Remark 8. Recall that for any $\pi \in \Pi$ there is a unique set I_t^π associated with each sequence $\mathcal{F}_t^\pi$. The oracle-based policy, however, might choose different sets I_t^* for the same sequence $\mathcal{F}_t^*$ because its actions are allowed to depend on G. For instance, consider the networks $G = (N, A, C)$ and $G' = (N, A', C')$ depicted in Figure 1(a) and (b), respectively, set $k = 2$, and assume that $A_0 = \{(1,7)\}$. For both G and G' the oracle-based policy yields $I_1^* = \{(1,7)\}$, which implies that $\mathcal{F}_2^* = (\mathcal{C}_0, \emptyset, P_0, (1,7), P_1)$, where $P_0 = 1-7$, and $P_1 = 1-3-6-7$. One can check that the oracle policy satisfies that $I_2^* = \{(1,7),(3,6)\}$ for $(G, \mathcal{C}_0)$, while $I_2^* = \{(1,7),(6,7)\}$ for $(G', \mathcal{C}_0)$. Hence, the oracle policy determines two different interdiction actions at time $t = 2$ for the same $\mathcal{F}_2^*$. ■

The oracle problem LB can be shown to be *NP*-hard using the reduction from the k-most vital arcs problem (Ball et al. 1989). Nevertheless, MIP formulation (2.11) can be effectively tackled by state-of-the-art solvers for small values of T and k (see Section 2.5). However, for larger values of T and k, a significant portion of the solver's running-time is invested into finding a feasible solution to LB. With this in mind, and considering that the total solution time typically depends on the quality of such solutions, we develop Algorithm 1, which constructs an initial feasible solution of (2.11).

The algorithm begins by finding a set of k-most vital arcs in G, and then solves a sequence of at most k shortest path problems. Thus, its practical complexity is that of the k-most vital

arcs problem, for which there exist effective solution algorithms (Israeli and Wood 2002). The intuition behind the algorithm is based on the following observation. Suppose $\mathcal{I}^*$ is a set of k-most vital arcs of G. Then, starting from the set $I_1 = A_0 \cap \mathcal{I}^*$, the evader's response each time period must reveal at least one arc in $\mathcal{I}^* \setminus I_t$. Thus, one can reconstruct $\mathcal{I}^*$ in at most k time periods simply by solving a shortest path in each period, and adding the newly revealed elements of $\mathcal{I}^*$ into the blocking action.

The pseudo-code of the approach is provided in Algorithm 1, where MostVitalArcs(G,k) returns a set of k most vital arcs in G and the length of the optimal solution, and Shortest–Path(G) returns the primal and dual solution to the LP formulation of the shortest path problem, as well as the optimal path length. Also, $\mathbf{1}\{\cdot\}$ denotes the indicator function.

Algorithm 1 Finding a feasible solution for $LB(G, \mathcal{C}_0, T)$

Require: $G = (N, A, C)$, A_0, k, and T

$[\mathcal{I}^*, z^*]$ =MostVitalArcs(G,k)

$I_0 = \emptyset$, $[p^0, y^0, z^0]$ =ShortestPath($G \setminus I_0$), $t = 0$

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

$t = t + 1$

$I_t = I_{t-1} \cup \left(\left\{(i,j) \in \mathcal{I}^* : \ p_{ij}^{t-1} = 1\right\}\right)$

$[p^t, y^t, z^t]$ =ShortestPath($G \setminus I_t$)

end while

Set $r_{ij}^s := \mathbf{1}\{(i,j) \in I_s\}$ $\forall s \leq t$, and $r^s = r^t$, $p^s = p^t$, $y^s = y^t$ $\forall s > t$.

return $\{(r^t, p^t, y^t) : \ t \in \mathcal{T}\}$

Feeding the initial feasible solution given by Algorithm 1 to the MIP solver decreases the overall solution time of LB (see, for example, the results in Section 2.5). However, for sufficiently large values of T any MIP-oriented solution approach is not effective.[6] Note, however, that Algorithm 1 provides an approach for finding a feasible solution of LB, which identifies a set of k-most vital arcs within k time periods, and then, these k-most vital arcs are successively repeated until time T. This observation suggests that it might be possible to solve LB for a relatively short time horizon and extend the solution to a larger time horizon.

[6]Increasing T by Δt increases the number of variables and constraints of LB by a factor of $\Theta(\Delta t \times |A|)$, which translates into an exponential increase in the worst-case time performance for any MIP solver based on branch-and-bound ideas.

(Note that this is not always possible, as there exist networks such that the optimal solution of LB does not involve discovering a set of k-most vital arcs sufficiently early, or even at all.)

Algorithm 2 incorporates the ideas above. There, T_0 corresponds to the time period in which a k-most vital arc solution is first discovered in Algorithm 1. The algorithm iterates from time $T' = T_0$ to time $T' = T$, solving $LB(G, \mathcal{C}_l, \mathcal{T}')$ at each time. If the solution of $LB(G, \mathcal{C}_l, \mathcal{T}')$ involves discovering a set of k-most vital arcs, then it is optimal to extend such set up to time T. Otherwise, the algorithm sets $T' = T' + 1$, and $LB(G, \mathcal{C}_l, \mathcal{T}')$ is solved again.

Algorithm 2 Solving $LB(G, \mathcal{C}_0, T)$

Require: $G = (N, A, C)$, A_0, k, T

Use Algorithm 1 to find z^* and T_0

if $T_0 \geq T$ **then**

Solve $LB(G, \mathcal{C}_l, \mathcal{T})$ via formulation (2.11)

else

Set $T' = T_0$

Solve $LB(G, \mathcal{C}_l, \mathcal{T}')$ via formulation (2.11) and denote by $\left\{(\tilde{r}^t, \tilde{p}^t, \tilde{y}^t) : \ \forall t \leq T'\right\}$ its solution

while $z^* > y_1^{T'} - y_n^{T'}$ and $T' < T$ **do**

Set $T' = T' + 1$

Solve $LB(G, \mathcal{C}_l, \mathcal{T}')$ via formulation (2.11) and denote by $\left\{(\tilde{r}^t, \tilde{p}^t, \tilde{y}^t) : t \leq T'\right\}$ its solution

end while

Set $(r^t, p^t, y^t) = (\tilde{r}^t, \tilde{p}^t, \tilde{y}^t)$ for all $t \leq T'$, and $(r^t, p^t, y^t) = (\tilde{r}^{T'}, \tilde{p}^{T'}, \tilde{y}^{T'})$ for all $t > T'$

end if

return $\{(r^t, p^t, y^t) : \ t \in \mathcal{T}\}$

(Note: efficiency of the algorithm is improved by providing an initial feasible solution of (2.11) to the MIP solver each time it is called. Such solutions can easily be constructed initially from Algorithm 1, and later from the solution of (2.11) in the previous iteration.)

Proposition 3. *Algorithm 2 correctly solves* $LB(G, \mathcal{C}_l, \mathcal{T})$.

Proof. See Appendix A.2. ∎

2.4.3 Lower Bound for Time-Stability

We extend the ideas in the previous section to the case of time-stability. In particular, for $\mathcal{C}_0$, $G \in \mathbb{G}(\mathcal{C}_0)$ and T, the oracle interdictor solves the following MIP:

$$
\begin{aligned}
TS(G, \mathcal{C}_0, T): \quad & \min \quad \sum_{t=0}^{T} w_t \\
& \text{s.t.} \quad z^*(1 - w^t) \leq y_1^t - y_n^t \quad \forall t \in \mathcal{T}, && \text{(2.13a)} \\
& \qquad w^t \in \{0, 1\} \qquad\qquad \forall t \in \mathcal{T}, && \text{(2.13b)} \\
& \qquad \text{and constraints (2.11a) to (2.11h).}
\end{aligned}
$$

In this formulation, w^t indicates whether the blocking decision in period t, $t \in \mathcal{T}$, is a set of k-most vital arcs for G or not. Note that constraints (2.13a) force $w^t = 1$ when the evader's path length at period t is lower than z^* (otherwise, $w^t = 0$ due to the objective function) for $t \in \mathcal{T}$. Let (r^*, p^*, y^*, w^*) denote an optimal solution to (2.13). As in the previous section, denote by

$$
I_t^*(G, \mathcal{C}_0, T) := \left\{(i,j) \in A : \; r_{ij}^{*t} = 1\right\} \text{ and } P_t^*(G, \mathcal{C}_0, T) := \left\{(i,j) \in A : \; p_{ij}^{*t} = 1\right\}, \; t \in \mathcal{T}
$$

the *oracle-based* policy (for time-stability). (Note that this policy is not necessarily in Π; recall Remark 8, which can be extended for the case of time-stability). Also, as in the previous section, we have that time-stability of the oracle based policy is a lower bound for time-stability of policies in Π. That is, for any $\mathcal{C}_0$, $G \in \mathbb{G}(\mathcal{C}_0)$ and T:

$$
\sum_{t \in \mathcal{T}} w^{*t} \leq \tau^{\pi}(G, \mathcal{C}_0), \qquad \pi \in \Pi. \tag{2.14}
$$

Solving (2.13) entails the same difficulties that are faced when solving LB. In this regard, Algorithm 1 also provides a feasible solution to the formulation above, provided one sets $w^t = 1$ for all $t < T_0$, and $w^t = 0$, otherwise. Note, however, that unlike in the case of regret minimization, T_0 provides an upper bound to the time-stability of the oracle-based policy. Thus, it is sufficient to solve $TS(G, \mathcal{C}_0, T_0)$ to generate an optimal solution for $TS(G, \mathcal{C}_0, T)$, where $T \geq T_0$. The solution procedure is summarized in Algorithm 3. Its correctness follows from the fact that T_0 is an upper bound for time-stability of the oracle-based policy.

Algorithm 3 Solving $TS(G, \mathcal{C}_0, T)$

Require: $G = (N, A, C)$, A_0, k, T

Use Algorithm 1 to find z^* and T_0

if $T_0 \geq T$ **then**

Solve $TS(G, \mathcal{C}_0, T)$ via formulation (2.13)

else

Solve $TS(G, \mathcal{C}_0, T_0)$ via formulation (2.13) and denote by $\{(\tilde{r}^t, \tilde{p}^t, \tilde{y}^t, \tilde{w}^t) : \ t \leq T_0\}$ its solution

Set $(r^t, p^t, y^t, w^t) = (\tilde{r}^t, \tilde{p}^t, \tilde{y}^t, \tilde{w}^t)$ for $t \leq T_0$, and $(r^t, p^t, y^t, w^t) = (\tilde{r}^{T_0}, \tilde{p}^{T_0}, \tilde{y}^{T_0}, 0)$ for $t > T_0$

end if

return $\{(r^t, p^t, y^t, w^t) : \ t \in \mathcal{T}\}$

2.5 COMPUTATIONAL STUDY

In this section, we study the practical performance of the proposed policies and algorithms. First, in Section 2.5.1 we describe our test instances, three additional benchmark policies and the implementation details. Then in Section 2.5.2 we briefly analyze the performance of Algorithms 1 and 2 for computation of the oracle-based policies introduced in Section 2.4. In Section 2.5.3 we compare the performance of the efficient policies discussed in Sections 2.3.1 and 2.3.2 (namely, Γ and Λ, respectively) against different benchmark policies (including oracle-based). We also conduct sensitivity analysis of the policies in Λ with respect to the amount and the *quality* of information initially available to the interdictor in Sections 2.5.4 and 2.5.5, respectively.

2.5.1 Test Instances, Benchmark Policies and Implementation Details

Network Structure and Arc Costs. We test our policies using the class of uniform random graphs (uniform graphs) of Erdös and Rényi (1959). The cost structure of each graph instance is generated as follows. First, for each arc $a \in A$, the bounds l_a and u_a are drawn

randomly (in sequence) from uniform distributions $U(0, 500)$ and $U(l_a, 500)$, respectively. Then cost c_a is set to $l_a + (u_a - l_a)x_a$, where x_a is drawn from a Beta(α, β) distribution. As policy performance might be sensitive to the relative location of c_a in $[l_a, u_a]$, we consider *left-skewed*, *symmetric* and *right-skewed* cost distributions by setting (α, β) to $(2, 10)$, $(10, 10)$ and $(10, 2)$, respectively.

Benchmark Policies. We consider three additional benchmark policies:

(i) The *lower bound* policy π_L interdicts a set of k-most vital arcs in the observed network, assuming that the cost of $a \in \widehat{A}_t$ is $c_a = l_a$. That is, the policy operates as a policy in Λ, but uses the lower bound l_a instead of u_a in equation (2.6).

(ii) The *mean bound* policy π_M interdicts a set of k-most vital arcs in the observed network, assuming that the cost of $a \in \widehat{A}_t$ is $c_a = (l_a + u_a)/2$.

(iii) The *random bound* policy π_R interdicts a set of k-most vital arcs in the observed network, assuming that the cost of $a \in \widehat{A}_t$ is either $c_a = l_a$ or $c_a = u_a$ with equal probability.

Initial Information. We design instances, where a fraction $p_a \in [0, 1]$ of the arcs from A is included into A_0, and a fraction $p_c \in [0, 1]$ of A_0 is included into $\widetilde{A}_0$. Specifically, for a given pair (p_a, p_c), starting from $A_0 = \emptyset$, arcs are selected randomly (uniformly) without replacement from A and added to A_0, until $|A_0| = \lfloor |A| \times p_a \rfloor$. Then, starting from $\widetilde{A}_0 = \emptyset$, arcs are selected randomly (uniformly) without replacement from A_0 and added to $\widetilde{A}_0$, until $|\widetilde{A}_0| = \lfloor |A_0| \times p_c \rfloor$. We construct these sets in a nested fashion, i.e., the set A_0 generated for p_a is a subset of that generated for $p'_a > p_a$. The same applies for $\widetilde{A}_0$.

Implementation Details. The algorithms are coded in Matlab R2012b and all the experiments are performed on a Windows PC with 3.7GHz CPU and 32GB RAM. We solve (2.11) and (2.13) using CPLEX 12.4. For finding a k-most vital arc solution we use the *basic covering decomposition algorithm* from Israeli and Wood (2002).

2.5.2 Computation of the Oracle-based Policy

In this section we demonstrate performance of Algorithms 1 and 2 by comparing three different approaches for solving LB. In the first approach, referred to as "MIP," we feed formulation (2.11) directly to a state-of-the-art MIP solver. In the second, which we denote

as "MIP i.s.," we use Algorithm 1 to generate an initial feasible solution to LB, which is then fed to the MIP solver together with (2.11). In the third, which we refer to as "Alg.," we use Algorithm 2. We test efficiency of these three solution procedures using 10 randomly generated uniform graphs, each considering $n = 40$, $p = 0.5$, $p_a = p_c = 0$ and $T = 15$. Costs are drawn from the symmetric distribution. Table 2 summarizes the running-time (in seconds) of each solution approach, where $k \in \{2, 4, 6, 8\}$. We set a time limit of an hour for all methods.

Table 2: Running times (in seconds) to solve LB. The entry "-" implies that an optimal solution was not found within one hour.

k=2			k=4			k=6			k=8		
Alg.	MIP i.s.	MIP	Alg.	MIP i.s.	MIP	Alg.	MIP i.s.	MIP	Alg.	MIP i.s.	MIP
1.5	17.9	29.9	5.4	59.0	124.5	35.9	707.4	-	-	-	-
0.6	4.8	40.1	2.2	22.1	63.3	4.7	169.4	205.6	-	-	-
1.5	9.6	33.3	3.1	122.1	641.1	22.5	1658.8	-	534.9	-	-
1.4	215.5	-	14.7	1799.5	2626.7	15.1	-	-	93.1	-	-
1.2	6.4	21.0	3.1	19.3	57.0	8.9	285.9	697.0	29.2	729.7	711.8
1.3	6.4	90.1	5.1	48.8	74.3	15.1	447.2	447.8	41.4	-	-
1.1	5.5	24.3	8.2	476.2	1742.1	36.7	-	-	-	-	-
1.0	4.5	20.7	1.8	12.9	95.7	3.1	115.4	591.3	130.6	-	2244.6
1.4	9.4	72.3	16.1	685.4	1425.1	33.5	-	-	190.4	-	-
1.3	35.3	2842.9	9.5	1104.3	-	9.4	1166.1	-	18.5	-	-

We observe that, in general, the solution time of Algorithm 2 is between one or two orders of magnitude less than that of the MIP solver. Also, we note that while providing an initial feasible solution to the solver improves its performance, Algorithm 2 is still significantly faster than the other methods. This suggests that the practical efficiency of Algorithm 2 can be mostly attributed to the idea of extending the solution of LB with a shorter time horizon (recall our discussion in Section 2.4.2).

2.5.3 Comparison to Benchmark Policies

In this section we compare the performance of $\lambda \in \Lambda$ and $\gamma \in \Gamma$ against the benchmark policies. For each pair (p_a, p_c), where $p_a, p_c \in \{0, 1/3, 2/3, 1\}$, we generate 20 random networks along with subsets A_0 and $\widetilde{A}_0$, and for each of them we generate three different cost vectors corresponding to each cost distribution. We also set $n = 40$, $p = 0.5$, $k = 6$ and $T = 21$.

First, we consider the settings from Section 2.3.1, where $\widehat{A}_0 = \emptyset$ (i.e., $p_c = 1$). (Note that in this case, classes Γ, Λ, π_L, π_M and π_R are equivalent.) Tables 3 and 4 summarize the regret and time-stability for each setting, respectively. There, γ denotes a policy in Γ, and each entry represents the average performance among all 20 instances and the mean absolute deviation (MAD), in parenthesis. To quantify the value of the initial information (in particular, the size of A_0), we include the performance of the oracle-based policy for the corresponding performance metrics (regret and time-stability), π^{oracle}.

Table 3: Average cumulative regret ($\times 10^2$) and MAD (in parenthesis) for $k = 6$.

	Left-Skewed		Symmetric		Right-Skewed	
p_a	γ	π^{oracle}	γ	π^{oracle}	γ	π^{oracle}
0	5.83 (2.12)	2.27 (0.64)	6.55 (2.57)	3.72 (1.20)	9.42 (3.32)	4.70 (1.46)
1/3	4.28 (1.19)	1.19 (0.49)	5.41 (1.94)	1.95 (0.82)	8.10 (2.72)	2.39 (1.03)
2/3	1.78 (1.03)	0.31 (0.34)	2.29 (1.04)	0.29 (0.32)	3.47 (1.49)	0.87 (0.76)

Table 4: Average time-stability and MAD (in parenthesis) for $k = 6$.

	Left-Skewed		Symmetric		Right-Skewed	
p_a	γ	π^{oracle}	γ	π^{oracle}	γ	π^{oracle}
0	12.55 (2.55)	4.95 (0.57)	9.60 (1.86)	5.40 (0.60)	10.50 (1.10)	5.05 (0.48)
1/3	9.75 (2.05)	3.15 (0.70)	8.30 (1.66)	3.65 (0.86)	8.95 (1.76)	3.70 (0.94)
2/3	5.15 (1.87)	1.80 (0.56)	4.30 (1.53)	1.65 (0.52)	5.10 (1.22)	2.2 (0.86)

We observe that the performance of policy $\gamma \in \Gamma$ is roughly between 2 to 4 times that of the oracle-based policy. Given that π^{oracle} has complete information about the network structure and arc costs, the difference in the performance of the policies is reasonably modest. (For example, in the experiments discussed below $\lambda \in \Lambda$ performs orders of magnitude better than the other benchmark policies, namely, π_L, π_M and π_R, in a number of test instances.) As one would expect, as the value of p_a increases (i.e., the size of the initially known arc subset A_0 increases), the performance of both policies improves; indeed, observe that the mean regret and time-stability as well as the regret and time-stability MAD decrease as p_a increases. Also, it is worth noting that the performance of the oracle policy differs significantly from zero, the value it would obtain if it was possible to signal availability of complete initial information.

Table 5 depicts the running-time statistics for computation of π^{oracle} (under the regret performance metric, thus, using Algorithm 2), which correspond to the results reported in Table 3. Similarly, in Table 6 we report the running-time statistics for computation of π^{oracle} (under the time-stability performance metric, thus, using Algorithm 3), which correspond to the results reported in Table 4. (The running time for computation of γ is not reported as for all instances it is less than 5 and 1 seconds for the regret and time-stability metrics, respectively.) It can be observed that, in general, the left-skewed case is significantly more time consuming. Furthermore, it is interesting to note that computing the oracle-based policy is more challenging for the regret performance metric than for time-stability.

Table 5: Average running times (in seconds) per replication and MAD (in parenthesis) for computing π^{oracle} using Algorithm 2 (regret performance metric), which correspond to the results reported in Table 3. Average times for computing γ are below 5 seconds across all configurations.

p_a	Left Skewed	Symmetric	Right Skewed
0	177.81 (255.77)	99.87 (153.27)	27.05 (21.24)
1/3	447.02 (756.74)	59.46 (32.97)	66.93 (30.91)
2/3	639.82 (1175.40)	22.88 (13.41)	48.92 (36.51)

Table 6: Average running times (in seconds) per replication and MAD (in parenthesis) for computing π^{oracle} using Algorithm 3 (time-stability performance metric), which correspond to the results reported in Table 4. Values of time-stability for γ are computed instantly given the regret.

p_a	Left Skewed	Symmetric	Right Skewed
0	41.86 (50.50)	13.53 (9.77)	12.16 (7.52)
1/3	25.95 (19.60)	29.57 (17.02)	32.32 (17.69)
2/3	10.69 (5.71)	10.32 (5.32)	18.52 (10.84)

Next, we consider instances with $\widehat{A}_0 \neq \emptyset$ and $p_c \in \{0, 1/3, 2/3\}$. Tables 7 and 8 summarize the regret and time-stability for each setting, respectively. There, γ and λ denote

policies in Γ and Λ, respectively. As before, each entry represents the average performance among all 20 instances and MAD, in parenthesis. Note that because the policies in Γ are defined for settings where $\widehat{A}_0 = \emptyset$, these type of policies discard any information in the set $\widehat{A}_0$.

Table 7: Average regret ($\times 10^2$) and MAD (in parenthesis) for $k = 6$. Among the entries denoting regret, the entry in bold is the best value, and the other entries indicate the difference with respect to the best value.

	Left-Skewed					Symmetric					Right-Skewed				
(p_a, p_c)	λ	π_L	π_M	π_R	γ	λ	π_L	π_M	π_R	γ	λ	π_L	π_M	π_R	γ
$(\frac{1}{3}, 0)$	0.54	0.93	**5.17**	0.87	0.67	0.32	7.70	**5.58**	2.98	0.97	**8.52**	17.26	3.03	6.98	0.90
	(1.42)	(3.54)	(1.49)	(1.93)	(2.12)	(2.04)	(4.80)	(1.88)	(2.54)	(2.57)	(3.10)	(9.85)	(4.40)	(3.49)	(3.32)
$(\frac{1}{3}, \frac{1}{3})$	0.27	0.26	**5.16**	0.27	0.35	0.52	3.96	**5.53**	1.61	1.37	**8.37**	6.97	1.61	3.43	1.15
	(1.51)	(2.39)	(1.46)	(1.79)	(1.81)	(2.11)	(4.12)	(1.95)	(2.05)	(2.29)	(2.95)	(5.93)	(3.92)	(4.20)	(3.62)
$(\frac{1}{3}, \frac{2}{3})$	0.72	**4.28**	0.48	0.62	1.01	0.13	1.76	**5.62**	0.93	0.44	**8.58**	3.37	0.81	2.29	0.63
	(1.44)	(1.09)	(1.29)	(1.53)	(1.82)	(1.71)	(2.17)	(1.94)	(1.74)	(1.96)	(3.14)	(5.13)	(3.81)	(3.78)	(3.72)
$(\frac{2}{3}, 0)$	0.62	5.51	**4.68**	1.49	1.16	1.63	23.18	**3.05**	11.05	3.50	**4.71**	35.69	6.58	23.53	4.71
	(1.59)	(5.05)	(1.70)	(2.06)	(2.12)	(1.47)	(9.64)	(1.35)	(5.12)	(2.57)	(1.49)	(14.14)	(5.93)	(5.44)	(3.32)
$(\frac{2}{3}, \frac{1}{3})$	0.61	3.44	**3.63**	0.90	1.51	1.22	15.13	**2.99**	7.77	3.49	**4.16**	28.47	4.73	15.76	4.74
	(1.57)	(3.63)	(1.44)	(1.75)	(1.77)	(1.63)	(5.00)	(1.46)	(3.25)	(2.21)	(1.73)	(9.88)	(5.57)	(5.21)	(3.21)
$(\frac{2}{3}, \frac{2}{3})$	0.53	1.82	**2.82**	0.90	0.87	0.59	9.09	**2.52**	3.69	1.82	**3.84**	16.24	1.49	8.74	3.05
	(1.45)	(2.77)	(1.33)	(2.09)	(1.51)	(1.31)	(5.87)	(1.16)	(2.83)	(2.04)	(1.76)	(8.20)	(2.80)	(4.68)	(3.04)
(1,0)	1.32	10.20	**3.27**	3.99	2.57	1.87	28.72	**1.30**	17.21	5.24	**0.89**	46.13	10.08	33.05	8.53
	(1.27)	(5.37)	(1.03)	(2.65)	(2.12)	(1.14)	(8.87)	(1.48)	(5.20)	(2.57)	(0.46)	(15.06)	(7.11)	(7.15)	(3.32)
$(1, \frac{1}{3})$	1.35	6.15	**2.38**	2.65	2.37	1.35	21.22	**1.04**	12.77	4.66	**0.45**	32.11	9.87	25.58	7.08
	(1.38)	(4.37)	(0.98)	(2.32)	(1.59)	(1.03)	(9.07)	(1.38)	(4.55)	(2.14)	(0.29)	(12.83)	(5.74)	(5.87)	(2.52)
$(1, \frac{2}{3})$	0.46	4.03	**1.05**	1.29	0.52	0.45	15.17	**0.47**	7.90	1.97	**0.25**	25.46	5.88	14.87	3.11
	(0.58)	(3.72)	(0.55)	(1.00)	(0.76)	(0.61)	(5.31)	(0.63)	(3.60)	(1.28)	(0.26)	(10.09)	(5.16)	(5.80)	(1.69)

Observe that, in general, policies λ and π_M yield the best results with respect to regret and time-stability, while the performance of π_L and π_R is significantly worse. For left-skewed and symmetric structures policy π_M is generally the best one, while performance of λ is not far behind and very close to π_M. For the right-skewed cost structure λ is the best policy, getting a significant difference for regret with respect to all other policies (including π_M). The difference is highly amplified when considering time-stability, as the other policies, except γ, get very close to the worst possible value ($\tau \approx 22$). Moreover, for many instances policies π_M, π_L and π_R fail to identify a set of k-most vital arcs. Recall that unless one assumes that arc costs are at their upper bounds, there is no guarantee that a policy achieves zero instantaneous regret (see Remark 7).

Our experimental observations corroborate our theoretical results: it is crucial for the interdictor to have a pessimistic attitude regarding the unobserved costs, i.e., it is better to overestimate the real costs of the arcs whose real costs are unknown rather than underestimate them. The results also suggest that policy λ is robust in the sense that it has a

Table 8: Average time-stability and MAD (in parenthesis) for $k = 6$. Among entries denoting time-stability, the entry in bold is the best value, and the other entries indicate the difference with respect to the best value. The entries in italic and "−" mean that the policy did not attain time-stability for some instances.

	Left-Skewed					Symmetric					Right-Skewed				
(p_a, p_c)	λ	π_L	π_M	π_R	γ	λ	π_L	π_M	π_R	γ	λ	π_L	π_M	π_R	γ
$(\frac{1}{3},0)$	1.1	3.35	**11.9**	3.25	0.65	0.5	*12.1*	**9.05**	9.1	0.55	**9.75**	*12.25*	7.8	*12.25*	0.75
	(1.60)	(6.83)	(1.73)	(3.68)	(2.55)	(1.31)	(-)	(2.27)	(4.04)	(1.86)	(1.85)	(-)	(6.23)	(-)	(1.10)
$(\frac{1}{3},\frac{1}{3})$	0.55	2.4	**11.7**	1.9	0.5	0.8	9.35	**8.95**	6.3	1.3	**9.35**	*12.65*	5.8	10.1	1.3
	(1.95)	(6.41)	(2.10)	(3.34)	(1.96)	(1.65)	(5.55)	(2.64)	(5.48)	(1.65)	(1.86)	(-)	(6.85)	(3.58)	(1.78)
$(\frac{1}{3},\frac{2}{3})$	1.05	**10.5**	0.5	1.3	1.35	0.3	6.05	**9.05**	4.55	0.5	**9.4**	10.35	4.4	9.25	0.85
	(2.05)	(3.40)	(1.80)	(2.04)	(2.28)	(1.49)	(6.90)	(2.37)	(4.68)	(1.96)	(1.78)	(3.83)	(6.58)	(5.03)	(2.08)
$(\frac{2}{3},0)$	0.8	9.55	**11.4**	6.8	1.15	2.35	*15.25*	**6.75**	*15.25*	2.85	**7.3**	*14.7*	*12.75*	*14.7*	3.2
	(1.72)	(2.00)	(2.20)	(4.36)	(2.55)	(1.23)	(-)	(2.58)	(-)	(1.86)	(1.29)	(-)	(-)	(-)	(1.10)
$(\frac{2}{3},\frac{1}{3})$	1.05	10.9	**9.2**	7	2.5	1.05	*14.9*	**7.1**	*14.2*	2.7	**6.6**	*15.4*	11.65	*15.4*	3.25
	(2.23)	(3.42)	(1.90)	(5.06)	(1.76)	(1.20)	(-)	(3.44)	(-)	(1.72)	(1.26)	(-)	(6.00)	(-)	(1.94)
$(\frac{2}{3},\frac{2}{3})$	0.55	9.8	**7.7**	4	0.9	1.15	*16*	**5**	11.65	2.3	**5.75**	*15.35*	6.65	*15.8*	2.15
	(1.75)	(6.75)	(1.77)	(5.84)	(1.92)	(1.60)	(-)	(1.50)	(6.76)	(1.66)	(1.45)	(-)	(8.64)	(-)	(1.51)
(1,0)	2.1	*12.8*	**9.2**	*11.1*	3.35	**7.8**	*14.2*	1.2	*13.6*	1.8	**4.4**	*17.6*	15.5	*17.6*	6.1
	(1.39)	(-)	(1.70)	(-)	(2.55)	(1.26)	(-)	(9.10)	(-)	(1.86)	(1.30)	(-)	(3.78)	(-)	(1.10)
$(1,\frac{1}{3})$	1.85	*14.7*	**7.3**	11.15	3.35	**6.3**	*15.7*	1.4	*15.7*	2.4	**3**	*17.95*	15.85	*19*	6.05
	(1.35)	(-)	(1.60)	(4.27)	(1.75)	(1.47)	(-)	(8.58)	(-)	(1.97)	(1.10)	(-)	(5.36)	(-)	(1.27)
$(1,\frac{2}{3})$	0.95	*17.25*	**3.7**	11.85	0.75	**3.45**	*18.55*	1.2	17.4	1.15	**1.8**	*20.2*	14.95	*20.2*	2.9
	(1.52)	(-)	(1.44)	(5.90)	(1.64)	(1.23)	(-)	(5.21)	(1.97)	(1.12)	(0.72)	(-)	(7.88)	(-)	(1.33)

consistently good performance across all cost settings, although not always yields the best result. We also note that policy γ has the same type of robust behavior, although it is typically outperformed by λ, which signals that it is valuable to exploit the cost bounds information.

2.5.4 Policy Performance: Sensitivity with Respect to $|\widetilde{A}_0|$

In this section we study the performance of policies in Λ as a function of the number of arcs for which the real cost is initially known (i.e., the amount of initial information). We also consider policy π_M as a benchmark, due to its consistent performance in the experiments in Section 2.5.3.

We set $p_a = 1/2$ and consider $p_c \in \{i/10 \colon 1 \leq i \leq 10\}$. As before, for every pair (p_a, p_c) we generate 20 networks with different cost structures (right-skewed, symmetric and left-skewed). Also, we set $n = 40$, $p = 0.7$ (giving us an average of 1089.8 arcs), $k = 8$ and $T = 28$. Figures 9 and 10 depict the results for the right-skewed and symmetric cost structure, respectively. The results for the left-skewed case are in Figure 18 in Appendix A.3.

For the right-skewed case we observe that λ is roughly constant at low values across all measures, indicating its good and consistent performance. On the other hand, performance of

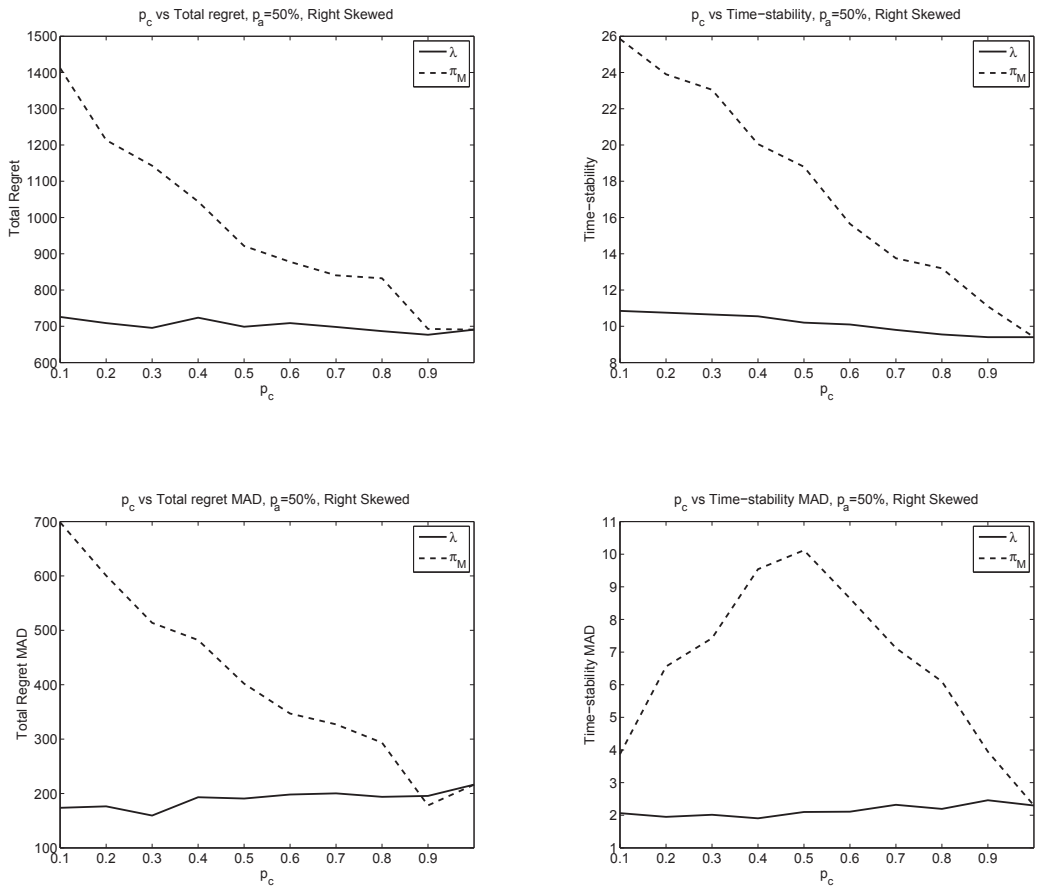


Figure 9: Behavior of the average time-stability, average total regret, time-stability MAD and total regret MAD as p_c increases. The cost distribution is right-skewed and $p_a = 1/2$.

π_M improves as there is more initial information of the network available, but is significantly worse than the one from λ. Also, note that time-stability MAD for π_M has a parabolic behavior, which points out that for low (high) amounts of initial information the time-stability of π_M is consistently high (low), while high variability is observed for intermediate values of initial information.

For the symmetric distribution, both the average regret and time-stability of λ and π_M decrease as more information is available (λ's regret has a subtle increase at one point due to the occurrence of an instance in which it performed significantly bad). In this setting π_M

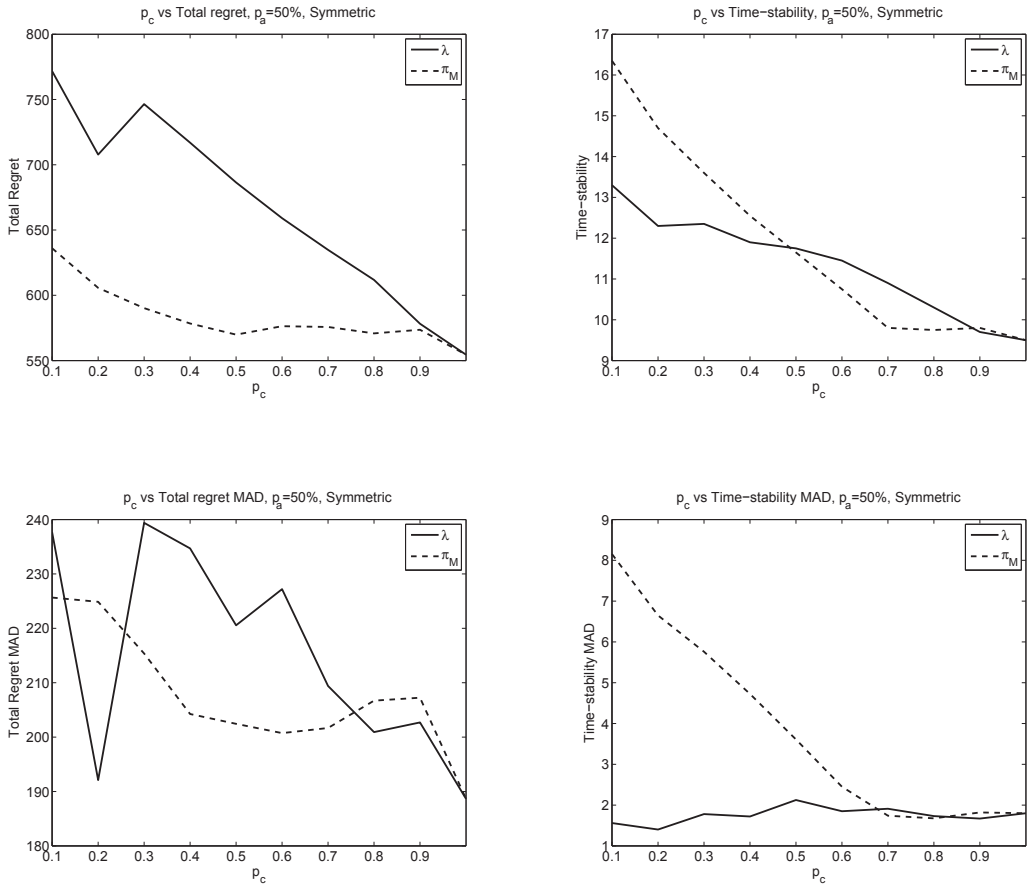


Figure 10: Behavior of the average time-stability, average total regret, time-stability MAD and total regret MAD as p_c increases. The cost distribution is symmetric and $p_a = 1/2$.

outperforms λ in terms of regret, however their difference is not so great when compared to the right-skewed case (observe the scale of the y-axis). Regarding time-stability, it can be seen that π_M is slightly better than λ for low amounts of initial information while the opposite is observed for large amounts. On the other hand, both λ's and π_M's regret MADs do not show any particular pattern (however, it can be considered relatively constant after noting that its scale is significantly smaller than the one in the right-skewed case), while λ's time-stability MAD is constant at low values and that of π_M tends to improve.

These observations confirm our previous conclusions regarding the robustness property of policy λ. They also suggest that different from π_M, performance of λ is highly consistent with respect to changes in the amount of initial information. Moreover, it is observed that the sensitivity of λ to the amount of initial information depends on the location of the true cost between the lower and upper bounds: if it is close to the upper bound (right-skewed), the performance tends to be fairly unsensitive, while for the other cases it tends to improve as more information is initially available.

It is important to note that, unlike for policies in Λ, the regret of the benchmark policies π_L, π_M and π_R may not converge in some instances (recall our Remark 7), thus in the long run (i.e., for sufficiently large values of T, e.g., $T \geq |A|$) the proposed policies might outperform the benchmark policies with certainty. However, such a feature does not rule out that: (i) the benchmark policies may converge for many instances and (ii) the regret of the benchmark policies may be smaller than the regret attained by the convergent policies, particularly in the case of finite horizons, as it can be observed for some instances in Tables 7 and 8 as well as in Figures 10 and 18.

2.5.5 Policy Performance: Sensitivity with Respect to Quality of Bounds in $\widehat{A}_0$

We conclude our numerical experiments by studying the performance of the policies in Λ and π_M as the quality of the initial information deteriorates, i.e., as $u_a - l_a$ increases for all $a \in \widehat{A}_0$. To this end, in this set of experiments we generate a cost vector by drawing c_a uniformly from $U(500, 1000)$ for each $a \in A$. We consider three sets of cost bounds: $(c_a - x_a, c_a + y_a)$, $(c_a - 5x_a, c_a + 5y_a)$ and $(c_a - 25x_a, c_a + 25y_a)$, where x_a and y_a are drawn uniformly from $[1, 20]$ for all $a \in A$. We refer to these three intervals as "I. #1," "I. #2," and "I. #3," respectively.

As in the previous experiments, for each pair (p_a, p_c) we generate 20 random networks along with subsets A_0 and $\widetilde{A}_0$, and for each of them we generate the cost vector $(c_a)_{a \in A}$ and x_a, y_a for all $a \in \widehat{A}_0$. We consider $n = 50$ nodes, $p = 0.5$ (the mean number of arcs is 1216.35.), $k = 15$ and $T = 53$. Table 9 summarizes the obtained results.

Table 9: Average regret ($\times 10^3$) and time-stability, and MAD (in parenthesis) for $k = 15$. The entries in bold denote the best value.

	Mean regret						Mean time-stability					
	Policy λ			Policy π_M			Policy λ			Policy π_M		
(p_a, p_c)	I. #1	I. #2	I. #3	I. #1	I. #2	I. #3	I. #1	I. #2	I. #3	I. #1	I. #2	I. #3
(0,0)	**49.95**	**49.95**	**49.95**	**49.95**	**49.95**	**49.95**	**20.65**	**20.65**	**20.65**	**20.65**	**20.65**	**20.65**
	(8.83)	(8.83)	(8.83)	(8.83)	(8.83)	(8.83)	(3.15)	(3.15)	(3.15)	(3.15)	(3.15)	(3.15)
$(\frac{1}{3},0)$	**44.52**	45.31	48.24	44.61	**44.61**	47.37	**19.40**	20.15	22.00	**19.00**	20.45	31.90
	(9.07)	(9.93)	(9.36)	(9.07)	(9.23)	(11.13)	(3.08)	(3.27)	(3.40)	(2.90)	(4.83)	(16.88)
$(\frac{1}{3},\frac{1}{3})$	**44.39**	44.80	46.69	44.71	**44.46**	47.39	**19.20**	19.80	20.70	**20.40**	**20.40**	31.75
	(8.96)	(9.46)	(9.67)	(9.16)	(9.26)	(11.64)	(3.26)	(3.18)	(3.24)	(4.52)	(4.86)	(17.00)
$(\frac{1}{3},\frac{2}{3})$	**44.58**	44.42	44.69	44.84	**44.70**	45.01	**19.20**	19.40	19.30	19.25	**19.05**	22.30
	(9.19)	(8.88)	(9.00)	(9.25)	(9.10)	(9.82)	(3.26)	(2.88)	(3.29)	(3.20)	(2.86)	(7.68)
$(\frac{1}{3},1)$	**44.74**	**44.74**	**44.74**	**44.74**	**44.74**	**44.74**	**19.15**	**19.15**	**19.15**	**19.15**	**19.15**	**19.15**
	(9.14)	(9.14)	(9.14)	(9.14)	(9.14)	(9.14)	(3.20)	(3.20)	(3.20)	(3.20)	(3.20)	(3.20)
$(\frac{2}{3},0)$	**25.99**	27.94	39.32	**25.49**	25.79	38.34	**12.55**	15.30	20.00	**13.40**	**13.40**	38.60
	(8.23)	(8.60)	(7.55)	(8.25)	(8.22)	(16.29)	(2.85)	(3.13)	(3.00)	(4.76)	(4.70)	(18.72)
$(\frac{2}{3},\frac{1}{3})$	**25.74**	26.58	34.02	**25.42**	26.97	37.29	**12.15**	13.30	16.80	**13.30**	17.60	36.30
	(8.65)	(7.99)	(8.35)	(8.19)	(9.53)	(13.69)	(2.95)	(2.84)	(2.60)	(4.62)	(10.76)	(20.04)
$(\frac{2}{3},\frac{2}{3})$	**25.67**	26.14	30.25	**25.35**	25.38	29.04	**11.75**	12.25	14.15	**11.15**	13.20	25.60
	(8.11)	(8.78)	(8.63)	(8.34)	(8.51)	(9.58)	(2.50)	(2.65)	(2.62)	(2.45)	(4.60)	(19.18)
$(\frac{2}{3},1)$	**25.31**	**25.31**	**25.31**	**25.31**	**25.31**	**25.31**	**11.00**	**11.00**	**11.00**	**11.00**	**11.00**	**11.00**
	(8.36)	(8.36)	(8.36)	(8.36)	(8.36)	(8.36)	(2.30)	(2.30)	(2.30)	(2.30)	(2.30)	(2.30)
(1,0)	**0.20**	3.78	25.61	**0.19**	1.58	32.84	**3.15**	7.60	17.05	**6.35**	12.10	43.35
	(0.17)	(1.59)	(3.69)	(0.33)	(2.25)	(21.15)	(1.47)	(2.04)	(2.07)	(9.33)	(16.36)	(15.44)
$(1,\frac{1}{3})$	**0.07**	1.91	17.90	**0.21**	2.34	23.60	**2.25**	5.15	12.65	**9.00**	19.45	38.05
	(0.06)	(1.00)	(3.53)	(0.35)	(3.17)	(19.03)	(1.10)	(1.90)	(2.22)	(13.20)	(23.49)	(20.93)
$(1,\frac{2}{3})$	**0.05**	0.93	8.66	**0.08**	0.63	8.33	**1.75**	3.60	7.05	**6.35**	11.65	22.40
	(0.05)	(0.63)	(3.44)	(0.14)	(0.97)	(10.91)	(0.68)	(1.44)	(1.85)	(9.33)	(16.54)	(24.48)

We observe that both λ and π_M are sensitive to changes in the quality of the information: as the intervals widen, performance deteriorates. Note that this effect is significantly more pronounced if the interdictor has more initial information available, i.e., for larger values of p_a and p_c. Policy π_M tends to be better than λ in regret for narrow cost intervals, i.e., in the cases with *good* information quality. However, λ is better (in particular, with respect to MAD values) as the intervals widen and the values of p_a increase. Similarly, π_M and λ are roughly similar for time-stability for narrow cost intervals, but λ is significantly better as these intervals widen and p_a increases. These results point out again at the *robust* behavior of the policies in Λ, i.e., they have a good performance across all instances, although they are not always the best.

In order to further validate the aforementioned conclusions, we design a similar experiment with more quality levels. Specifically, we pick $p_a = 2/3$, $p_c \in \{0, 1/3, 2/3\}$ and generate the cost vector by drawing c_a from $U(100, 200)$. We consider 10 sets of costs bounds $(c_a - mx_a, c_a + my_a)$, where m is referred to as the interval-width multiplier. We change the value of m from 1 to 10, while x_a and y_a are drawn uniformly from $[1, 10]$.

For each pair (p_a, p_c) we generate 30 different networks along with A_0, $\widetilde{A}_0$, the cost vector $(c_a)_{a\in A}$ and x_a, y_a for all $a \in \widehat{A}_0$. We set $n = 40$, $p = 0.8$ (the mean number of arcs of 1248.73), $k = 10$ and $T = 36$. Figure 11 depicts the results for the case $p_c = 1/3$. Additional results for other values of p_a and p_c are available in Appendix A.3.

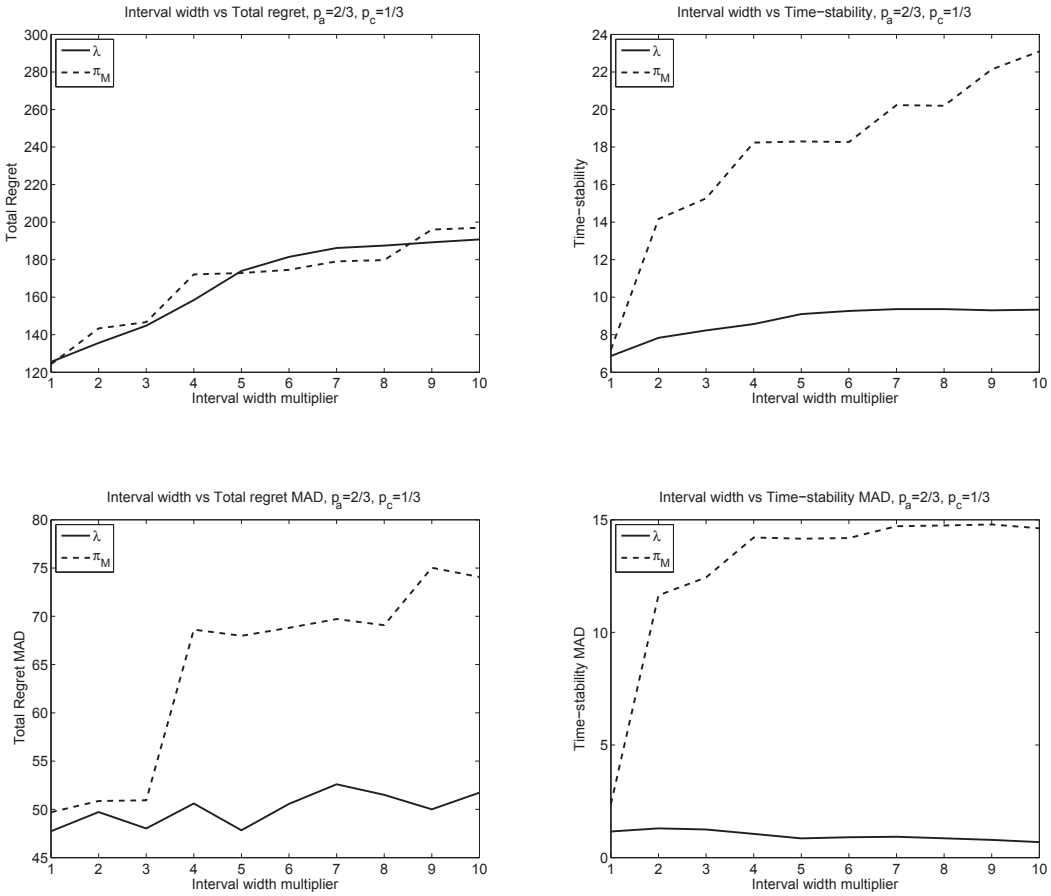


Figure 11: Behavior of the average time-stability, average total regret, time-stability MAD and total regret MAD as the cost intervals widen for the case of $p_a = 2/3$ and $p_c = 1/3$. Given the interval-width multiplier m, the lower and upper bounds of the arc costs in $\widehat{A}_0$ are $l_a = c_a - mx_a$ and $u_a = c_a + my_a$, respectively.

The results in Figure 11 illustrate in greater detail the performance of λ and π_M as the quality of cost information worsens. In particular, one observes that, with respect to the total regret, the performance of both policies degrades at a similar rate. This conclusion is

not true, however, when considering the total regret MAD as it tends to increase for π_M, while remaining relatively constant for λ. With respect to the time-stability metric, it is observed that although the performance of λ deteriorates, it does so at a much slower rate than π_M, and that its time-stability MAD remains virtually constant as the intervals widens. These results reinforce the conclusions of our previous experiments regarding robustness of the proposed policies. Moreover, they show that λ can be considered somewhat insensitive to the quality of the initial information, while the performance of π_M can significantly degrade if the quality of the initial information is not sufficiently good.

2.6 CONCLUDING REMARKS

In this chapter we study sequential interdiction of a directed network when the interdictor has incomplete initial information about the network. By observing the evader's actions (who travels along shortest paths in each time period), the interdictor learns about the structure and costs of the network and adjusts its actions so as to maximize the cumulative cost incurred by the evader. We formally define the concept of efficient interdiction policies and propose a class of simple interdiction policies that are efficient both with respect to regret and time-stability.

Our theoretical results are supported by numerical experiments which suggest that the proposed policies are robust, in the sense that they yield good results across various levels of the initial information. Aligned with intuition, our interdiction policies get better results as the quality of the information improves. One important conclusion of our work is that it is crucial for the interdictor to have a pessimistic attitude regarding unobserved arc costs, i.e., it is better to overestimate the real costs of the arcs whose real costs are unknown rather than underestimate them. Otherwise, as we demonstrate both theoretically and computationally, the interdictor is not guaranteed to converge to an optimal k-most vital arc solution, i.e., an interdiction solution with an instantaneous regret of zero. Finally, we propose a semi-oracle benchmark policy that serves as a lower bound on the performance of any feasible interdiction policy. We formulate it as an MIP and describe an algorithmic approach for its computation.

Our work considers *myopic* evaders (recall assumption **A2**), who always traverse along shortest paths of the interdicted network. Consider now a *strategic* evader who does not necessarily travel via shortest paths in each time period, but rather desires to minimize the total costs of moving through the network over time horizon $\mathcal{T}$, i.e., $\sum_{t\in\mathcal{T}} \ell(P_t)$, and suppose that $\widehat{A}_0 = \emptyset$ and that the interdictor uses policies in Γ (similar arguments apply for the more general case). The same type of analysis can be applied to the setting in which the evader does not observe the actions of the interdictor upfront, and needs to learn them in real time.

Using an approach similar to the one used in the proof of Lemma 2, it can be shown that if the evader desires to incur a cost less than $z(G[A_t^\gamma \setminus I_t^\gamma])$ at time period t, then at least one new arc must be revealed to the interdictor in P_t^γ. Thus, we conclude that Lemma 2 holds for any reasonable decision-making process of the evader, i.e., the evader cannot avoid the convergence of policies in Γ.

Therefore, our initial assumption that the evader moves through shortest paths in $G[A\setminus I_t^\gamma]$ turns out to be not too restrictive (at least for the property **C1** to hold for the proposed policies). Moreover, for many instances, this myopic approach might yield the best performance. However, in general there may be a trade-off for the evader between using shortest paths and using alternative paths that, although are not the shortest ones, might improve the performance over the whole time horizon. While the evader's decision-making problem can be casted as sequential mixed-integer bilevel program, as such, it might be intractable in practice. Nevertheless, it presents an interesting avenue for future research.

There are several other research directions that remain open at this point. An immediate one relates to whether our results can be extended or serve as the basis for studying settings, where we relax assumptions **A1** and **(A3)**. In particular, relaxing the former one results in decision-making problems with alternative feedbacks, where, for example, only a noisy signal of the evader cost is revealed in each time period, or only some components of the evader action are revealed.

Note that if the source and destination nodes are not known initially to the interdictor, then they would be inferred immediately from the evader's actions due to our assumptions of the perfect feedback. Thus, our model also accommodates the setting where the interdictor is not initially aware of such information. Similarly, our model would directly extend to

settings, where source and destination nodes are chosen from a given set. In this regard, an interesting extension would be cases, where the source and destination nodes are chosen randomly in each time period according to some distribution. In particular, we believe that our methods are not trivially extendable to problems, where the source and destination nodes location distribution is initially unknown (observe that the interdictor would not be able to state the objective function upfront under this scenario).

We should also note that our methods can be extended to the setting, where each blocking action does not fully eliminate an arc but rather increases its cost. In such a case, the interdictor might have a budget that must be allocated across arcs. Consequently, in each time period our policies would solve an instance of the *Maximizing the Shortest Path* (MXSP) problem from Israeli and Wood (2002).

Finally, with regard to more general interdiction models in the literature, it is possible to device computational procedures similar to the one proposed in this work. In the interdiction literature, such setups might include evaders that aim to maximize the flow of illegal materials (or desire to minimize their transportation costs) in the network with capacity constraints on its arcs. Ultimately, the suitability of our approach would rest on the nature of each specific interdiction model (e.g., whether it admits a tractable solution approach in the case of complete information), and, more importantly, the type of the feedback obtained by the interdictor. Generalizing our approach for generic bilevel interdiction models is outside the scope of this work, and constitutes an interesting line for future research.

3.0 SEQUENTIAL MAX-MIN BILEVEL LINEAR PROGRAMMING WITH INCOMPLETE INFORMATION AND LEARNING

3.1 INTRODUCTION

An important class of bilevel programs, known as *max-min* problems, deals with settings where the leader and follower are adversaries and 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 (Fulkerson and Harding 1977, Corley and Chang 1974, Israeli and Wood 2002, Wollmer 1964, McMasters and Mustin 1970, Ghare et al. 1971, Corley and Chang 1974, Ratliff et al. 1975, Wood 1993, 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. Depending on the specific application, the objective of the leader 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. 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 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. (2006), Zenklusen (2010), Shen et al. (2012*b*), 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., Hemmecke et al. (2003), Held et al. (2005), Held and Woodruff (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 chapter 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 their 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.

In SMPIs 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 3.4 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*, i.e., the first time period by which the costs the follower incurs coincides with the

maximum possible cost an *oracle* leader with complete knowledge of the bilevel problem attains. Time-stability is closely related to the *regret* (in particular, any upper bound on the time-stability of a policy implies an upper bound in the regret on that policy), a more common measure of performance in online optimization settings (Bubeck and Cesa-Bianchi 2012, Hazan 2015).

In this chapter 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 leader. For these reasons, implementing the policies in Λ involve solving at each time a max-min bilevel problem with lower-level robustness constraints, and as such their computation involves 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 the 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 cannot signal it through her actions. As such, the semi-oracle combines the knowledge of the standard oracle with the practical limitations of the leader. Our numerical results show that the policies in Λ consistently outperform reasonable benchmark, and perform reasonably close to the semi-oracle.

The remainder of the chapter is organized as follows. In Section 3.2 we provide a mathematical formulation of the problem, and illustrate it with examples of the minimum-cost flow interdiction and the Attacker-Defender knapsack problems. Section 3.3 discusses greedy and robust policies along with their main properties, while Section 3.4 extends most of the results of greedy and robust policies for the case of uncertainty in the lower-level constraint

matrix. Section 3.5 discusses the semi-oracle benchmark and Section 3.6 presents numerical experiments. In Section 3.7 we give conclusions and directions for future work. Most proofs and supporting material are relegated to Appendix B.

3.2 BASIC MODEL: COST UNCERTAINTY

We consider a sequential and adversarial decision-making process where at each time $t \in \mathcal{T} \coloneqq \{0, 1, \cdots, T\}$ a *leader* and a *follower* interact. At the beginning of time $t \in \mathcal{T}$, in the *complete information setting*, 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 $c_a y_a$, 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, at time t the follower selects vector $y(x)$, where for any $x \in X$

$$y(x) \in \arg\min\{\boldsymbol{c}^\top y \ : \ y \in Y(x)\},$$

$\boldsymbol{c} \coloneqq (c_a \colon a \in A) \in \mathbb{R}^{|A|}$, and where for any $x \in \mathbb{Z}_+^k \times \mathbb{R}_+^{|I|-k}$ the set $Y(x)$ denotes the follower's set of feasible actions given the leader decision x. We assume that

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

In the above, $\boldsymbol{F} \coloneqq (F_{da} \colon d \in C_F, a \in A)$ 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 $\boldsymbol{f} \coloneqq (f_d, d \in C_F) \in \mathbb{R}^{|C_F|}$. The objective of the leader is to choose $x \in X$ so as to maximize the cumulative cost the follower faces through $\mathcal{T}$. Note that, *had the leader full information about the problem*, at each time $t \in \mathcal{T}$ she would implement a solution to the bilevel problem

$$z^* \coloneqq \max\{z(x) \ : \ x \in X\}, \tag{3.1}$$

where for any $x \in X$,

$$z(x) \coloneqq \min\{\boldsymbol{c}^\top y \ : \ y \in Y(x)\}.$$

Throughout this chapter, we assume 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 $Y(x)$. Moreover, 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 the subset of the follower's activities $A^t \subseteq A$, the subset of the leader's resources $I^t \subseteq I$, the upper-level constraints $C_L^t \subseteq C_L$ and the lower-level constraints $C_F^t \subseteq C_F$. 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 $\boldsymbol{F}^t \coloneqq (F_{da} \colon a \in C_F^t, a \in A^t)$ and $\boldsymbol{f}^t \coloneqq (f_d \colon d \in C_F^t)$. 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 $\boldsymbol{H}^t \coloneqq (H_{di} \colon d \in C_L^t, i \in I^t)$, $\boldsymbol{h}^t \coloneqq (h_d \colon d \in C_L^t)$ and $\boldsymbol{L}^t \coloneqq (L_{di} \colon d \in C_F^t, i \in I^t)$.

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

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

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 (3.1) 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 3.4.

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 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.

Given the framework above, at any given time $t \in \mathcal{T}$ the following sequence of events takes place:

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

$$X^t \coloneqq \{x \in \mathbb{R}_+^{|I^t|} \ : \ \boldsymbol{H}^t x \leq \boldsymbol{h}^t, \ x \geq 0\}. \tag{3.2}$$

2. The follower solves the linear program $z(\bar{x}^t)$, where $\bar{x}^t$ is defined as $\bar{x}_i^t := x_i^t$ if $i \in I^t$, and $\bar{x}_i^t := 0$ if $i \in I \setminus I^t$. That is, he solves

$$z(\bar{x}^t) = \min_{y \geq 0} \boldsymbol{c}^\top y \tag{3.3}$$
$$\text{s.t. } \boldsymbol{F}y + \sum_{i \in I^t} \boldsymbol{L}_i x_i^t \leq \boldsymbol{f},$$

where $\boldsymbol{L}_i$ is the i-th column of L. For notational convenience, we set $y^t := y(\bar{x}^t)$ and $z^t := z(\bar{x}^t)$.

3. The response of the follower generates *feedback* $\mathcal{F}^t$. The leader observes the information in $\mathcal{F}^t$ and exploits it to update her current knowledge to I^{t+1}, C_L^{t+1}, A^{t+1}, C_F^{t+1} and $\mathcal{U}^{t+1}$ (thus, potentially updating $\boldsymbol{H}^{t+1}$, $\boldsymbol{h}^{t+1}$, $\boldsymbol{F}^{t+1}$, $\boldsymbol{L}^{t+1}$ and $\boldsymbol{f}^{t+1}$ as well as c_a for any new activity learned).

The next section elaborates on the information update in $\mathcal{F}^t$. Before that, we illustrate the assumptions above and the flexibility of the framework through the following examples.

Example 1. Consider a smuggling interdiction problem where a smuggler (the follower) operates over a directed network $G = (V, E)$ and is required to satisfy the demand for illegal goods across different locations. At each time period, the smuggler moves goods from supply vertices $V_S \subseteq V$ to demand vertices $V_D \subseteq V$. Some of the vertices are temporary depots (i.e., transshipment vertices) and we denote them by V_N. Denote by $b(v)$ the amount of goods that vertex v supplies/demands, where $b(v) > 0$ for $v \in V_S$ and $b(v) < 0$ for $v \in V_D$. We assume that $b(v) = 0$ for vertices in V_N, and that $\sum_{v \in V_S} b(v) = \sum_{v \in V_D} b(v)$.

For any $e = (v, w) \in E$, it costs the smuggler c_e to ship one unit of the illegal good from vertex v to vertex w through link e, and due to the transportation limitations (e.g., the fleet or infrastructure size) he can move at most u_e units from v to w at any given time. The smuggler's objective is to ship the goods across the network at each time period in order to minimize the shipment costs, subject to the requirement of supplying all demand.

Consider the follower's minimum cost flow problem over G. Let $\boldsymbol{M}$ be the node-arc adjacency matrix of G, so $\boldsymbol{M}$ is a $|V| \times |E|$ matrix, where for any $v \in V$, $M_{ve} = 1$ for all $e \in E$ such that $e = (v, w)$ for some $w \in V$, and $M_{ve} = -1$ for all $e \in E$ such that $e = (w, v)$

for some $w \in V$. Let $\boldsymbol{b}$ be the vector given by $b_v = b(v)$ for all $v \in V$, $\boldsymbol{c}$ and $\boldsymbol{u}$ be the cost and upper-bound vectors, respectively. For any $e \in E$ define y_e as the amount of goods the smuggler ships through edge e. Then, *without the leader's intervention* the follower would solve the min-cost flow problem of the form:

$$y^* \in \arg\min_y \{\boldsymbol{c}^\top y \,:\, \boldsymbol{M}y \leq \boldsymbol{b}, -\boldsymbol{M}y \leq -\boldsymbol{b}, \boldsymbol{I}y \leq \boldsymbol{u}, y \in \mathbb{R}_+^{|E|}\},$$

where $\boldsymbol{I}$ is a $|E| \times |E|$ identity matrix. Observe that y^* above can be thought as always taking only integer values (as long as $\boldsymbol{u}$ and $\boldsymbol{b}$ are integers) as the constraint matrix is totally unimodular, see Wolsey and Nemhauser (2014).

Law enforcement, on the other hand, acts as the leader. She assigns patrolling and interdicting vehicles to links in G. We assume that there are K types of vehicles capable of interdicting any edge. The leader controls r_k units of type k vehicles, each of which reduces the shipment capacity of arc e by d_{ke} units when assigned to the said arc. For any $k \in K$ and $e \in E$, define x_{ke} as the number of vehicles type k the leader sends to edge e. We assume that the values of x should satisfy the constraints $\sum_{e \in E} x_{ke} \leq r_k$ for all $k \in K$, $\sum_{k \in K} d_{ke} x_{ke} \leq u_e$ for all $e \in E$, and $x_{ke} \in \mathbb{Z}_+$ for all $k \in K$ and $e \in E$. Observe that this problem can be viewed as a generalization of the typical minimum cost flow interdiction problem (Chern and Lin 1995, Smith and Lim 2008).

We can model the setting above within our framework as follows. The set of the follower's activities corresponds to E (i.e., $A = E$). For each vertex there are two flow constraints and for each edge of E there is an upper bounding constraint. Thus, $|C_F| = 2|V| + |E|$. Matrix $\boldsymbol{F}$ is given by $\boldsymbol{F} = (\boldsymbol{M}; -\boldsymbol{M}; \boldsymbol{I})$, the right-hand side vector is $\boldsymbol{f} = (\boldsymbol{b}; -\boldsymbol{b}; \boldsymbol{u})$, and the cost vector $\boldsymbol{c}$ is precisely the cost vector of the network. On the other hand, we model the set of the leader's resources by $I = K \times E$, where each leader resource is represented by a vehicle type and an edge. Note that there is constraint associated with each vehicle type and each edge in the leader's problem, hence $|C_L| = |K| + |E|$. Henceforth, if the leader has all the information of the network, we have $\boldsymbol{H} = (\boldsymbol{O}\ \boldsymbol{D})$, where $\boldsymbol{O}$ is the $|K| \times |K||E|$ matrix given by $O_{k,(k,e)} = 1$ for all $e \in E$, and zero otherwise; and $\boldsymbol{D}$ is the $|E| \times |K||E|$ matrix defined by $D_{e,(k,e)} = d_{ke}$ for all $k \in K$, and zero otherwise. Vector $\boldsymbol{h}$, on the other hand, is given by $\boldsymbol{h} = (\boldsymbol{r}; \boldsymbol{u})$, where the vector $\boldsymbol{r}$ is defined by $\boldsymbol{r} = (r_k \colon k \in K)$. Finally, observe that by

the definition of the interdiction activities, matrix $\boldsymbol{L}$ is given by $\boldsymbol{L} = (\mathbf{0}; \boldsymbol{D})$, where $\mathbf{0}$ is a matrix of zeros of size $2|V| \times |K||E|$ as we assume that the leader cannot interdict nodes.

Assume next that at time $t = 0$ the leader does not know all the edges nor all the nodes in G. For each vertex she observes, she knows whether it is supply or demand node, and knows with certainty the value of $b(v)$, while for each edge $e \in A^0$ she knows with certainty its shipment capacity u_e, however she does not know its shipment cost c_e. For each $e \in A^0$ she estimates the cost to be in the interval $[\ell_e, m_e]$, $\ell_e \leq m_e$, and hence $\mathcal{U}^0 = \{\hat{\boldsymbol{c}}^0 \in \mathbb{R}^{|A^0|} \colon \ell_e \leq \hat{c}^0_e \leq m_e \ \forall e \in A^0\}$, so $\boldsymbol{G}^0 = [\boldsymbol{I}; -\boldsymbol{I}]$ and $\boldsymbol{g}^0 = (\boldsymbol{m}; \boldsymbol{\ell})$, with $\boldsymbol{m} = (m_e \colon e \in A^0)$ and $\boldsymbol{\ell} = (\ell_e \colon e \in A^0)$. ■

Example 2. 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 $\ell_e \leq c_e \leq m_e$ for the profits, which implies that $\mathcal{U}^0 = \{\hat{\boldsymbol{c}}^0 \in \mathbb{R}^{|A^0|} \colon \ell_e \leq \hat{c}^0_e \leq m_e \ \forall e \in A^0\}$. Thus, $\boldsymbol{G}^0 = [\boldsymbol{I}; -\boldsymbol{I}]$ and $\boldsymbol{g}^0 = (\boldsymbol{m}; \boldsymbol{\ell})$, with $\boldsymbol{m} = (m_e \colon e \in A^0)$ and $\boldsymbol{\ell} = (\ell_e \colon e \in A^0)$. ■

Example 3. See Section B.3.3 of the Appendix for an example in assignment interdiction. ■

3.2.1 Feedback

Depending on the particular application, the feedback $\mathcal{F} := (\mathcal{F}^t, t \in \mathcal{T})$ 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 2. Let time $t \in \mathcal{T}$ be given and consider the bilevel problem (3.1).

- We say that the follower *performs* activity $a \in A$ (leader *uses* resource $i \in I$) at time t if and only if $y^t_a > 0$ ($x^t_i > 0$).
- 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 exists 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 have an effect in his cost or constraints, and 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 this example, the follower performs activity $e \in A$ as long as he ships goods through edge e. Similarly, the leader uses resource $(k, e) \in I$ as long as she sends a vehicle type k to interdict edge e. Associated with activity (edge) $e = (v, w) \in A$ there are five constraints in $C_F(e)$. The first four constraints correspond to the supply/demand restrictions at v and w, while the additional constraint corresponds to the maximum shipment capacity constraint of edge e. Additionally, for any resource $(k, e) \in I$ we have that $C_L(k, e)$ consists of two constraints. One of them restricts the amount of vehicles type k that can be used across all edges (i.e., $\sum_{e \in E} x_{ke} \leq r_k$), and the other one corresponds to the maximum interdiction allowed across all vehicle types on edge e (i.e., $\sum_{k \in K} d_{ke} x_{ke} \leq u_e$). Finally, for each edge $e \in A$, we have $I(e) = \{(1, e), (2, e), \cdots, (|K|, e)\}$, that is, $I(e)$ consists of $|K|$ leader resources, one for each type of vehicle that can interdict edge $e \in A$. ■

Example 2 (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 3. 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 and that the above conditions also hold for the initial information known by the leader before any interaction takes place (see also Section 3.2.2). Therefore, at any given time $t \in \mathcal{T}$ the matrices $\boldsymbol{F}$, $\boldsymbol{L}$ and $\boldsymbol{H}$ can be partitioned in submatrices as follows:

$$\boldsymbol{F} = \begin{array}{c} \\ C_F^t \\ C_F \setminus C_F^t \end{array} \begin{array}{c} \begin{array}{cc} A^t & A \setminus A^t \end{array} \\ \begin{pmatrix} \boldsymbol{F}_1 & \boldsymbol{F}_2 \\ \boldsymbol{0} & \boldsymbol{F}_3 \end{pmatrix} \end{array}, \quad \boldsymbol{L} = \begin{array}{c} \\ C_F^t \\ C_F \setminus C_F^t \end{array} \begin{array}{c} \begin{array}{cc} I^t & I \setminus I^t \end{array} \\ \begin{pmatrix} \boldsymbol{L}_1 & \boldsymbol{0} \\ \boldsymbol{L}_2 & \boldsymbol{L}_3 \end{pmatrix} \end{array}, \tag{3.4a}$$

$$\boldsymbol{H} = \begin{array}{c} \\ C_L^t \\ C_L \setminus C_L^t \end{array} \begin{array}{c} \begin{array}{cc} I^t & I \setminus I^t \end{array} \\ \begin{pmatrix} \boldsymbol{H}_1 & \boldsymbol{H}_2 \\ \boldsymbol{0} & \boldsymbol{H}_3 \end{pmatrix} \end{array}, \tag{3.4b}$$

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*a*) 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 (3.1). 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). Consider standard feedback in the smuggling example, which implies that the leader observes the total cost incurred by the smuggler at each period. In addition, if the smuggler ships goods through an edge $e = (v, w) \in A$ that the leader was not aware of, then the leader learns about the existence of that edge. Moreover, as she learns $C_F(e)$, she becomes aware about the existence of vertices v and w as well as the value of the supply/demand b_v and b_w. She also learns the maximum shipment capacity u_e of edge e.

On the other hand, the leader also observes $I(e)$. Thus, she can now send vehicles to interdict edge e. Consequently, she learns about all the vehicle resources/capacity constraints in $C_L(k, e)$ for all $(k, e) \in I(e)$; and as she learns L_{di} for all $d \in C_F(a)$ and all $i \in I(a)$, then she also learns about the value of the capacity restrictions d_{ke} for all vehicle types $k \in K$. ■

Example 2 (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 to guarantee that the leader learns the follower's data in $\boldsymbol{c}$ or his response y^t with more accuracy. In this chapter we consider the following two cases.

Definition 4. Let $\mathcal{F}$ be standard. We say $\mathcal{F}$ is:

- *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 $y_a^t > 0$.
- *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 smuggling setting, Value–Perfect feedback means that if at period $t \in \mathcal{T}$ the smuggler ships goods through edge e, then the leader learns c_e, the cost of shipping one unit of the illegal good through e. On the other hand, Response–Perfect feedback means that the leader observes y_e^t, the amount of goods shipped by the smuggler through e at time $t \in \mathcal{T}$. ■

Example 2 (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. ■

3.2.2 Optimality Criteria

In this section we define what constitutes a 'good' decision-making policy for the leader. In contrast with most work in online optimization, we measure the performance of a policy in terms of its *time-stability* rather than of its regret. The time-stability of a policy corresponds to the first time period by which the actions prescribed by the policy coincide with the actions of an oracle decision-maker. This implies that from the this time period onwards, the policy

prescribes the same decision that the oracle prescribes. Recall that the oracle has all the information about the problem and thus, implements the best possible decision starting at time $t = 0$. As it will be seen below, any upper bound on time-stability implies an upper bound on regret.

To formally introduce time-stability and the concept of optimality, 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

$$\mathcal{D}^0 \coloneqq (A^0, I^0, C_F^0, C_L^0, \mathcal{U}^0, \boldsymbol{H}^0, \boldsymbol{h}^0, \boldsymbol{F}^0, \boldsymbol{L}^0, \boldsymbol{f}^0).$$

Note that given some initial information $\mathcal{D}^0$, there might be several different bilevel problems of the form (3.1) 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 $\mathcal{D}^0$:

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

where

C1: $A^0 \subseteq A$, $I^0 \subseteq I$, $C_F^0 \subseteq C_F$, $C_L^0 \subseteq C_L$.
C2: $I^0 = \cup_{a \in A^0} I(a)$, $C_L^0 = \cup_{i \in I^0} C_L(i)$, $C_F^0 = \cup_{a \in A^0} C_F(a)$.
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 $\boldsymbol{H}$, $\boldsymbol{h}$, $\boldsymbol{F}$, $\boldsymbol{L}$, $\boldsymbol{f}$.

Note that conditions **C2**-**C3** state that the information that the leader initially knows satisfies the standard feedback conditions at time $t = 0$. 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:

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

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 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.,

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

There is a clear connection between time-stability and regret. Indeed, the *regret* $R^\pi_{T_0}(\mathcal{D}^0, \mathcal{D})$ of policy π on the pair $(\mathcal{D}^0, \mathcal{D})$ until time $T_0 \geq 0$ is defined as

$$R^\pi_{T_0}(\mathcal{D}^0, \mathcal{D}) := \sum_{0 \leq t \leq T_0} (z^* - c^\top y^{t,\pi}).$$

If U is an upper bound on the value of $(z^* - c^\top y^{t,\pi})$ for any $t \in \mathcal{T}$, then it immediately follows that

$$R^\pi_{T_0}(\mathcal{D}^0, \mathcal{D}) \leq U \cdot \tau^\pi(\mathcal{D}^0, \mathcal{D}),$$

for any $T_0 \leq T$ as long as the time-stability is finite. Consequently, any finite upper bound on the time-stability provides a finite upper bound on the regret.

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 an example in our general setting.

Henceforth, we study an alternative optimality notion referred to as *weak optimality*. Roughly speaking, π is weakly better than π' if the worst-case time-stability of π across all possible instances is at most the worst-case time-stability of π' across all possible instances, that is, if:

$$\sup_{(\mathcal{D}^0, \mathcal{D}) \in \mathbb{G}} \tau^\pi(\mathcal{D}^0, \mathcal{D}) \leq \sup_{(\mathcal{D}^0, \mathcal{D}) \in \mathbb{G}} \tau^{\pi'}(\mathcal{D}^0, \mathcal{D}). \tag{3.5}$$

A policy π would be weakly optimal if it is weakly better than any other policy. It turns out, however, that the above definition is not meaningful as the suprema in (3.5) are infinity. Certainly, for any policy it can be readily checked that there are instances where the time-stability increases linearly with $|A|$, see e.g., Proposition 5 in Section 3.3.2.

In order to address this issue, we take the suprema in equation (3.5) over instances of a fixed size, which we assume is given in terms of the follower's problem. Specifically, we 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 $n \geq n^0$):

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

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

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

It should be clear that the notion of weak optimality is an adaptation of the notion on 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).

3.3 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.

3.3.1 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

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

Observe that, in contrast to $Y(x)$, the leader completely knows $Y^t(x)$ at time t. More importantly, $Y^t(x)$ can be considered as the "best guess" the leader makes about the follower's feasible region, given that she decides x. For any $x \in X^t$, define $z_R^t(x)$ as the value of the robust linear program

$$z_R^t(x) \coloneqq \min_y \Big\{ \max_{\hat{\boldsymbol{c}}^t \in \mathcal{U}^t} \{ (\hat{\boldsymbol{c}})^\top y \} \ : \ y \in Y^t(x) \Big\}.$$

Note that $z_R^t(x)$ is the cost that the leader expects the follower would incur if she chooses vector x and if the follower's worst-case scenario over $\mathcal{U}^t$ is realized. 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,

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

Finally, for any policy π define $\xi^\pi := \xi^\pi(\mathcal{D}^0, \mathcal{D})$ as,

$$\xi^\pi := \min\{t \in \mathcal{T} \ : \ z_R^{t,*} = z^{t,\pi}\}.$$

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

$$x^{t,\lambda} \in \arg\max\{z_R^t(x) \ : \ x \in X^t\} \qquad \forall t \leq \xi^\lambda, \tag{3.6}$$

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

It is important to note that policies in Λ can be computed by standard mixed integer programming (MIP) solvers as robust bilevel problem (3.6) can be reduced to a single-level MIP, see Appendix B.3.2 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 that the follower faces at t ($z^{t,\lambda}$), and the cost the leader expects that the follower incurs ($z_R^{t,*}$). In addition, it reveals the importance that time period ξ^λ has for time-stability.

Theorem 3. *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 $\tau^\lambda \leq \xi^\lambda$.*

Theorem 3 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 3 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 in general by only assuming standard feedback. However, if the stronger notions of either Value–Perfect or Response–Perfect feedback are assumed, it is shown in the following sections that the leader must learn new information of the follower's problem.

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. The condition is given in terms of the polyhedral dimension $\dim(\mathcal{U}^t)$ of $\mathcal{U}^t$, which is the maximum number of affine independent points within $\mathcal{U}^t$. In particular, if $\dim(\mathcal{U}^t) = 0$, then it consists only of one point. That is, if $\dim(\mathcal{U}^t) = 0$, then $\mathcal{U}^t = \{\boldsymbol{c}^t\}$.

Proposition 4. *Suppose $t \in \mathcal{T}$ satisfies that $\dim(\mathcal{U}^t) = 0$ and assume that $y_a^t = 0$ for all $a \notin A^t$. Then $\xi^\lambda \leq t$, and, in particular, $\tau^\lambda \leq t$.*

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. We use this result in the following sections to establish upper bounds on ξ^λ (and hence, on τ^λ) under Value–Perfect and Response–Perfect feedbacks.

3.3.2 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 7. *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 $\dim(\mathcal{U}^{t+1}) < \dim(\mathcal{U}^t)$.*

A direct consequence of the above result is that, in conjunction with Proposition 4, it provides an upper bound for the time-stability for any policy in Λ:

Theorem 4. *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 7 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 4 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 exploitation (i.e., deciding greedily) always implies exploration (i.e., discovering new information). We emphasize that the key is robustness, as if the leader uses 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 4 is tight across all instances and, more importantly, across all policies. In other words, we establish that policies in Λ are weakly optimal.

Proposition 5. *Consider $\lambda \in \Lambda$ and suppose that $\mathcal{F}$ is Value–Perfect. Then, for any $\boldsymbol{s} = (n, n^0) \in S$*

$$\max_{(\mathcal{D}^0, \mathcal{D}) \in \mathbb{G}_{\boldsymbol{s}}} \tau^{\lambda}(\mathcal{D}^0, \mathcal{D}) \leq n. \tag{3.7}$$

Moreover, λ is weakly optimal.

Proof. First, observe that equation (3.7) is an immediate consequence of Theorem 4. 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 $\mathcal{U}^0 = \{\hat{\boldsymbol{c}}^0 \in \mathbb{R}^{n^0} \colon \ell \leq \hat{c}^0_j \leq u, \ j = 1, \ldots, n^0\}$, 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 $t = t_0 + 1, \cdots, t_n$,

$$z^{t,\pi} \leq c_{j^{\pi,t}} < z^* \tag{3.8}$$

(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 (3.8) 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. ■

3.3.3 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^t_a for all $a \in A$ such that $y^t_a > 0$. In this setting, the leader should update the uncertainty set $\mathcal{U}^t$ to $\mathcal{U}^{t+1}$ by including the linear equality

$$\sum_{a \in A^{t+1}} y^{t,\lambda}_a \hat{c}_a = z^{t,\lambda}.$$

That is,

$$\mathcal{U}^{t+1} = \Big\{ \hat{\boldsymbol{c}} \in \mathbb{R}^{|A^{t+1}|} \; : \; (\hat{c}_a)_{a \in A^t} \in \mathcal{U}^t, \sum_{a \in A^{t+1}} y^{t,\lambda}_a \hat{c}_a = z^{t,\lambda} \Big\}. \tag{3.9}$$

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 (3.9) 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 $z^{t,\lambda} < z^{t,*}_R$. Specifically, if $y^t_a = 0$ 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 7:

Lemma 8. *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*

$$\dim(\mathcal{U}^{t+1}) < \dim(\mathcal{U}^t).$$

On the other hand, 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 8 immediately provides the following upper bound:

Theorem 5. *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 5. Thus, its proof is omitted.

Proposition 6. *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.

3.4 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., see Remark 9 below. 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 3.2 (which, in view of the current discussion, can be referred to as simply the *cost model*) are also valid. Moreover, we show that for the Value–Perfect feedback case, the time-stability upper bound of Theorem 4 also holds, while for Response–Perfect feedback, an extension of the upper bound in Theorem 5 holds under certain assumptions.

Remark 9. Consider the case where there is uncertainty regarding the cost function c. In this case the problem $\min\{\boldsymbol{c}^\top y\colon y \in Y(x)\}$ can be equivalently posed as $\min\{y_0\colon (y_0, y) \in Y'(x)\}$ where

$$Y'(x) := \{(y_0, y) \in \mathbb{R}^{|A|+1} \ : \ \boldsymbol{F}'(y_0; y) + \boldsymbol{L}'x \leq (0; \boldsymbol{f})\},$$

$$\boldsymbol{F}' := \begin{pmatrix} -1 & \boldsymbol{c}^\top \\ \boldsymbol{0} & \boldsymbol{F} \end{pmatrix}, \qquad \boldsymbol{L}' := \begin{pmatrix} \boldsymbol{0}^\top \\ \boldsymbol{L} \end{pmatrix},$$

and in each case $\boldsymbol{0}$ is a column vector of zeros of appropriate dimensions. Observe that this new formulation has an additional variable but there is no uncertainty regarding any cost coefficient (it is always one for the new variable and zero for the rest).

3.4.1 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

$$n_d^t := |\{a \in A^t\colon d \in C_F(a)\}|.$$

We replace assumption **(A2)** from Section 3.2 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 6. In the context of 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

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

where

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

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

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 3.2. Using them, we extend most of the results in the next sections.

3.4.2 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

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

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 7. We say that $\lambda \in \Lambda_E \subseteq \Pi$ if and only if

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

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.

3.4.2.1 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 3.3.1, (i.e., Theorem 3 and Proposition 4) also hold for Λ_E.

Proposition 7. *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) $\tau^\lambda \leq \xi_E^\lambda$.

(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 7 and Theorem 4 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,

$$\widetilde{A}_E^t := \{a \in A^t \ : \ \forall \hat{\boldsymbol{F}} \in \mathcal{U}^t \ \hat{F}_{da} = F_{da} \ \forall d \in C_F^t\}.$$

Proposition 8. *Suppose $\lambda \in \Lambda_E$ and that feedback $\mathcal{F}$ is Value–Perfect. Then,*

(i) If $z^{t,\lambda} < z_E^{t,*}$ then $\widetilde{A}_E^{t+1} \setminus \widetilde{A}_E^t \neq \emptyset$.

(ii) $\tau^\lambda \leq \xi_E^\lambda \leq |A \setminus \widetilde{A}_E^0|$.

3.4.2.2 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 9. *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*

$$\left(\widetilde{\boldsymbol{F}}_d^t\right)^\top y^{t,\lambda} > f_d - \left(\boldsymbol{L}_d^t\right)^\top x^{t,\lambda}. \tag{3.10}$$

The above result implies that the leader can remove matrix $\widetilde{\boldsymbol{F}}^t$ from the uncertainty set at time t, as equation (3.10) 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 (3.10) holds at time t, that is

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

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 (3.10) is to define $\mathcal{U}^{t+1}$ as

$$\mathcal{U}^{t+1} = \{ \hat{\boldsymbol{F}}^t \in \mathcal{U}^t \ : \ \left(\hat{\boldsymbol{F}}_d^t\right)^\top y^{t,\lambda} \leq f_d - \left(\boldsymbol{L}_d^t\right)^\top x^{t,\lambda} \ \forall d \in D^{t,\lambda} \}, \tag{3.11}$$

where we note that $\mathcal{U}^{t+1} \subset \mathcal{U}^t$ by Lemma 9. 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 (3.10) holds cannot be guaranteed, and hence the update in equation (3.11) 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 (3.11) does not necessarily reduce the dimension of $\mathcal{U}^t$, and hence an upper bound similar to that of Theorem 5 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 9. *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.

Proposition 10. *Let $\epsilon > 0$, $\lambda \in \Lambda_E$, and $t \in \mathcal{T}$ be given. Assume that $y_a^t = 0$ for all $a \notin A^t$ and define $\Delta^t := \mathcal{U}^t \setminus \mathcal{U}^{t+1}$. If $z_E^* - z^{t,\lambda} > \epsilon$, then there exist $K > 0$ (independent of ϵ) such that*

$$diam(\Delta^t) > \frac{-\|y^{t,\lambda}\| + \sqrt{\|y^{t,\lambda}\|^2 + 4K\epsilon^2\|\boldsymbol{c}^t\|^{-2}}}{2K},$$

where $diam(\Delta^t)$ denotes the diameter of polyhedron Δ^t, i.e., $diam(\Delta^t) = \max_{\boldsymbol{F}', \boldsymbol{F}'' \in \Delta^t} \|\boldsymbol{F}' - \boldsymbol{F}''\|$.

3.5 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*a*). Such an oracle faces no uncertainty and is able to make the best possible decision. In our problem setting, the oracle solves problem (3.1) 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:

$$\min \sum_{t \in \mathcal{T}} \mathbb{1}_{\{\boldsymbol{c}^\top y^t < z^*\}} \tag{3.12a}$$

$$\text{s.t. } x^t \in X \qquad t \in \mathcal{T} \tag{3.12b}$$

$$y^t \in \arg\min\{\boldsymbol{c}^\top y \colon y \in Y(x^t)\} \qquad t \in \mathcal{T} \tag{3.12c}$$

$$x_i^t = 0 \qquad i \in I \setminus I^t, t \in \mathcal{T} \tag{3.12d}$$

$$I^{t+1} = I^t \cup \bigcup_{a \colon y_a^t > 0} I(a) \qquad t \in \mathcal{T} \setminus \{T\}, \tag{3.12e}$$

where constraint (3.12d) 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 (3.12b) and (3.12c), 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 (3.12) 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 (3.12), 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 (3.12).

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 (3.12) 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 3.2.2).

It can be readily seen that the semi-oracle optimization problem (3.12) 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 (3.12) is given by:

$$\min_{u,v,w,y,x,\theta} \sum_{t\in\mathcal{T}} w^t \tag{3.13a}$$

$$\text{s.t. } \boldsymbol{H}x^t \le \boldsymbol{h} \qquad t\in\mathcal{T} \tag{3.13b}$$

$$\boldsymbol{F}y^t + \boldsymbol{L}x^t \le \boldsymbol{f},\ -\boldsymbol{F}^\top\theta^t \le \boldsymbol{c} \qquad t\in\mathcal{T} \tag{3.13c}$$

$$\theta^t \le \boldsymbol{M}^{\theta^t}u^t,\ y^t \le \boldsymbol{M}^{y^t}v^t \qquad t\in\mathcal{T} \tag{3.13d}$$

$$\boldsymbol{f} - \boldsymbol{F}y^t - \boldsymbol{L}x^t \le \boldsymbol{M}^{p^t}(\boldsymbol{1}-u^t) \qquad t\in\mathcal{T} \tag{3.13e}$$

$$\boldsymbol{c} + \boldsymbol{F}^\top\theta^t \le \boldsymbol{M}^{q^t}(\boldsymbol{1}-v^t) \qquad t\in\mathcal{T} \tag{3.13f}$$

$$x_i^t \le M^{x_i}\sum_{s=0}^{t-1}\sum_{a\in A(i)\setminus A^0} y_a^s \qquad t\in\mathcal{T},\ i\in I\setminus I^0 \tag{3.13g}$$

$$z^*(1-M^w w^t) \le \boldsymbol{c}^\top y^t \qquad t\in\mathcal{T} \tag{3.13h}$$

$$u^t\in\{0,1\}^{|C_F|}, v^t\in\{0,1\}^{|A|}, w^t\in\{0,1\} \qquad t\in\mathcal{T} \tag{3.13i}$$

$$y^t\in\mathbb{R}_+^{|A|}, x^t\in\mathbb{R}_+^{|I|-k}\times\mathbb{Z}_+^k, \theta^t\in\mathbb{R}_+^{|C_F|} \qquad t\in\mathcal{T}, \tag{3.13j}$$

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 (3.13c) (primal and dual feasibility) and (3.13d), (3.13e), and (3.13f) (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 (3.13h). 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 (3.13g) 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 3.6 for an example.

We close this section by noting that although MIP problem (3.13) 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 B.3.1.

3.6 COMPUTATIONAL STUDY

In this section we demonstrate the numerical performance of the policies in Λ. For this, we use the AD Knapsack problem of Example 2. We consider both Value–Perfect and Response–Perfect feedbacks as well as two different models for the initial uncertainty set. 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 (3.6), see Section B.3.2 of the Appendix for further details. Generally speaking, the transformation of optimization problem (3.6) 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 (3.6) is NP-hard, as bilevel linear optimization is its special case.

Test Instances. We consider the AD knapsack problem from Example 2, where the defender has $n = 12$ assets, $\boldsymbol{b} = \boldsymbol{r} = \mathbf{1}$, and $B = R = 4$. We consider two models of initial uncertainty sets, namely, *hypercube* uncertainty and *general* uncertainty:

- In the hypercube model the defender's profits satisfy $p_a \in [\ell_a, \ell_a + m_a]$, $a \in A$, where ℓ_a is drawn at random from uniform discrete $U(1, 5)$ distribution and m_a is drawn from $U(1, 15)$ distribution.
- For the general uncertainty model we generate a non-negative polytope with $C_U = 3$ inequalities. The polytope is given by $\mathcal{P} := \{p : \ G\,p \leq g, p \geq 0\}$, where $G_{u,j}$ is drawn at random from $U(1, 10)$ distribution for $j \in \{4(u-1)+1, 4u\}$, and $G_{u,j} = 0$ otherwise, while g_u is drawn at random from $U(1, 20)$ distribution, for $u = 1, 2, 3$.

Given a polytope $\mathcal{P}$, we generate the follower's profit vector $\boldsymbol{p}$ by using the following approach. First, we compute the barycenter (or analytical center) of the polytope by solving the following convex problem (see, e.g., Bertsimas and Tsitsiklis (1997)):

$$\boldsymbol{p}_b \in \arg\min_{(\hat{p},\hat{q})\geq 0} \Big\{ -\sum_{j=1}^{n} \log(\hat{p}_j) - \sum_{u=1}^{|C_U|} \log(\hat{q}_u) \ : \ \boldsymbol{G}\hat{p} + \hat{q} = \boldsymbol{g},\, \hat{p} \geq 0 \Big\}.$$

Next, we randomly construct an extreme point of $\mathcal{P}$ by first generating a vector $\boldsymbol{\ell}$ of size n (where each entry is zero or one with the same probability) and then solving an LP of the form:

$$\boldsymbol{p}_e \in \arg\max_{\hat{\boldsymbol{p}}} \{\boldsymbol{\ell}^\top \hat{\boldsymbol{p}} \ : \ \boldsymbol{G}\hat{p} \leq \boldsymbol{g},\, \hat{p} \geq 0\}.$$

Finally, we combine the barycenter with the obtained extreme point by $\boldsymbol{p} = (\boldsymbol{p}_b + 7\boldsymbol{p}_e)/8$, to generate an interior point of the polytope $\mathcal{P}$, and hence ensuring that $p_a > 0$ for all $a \in A$.

For each uncertainty model, we generated at random $N = 30$ instances, considering both Value–Perfect and Response–Perfect feedbacks. We consider three sets of initial information A^0: in the first, the leader knows four activities of the follower; in the second, she knows eight activities; and in the last, she knows all activities. Finally, we set $T = 24$.

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

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

$$x^{t,\pi_b} \in \arg\max_{x \in X^t} \{(\boldsymbol{c}^t)_b^\top y \ : \ y \in Y^t(x)\}, \tag{3.14}$$

  where $\boldsymbol{c}_b^t = -\boldsymbol{p}_b^t$, and $\boldsymbol{p}_b^t$ is the barycenter of the polytope $\mathcal{U}^t$.
- The *random policy* π_r: At each time $t \in \mathcal{T}$ the policy computes x^{t,π_r} by solving problem (3.14) with $\boldsymbol{c}_r^t$ used instead of $\boldsymbol{c}_b^t$. We have $\boldsymbol{c}_r^t = -\boldsymbol{p}_r^t$, and $\boldsymbol{p}_r^t$ is a randomly generated extreme point of $\mathcal{U}^t$ that is obtained by solving the linear program $\boldsymbol{p}_r^t \in \arg\max\{\boldsymbol{\ell}^{t,T}\hat{p} \colon \hat{p} \in \mathcal{U}^t\}$. 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 leader observes a follower's response that she has observed in the earlier time periods, then the policy keeps using the same solution thereafter. That is, if time t' is the earliest period such that $z^{t',\pi_s} = z^{t,\pi_s}$ for some $t < t'$, then $x^{t,\pi_s} = x^{t',\pi_s}$ for all $t \geq t'$.
- We also consider the *lower bound* provided by the semi-oracle approach discussed in Section 3.5. While it is not an admissible policy, with a slight abuse of notation we denote it by π^* hereafter.

Results and Discussion. For each uncertainty model we compute its time-stability across $N = 30$ replications by using each of the policies described above. Tables 10 and 11 report the mean time-stability and mean absolute deviation (MAD) for the hypercube uncertainty model under Value–Perfect and Response–Perfect feedbacks, respectively. Similarly, Tables 12 and 13 show the same results for the general uncertainty model. For the sake of reporting averages, *policies that do not find an optimal solution within the first 24 periods of an instance are assigned the value* $\tau^\pi = T = 24$.

Table 10: Time-stability mean and MAD for the hypercube uncertainty model and Value Perfect feedback.

(a) Value–Perfect: time-stability mean

A^0	λ	π_b	π_r	π_s	π^*
$\{1, \cdots, 4\}$	2.13	20.17	21.70	21.70	1
$\{1, \cdots, 8\}$	2.93	21.73	23.20	23.20	0.93
A	3.03	21.77	24.00	24.00	0

(b) Value–Perfect: time-stability MAD

A^0	λ	π_b	π_r	π_s	π^*
$\{1, \cdots, 4\}$	0.29	6.39	4.14	4.14	0.00
$\{1, \cdots, 8\}$	0.50	4.08	1.55	1.55	0.12
A	0.59	4.02	0.00	0.00	0

We observe that the proposed policies $\lambda \in \Lambda$ consistently outperform the benchmark except for the semi-oracle lower bound π^*, which is expected. For most instances in the hypercube model, policies π_b, π_r and π_s yield very poor time-stability results, not being able to find an optimal solution for most cases within the time horizon. Their performance improves, however, for the general uncertainty model. 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.

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, by a factor of at least two, under both uncertainty models. 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. Although this behavior is rather counter-intuitive, it might stem from the fact

Table 11: Time-stability mean and MAD for the hypercube uncertainty model and Response Perfect feedback.

(a) Response–Perfect: time-stability mean

A^0	λ	π_b	π_r	π_s	π^*
$\{1, \cdots, 4\}$	7.77	23.23	21.10	21.87	1
$\{1, \cdots, 8\}$	8.00	23.20	23.20	23.20	0.93
A	7.93	24.00	24.00	24.00	0

(b) Response–Perfect: time-stability MAD

A^0	λ	π_b	π_r	π_s	π^*
$\{1, \cdots, 4\}$	2.21	1.48	4.64	3.84	0.00
$\{1, \cdots, 8\}$	1.87	1.55	1.55	1.55	0.12
A	1.74	0.00	0.00	0.00	0

Table 12: Time-stability mean and MAD for the general uncertainty model and Value Perfect feedback.

(a) Value–Perfect: time-stability mean

A^0	λ	π_b	π_r	π_s	π^*
$\{1, \ldots, 4\}$	2	14.20	4.43	3.07	1
$\{1, \ldots, 8\}$	1	10.27	5.20	4.30	0.97
A	1	9.37	7.67	6.67	0

(b) Value–Perfect: time-stability MAD

A^0	λ	π_b	π_r	π_s	π^*
$\{1, \ldots, 4\}$	0	11.11	2.77	1.52	0.00
$\{1, \ldots, 8\}$	0	10.99	3.73	3.94	0.06
A	0	10.73	6.56	6.93	0

that in this particular bilevel setting, the follower's activities are fairly independent of each other (they are interrelated only through the follower's budget constraint). Hence, partial knowledge of the follower activities does not implicitly reveal much information about the remaining unknown activities. We note that in more complex bilevel settings the amount of initial information does have a very important effect (see, e.g., discussion in Borrero et al.

Table 13: Time-stability mean and MAD for the general uncertainty model and Response Perfect feedback.

(a) Response–Perfect: time-stability mean

A^0	λ	π_b	π_r	π_s	π^*
$\{1,\ldots,4\}$	6.90	14.23	9.10	11.37	1
$\{1,\ldots,8\}$	7.50	17.30	14.53	12.97	0.97
A	7.73	16.77	17.73	11.93	0

(b) Response–Perfect: time-stability MAD

A^0	λ	π_b	π_r	π_s	π^*
$\{1,\cdots,4\}$	0.30	11.07	3.87	8.42	0.00
$\{1,\cdots,8\}$	0.53	9.38	7.37	9.56	0.06
A	0.41	9.64	7.14	8.85	0

(2016) for an example in the context of the shortest path interdiction).

An important feature of the policies in Λ is their low variability. This is especially true in the general uncertainty model, where the policies yield no variability in the Value–Perfect setting, and a very low variability in the Response–Perfect setting. In contrast, the benchmark policies are orders of magnitude more variable in the general uncertainty setting. While these policies have low MAD values in Tables 10(b) and 11(b) for the hypercube uncertainty, 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 24 periods of an instance are assigned the value $\tau^\pi = T = 24$).

3.7 CONCLUDING REMARKS

This chapter 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 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. Also, 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.

4.0 SEQUENTIAL ASYMMETRIC BILEVEL LINEAR PROGRAMMING WITH INCOMPLETE INFORMATION AND LEARNING

4.1 INTRODUCTION

Several applications of bilevel programming involve non-adversarial decision-makers where the follower's objective function might be unrelated to the objective function of the leader. Therefore, by optimizing her objective the leader does not necessarily seeks to degrade the follower's performance, see Colson et al. (2007*a*), Dempe (2002), Saharidis et al. (2013). Moreover, similar to the adversarial problems studied in the previous chapters, there are examples of this class of *asymmetric* bilevel problems where the assumption that the leader has complete information about the follower's problem does not hold. As a consequence, the leader is forced to learn about the structure and data of the follower's problem from her interactions with him.

As an example, consider a plant selection problem in a decentralized manufacturing environment (Cao and Chen 2006). Here, the leader is a principal firm that has to hire auxiliary plants (managed by the follower) to manufacture a set of given products. The leader's objective is to minimize the costs of opening the plants and of unused utilization. Given the plants selected by the leader, the follower must configure their operation in order to minimize the operational cost of each plant. Since the auxiliary plants are independent of the principal firm, the leader might not know the precise information the follower uses to operate them, and moreover, the follower might not have any incentive to reveal this information to the principal firm. Hence, the leader must learn this undisclosed information by observing the follower's reactions to her decisions each time a contract renegotiation takes place.

Alternatively, recall the repeated version of a network pricing (or tariff) problem studied in Bouhtou et al. (2007) that is discussed in the Introduction. There, the leader has no access to the follower's problem information as it involves information on competing firms that might not be publicly available. Moreover, the follower, being the network user, has no incentive to disclose this information to the leader; such disclosure might impair his ability to optimize his objective function.

In this chapter we study such sequential asymmetric bilevel linear problems with incomplete information within the framework developed for the max-min bilevel case in Chapter 3. We assume that the leader and the follower interact across a set of given time periods, that the problems are given in terms of selecting the levels of resources to be used and of activities to be performed, and that the leader knows that the follower's cost vector belongs to a certain uncertainty set. In addition, we assume that the decision-making process of the leader is given in terms of policies, and we measure their performance by using time-stability. Particularly, as we discuss in Section 4.2, our focus is to find weakly optimal time-stability policies, that is, we seek to find the leader's decision-making policies that attain the best possible worst-case performance across all instances.

In the present study we consider two classes of policies. The first one, presented in Section 4.3, is the adaptation of the greedy and robust policies of Chapter 3 to the asymmetric setting. We show that these policies do not retain the properties that lead to weak optimality. Moreover, they can get stuck in sub-optimal solutions and they might be unable to provide certificates of optimality in real time. On the other hand, in Section 4.4 we discuss a class of *greedy and 'best'-case policies*. We show that these policies can provide a finite upper bound on the time-stability, and always provide a certificate of optimality in real time.

Importantly, the greedy and 'best'-case policies allow for more flexibility (compared to the greedy and robust policies) in the way that the uncertainty set can be updated. As such, we study different *updating mechanisms* for this class of policies under the different feedback modes, and show how the decisions prescribed by these policies can be computed using mixed integer programming (MIP) formulations. Particularly, these formulations can be viewed as the extensive form of a two-stage stochastic MIP. Hence, different decomposition techniques from the stochastic programming literature might be used to compute larger instances.

In Section 4.5 we perform computational experiments that study the time-stability of the greedy and best-case policies under various configurations of the feedback and the updating mechanisms. We show that for most of the configurations these policies greatly outperform the worst-case theoretical time-stability upper bound. Moreover, the results show that for most cases the time-stability is upper bounded by the number of actions of the follower, which suggests that under certain assumptions on the feedback and the updating mechanisms, the greedy and best-case policies might be weakly optimal. Finally, in Section 4.6 we provide conclusions and possible directions of future research.

4.2 PROBLEM FORMULATION

Consider the full-information leader's problem given by

$$w^* = \max\{w(x)\colon x \in X\}, \tag{4.1}$$

where

$$X = \{x \in \mathbb{R}_+^{|I|-k} \times \mathbb{Z}_+^k\colon \boldsymbol{H}x \leq \boldsymbol{h}\}, \tag{4.2}$$

and for any $x \in X$

$$w(x) = \max\{\boldsymbol{d}^\top y\colon y \in Z(x)\}, \tag{4.3}$$

with $\boldsymbol{d} \in \mathbb{R}^{|A|}$ being a given vector. For any given $x \in X$ the set $Z(x)$ is defined as

$$Z(x) = \arg\min\{\boldsymbol{c}^\top y\colon y \in Y(x)\}, \tag{4.4}$$

where for any $x \in X$ we have

$$Y(x) = \{y \in \mathbb{R}_+^{|A|}\colon \boldsymbol{F}y + \boldsymbol{L}x \leq \boldsymbol{f}\}. \tag{4.5}$$

We make the assumption that the leader does not know vector $\boldsymbol{c}$ with certainty, but knows that it lies within an uncertainty set $\mathcal{U}^0$, where $\mathcal{U}^0$ is assumed to be a polyhedron given by

$$\mathcal{U}^0 = \{\hat{c} \in \mathbb{R}^{|A|}\colon \boldsymbol{G}^0\hat{c} \leq \boldsymbol{g}^0\}. \tag{4.6}$$

In contrast with the max-min model, we make the simplifying assumption that the leader knows all the variables and constraints and all the remaining data. Thus, $A^0 = A$, $I^0 = I$, $C_F^0 = C_F$, and $C_L^0 = C_L$. Importantly, we also make the assumption of a bilevel *optimistic approach*, that is, the follower cooperates with the leader if there are multiple optimal solutions for the follower's problem, see, e.g., Dempe (2002).

We assume that the leader knows the upper-level vector $\boldsymbol{d}$ with certainty; which reflects the fact that she is aware of her resources and capabilities. A further generalization would assume that her knowledge of $\boldsymbol{d}$ is also subject to uncertainty, i.e., $\boldsymbol{d} \in \widetilde{\mathcal{U}}^0$ for some uncertainty set $\widetilde{\mathcal{U}}^0$. Such model generalizes the max-min problem studied in Chapter 3 (by setting $\boldsymbol{d} = \boldsymbol{c}$ and $\widetilde{\mathcal{U}}^0 = \mathcal{U}^0$) as well as the model we study in this chapter (by setting $\widetilde{\mathcal{U}}^0 = \{\boldsymbol{d}\}$). We leave the study of such model outside this thesis as the subject of further research.

Given this set-up, at each time $t \in \mathcal{T} = \{1, 2, \ldots, T\}$ the following sequence of events take place:

1. The leader chooses $x^t \in X$.
2. The follower solves the following linear program:

$$z(x^t) = \min\{\boldsymbol{c}^\top y \colon y \in Y(x^t)\}. \tag{4.7}$$

   For notational convenience, we set $y^t \coloneqq y(x^t)$ and $z^t \coloneqq z(x^t)$, where we recall that $y(x^t)$ is the vector the follower chooses at time t if the leader implements x^t.
3. The leader receives the *profit* $w^t = \boldsymbol{d}^\top y^t$.
4. The response of the follower generates a *feedback* $\mathcal{F}^t$. The leader observes the information in $\mathcal{F}^t$ and exploits it to update her current knowledge of the uncertainty set to $\mathcal{U}^{t+1}$.

In this setting, the Standard, Value–Perfect, and Response–Perfect feedbacks are defined as in Section 3.2.1. We make the convention that for this problem all these feedback types also reveal the value of w^t to the leader at each time $t \in \mathcal{T}$.

The leader decides in terms of policies (cf. Section 3.2.2; in particular we use a superscript π to discuss vectors and quantities associated with policy π), and her objective is to find

a weakly optimal time-stability policy (see Section 3.2.2). The time-stability of policy π is defined as τ^π, where

$$\tau^\pi = \min\{t \in \mathcal{T} \colon w^{s,\pi} = w^* \text{ for all } s \geq t\}. \tag{4.8}$$

Observe that time-stability has the same interpretation as before, being the first time period by which the leader implements the optimal full-information bilevel solution from there on.

4.3 GREEDY AND ROBUST POLICIES

Following the developments in Sections 2.3 and 3.3 for the shortest-path interdiction and max-min bilevel problems, in this section we consider *greedy and robust* policies. If Δ is such set of policies, then we say $\delta \in \Delta$ if and only if

$$x^{t,\delta} \in \arg\max\Big\{\boldsymbol{d}^\top y \colon y \in \arg\min\big\{\max\{\hat{\boldsymbol{c}}^\top y' \colon \hat{\boldsymbol{c}} \in \mathcal{U}^t\} \colon y' \in Y(x)\big\},\ x \in X\Big\}. \tag{4.9}$$

When the leader decides using a policy $\delta \in \Delta$, then she *expects* her profit at period t to be

$$w_R^t(x^{t,\delta}) = \boldsymbol{d}^\top y, \quad y \in W_R^t(x^{t,\delta}), \tag{4.10}$$

where for any $x \in X$ we denote

$$W_R^t(x) = \arg\max\{\boldsymbol{d}^\top y \colon y \in Z_R^t(x)\}, \tag{4.11}$$

and for any $x \in X$ the set $Z_R^t(x)$ is defined as:

$$Z_R^t(x) = \arg\min\big\{\max\{\hat{\boldsymbol{c}}^\top y \colon \hat{c} \in \mathcal{U}^t\} \colon y \in Y(x)\big\}. \tag{4.12}$$

Also, if the leader is using policy $\delta \in \Delta$, then the cost the leader *expects* the follower incurs at t is given by $z_R^t(x^{t,\delta})$, where for any $x \in X$ (c.f. Section 3.3.1)

$$z_R^t(x) = \max\{\hat{\boldsymbol{c}}^\top y \colon \hat{c} \in \mathcal{U}^t\}, \quad y \in Z_R^t(x). \tag{4.13}$$

One the main properties of the greedy and robust policies in the max-min context is that whenever the leader's expectations are different from what she observes, then the follower

must reveal new information to the leader. This result holds here as well as the following lemma shows. (For notational simplicity let us write $w_R^{t,\delta}$ for $w_R^t(x^{t,\delta})$ from here on and let $w^{t,\delta}$ be the profit the leader observes at period t is she is using policy δ.)

Lemma 10. *Suppose that $\delta \in \Delta$ and that feedback is Value–Perfect or Response–Perfect. If $w^{t,\delta} \neq w_R^{t,\delta}$ then $\dim(\mathcal{U}^{t+1}) < \dim(\mathcal{U}^t)$.*

Proof. First, for Value–Perfect feedback, assume that $w^{t,\delta} > w_R^{t,\delta}$. This implies that $y^{t,\delta} \notin W_R^t(x^{t,\delta})$, and this implies that either (i) $y^{t,\delta} \notin Z_R^t(x^{t,\delta})$, or that (ii) $y^{t,\delta} \in Z_R^t(x^{t,\delta})$ and that there exist $y \in Z_R^t(x^{t,\delta})$ such that $\boldsymbol{d}^\top y > \boldsymbol{d}^\top y^{t,\delta}$.

Suppose that (i) holds. As $y^{t,\delta} \in Y(x^{t,\delta})$, then it must be the case that there exist $\tilde{y} \in Y(x^{t,\delta})$ such that $\max\{\hat{\boldsymbol{c}}^\top \tilde{y} \colon \hat{c} \in \mathcal{U}^t\} < \max\{\hat{\boldsymbol{c}}^\top y^{t,\delta} \colon \hat{c} \in \mathcal{U}^t\}$. Suppose that $y_a^{t,\delta} = 0$ for all $a \in A \setminus \widetilde{A}^t$. Then $\max\{\hat{\boldsymbol{c}}^\top y^{t,\delta} \colon \hat{c} \in \mathcal{U}^t\} = \sum_{a \in \widetilde{A}^t} c_a y_a^{t,\delta}$, and hence

$$\sum_{a \in A} c_a \tilde{y}_a \leq \max\{\hat{\boldsymbol{c}}^\top \tilde{y} \colon \hat{c} \in \mathcal{U}^t\} < \sum_{a \in \widetilde{A}^t} c_a y_a^{t,\delta} = \sum_{a \in A} c_a y_a^{t,\delta}. \tag{4.14}$$

Hence, $y^{t,\delta} \notin Z(x^{t,\delta})$, which contradicts the definition of $y^{t,\delta}$. So it follows that $y_a^{t,\delta} > 0$ for some $a \in A \setminus \widetilde{A}^t$, and the result follows. Finally, observe that (ii) cannot hold as it immediately contradicts the definition of $w_R(x^{t,\delta})$.

Now, suppose that $w^{t,\delta} < w_R^{t,\delta}$, then there are two possibilities: (i) as before and (ii') $y^{t,\delta} \in Z_R^t(x^{t,\delta})$ and there exist $y \in Z_R^t(x^{t,\delta})$, such that $\boldsymbol{d}^\top y^{t,\delta} < \boldsymbol{d}^\top y$. If (i) holds then the results holds for the same arguments. Finally, (ii') cannot hold in this setting as it would contradict the optimistic assumption of the bilevel problem.

For Response–Perfect feedback, first assume that $w^{t,\delta} > w_R^{t,\delta}$. Then, by the same reasoning either (i) $y^{t,\delta} \notin Z_R^t(x^{t,\delta})$, or (ii) $y^{t,\delta} \in Z_R^t(x^{t,\delta})$ and that there exist $y \in Z_R^t(x^{t,\delta})$ such that $\boldsymbol{d}^\top y > \boldsymbol{d}^\top y^{t,\delta}$. For (i) suppose that $[y^{t,\delta}; z^{t,\delta}]$ is linearly dependent with all the rows of $[\boldsymbol{G}^{t,=}, \boldsymbol{g}^{t,=}]$ (see Lemma 8 in Section 3.3.3). This implies that $\{\hat{c} \colon \boldsymbol{G}^{t,=}\hat{c} = g^{t,=}\} = \{\hat{c} \colon \boldsymbol{G}^{t,=}\hat{c} = g^{t,=}, y^{t,\delta,\top}\hat{c} = z^{t,\delta}\}$. Thus, for any $\hat{c} \in \mathcal{U}^t$ it follows that $y^{t,\delta,\top}\hat{c} = z^{t,\delta}$. In particular, $\max\{\hat{c}^\top y^{t,\delta} \colon \hat{c} \in \mathcal{U}^t\} = z^{t,\delta}$, and we can conclude that $y^{t,\delta} \notin Z(x^{t,\delta})$ by using the same chain of inequalities in (4.14). This gives the desired contradiction. Also, as before (ii) cannot hold from the definition of $w_R(x^{t,\delta})$, and it is also readily seen that the same arguments of the proof for Value–Perfect feedback apply for the case where $w^{t,\delta} < w_R^{t,\delta}$. ∎

Remark 10. Note that the proof on Lemma 10 does not use the fact that the δ policies are greedy. Hence, the lemma holds for a broader class of robust policies. ■

In contrast with the max-min setting, in the asymmetric case the 'sandwich' theorem does not hold (cf. Theorem 3 in Section 3.3), i.e, we do not have that the chain of inequalities $w^{t,\delta} \leq w^* \leq w_R^{t,\delta}$ holds in general. In fact, it is possible to come with examples, see Remark 11, where

- $w^{t,\delta} > w_R^{t,\delta}$ (and thus $w^* > w_R^{t,\delta}$ as well),
- $w^{t,\delta} < w_R^{t,\delta}$,
- $w^* < w_R^{t,\delta}$,
- If $w^{t,\delta} = w_R^{t,\delta}$, then it might happen that $x^{t,\delta}$ is not the optimal solution of the asymmetric bilevel problem.

(However, note that it is straightforward that $w^{t,\delta} \leq w^*$ for all $t \in \mathcal{T}$). Moreover, if the leader ignores the upper-level feedback and focuses only on the follower's costs, the fact that $z^{t,\delta} = z_R^{t,\delta}$ does not give any information about the optimality of $x^{t,\delta}$ in general, see Remark 12.

The above observations imply that nothing can be said in general about the optimality of $x^{t,\delta}$ whenever the profit the leader observes is the same as the one she was expecting. The same holds true for the expected follower's costs versus the actual costs he incurs. This behavior is very troublesome as it implies that a policy $\delta \in \Delta$ might not find the optimal solution, and might not provide a certificate of optimality in real time.

Moreover, if $z^{t,\delta} = z_R^{t,\delta}$ and the solution used by the follower does not reveal any new information (e.g., $y_a^t = 0$ for all $a \in A \setminus \widetilde{A}^t$), then the leader does not learn anything new (see Remark 12) and hence at time $t+1$ the decision of time t is going to be repeated. In other words, policies in Δ might *stall*, i.e., might repeat a suboptimal solution at all time periods indefinitely without forcing the follower to reveal any new information.

Remark 11. In this remark we give various counterexamples that show that, in contrast to the max-min problem, in the asymmetric case the greedy and robust policies do not yield the relationships of the 'sandwich' theorem. The bilevel problem we use is the asymmetric

shortest path interdiction problem (ASPI), see Bayrak and Bailey (2008). Here, the follower's objective is to move between two fixed nodes at a minimum cost, while the objective of the leader is to maximize the profit of the shortest path that the follower uses. The follower incurs a cost of c_a if he uses arc a, while the leader gets a profit of d_a if the follower uses uses arc a. The leader does not know the real costs $\boldsymbol{c}$ the follower uses to decide, but she knows that the cost of arc a lies in the interval $c_a \in [\ell_a, u_a]$.

Figure 12 shows an example of an instance of the ASPI where $w^{t,\delta} > w_R^{t,*}$ and where $w^* > w_R^{t,*}$. Here (and in the remaining counterexamples), the objective of the follower is to move between nodes 1 and 7. Observe that if the leader is deciding robustly, then she assumes that the cost that the follower incurs by traversing an arc is given by u_a. Hence, the solution for any policy $\delta \in \Delta$ is to block arcs $(1,2)$ and $(1,3)$ (or more generally to block the two upper-most paths). Given this, the leader expects that the follower uses path 1–4–7, and hence she expects a profit of $w_R^{t,\delta} = 60$. However, observe that if the leader blocks $(1,2)$ and $(1,3)$, then the path that the follower uses is 1–6–7, which yields a profit of $w^{t,\delta} = 80$ to the leader. Observe moreover, that $w^* = w^{t,\delta}$, so this example also shows that $w_R^{t,\delta} < w^*$.

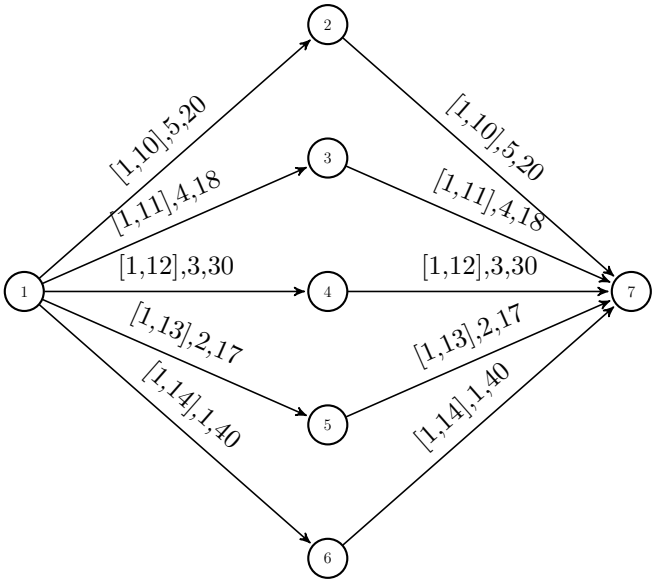


Figure 12: Example of an instance when $w^{t,\delta} > w_R^{t,*}$. The labeling of the arcs is given by $[\ell_a, u_a], c_a, d_a$.

Conversely, Figure 13 shows an example of an instance of the ASPI where $w^{t,\delta} < w_R^{t,*}$. Here, the solution for any policy $\delta \in \Delta$ is to block again arcs $(1,2)$ and $(1,3)$, and as before, the leader expects that the follower uses path 1–4–7. This yields an expected profit of $w_R^{t,\delta} = 60$. However, observe that if the leader blocks $(1,2)$ and $(1,3)$, the path that the follower uses is 1–5–7, which yields a profit of $w^{t,\delta} = 34$ to the leader.

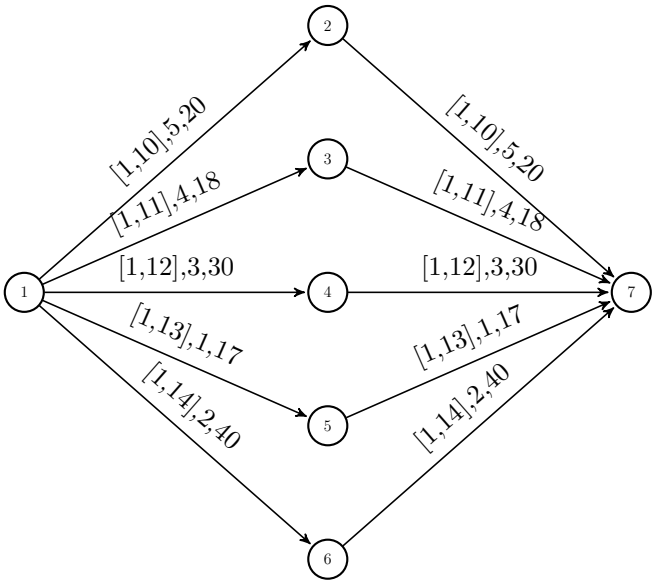


Figure 13: Example of an instance when $w^{t,\delta} < w_R^{t,*}$. The labeling of the arcs is given by $[\ell_a, u_a], c_a, d_a$.

Figure 14 shows an example of an instance of the ASPI where $w^* < w_R^{t,*}$. Here, the solution for any policy $\delta \in \Delta$ is to block again arcs $(1,2)$ and $(1,3)$, and as before, the leader expects that the follower uses path 1–4–7. This yields an expected profit of $w_R^{t,\delta} = 60$. The full-information optimal solution, however, is to remove $(1,3)$ and $(1,5)$. This makes the follower use path 1–2–7, and yields a profit of $w^* = 40$, hence $w^* < w_R^{t,\delta}$.

Finally, Figure 15 shows an example of an instance of the ASPI where the main conclusion of 'sandwich' theorem fails to hold, i.e., where the fact that $w^{t,\delta} = w_R^{t,*}$ does not imply that $w^{t,\delta} = w^*$. Here, it is readily checked that the solution for any policy $\delta \in \Delta$ is to block arcs $(1,2)$ and $(1,3)$. The leader expects that the follower uses path 1–4–7, which yields an expected profit of $w_R^{t,\delta} = 60$. In this case, it is seen that the the response of the follower is indeed as expected by the leader, that is, to use 1–4–7, and hence $w^{t,\delta} = 60 = w_R^{t,\delta}$. However,

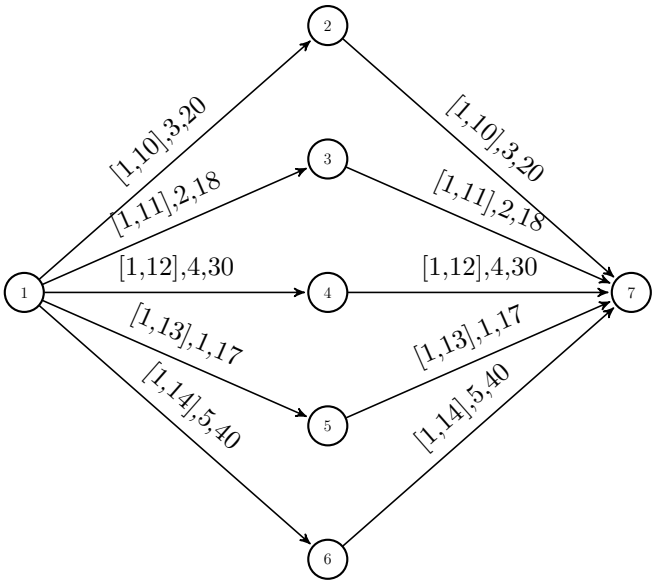


Figure 14: Example of an instance when $w^* < w_R^{t,*}$. The labeling of the arcs is given by $[\ell_a, u_a], c_a, d_a$.

note that the optimal full-information solution for the leader is to remove the arcs $(1,4)$ and $(1,5)$, as this implies that the follower would use path 1–6–7, which gives an optimal profit of $w^* = 80$. ■

Remark 12. Figure 16 shows an example in the ASPI where the fact that $z^{t,\delta} = z_R^{t,\delta}$ does not imply that $w^{t,\delta} = w^*$. Observe that any policy in δ blocks any two arcs among $(1,2)$, $(1,3)$, (2,7) and (3,7), and hence she expects that the follower use path 1–4–7 at a cost of $z_R^{t,\delta} = 14$. Given the real costs of arcs (1,4) and (4,7), then the follower will use the same path 1–4–7, with $z^{t,\delta} = 14$ as well, which yields a profit of $w^{t,\delta} = 6$ for the leader. However, it is seen that the optimal full-information solution is to remove (1,3) and (1,4), which makes the follower use path 1–5–7, and yields a profit of $w^* = 8$ to the leader. ■

Regardless of the fact that in general the 'sandwich' theorem does not hold, it is possible to characterize whenever the equation $w^{t,\delta} = w_R^{t,\delta}$ implies that the optimal solution of the full-information problem has been found. To this end, define $X_P^{t,\delta}$ as the following set of leader's solutions: $x \in X_P^{t,\delta}$ if and only if

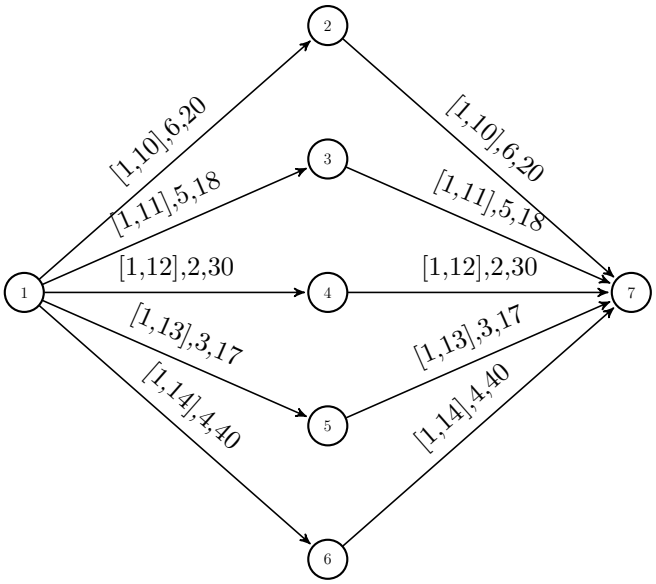


Figure 15: Example of an instance when $w^{t,\delta} = w_R^{t,*}$ does not imply that $w^{t,\delta} = w^*$, and where $w_R^{t,\delta} < w^*$. The labeling of the arcs is given by $[\ell_a, u_a], c_a, d_a$.

C.1. There exist $\hat{y} \in Z(x) \cap (Z_R^t(x))^c$, where A^c denotes the complement of A.

C.2. The inequality $\boldsymbol{d}^\top y < \boldsymbol{d}^\top \hat{y}$ holds for any $y \in Z_R^t(x^{t,\delta})$.

Simply speaking, $x \in X_P^{t,\delta}$ if, after implementing x, then the follower can implement a solution $\hat{y}$ that has a better objective value than any solution that the leader expects by using δ (at time t). Let us define the set of all those solutions $\hat{y}$ as $Y_P^{t,\delta}(x)$, i.e.,

$$Y_P^{t,\delta}(x) = \{\hat{y} \in Z(x) \colon \hat{y} \notin Z_R^t(x),\, \boldsymbol{d}^\top y < \boldsymbol{d}^\top \hat{y}\ \forall y \in Z_R^t(x^{t,\delta})\}.$$

Finally, let $Y_P^{t,\delta}$ be the set of all follower's responses associated with the solutions in $X_P^{t,\delta}$:

$$Y_P^{t,\delta} = \bigcup_{x \in X_P^{t,\delta}} Y_P^{t,\delta}(x). \tag{4.15}$$

We have the following result:

Lemma 11. *Assume that $w^{t,\delta} = w_R^{t,\delta}$. Then, $x^{t,\delta}$ is an optimal solution of the full information problem (i.e., $w^{t,\delta} = w^*$) iff $Y_P^{t,\delta} = \emptyset$.*

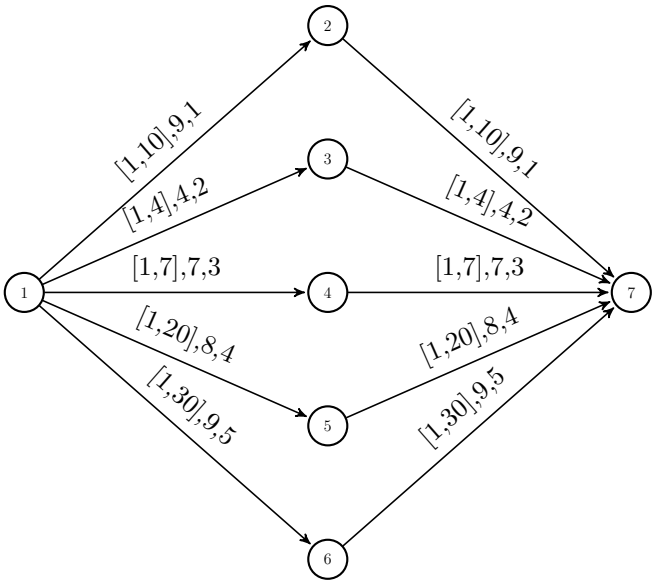


Figure 16: Example of an instance when $z^{t,\delta} = z_R^{t,*}$ does not imply that $w^{t,\delta} = w^*$. The labeling of the arcs is given by $[\ell_a, u_a], c_a, d_a$.

Proof. Assume that $Y_P^{t,\delta} \neq \emptyset$, then there exist x such that $Y_P^{t,\delta}(x) \neq \emptyset$, and this immediately implies that $w_R^{t,\delta} < w^*$. Hence, since by assumption $w^{t,\delta} = w_R^{t,\delta}$, then $w^{t,\delta} < w^*$ and $x^{t,\delta}$ cannot be an optimal solution of the full-information problem. Conversely, suppose $x^{t,\delta}$ is not an optimal full information solution. Then, as $w^{t,\delta} = w_R^{t,\delta}$, it follows that $w_R^{t,\delta} < w^*$, and hence there exist $\hat{y} \in Z(x^*)$ such that $\boldsymbol{d}^\top y < \boldsymbol{d}^\top \hat{y}$ for all $Z_R^t(x^{t,\delta})$. In addition, it must be the case that $\hat{y} \notin Z_R^t(x^*)$; as if $\hat{y} \in Z_R^t(x^*)$, then $x^{t,\delta}$ is not a greedy solution at time t, contradicting the fact that $\delta \in \Delta$. It can be concluded that $\hat{y} \in Y_P^{t,\delta}(x^*)$, as desired. ∎

Although Lemma 11 gives necessary and sufficient conditions for the 'sandwich' theorem to hold, identifying when this conditions hold (i.e., whether $Y_P^{t,\delta}$ is empty or not) requires the knowledge of $Z(x)$, which in turn requires a precise knowledge of $\boldsymbol{c}$. Hence, in general, determining whether $w^{t,\delta} = w_R^{t,\delta}$ implies convergence to the optimal solution has no straightforward answer. Moreover, even finding sufficient conditions for $Y_P^{t,\delta}$ to be empty (or non-empty) is not straightforward, see Remark 13. Nevertheless, if it turns out that for

a particular instance or for a particular class of problems it is possible to compute $Y_P^{t,\delta}$ at each time $t \in \mathcal{T}$, then one can obtain a policy that has a time-stability that is linearly upper bounded, see Remark 14.

Remark 13. A sufficient condition for the set $Y_P^{t,\delta}$ to be empty can be given by determining whether a finite (although exponentially large) sequence of sets are empty. To this end, assume that $\mathcal{U}^t$ is bounded, let the set of its extreme points be $\text{ext}(\mathcal{U}^t) = \{\boldsymbol{c}^{(1)}, \dots, \boldsymbol{c}^{(U)}\}$ and let $Y_P^{t,\delta}(\boldsymbol{c}^{(j)})$ be defined as $Y_P^{t,\delta}$ by replacing $\boldsymbol{c}$ by $\boldsymbol{c}^{(j)}$ in the definition of $Z(x)$. Then, since $\mathcal{U}^t$ is a bounded polytope, $Y_P^{t,\delta} = \emptyset$ if $\bigcup_{j=1}^{U} Y_P^{t,\delta}(\boldsymbol{c}^{(j)}) = \emptyset$. Therefore, if $Y_P^{t,\delta}(\boldsymbol{c}^{(j)}) = \emptyset$ for all $j = 1, \dots, U$, then $Y_P^{t,\delta} = \emptyset$.

Remark 14. Assume that for certain instance or class of problem it is possible to determine whether or not $Y_P^{t,\delta} = \emptyset$. Then the following result holds:

Lemma 12. *Let $\delta \in \Delta$ and $t \in \mathcal{T}$ be given, assume that $w^{t,\delta} = w_R^{t,\delta}$, and suppose that $Y_P^{t,\delta} \neq \emptyset$. Let $\pi \in \Pi$ be a policy such that $\pi^s = \delta^s$ for all $s \leq t$ and $x^{t+1,\pi} \in X_P^{t,\delta}$. Then,* $\dim(\mathcal{U}^{t+2}) < \dim(\mathcal{U}^{t+1})$.

Proof. Observe that if the leader implements $x^{t+1,\pi}$ then the follower, by the optimistic assumption, implements at time $t+1$ a solution $\hat{y} \in Y_P^{t,\delta}$. By **C.1**, $\hat{y} \notin Z_R^t(x^{t+1,\delta})$, which implies, by the same arguments of Lemma 10, that $\hat{y}_a > 0$ for some $a \notin \widetilde{A}^{t+1}$. This gives the desired result. ∎

Let $t \in \mathcal{T}$ be such that a policy $\delta \in \Delta$ stalls. Two things can occur: First, if $Y_P^{t,\delta} = \emptyset$, then $x^{t,\delta}$ is the optimal solution, and the leader can keep implementing that solution from there on. Otherwise, the leader can implement a solution in $X_P^{t,\delta}$ and the follower is forced to reveal new information.

These observations motivate us to define a new set of policies Φ. We say that $\phi \in \Phi$ if and only if

$$x^{t,\phi} \in \begin{cases} X_P^{t-1,\phi}, & \text{if } \widetilde{A}^t \setminus \widetilde{A}^{t-1} = \emptyset \text{ and } X_P^{t-1,\phi} \neq \emptyset, \\ \{x^{t-1,\phi}\}, & \text{if } \widetilde{A}^t \setminus \widetilde{A}^{t-1} = \emptyset \text{ and } X_P^{t-1,\phi} = \emptyset\ , \\ \arg\max\{w_R^t(x) \colon x \in X\}, & \text{otherwise.} \end{cases} \tag{4.16}$$

In the above definition we make the convention that $\widetilde{A}^0 \setminus \widetilde{A}^{-1} \neq \emptyset$ and that $X^{t,\phi}$ is defined in the same way as $X^{t,\delta}$ for any $t \in \mathcal{T}$. We note that the set of policies ϕ is well-defined as it is readily seen that if the leader implements a solution of $X_P^{t-1,\phi}$ at time t, then $\widetilde{A}^t \setminus \widetilde{A}^{t-1} \neq \emptyset$ by Lemma 12.

From Lemmas 10 and 12, and the usual sufficiency condition when $\dim(\mathcal{U}^t) = 0$ (cf. Proposition 4) we have the following result:

Theorem 6. *Let $\phi \in \Phi$ be given. Then $\tau^\phi \leq 2|A|$. Moreover, policies in ϕ provide a real time certificate of optimality.*

Therefore, policies in Φ have a linear time-stability performance. ■

4.4 GREEDY AND BEST–CASE POLICIES

In this section we present a class of deterministic policies that are able to provide a certificate of optimality in real time, and can provide a finite upper bound on the time-stability. Although in general the bound is worst-case exponential, in practice the policies find an optimal solution in far less time periods. In addition, these policies can handle a broad class of uncertainty sets without compromising the tractability of the model. In particular, it can be assumed that the uncertainty set is non-convex and still get tractable mixed–integer programming formulations. This flexibility contrasts with robust (i.e., worst-case) approaches, where uncertainty sets must be convex (Ben-Tal et al. 2009).

4.4.1 Definition and General Convergence Results

Consider the following mathematical program

$$w^{t,E} = \max_{x,y,\hat{c}} \boldsymbol{d}^\top y \tag{4.17a}$$

$$\text{s.t. } x \in X \tag{4.17b}$$

$$\hat{c} \in \mathcal{U}^t \tag{4.17c}$$

$$y \in \arg\min\{\hat{c}^\top y' : y' \in Y(x)\}, \tag{4.17d}$$

and let $\mathcal{S}^t$ be the set of optimal solutions, i.e., $\mathcal{S}^t = \arg\max\{\boldsymbol{d}^\top y\colon$ (4.17b)–(4.17d) hold$\}$. We define the time $\xi \in \mathcal{T}$ as the first time that the best observed solution coincides with the solution the leader 'expects' if she solves the above mathematical program:

$$\xi = \min\{t \in \mathcal{T}\colon W^t = w^{t,E}\} \tag{4.18}$$

where $W^t = \max\{w^s\colon s \leq t\}$ (with $W^0 = -\infty$), and we let $s(\xi) \leq \xi$ be the time attaining the maximum in W^ξ. We define the set of policies Ψ as follows:

Definition 8. We say that $\psi \in \Psi$ if and only if $x^{t,\psi} \in \text{Proj}_{y,\hat{c}}(\mathcal{S}^t)$ for all $t \leq \xi$, and $x^{t,\psi} = x^{s(\xi),\psi}$ for all $t > \xi$.

Observe that if the leader implements $x^{t,\psi}$, then there exists a point $(x^{t,\psi}, y^{t,E}, c^{t,E}) \in \mathcal{S}^t$ such that $w^{t,E} = \boldsymbol{d}^\top y^{t,E}$. We call $y^{t,E}$ the response that the leader *expects* the follower will use at time t, and $c^{t,E}$ the *estimated* cost vector at time t. In addition, we define $z^{t,E} = (\boldsymbol{c}^{t,E})^\top y^{t,E}$ as the cost of the follower's problem the leader expects at time t. Note that there might be many optimal solutions of (4.17a)–(4.17d); hence, we make the assumption that when this is the case the leader selects $(x^{t,\psi}, y^{t,E}, c^{t,E})$ according to some fixed rule.

Policies in the set Ψ can be considered as greedy and best-case policies. They are greedy, since as with policies in Δ, they seek to optimize the leader's immediate performance. They are best-case, as the leader assumes that the follower's cost vector realizes its best possible value from her point of view. We note that this is similar to what policies in Λ do in the max-min setting. Specifically, by using any policy in λ the leader implicitly assumes that the follower's cost vector realizes its worst case for the follower. Given the max-min relationship between the leader's and follower's objective function, this is the same as saying that the cost vector realizes its best-case for the leader.

Theorem 7. *Suppose $\psi \in \Psi$ is given, let $W^{t,\psi} = \max\{w^{t',\psi}\colon t' \leq t\}$ for all $t \geq 1$ (with $W^0 = -\infty$), and let s be the time-period where the maximum in $W^{t,\psi}$ is attained. If $w^{t,E} \leq W^{t,\psi}$, then $x^{s,\psi}$ is an optimal solution of the full-information problem (4.1).*

Proof. Let $x^* \in X$ and $y^* \in \arg\min\{\boldsymbol{c}^\top y\colon y \in Y(x^*)\}$ be such that $\boldsymbol{d}^\top y^* = w^*$. That is, (x^*, y^*) is an optimal solution of the full-information bilevel problem. Since $\boldsymbol{c} \in \mathcal{U}^t$ for all $t \in \mathcal{T}$, then $(x^*, y^*, \boldsymbol{c}) \in \mathcal{S}^t$, and thus, $w^* \leq w^{E,t}$ for all $t \in \mathcal{T}$. In addition, since $x^{t',\psi} \in X$

for all $t' \in \mathcal{T}$, then, from the definition of x^*, we have that $w^{t',\psi} \leq w^*$ for any given $t' \in \mathcal{T}$. These observations imply that for any $t, t' \in \mathcal{T}$ we have the following chain of inequalities (or 'sandwich' result)

$$w^{t',\psi} \leq w^* \leq w^{E,t}. \tag{4.19}$$

Hence, since $w^{E,t} \leq W^{t,\psi}$, then $w^{E,t} \leq w^{s,\psi}$, and Equation (4.19) implies that $w^{s,\psi} = w^*$, which gives the desired result. ■

Theorem 7 can be viewed as an analogous of the 'sandwich' theorem for the max-min case (cf. Theorem 3 in Section 3.3.1). Importantly, it implies that by using policies in Ψ the leader can get a certificate of optimality in real time, as by Standard Feedback, she is aware of the value of $W^{t,\psi}$ for all $t \in \mathcal{T}$. Hence, as soon as time ξ happens, the leader can be made sure that $x^{s(\xi),\psi}$ is an optimal solution of the full-information problem.

One of the main disadvantages of using a policy in Ψ is that a suboptimal solution $x^{s,\psi}$ can be repeated at a later point in time. From the leader's standpoint, it makes sense to repeat $x^{s,\psi}$ at time $t > s$ as long as $w^{t,E} = w^{s,\psi}$. If this is the case, then $w^{t,E} = W^{t,\psi}$ and convergence can be assured. Unfortunately, in general it is easy to come with examples where $x^{s,\psi}$ is repeated at a point t, and it holds that both $w^{t,E} > w^{s,\psi}$ and $w^{t,E} > W^{t,\psi}$. Hence, in such settings it is reasonable for the leader to avoid implementing $x^{s,\psi}$ at time t.

In order for policies in ψ to avoid repeating solutions in the way described above, we include the following constraint in the definition of the policies in Ψ for $t \geq 1$:

$$\boldsymbol{d}^\top y - w^{s,\psi} \leq D\|x - x^{s,\psi}\|_1 \qquad \forall\ s = 1, \ldots, t-1. \tag{4.20}$$

In this constraint, D is a sufficiently large constant such that $D\|x - x^{s,\psi}\|_1$ upper bounds $\boldsymbol{d}^\top y - w^{s,\psi}$. Constraint (4.20) can be enforced without adding any new variables if X is a binary set; in any other case it can be enforced using typical integer programming modeling techniques by adding (at most $|A|$ of each) constraints and binary variables.

We note that constraint (4.20) is *valid* as it does not remove the real cost vector $\boldsymbol{c}$ from the feasible region $\mathcal{S}^t$. To see why, note that if $x \neq x^{s,\psi}$, then $(x, y, \hat{c})$ satisfies (4.17a)–(4.17d) if and only if it satisfies (4.17a)–(4.17d) and (4.20). On the other hand, if $x = x^{s,\psi}$, then by the optimistic assumption and the definition of $w^{s,\psi}$, it must be the case that $\boldsymbol{d}^\top y - w^{s,\psi} \leq 0$ for all $y \in \arg\min\{\boldsymbol{c}^\top y' \colon y' \in Y(x)\}$.

Importantly, it can be readily seen that if a policy in ψ repeats a previous solution at time $t \leq \xi$, then it must be the case that $t = \xi$, and optimality can be guaranteed. For this reason, from now on, we assume that the constraints defined in (4.20) are included in the mathematical program (4.17a)–(4.17d).

Clearly, the optimization problem (4.17a)–(4.17d) is a non-convex, NP-hard problem, as it is a bilinear bilevel problem. However, if the lower-level problem $Y(x)$ is continuous, and the uncertainty sets $\mathcal{U}^t$ can be represented via linear constraints, then the decisions of policy ψ can be computed by solving a mixed-integer problem (MIP).

Proposition 11. *Assume that the follower's variables are continuous and that for any $x \in X$ and $\hat{c} \in \mathcal{U}^0$ the follower's problem has an optimal solution. In addition, suppose that for any $t \in \mathcal{T}$ the set $\mathcal{U}^t$ is polyhedral, that is,*

$$\mathcal{U}^t = \{\hat{c} \in \mathbb{R}^{|A|} \colon \exists v \in \mathbb{R}^{b^t - n^t} \times \mathbb{Z}^{n^t} \ \ s.t. \ \boldsymbol{G}^t \hat{c} + \boldsymbol{J}^t v \leq \boldsymbol{g}^t\}, \tag{4.21}$$

where $b^t, n^t \geq 0$, $\boldsymbol{G}^t$ and $\boldsymbol{J}^t$ are matrices of sizes $u^t \times |A|$ and $u^t \times b^t$, respectively. Then the mathematical program (4.17a)–(4.17d) (along with (4.20)) can be formulated as the following MIP:

$$\max \ \boldsymbol{d}^\top y \tag{4.22a}$$

$$s.t. \ \boldsymbol{H}x \leq \boldsymbol{h} \tag{4.22b}$$

$$\boldsymbol{G}^t \hat{c} + \boldsymbol{J}^t v \leq \boldsymbol{g}^t \tag{4.22c}$$

Equation (4.20)

$$\boldsymbol{F}y + \boldsymbol{L}x \leq \boldsymbol{f} \tag{4.22d}$$

$$-\boldsymbol{F}^\top p - \hat{c} \leq \boldsymbol{0} \tag{4.22e}$$

$$\boldsymbol{f} - \boldsymbol{L}x - \boldsymbol{F}y \leq \boldsymbol{M}^p \hat{u} \tag{4.22f}$$

$$p \leq \boldsymbol{M}^p(\boldsymbol{1} - \hat{u}) \tag{4.22g}$$

$$\hat{c} + \boldsymbol{F}^\top p \leq \boldsymbol{M}^y \hat{v} \tag{4.22h}$$

$$y \leq \boldsymbol{M}^y(\boldsymbol{1} - \hat{v}) \tag{4.22i}$$

$$y \in \mathbb{R}^{|A|}_+, p \in \mathbb{R}^{|C_F|}_+, x \in \mathbb{R}^{|I|-k}_+ \times \mathbb{Z}^k_+, \hat{c} \in \mathbb{R}^{|A|} \tag{4.22j}$$

$$\hat{u} \in \{0,1\}^{|C_F|}, \hat{v} \in \{0,1\}^{|A|}, \ell \in \mathbb{R}^{b^t - n^t} \times \mathbb{Z}^{n^t}. \tag{4.22k}$$

In this formulation $\boldsymbol{M}^p$ *is a diagonal matrix of the appropriate dimensions, where* $\max\{p_d, (\boldsymbol{f} - \boldsymbol{L}x - \boldsymbol{F}y)_d\} \leq \boldsymbol{M}^p_d$ *for all* $d \in C_F$ *and all feasible* p, y *and* x. *Similarly,* $\boldsymbol{M}^y$ *is a diagonal matrix of the appropriate dimensions, where* $\max\{y_a, (\hat{c} + \boldsymbol{F}^\top p)_a\} \leq \boldsymbol{M}^y_a$ *for all* $a \in A$ *and all feasible* p *and* $\hat{c}$.

Proof. The result follows by replacing the follower's problem by its Karush-Kuhn-Tucker (KKT) conditions, i.e., by enforcing the lower-level primal feasibility, dual feasibility, and complementary slackness, see, e.g., Dempe (2002). ∎

Under the assumptions of Proposition 11, and from the previous discussion regarding the repetition of solutions, we have the following finite upper bound on the time-stability of policies in Ψ:

Theorem 8. *Suppose* $\psi \in \Psi$ *and that the assumptions of Proposition 11 hold. Then* $\tau^\psi \leq |ext(conv(X))|$, *where* $ext(A)$ *is the set of extreme points of set* A.

It is important to observe that the derivation of this bound is largely independent of the way the uncertainty set is updated. In the following sections, we discuss how more specific updating mechanisms can lead to improvements in the time-stability of the policies in Ψ.

In this sense, we begin our discussion with what we call the *basic update*, where at each time the set $\mathcal{U}^t$ is updated by adding to it a linear constraint. Subsequently we discuss more general *convex* and *non-convex updates*, where $\mathcal{U}^t$ is updated using more complicated mechanisms.

4.4.2 The Basic Uncertainty Set Update

Assume that the leader uses a policy $\psi \in \Psi$ and suppose that she has only Standard Feedback available, thus she only observes the values of $w^{t,\psi}$ and $z^{t,\psi}$ at each time t. We define the basic update at the end of time t by

$$\mathcal{U}^{t+1} = \mathcal{U}^t \cap \mathcal{L}^t, \qquad \text{where } \mathcal{L}^t = \{\hat{c} \in \mathbb{R}^{|A|} \colon (y^{t,E})^\top \hat{c} > z^{t,\psi}\}. \tag{4.23}$$

We have the following simple observation.

Lemma 13. *Let* $t \in \mathcal{T}$ *be given such that* $t < \xi$ *(thus,* $W^{t,\psi} < w^{t,E}$*), then* $\boldsymbol{c} \in \mathcal{L}^t$. *Moreover, if* $z^{t,E} \leq z^{t,\psi}$, *then* $c^{t,E} \notin \mathcal{L}^t$.

Proof. Observe that as $w^{t,\psi} < w^{t,E}$, then $y^{t,E} \notin Z(x^{t,\psi})$. This follows by contradiction: If $y^{t,E} \in Z(x^{t,\psi})$, then by the optimistic assumption it would follow that $w^{t,E} = w^{t,\psi}$, yielding a contradiction. Hence, it can be concluded that $\boldsymbol{c}^\top y^{t,E} > z^{t,\psi}$ as desired. On the other hand, if $z^{t,E} \leq z^{t,\psi}$, then by definition $(c^{t,E})^\top y^{t,E} \leq z^{t,\psi}$ and it is clear that this implies that $c^{t,E} \notin \mathcal{L}^t$. ∎

The previous lemma implies that the basic update is valid (as it does not remove the real cost vector $\boldsymbol{c}$), and it 'shrinks' the leader's uncertainty by removing $c^{t,E}$ from the set of possible cost vectors, as long as $z^{t,E} \leq z^{t,\psi}$. Unfortunately, it is readily seen that it does not remove $c^{t,E}$ whenever $z^{t,E} > z^{t,\psi}$.

One potential drawback of using the basic update is that it is given by an open inequality, and as such, it cannot be handled directly by most optimization solvers. This issue can be overcome easily if $z^{t,E} < z^{t,\psi}$. In this case putting a '$\geq$' sign in the definition of $\mathcal{L}^t$ makes the update equally valid and it also removes $c^{t,E}$. On the other hand, in general, one can put a '$\geq$' sign in the definition of $\mathcal{L}^t$ by adding a small-enough term ϵ^t in the right-hand side. The value of such ϵ^t depends on the specific problem and might be easy to obtain. For instance, if it is known from the structure of the problem at hand that $z^{t,\psi}$ is integer for any possible values of t, then $\epsilon^t = 1$ for all $t \in \mathcal{T}$.

Now consider that the leader has access to Value–Perfect or Response–Perfect feedback. In such settings, the uncertainty set can be further updated by adding the following linear constraints in addition to $\mathcal{L}^t$

$$\mathcal{U}^{t+1} = \mathcal{U}^t \cap \mathcal{L}^t \cap \mathcal{V}^t, \qquad \text{(for Value–Perfect feedback)}, \tag{4.24}$$

$$\mathcal{U}^{t+1} = \mathcal{U}^t \cap \mathcal{L}^t \cap \mathcal{R}^t, \qquad \text{(for Response–Perfect feedback)}, \tag{4.25}$$

where we define $\mathcal{V}^t$ and $\mathcal{R}^t$ as

$$\mathcal{V}^t = \{\hat{c} \in \mathbb{R}^{|A|} \colon \hat{c}_a = c_a \quad \forall a \text{ s.t. } y^t_a > 0\}, \tag{4.26}$$

$$\mathcal{R}^t = \{\hat{c} \in \mathbb{R}^{|A|} \colon (y^t)^\top \hat{c} = z^{t,\psi}\}. \tag{4.27}$$

The additional feedback, either Value–Perfect or Response–Perfect, assures that $\mathcal{U}^{t+1} \setminus \mathcal{U}^t \neq \emptyset$, i.e., the leader's uncertainty always shrink independent of the value of $z^{t,\psi}$. These assertions are proven below.

Lemma 14. *Let $\psi \in \Psi$, suppose that $t < \xi$, and that feedback is Value–Perfect. Define $\widetilde{A}^t = \{a \in A\colon \hat{c}_a = c_a \text{ for all } \hat{c} \in \mathcal{U}^t\}$. If $z^{t,E} > z^{t,\psi}$ then there exist an $a \notin \widetilde{A}^t$ such that $y^t_a > 0$. Moreover, $c^{t,E} \notin \mathcal{V}^t$.*

Proof. The proof is by contradiction. Suppose that $y^t_a = 0$ for all $a \notin \widetilde{A}^t$. Then, for any $\hat{c} \in \mathcal{U}^t$, $\hat{c}^\top y^t = \boldsymbol{c}^\top y^t = z^{t,\psi}$. In particular $c^{E,t} \in \mathcal{U}^t$, hence we would have that $(c^{t,E})^\top y^t = z^{t,\psi} < z^{t,E} = (c^{t,E})^\top y^{t,E}$. This implies that, $y^{t,E} \notin Z(x^{t,\psi})$, which is a contradiction. Finally, observe that the same argument shows that the fact that $c^{t,E} \in \mathcal{V}^t$ yields a contradiction. ∎

Simply speaking, the previous lemma states that if $z^{t,\psi} < z^{t,E}$, then the follower must reveal the cost of an activity with his response of time t. Unfortunately, it is easy to come with examples where this result does not hold when $z^{t,E} \leq z^{t,\psi}$. This implies that Lemma 7 in Section 3.3.2 of the max-min setting does not apply for this class of bilevel problems.

Lemma 15. *Let $\psi \in \Psi$, suppose that $t < \xi$, and that feedback is Response–Perfect. If $z^{t,E} > z^{t,\psi}$, then $\dim(\mathcal{U}^{t+1}) < \dim(\mathcal{U}^t)$. Moreover, $c^{t,E} \notin \mathcal{R}^t$.*

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

$$\{\hat{c} \in \mathbb{R}^{|A|}\colon (y^s)^\top \hat{c} = z^{s,\psi},\ s \leq t-1\} = \{\hat{c} \in \mathbb{R}^{|A|}\colon (y^s)^\top \hat{c} = z^{s,\psi},\ s \leq t\}. \tag{4.28}$$

In particular, since $c^{t,E} \in \mathcal{U}^t$, then $\hat{c} \in \{\hat{c} \in \mathbb{R}^{|A|}\colon (y^s)^\top \hat{c} = z^{s,\psi},\ s \leq t-1\}$, and by Equation (4.28) it follows that $(c^{t,E})^\top y^t = z^{t,\psi}$. Now, as we are assuming that $z^{t,E} > z^{t,\psi}$, it follows that $(c^{t,E})^\top y^{t,E} > c^{t,E,\top} y^t$, i.e., that $y^{t,E} \notin \arg\min\{(c^{t,E})^\top y\colon y \in Y(x^{t,\psi})\}$, which is a contradiction. Finally, observe that the above arguments imply that $c^{t,E} \notin \mathcal{R}^t$. ∎

Observe that as with Value–Perfect feedback update, if $z^{t,E} \leq z^{t,\psi}$, then nothing can be said with respect to the reduction of dimension of the uncertainty set.

4.4.3 The Convex Uncertainty Set Update

In this section we discuss an additional updating mechanism that can potentially improve the efficacy of the linear update by better exploiting the information that $z^{t,\psi}$ gives. Specifically, note that if the leader observes the value of $z^{t,\psi}$, then she can be made sure that $z^{t,\psi} \leq \min\{\boldsymbol{c}^\top y\colon y \in Y(x^{t,\psi})\}$ (in fact she can be made sure that the inequality holds as an equality,

this is the subject of the next section). This observation motivates the following updating procedure, which we refer to as the *convex update*:

$$\mathcal{U}^{t+1} = \mathcal{U}^t \cap \mathcal{L}^t \cap \mathcal{C}^t, \tag{4.29}$$

where

$$\mathcal{C}^t = \{\hat{c} \in \mathbb{R}^{|A|} \colon z^{t,\psi} \leq \hat{c}^\top y \ \forall \ y \in Y(x^{t,\psi})\}. \tag{4.30}$$

Clearly, $\mathcal{C}^t$ is a convex set as it is the intersection of (possibly infinite number of) linear inequalities.

In general, the set $\mathcal{C}^t$ can be considered as a system of *semi-infinite* linear constraints, as $Y(x^{t,\psi})$ might be a set of infinite cardinality. We note that $\mathcal{C}^t$ is a polyhedron (as one can only consider the extreme points of $Y(x^{t,\psi})$ with out loss of generality). However, the set of extreme points of $Y(x^{t,\psi})$ is exponentially large in $|A|$ and $|C_F|$ in general.

It turns out that the set $\mathcal{C}^t$ can be represented by only considering $1+|A|$ linear constraints, by adding $|C_F|$ new continuous variables:

Lemma 16. *Let $t \in \mathcal{T}$ be given and suppose that the follower's problem has an optimal solution for any $x \in X$. Then*

$$\mathcal{C}^t = \{\hat{c} \in \mathbb{R}^{|A|} \colon \exists q^t \in \mathbb{R}_+^{|C_F|} \ s.t. \ (\boldsymbol{L}x^{t,\psi} - \boldsymbol{f})^\top q^t \geq z^{t,\psi}, -\boldsymbol{F}^\top q^t - \hat{c} \leq 0\}. \tag{4.31}$$

Proof. Observe that $\hat{c} \in \mathcal{C}^t$ if and only if

$$z^{t,\psi} \leq \min\{\hat{c}^\top y \colon y \in Y(x^{t,\psi})\}, \tag{4.32}$$

and by strong duality, this is equivalent to

$$z^{t,\psi} \leq \max\{(\boldsymbol{L}x^{t,\psi} - \boldsymbol{f})^\top q^t \colon \ -\boldsymbol{F}^\top q^t - \hat{c} \leq 0, q^t \geq 0\}. \tag{4.33}$$

Moreover, the previous equation holds if and only if there exists a point q^t feasible in the dual program such that $z^{t,\psi} \leq (\boldsymbol{L}x^{t,\psi} - \boldsymbol{f})^\top q^t$. ∎

From the above result, it follows that if the uncertainty set is updated by using the convex update, then $\mathcal{U}^t$ is polyhedral for all t. Particularly, its representation in terms of

matrixes $\boldsymbol{G}^t$ and $\boldsymbol{J}^t$, as given by Equation (4.21) yields that $\boldsymbol{G}^t$ is a $(u^0 + t(2 + |A|)) \times |A|$ matrix given by

$$\boldsymbol{G}^t = \big(\boldsymbol{G}^0; -(y^{1,E})^\top; -(y^{2,E})^\top; \ldots; -(y^{t,E})^\top; -\boldsymbol{I}'; -\boldsymbol{I}'; \ldots; -\boldsymbol{I}'\big) \tag{4.34}$$

with $\boldsymbol{I}' = (\boldsymbol{0}^\top; \boldsymbol{I})$, where $\boldsymbol{0}^\top$ is a $1 \times |A|$ vector of zeros and $\boldsymbol{I}$ is a $|A| \times |A|$ identity matrix. In addition, matrix $\boldsymbol{J}^t$ is given by $\boldsymbol{J}^t = (\boldsymbol{0}; \tilde{\boldsymbol{J}}^t)$, where $\boldsymbol{0}$ is a $(u^0 + t) \times t|C_F|$ matrix of zeros, and $\tilde{\boldsymbol{J}}^t$ is a block-diagonal matrix with t blocks, where the s-th block, $1 \leq s \leq t$, is a $(1 + |A|) \times |C_F|$ matrix given by $\big(-(\boldsymbol{L}x^{s,\psi} - \boldsymbol{f})^\top; -\boldsymbol{F}^\top\big)$. The right-hand side vector of the representation satisfies

$$\boldsymbol{g}^t = (\boldsymbol{g}^0; -z^{1,\psi} - \epsilon^1; -z^{2,\psi} - \epsilon^2; \ldots; -z^{t,\psi} - \epsilon^t; \boldsymbol{0}^1; \boldsymbol{0}^2; \ldots; \boldsymbol{0}^t) \tag{4.35}$$

where $\boldsymbol{0}^s = (-z^{s,\psi}; \boldsymbol{0})$, with $\boldsymbol{0}$ being a $|A| \times 1$ vector of all zeros.

Interestingly, given the structure of the above linear representation of the uncertainty set, the MIP (4.22a)–(4.22k) can be viewed as the extensive form a two-stage stochastic mixed-integer problem (SMIP) with continuous second-stage variables, and mixed-integer first-stage variables. Indeed, note that (4.22a)–(4.22k) can be formulated as

$$\max\ \boldsymbol{d}^\top y + E[g(\hat{c}, \boldsymbol{\xi})] \tag{4.36a}$$

$$\text{s.t. } \boldsymbol{H}x \leq h,\ \boldsymbol{G}^0\hat{c} \leq \boldsymbol{g}^0,\ \boldsymbol{F}y + \boldsymbol{L}x \leq \boldsymbol{f},\ -\boldsymbol{F}^\top p - \hat{c} \leq 0 \tag{4.36b}$$

$$\boldsymbol{f} - \boldsymbol{L}x - \boldsymbol{F}y \leq \boldsymbol{M}^p\hat{u},\ p \leq \boldsymbol{M}^p(\boldsymbol{1} - \hat{u}) \tag{4.36c}$$

$$\hat{c} + \boldsymbol{F}^\top p \leq \boldsymbol{M}^y\hat{v},\ y \leq \boldsymbol{M}^y(\boldsymbol{1} - \hat{v}) \tag{4.36d}$$

$$\hat{u} \in \{0,1\}^{|C_F|}, \hat{v} \in \{0,1\}^{|A|}, y \in \mathbb{R}_+^{|A|}, p \in \mathbb{R}_+^{|C_F|} \tag{4.36e}$$

$$x \in \mathbb{R}_+^{|I|-k} \times \mathbb{Z}_+^k, \hat{c} \in \mathbb{R}^{|A|}, \tag{4.36f}$$

where $\boldsymbol{\xi}$ is a 'random' vector that takes values on the discrete set

$$\boldsymbol{\xi} \in \big\{(x^{s,\psi}, z^{s,\psi}, y^{s,E})\colon s \leq t-1\big\}. \tag{4.37}$$

For a given realization ξ^s of the random vector, the second-stage problem function $g(\hat{c}, \xi^s)$ is given by

$$g\big(\hat{c}, \xi^s\big) = \max\ 0 \tag{4.38a}$$

$$\text{s.t. } (\boldsymbol{L}x^{s,\psi} - \boldsymbol{f})^\top q = z^{s,\psi} \tag{4.38b}$$

$$-\boldsymbol{F}^\top q - \hat{c} \leq 0 \tag{4.38c}$$

$$\hat{c}^\top y^{s,E} \geq z^{s,\psi} + \epsilon^s \tag{4.38d}$$

$$q \in \mathbb{R}_+^{|C_F|}. \tag{4.38e}$$

Note that from the standpoint of the SMIP formulation, the distribution function of $\boldsymbol{\xi}$ is irrelevant as the second-stage value is always zero or $-\infty$ (in case of infeasibility) for any realization of $\boldsymbol{\xi}$.

We note that the SMIP equivalence suggests that SMIP techniques can be used to solve the MIP (4.22a)–(4.22k). In particular, decomposition techniques, such as Bender's, can provide algorithmic advantages over directly feeding the extensive form formulation into an MIP solver.

To close this section, observe that we can enhance the convex update by assuming Value–Perfect or Response–Perfect feedbacks. If such feedbacks are assumed, then the update becomes

$$\mathcal{U}^{t+1} = \mathcal{U}^t \cap \mathcal{L}^t \cap \mathcal{C}^t \cap \mathcal{V}^t \text{ and } \mathcal{U}^{t+1} = \mathcal{U}^t \cap \mathcal{L}^t \cap \mathcal{C}^t \cap \mathcal{R}^t, \tag{4.39}$$

for Value–Perfect and Response–Perfect feedback, respectively, where $\mathcal{V}^t$ and $\mathcal{R}^t$ are defined as in equations (4.26)–(4.27).

4.4.4 The Non-Convex Uncertainty Set Update

The non-convex update generalizes the convex update discussed in the previous section. Note that at any given time $t \in \mathcal{T}$, the leader knows by definition that the real cost vector $\boldsymbol{c}$ satisfies $z^{t,\psi} = \min\{\boldsymbol{c}^\top y \colon y \in Y(x^{t,\psi})\}$. Hence, $\boldsymbol{c}$ belongs to a set given by

$$\tilde{\mathcal{N}}^t = \left\{\hat{c} \in \mathbb{R}^{|A|} \colon z^{t,\psi} = \min\{\hat{c}^\top y \colon y \in Y(x^{t,\psi})\}\right\}. \tag{4.40}$$

Equivalently, this set can be represented as

$$\tilde{\mathcal{N}}^t = \left\{\hat{c} \in \mathbb{R}^{|A|} \colon \exists y \in \arg\min\{\hat{c}^\top y' \colon y' \in Y(x^{t,\psi})\} \text{ s.t. } z^{t,\psi} = \hat{c}^\top y\right\}. \tag{4.41}$$

Moreover, given that it is known that $\boldsymbol{d}^\top y^t = w^{t,\psi}$, the above set can be further restricted to

$$\mathcal{N}^t = \left\{\hat{c} \in \mathbb{R}^{|A|} \colon \exists y \in \arg\min\{\hat{c}^\top y' \colon y' \in Y(x^{t,\psi})\} \text{ s.t. } z^{t,\psi} = \hat{c}^\top y, \boldsymbol{d}^\top y = w^{t,\psi}\right\}. \tag{4.42}$$

The *non-convex* update is defined as

$$\mathcal{U}^{t+1} = \mathcal{U}^t \cap \mathcal{L}^t \cap \mathcal{N}^t \qquad \forall t \in \mathcal{T}. \tag{4.43}$$

Clearly, the update defined by Equation (4.43) is valid and generalizes the convex update in the sense that $\mathcal{N}^t \subseteq \mathcal{C}^t$. Unfortunately, it makes the computation of $\mathcal{U}^t$ significantly more challenging, as determining $\mathcal{U}^t$ via this update requires solving a non-convex optimization problem (hence the name). Indeed, assuming that the follower's problem has an optimal solution for any $\hat{c} \in \mathcal{U}^0$ and $x \in X$, and using strong duality, we have that $\hat{c} \in \tilde{\mathcal{N}}^t$ if and only if

$$z^{t,\psi} = \max\{(\boldsymbol{L}x^{t,\psi} - \boldsymbol{f})^\top q^t \colon -\boldsymbol{F}^\top p - \hat{c} \leq 0\}. \tag{4.44}$$

In contrast with Equation (4.33), the above assertion cannot be replaced by one assuring the existence of one q^t that is dual feasible due to the equality sign. Hence, in order to remove the maximization in equation (4.44) we can use the KKT conditions. As such, the set $\mathcal{N}^t$ can be represented as

$$\mathcal{N}^t = \bigcup_{E^t \subseteq A, D^t \subseteq C_F} \mathcal{N}^t(E^t, D^t), \tag{4.45}$$

where for any given $E^t \subseteq A$, $D^t \subseteq C_F$, the set $\mathcal{N}^t(E^t, D^t)$ is defined by

$$\begin{aligned}
\mathcal{N}^t(E^t, D^t) = \Big\{\hat{c} \in \mathbb{R}^{|A|} \colon \exists r^t \in \mathbb{R}_+^{|A|}, q^t \in \mathbb{R}_+^{|C_F|} \text{ s.t.} \\
& \boldsymbol{F}r^t \leq \boldsymbol{f} - \boldsymbol{L}x^{t,\psi} \\
& -\boldsymbol{F}^\top q^t - \hat{c} \leq 0 \\
& (\boldsymbol{L}x^{t,\psi} - \boldsymbol{f})^\top q^t = z^{t,\psi} \\
& \boldsymbol{d}^\top r^t = w^{t,\psi} \\
& r_a^t = 0 && \forall\, a \in E^t \\
& \hat{c}_a + \boldsymbol{F}_a^\top q^t = 0 && \forall\, a \in A \setminus E^t \\
& q_d^t = 0 && \forall\, d \in D^t
\end{aligned}$$

$$\boldsymbol{F}_d r^t = f_d - \boldsymbol{L}_d x^{t,\psi} \qquad \forall\, d \in C_F \setminus D^t \Big\}.$$

In the above definition, $\boldsymbol{F}_a^\top$ is the a-th row of the matrix $\boldsymbol{F}^\top$, while $\boldsymbol{F}_d$ and $\boldsymbol{L}_d$ are the d-th row of the matrices $\boldsymbol{F}$ and $\boldsymbol{L}$ respectively. Importantly, in general it is not possible to provide a polynomially-sized convex representation of the non-convex update, unless $NP = P$. This follows from the fact that the *inverse optimal value problem* is NP–complete, see Ahmed and Guan (2005).

The computation of the ψ policies via the non-convex update can be formulated as the following mixed-integer problem

$$w^{t,E} = \max\ \boldsymbol{d}^\top y \tag{4.46a}$$

$$\text{s.t. } \boldsymbol{H}x \le h,\ \boldsymbol{G}\hat{c} \le \boldsymbol{g},\ \boldsymbol{F}y + \boldsymbol{L}x \le \boldsymbol{f},\ -\boldsymbol{F}^\top p - \hat{c} \le 0 \tag{4.46b}$$

$$\boldsymbol{f} - \boldsymbol{L}x - \boldsymbol{F}y \le \boldsymbol{M}^p \hat{u},\ p \le \boldsymbol{M}^p(\mathbf{1} - \hat{u}) \tag{4.46c}$$

$$\hat{c} + \boldsymbol{F}^\top p \le \boldsymbol{M}^y \hat{v},\ y \le \boldsymbol{M}^y(\mathbf{1} - \hat{v}) \tag{4.46d}$$

$$\boldsymbol{F}r^s \le \boldsymbol{f} - \boldsymbol{L}x^{s,\psi},\ -\boldsymbol{F}^\top q^s - \hat{c} \le 0 \qquad \forall s \le t-1 \tag{4.46e}$$

$$\boldsymbol{f} - \boldsymbol{F}r^s \le \boldsymbol{M}^p u^s + \boldsymbol{L}x^{s,\psi},\ q^s \le \boldsymbol{M}^p(\mathbf{1} - u^s) \qquad \forall s \le t-1 \tag{4.46f}$$

$$\hat{c} + \boldsymbol{F}^\top q^s \le \boldsymbol{M}^y v^s,\ r^s \le \boldsymbol{M}^y(\mathbf{1} - v^s) \qquad \forall s \le t-1 \tag{4.46g}$$

$$\boldsymbol{d}^\top r^s = w^{s,\psi} \qquad \forall s \le t-1 \tag{4.46h}$$

$$(\boldsymbol{L}x^{s,\psi} - \boldsymbol{f})^\top r^s = z^{s,\psi} \qquad \forall s \le t-1 \tag{4.46i}$$

$$\hat{c}^\top y^{s,E} \ge z^{s,\psi} + \epsilon^s \qquad \forall s \le t-1 \tag{4.46j}$$

$$\hat{u}, u^s \in \{0,1\}^{|C_F|}, \hat{v}, v^s \in \{0,1\}^{|A|},\ \forall s \le t-1 \tag{4.46k}$$

$$y, r^s \in \mathbb{R}_+^{|A|}, p, q^s \in \mathbb{R}_+^{|C_F|},\ \forall s \le t-1 \tag{4.46l}$$

$$x \in \mathbb{R}_+^{|I|-k} \times \mathbb{Z}_+^k, \hat{c} \in \mathbb{R}^{|A|}, \tag{4.46m}$$

where $\boldsymbol{M}^p$ is a diagonal matrix of the appropriate dimensions, and $\max\{p_d, (\boldsymbol{f} - \boldsymbol{L}x - \boldsymbol{F}y)_d\} \le \boldsymbol{M}_d^p$ for all $d \in C_F$ and all feasible p, y and x. Similarly, $\boldsymbol{M}^y$ is a diagonal matrix of the appropriate dimensions where $\max\{y_a, (\hat{c} + \boldsymbol{F}^\top p)_a\} \le \boldsymbol{M}_a^y$ for all $a \in A$ and all feasible p and $\hat{c}$.

Similarly with the convex case, the MIP formulation of the non-convex update can be viewed as a two-stage mixed-integer stochastic program (SMIP), and hence, it can be solved

via its *extensive form*, i.e., by solving the MIP directly, or by more specific decomposition and branch and cut algorithms for SMIPs, see e.g., Sen and Sherali (2006). Indeed, problem (4.46) is given by:

$$\max \boldsymbol{d}^\top y + E[g(\hat{c}, \xi)] \tag{4.47a}$$

$$\text{s.t. } \boldsymbol{H}x \leq h,\ \boldsymbol{G}\hat{c} \leq \boldsymbol{g},\ \boldsymbol{F}y + \boldsymbol{L}x \leq \boldsymbol{f},\ -\boldsymbol{F}^\top p - \hat{c} \leq 0 \tag{4.47b}$$

$$\boldsymbol{f} - \boldsymbol{L}x - \boldsymbol{F}y \leq \boldsymbol{M}^p \hat{u},\ p \leq \boldsymbol{M}^p(\mathbf{1} - \hat{u}) \tag{4.47c}$$

$$\hat{c} + \boldsymbol{F}^\top p \leq \boldsymbol{M}^y \hat{v},\ y \leq \boldsymbol{M}^y(\mathbf{1} - \hat{v}) \tag{4.47d}$$

$$\hat{u} \in \{0,1\}^{|C_F|}, \hat{v} \in \{0,1\}^{|A|}, y \in \mathbb{R}_+^{|A|}, p \in \mathbb{R}_+^{|C_F|} \tag{4.47e}$$

$$x \in \mathbb{R}_+^{|I|-k} \times \mathbb{Z}_+^k, \hat{c} \in \mathbb{R}^{|A|}, \tag{4.47f}$$

where the $g(\hat{c}, \xi)$ is the second-stage value function given the first-stage decision variable $\hat{c}$. Here, the 'random vector' $\boldsymbol{\xi}$ takes its values on the discrete set

$$\boldsymbol{\xi} \in \big\{(x^{s,\psi}, z^{s,\psi}, w^{s,\psi}, y^{s,E})\colon s \leq t-1\big\}. \tag{4.48}$$

For any given realization of $\xi^s = (x^{s,\psi}, z^{s,\psi}, w^{s,\psi}, y^{s,E})$, we have that $g(\hat{c}, \xi^s)$ is the value of the following MIP

$$g\big(\hat{c}, \xi^s\big) = \max\ 0 \tag{4.49a}$$

$$\text{s.t. } \boldsymbol{F}r \leq \boldsymbol{f} - \boldsymbol{L}x^{s,\psi},\ -\boldsymbol{F}^\top q - \hat{c} \leq 0 \tag{4.49b}$$

$$\boldsymbol{f} - \boldsymbol{F}r \leq \boldsymbol{M}^p u + \boldsymbol{L}x^{s,\psi},\ p \leq \boldsymbol{M}^p(\mathbf{1} - u) \tag{4.49c}$$

$$\hat{c} + \boldsymbol{F}^\top q \leq \boldsymbol{M}^y v,\ r \leq \boldsymbol{M}^y(\mathbf{1} - v) \tag{4.49d}$$

$$\boldsymbol{d}^\top r = w^{s,\psi} \tag{4.49e}$$

$$(\boldsymbol{L}x^{s,\psi} - \boldsymbol{f})^\top r = z^{s,\psi} \tag{4.49f}$$

$$\hat{c}^\top y^{s,E} \geq z^{s,\psi} + \epsilon^s \tag{4.49g}$$

$$u \in \{0,1\}^{|C_F|}, v \in \{0,1\}^{|A|}, r \in \mathbb{R}_+^{|A|}, q \in \mathbb{R}_+^{|C_F|}. \tag{4.49h}$$

As before, note that from the standpoint of the SMIP formulation, the distribution function of ξ is irrelevant as the second-stage value is always zero or $-\infty$ (in case of infeasibility) for any realization of ξ.

Finally, observe that we can enhance the convex update by assuming Value–Perfect or Response–Perfect feedback. If such feedbacks are assumed, then the update becomes

$$\mathcal{U}^{t+1} = \mathcal{U}^t \cap \mathcal{L}^t \cap \mathcal{N}^t \cap \mathcal{V}^t \text{ and } \mathcal{U}^{t+1} = \mathcal{U}^t \cap \mathcal{L}^t \cap \mathcal{N}^t \cap \mathcal{R}^t, \tag{4.50}$$

for Value–Perfect and Response–Perfect feedback, respectively, where $\mathcal{V}^t$ and $\mathcal{R}^t$ are defined as in equations (4.26)–(4.27).

4.5 COMPUTATIONAL STUDY

In this section we perform preliminary computational experiments to observe the performance of the time-stability, as well as of the regret, for the policies in ψ under the various feedback and update scenarios. The bilevel problem we use is the Asymmetric Shortest Path Interdiction Bilevel Problem (ASPI), as described in Remark 11. The full-information formulation of this problem, under the optimistic assumption, is given by

$$\max \boldsymbol{d}^\top y \tag{4.51a}$$

$$\text{s.t. } \mathbf{1}^\top x = k \tag{4.51b}$$

$$y \in \arg\min\{\boldsymbol{c}^\top y' \colon \boldsymbol{M} y' = \boldsymbol{b}, y' \leq \mathbf{1} - x, y' \geq 0\} \tag{4.51c}$$

$$x \in \{0,1\}^{|A|}. \tag{4.51d}$$

In this formulation, matrix $\boldsymbol{M}$ is the node-arc adjacency matrix of the directed network $G = (N, A)$. Vector $\boldsymbol{b} \in \mathbb{R}^{|C_F|}$ satisfies that $b_1 = 1$ and $b_m = -1$, where node $d = 1$ is the source node and node $d = m$ is the sink node, and $\mathbf{1}$ is a $|A| \times 1$ vector of ones. The leader's decision variables x are binary valued, and x_a takes the value 1 if and only if arc a is interdicted.

We note that in the specific case of the ASPI, the follower's problem in Equations (4.51a)–(4.51d) has an alternative formulation given by (see Israeli and Wood (2002))

$$y \in \arg\min\{(\boldsymbol{c} + Qx)^\top y' \colon \boldsymbol{M} y' = \boldsymbol{b}, y' \geq 0\}, \tag{4.52}$$

where Q is a sufficiently large positive constant (e.g., $Q = |A| \max\{c_a \colon a \in A\}$). Using the formulation given in (4.52) instead of the one in (4.51a)–(4.51d) is slightly better from the computational point of view. Specifically, note that if the formulation (4.51a)–(4.51d) is used, then $|C_F| = |A| + |N|$, and hence there are $|A| + |N|$ dual variables that have to be introduced in the formulation (4.22a)–(4.22k), as well as at each time t, for the convex and non-convex updates. This implies that $(|A|+|N|)(t+1)$ binary variables are introduced due to the follower's problem constraints in this approach.

In contrast, if formulation (4.52) is used, then $|C_F| = |N|$. Here, $\hat{c}$ becomes $\hat{c} + Qx$ in all formulations (which does not introduce any non-linearity), and only $|N|$ dual variables are introduced in (4.22a)–(4.22k), as well as each time t, for the convex and non-convex updates. This obviously yields a reduction in the continuous variables, but more importantly, results in a reduction of $(1 + t)|A|$ binary variables for the non-convex update approach.

Description of the Instances: We consider layered networks with two layers and four nodes per layer. The generic structure is depicted in Figure 17. The values of the profit vector $\boldsymbol{d}$ are drawn at random from a Uniform(1,80) discrete random variable. We assume that the initial uncertainty set $\mathcal{U}^0$ is an hypercube, hence $\boldsymbol{G}^0 = [\boldsymbol{I}; -\boldsymbol{I}]$, where $\boldsymbol{I}$ is a $|A| \times |A|$ identity matrix, and $\boldsymbol{g}^0 = [\boldsymbol{u}; -\boldsymbol{\ell}]$, with $\boldsymbol{u} = (u_a \colon a \in A)$ is the vector of upper bounds for $\boldsymbol{c}$ and $\boldsymbol{\ell} = (\ell_a \colon a \in A)$ is the vector of lower bounds for $\boldsymbol{c}$.

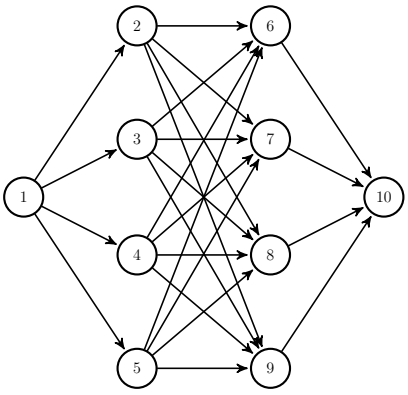


Figure 17: A layered network with two layers and four nodes per layer. It has $|N| = 10$ nodes and $|A| = 24$ directed arcs.

For each $a \in A$, the values of u_a, ℓ_a and c_a are drawn independently at random from a Uniform(1,40) distribution, and then organized accordingly. For the experiments we select a time horizon of $T = 15$, and we select the value of k, the number of arcs to be interdicted, to belong in $\{1, 2, 3\}$.

We compute the decisions of a policy $\psi \in \Psi$ under five different updating mechanisms. Specifically, we choose the basic update, the convex update, and the non-convex update, as described in sections 4.4.2–4.4.4. In addition, we consider the *weak convex* update and the *weak non-convex* update. The weak convex update is defined as the regular convex update, with the exception that the basic update is no longer included, i.e., $\mathcal{C}^t$ replaces $\mathcal{L}^t \cap \mathcal{C}^t$. The weak non-convex update is defined in a similar way.

For each of the updating mechanisms discussed above, we consider four different feedback types. Those are Standard, Value–Perfect, Response–Perfect, and Value–Perfect plus Response–Perfect feedback. As the name suggests, in this latter feedback type we consider both Value–Perfect and Response–Perfect feedback simultaneously (hence, e.g., $\mathcal{U}^{t+1} = \mathcal{U}^t \cap \mathcal{L}^t \cap \mathcal{V}^t \cap \mathcal{R}^t$ for the basic update).

Results and Discussion: For each updating mechanisms and feedback type we generate 30 different independent replications. The resulting mean time-stability and regret are given in Tables 14, 15, and 16, for $k = 1, 2, 3$ respectively. We measure dispersion using the mean absolute deviation (MAD), and the resulting MAD for time-stability and regret are given in Tables 17, 18, and 19, for $k = 1, 2, 3$ respectively. In reporting the results we make the convention that if time-stability cannot be guaranteed before time T, then $\tau^\psi = T$.

Table 14: Mean for the time-stability (τ^ψ) and regret (R_T^ψ) for a policy $\psi \in \Psi$ when $k = 1$.

	Standard Feedback		RP Feedback		VP Feedback		RP+VP Feedback	
	τ^ψ	R_T^ψ	τ^ψ	R_T^ψ	τ^ψ	R_T^ψ	τ^ψ	R_T^ψ
Basic	12.10	410.90	4.40	99.50	4.13	88.93	3.77	76.33
Convex	10.90	347.90	3.27	67.57	3.07	59.63	3.07	59.63
Non-convex	3.70	69.47	3.40	65.70	3.13	59.70	3.13	59.70
Weak convex	10.67	346.67	3.90	92.30	4.03	104.30	4.03	104.30
Weak non-convex	4.57	98.97	4.27	99.07	4.07	108.33	4.07	108.33

Table 15: Mean for the time-stability (τ^{ψ}) and regret (R_T^{ψ}) for a policy $\psi \in \Psi$ when $k = 2$.

	Standard Feedback		RP Feedback		VP Feedback		RP+VP Feedback	
	τ^{ψ}	R_T^{ψ}	τ^{ψ}	R_T^{ψ}	τ^{ψ}	R_T^{ψ}	τ^{ψ}	R_T^{ψ}
Basic	13.67	589.53	6.80	218.83	6.53	209.50	5.83	177.53
Convex	13.53	548.80	5.43	175.07	5.13	157.27	5.13	157.27
Non-convex	8.37	304.57	5.33	162.80	5.07	161.20	5.07	161.20
Weak convex	13.57	559.43	8.80	308.20	8.00	240.20	8.00	240.20
Weak non-convex	7.60	254.40	7.83	244.30	6.73	223.27	6.73	223.27

We make the following observations regarding the results for mean time-stability and regret:

- For standard feedback, the non-convex update clearly outperforms the other updating mechanisms across all configurations.
- For the other specialized feedback types (Value–Perfect, Response–Perfect, and Value–Perfect+Response–Perfect), both the convex and non-convex updates slightly improve the performance of the basic update. In general, the non-convex update usually has better performance than the convex update.
- The introduction of specialized feedbacks greatly improves the performance of the policy. In particular, Value–Perfect plus Response–Perfect is slightly better than Value–Perfect feedback. In turn, Value–Perfect is slightly better than the Response–Perfect feedback.

Table 16: Mean for the time-stability (τ^{ψ}) and regret (R_T^{ψ}) for a policy $\psi \in \Psi$ when $k = 3$.

	Standard Feedback		RP Feedback		VP Feedback		RP+VP Feedback	
	τ^{ψ}	R_T^{ψ}	τ^{ψ}	R_T^{ψ}	τ^{ψ}	R_T^{ψ}	τ^{ψ}	R_T^{ψ}
Basic	11.43	523.37	5.70	211.93	5.27	189.93	5.13	184.30
Convex	10.63	470.17	4.90	165.63	4.77	170.43	4.77	170.43
Non-convex	5.27	182.13	5.03	178.07	4.43	151.07	4.43	151.07
Weak convex	10.93	489.60	8.70	339.80	7.60	287.17	7.60	287.17
Weak non-convex	6.97	236.67	7.20	241.60	6.80	261.10	6.80	261.10

- When the basic update is not included into the convex and non-convex updates, these mechanisms produce slightly worse results under the specialized feedbacks than the basic update. For standard feedback, the non-convex update remains the best.
- Mean time-stability and mean regret are clearly correlated. In general, the higher the time-stability the higher the regret, and viceversa.

These observations suggest the following conclusions. In the presence of specialized feedback it is not necessary to use any sophisticated updating mechanism, i.e., the basic update is a *simple and sufficient* mechanism to incorporate new information. However, if only standard feedback is available, then the non-convex update is the clear updating mechanism of choice. On the other hand, regarding dispersion, we note that the observations and conclusions for the means largely apply to the MAD as well.

Table 17: MAD for the time-stability (τ^ψ) and regret (R_T^ψ) for a policy $\psi \in \Psi$ when $k = 1$.

	Standard Feedback		RP Feedback		VP Feedback		RP+VP Feedback	
	τ^ψ	R_T^ψ	τ^ψ	R_T^ψ	τ^ψ	R_T^ψ	τ^ψ	R_T^ψ
Basic	3.67	292.83	1.79	61.07	1.57	48.39	1.42	37.38
Convex	4.45	270.76	0.94	30.87	0.88	28.94	0.88	28.94
Non-convex	1.25	31.60	1.08	29.55	0.98	28.35	0.98	28.35
Weak convex	4.73	274.31	1.18	60.93	1.65	92.06	1.65	92.06
Weak non-convex	1.32	55.63	1.11	61.02	1.23	81.91	1.23	81.91

Table 18: MAD for the time-stability (τ^ψ) and regret (R_T^ψ) for a policy $\psi \in \Psi$ when $k = 2$.

	Standard Feedback		RP Feedback		VP Feedback		RP+VP Feedback	
	τ^ψ	R_T^ψ	τ^ψ	R_T^ψ	τ^ψ	R_T^ψ	τ^ψ	R_T^ψ
Basic	1.98	360.37	2.85	122.86	2.77	125.13	2.34	101.94
Convex	2.21	335.48	1.70	85.54	1.38	77.87	1.38	77.87
Non-convex	3.53	202.16	1.64	84.11	1.42	89.65	1.42	89.65
Weak convex	2.19	333.55	3.03	164.99	3.40	152.92	3.40	152.92
Weak non-convex	2.67	140.79	3.29	146.23	2.58	133.52	2.58	133.52

Finally, we compare the updating mechanisms and the feedback types by counting in how many replications optimality can be guaranteed via the 'sandwich' Theorem (cf. Theorem 7). These results are shown in Tables 20, 21 and 22 for $k = 1, 2, 3$, respectively.

Table 19: MAD for the time-stability (τ^{ψ}) and regret (R_T^{ψ}) for a policy $\psi \in \Psi$ when $k = 3$.

	Standard Feedback		RP Feedback		VP Feedback		RP+VP Feedback	
	τ^{ψ}	R_T^{ψ}	τ^{ψ}	R_T^{ψ}	τ^{ψ}	R_T^{ψ}	τ^{ψ}	R_T^{ψ}
Basic	5.23	341.48	2.81	146.79	2.60	126.39	2.48	130.89
Convex	5.82	335.52	2.38	105.83	2.20	115.16	2.20	115.16
Non-convex	2.75	121.37	2.63	120.22	2.13	103.48	2.13	103.48
Weak convex	5.42	332.01	5.19	268.55	4.51	229.82	4.51	229.82
Weak non-convex	4.23	177.11	4.07	187.49	3.71	194.40	3.71	194.40

Table 20: Number of replications for which optimality is guaranteed, $k = 1$

	SF	RPF	VPF	RPF+VPF
Basic	12	30	30	30
Convex	15	30	30	30
Non-convex	30	30	30	30
Weak convex	15	30	30	30
Weak non-convex	30	30	30	30

Table 21: Number of replications for which optimality is guaranteed, $k = 2$

	SF	RPF	VPF	RPF+VPF
Basic	8	30	29	30
Convex	8	30	30	30
Non-convex	24	30	30	30
Weak convex	8	27	27	27
Weak non-convex	29	25	28	28

Table 22: Number of replications for which optimality is guaranteed, $k = 3$

	SF	RPF	VPF	RPF+VPF
Basic	8	30	30	30
Convex	10	30	30	30
Non-convex	30	30	30	30
Weak convex	10	21	25	25
Weak non-convex	27	27	26	26

These results confirm the conclusions of the analysis for the mean time-stability and regret. In particular, it is remarkable that under standard feedback both the basic and convex updating mechanisms can guarantee optimality in less than a third of the cases, while the non-convex update guarantees optimality in virtually all cases. Moreover, under a specialized feedback, optimality can be guaranteed in all cases. However, surprisingly, this is no longer true in the weak convex and non-convex updating mechanisms, where it can be seen that optimality cannot be guaranteed in 10 to 20% of the replications for $k = 2, 3$.

In addition, we observe that under the specialized feedback, or under Standard feedback with the non-convex update, the time-stability is, on average, far less than the upper bound of Theorem 8. Indeed, for the instances we consider in the experiments we have that $|ext(conv(X))|$ is 24, 274 and 2024 for $k = 1, 2, 3$, respectively. Such behavior suggests that the theoretical upper bound for the time-stability for the policies in Ψ can be significantly tightened for the specialized feedback, and perhaps more interestingly, for the Standard feedback with the non-convex update.

4.6 CONCLUDING REMARKS

In this chapter we study the asymmetric bilevel linear problem with incomplete information and learning. In this problem, in contrast with the **SMPI** problem of Chapter 3, and the

shortest path interdiction problem with incomplete information of Chapter 2, the follower's objective function is independent of the objective function of the leader. We show how the greedy and robust policies, that are weakly optimal and efficient for these problems, respectively, no longer have a desirable performance in this setting. Importantly, greedy and robust policies might stall, and might not provide a certificate of optimality in real time.

We then consider a class of greedy and 'best'–case policies and show that they can guarantee a finite (albeit exponentially large) upper bound on time-stability. In addition, they provide a certificate of optimality in real time, and can be computed using mixed-integer programming. For this class of policies we consider different updating mechanisms, and show that there is an interesting connection between the resulting MIP formulations and certain types of two-stage stochastic programming problems. Such a connection is highly important because stochastic programming algorithms can be employed to compute the decisions that the policies yield.

Our numerical experience shows that the performance of the greedy and 'best'–case policies is highly dependent on the updating mechanisms and the type of feedback that is used. Remarkably, the non-convex updating mechanism gives a very good performance under Standard feedback, which is the most complicated feedback type. In contrast, both the non-convex and convex mechanisms, do not produce significant improvements over the basic linear updating mechanism when considering Value–Perfect or Response–Perfect feedbacks.

At this point, the question of whether the policies can provide a linear upper bound on the time-stability (in the parameters of the full-information bilevel problem) remains open. Also, the question of whether these policies are weakly optimal for the asymmetric setting is also open, and further research is required to determine the conditions (if any) under which this type of optimality can be guaranteed.

5.0 CONCLUSIONS

This dissertation considered SBPI, a class of sequential hierarchical problems where the leader decision-maker has incomplete knowledge about the information that the follower decision-maker uses to decide. As a consequence, in order to improve the quality of her decisions, the leader has to learn the information of the follower's problem by observing his reactions to her actions. In the dissertation we study three particular models within this framework: the shortest path interdiction problem, the max-min bilevel linear problem, and the asymmetric bilevel linear problem.

We show that whenever the leader's goal is to maximally degrade the follower's performance, as in shortest path interdiction or max-min bilevel linear programming, the leader should base her decisions on greedy and robust decision-making policies. These policies are greedy as they seek to maximize the immediate disruption to the follower, regardless of the future, and they are robust, as they assume that the uncertain data of the follower's problem realizes its worst case scenario from the follower's perspective.

The greedy and robust policies have many important theoretical properties under specific types of feedback such as Value–Perfect or Response–Perfect. On the one hand, for the shortest path interdiction problem, we show that they are efficient, which means that they find the full-information optimal solution within a finite time horizon, are homogeneous between them, and are not dominated by any non-robust or non-greedy policy. We refine this concept for the max-min bilevel setting, where we show that the policies are optimal in the weak sense, i.e., they have the best possible worst-case time-stability performance across all possible instances. The time-stability of these policies, moreover, is bounded by the number of variables of the follower's problem.

In addition, the greedy and robust policies have two important practical advantages. First, the decisions that they prescribe involve solving max-min bilevel linear programming problems with robust lower-level constraints. These problems can be formulated as mixed-integer programs, which implies that the policies can be readily computed using off-the-shelf MIP solvers. Second, the policies provide a certificate of optimality in real time, namely, at any given time period the leader can determine whether an optimal full-information solution has been found by comparing the feedback she gets from the follower with the cost that the policies predict.

For the asymmetric bilevel linear problem the leader's decisions do not necessarily seek to maximize the disruption of the follower's problem. In fact, in these problems the leader's costs (or profits) are measured independently from the performance of the follower. In this setting, the greedy and robust policies no longer have all the desirable theoretical and practical advantages they have for the symmetric cases. Thus, to address this issue, we study a class of greedy policies and 'best'–case policies.

We show how these best-case policies retain most of the important features of the greedy and robust policies for the asymmetric case. Particularly, they have bounded time-stability upper bounds, provide certificates of optimality in real-time, and can be computed using MIPs. In addition, they allow more generality for the structure of the uncertainty set of the follower's data, which leads to the concepts of convex and non-convex updating mechanisms.

Besides the above considerations, the research in this dissertation shows important connections between three important classes of optimization problems under uncertainty. The SBPI belongs to the class of online optimization models, as the leader has very limited knowledge on the uncertain information and cannot make reliable estimates of future outcomes. Nevertheless, if she is dealing with max-min problems and assumes a robust view on the uncertainty, as in the class of robust optimization models, she is guaranteed to make the best decision possible, as per weak-optimality. In contrast, in asymmetric problems, she has to assume a different 'anti'-robust view on uncertainty, and this leads to MIP formulations that fit within the two-stage stochastic programming paradigm.

Several important questions arise as a product of this research. The most immediate one relates to be able to guarantee weak optimality for models where the there is uncertainty in

the follower's constraints. This question also arises for the asymmetric bilevel linear model. Subsequently, it is natural to consider a broader class of bilevel problems under uncertainty, where the leader might also have incomplete information regarding her own objective function and constraints. From a different perspective, the problem we study can be viewed as a non-convex, combinatorial, online optimization problem. As such it belongs to a class of problems scarcely studied in the literature, and it is a matter of future research if the methods developed in this dissertation, can be generalized to this broader class of decision-making models.

APPENDIX A

SUPPLEMENT FOR CHAPTER 2

This appendix contains some proofs and additional numerical results for Chapter 2.

A.1 BASIC PROPERTIES OF K-MOST VITAL ARCS

Let $G = (N, A, C)$ be a directed network. Recall that a set of k-most vital arcs of G is a subset $L \subseteq A$ that satisfies

$$L \in \underset{\{L \subseteq A:\ |L| \leq k\}}{\arg\max}\ z(G[A \setminus L]).$$

Proposition 12. *Given $G = (N, A, C)$, let $Y \subseteq X \subseteq A$. If L_X and L_Y are sets of k-most vital arcs of $G[X]$ and $G[Y]$, respectively, then $z(G[X \setminus L_X]) \leq z(G[Y \setminus L_Y])$.*

Proof. Let $U = Y \cap L_X$. Then by the definition of L_X and L_Y:

$$z(G[Y \setminus L_Y]) \geq z(G[Y \setminus U]) = z(G[Y \setminus L_X]) \geq z(G[X \setminus L_X]), \tag{A.1}$$

which concludes the proof. ∎

Proposition 13. *Let $G = (N, A, C)$ and $G' = (N, A, C')$ be networks such that $c_a \leq c'_a$ for all $a \in A$. If L_A and L'_A are sets of k-most vital arcs of G and G', respectively, then*

$$z(G[A \setminus L_A]) \leq z(G'[A \setminus L'_A]). \tag{A.2}$$

Proof. Since $c_a \leq c'_a$ for all $a \in A$, then $z(G[A \setminus L_A]) \leq z(G'[A \setminus L_A])$. Then by the definition of L'_A, it must hold that $z(G'[A \setminus L_A]) \leq z(G'[A \setminus L'_A])$, which concludes the proof. ∎

A.2 ADDITIONAL PROOFS

In this appendix we provide proofs for some of the lemmas and for Propositions 2 and 3.

Lemma 1. *Given $G = (N, A, C)$, let L and A' be such that $L \subseteq A' \subseteq A$ and $(G[A'], L)$ is k-complete. Then L is a set of k-most vital arcs of $G[U]$ for all U such that $A' \subseteq U \subseteq A$.*

Proof. Since $G[A']$ is L-spare, we have that

$$z(G[A' \setminus L]) = z(G[A \setminus L]). \tag{A.3}$$

Let L_U be a set of k-most vital arcs of $G[U]$, where $A' \subseteq U \subseteq A$. Then

$$z(G[A \setminus L]) \leq z(G[U \setminus L]) \leq z(G[U \setminus L_U]) \leq z(G[A' \setminus L]). \tag{A.4}$$

The first inequality is due to the fact that $(U \setminus L) \subseteq (A \setminus L)$, the second from the definition of L_U, and the last one from Proposition 12 (in Appendix A.1) and the fact that L is a set of k-most vital arcs of $G[A']$. Therefore, from equations (A.3) and (A.4) we have that $z(G[U \setminus L]) = z(G[U \setminus L_U])$, which implies that L is a set of k-most vital arcs of $G[U]$. ∎

Lemma 2. *Let $\gamma \in \Gamma$. Then for any $\mathcal{C}_0$ and $G \in \mathbb{G}(\mathcal{C}_0)$:*

1. *$\tau^\gamma(G, \mathcal{C}_0) \leq x^\gamma(G, \mathcal{C}_0)$;*
2. *if $T > |A|$ then $\tau^\gamma(G, \mathcal{C}_0) \leq |A|$.*

Proof. To simplify the notation, let $x \equiv x^\gamma$. Note that by the definition of x, $G[A^\gamma_x]$ is a I^γ_x-spare network. As I^γ_x is also a set of k-most vital arcs of $G[A^\gamma_x]$, it follows that $(G[A^\gamma_x], I^\gamma_x)$ is k-complete and $z(G[A^\gamma_x \setminus I^\gamma_x]) = z^*(G)$ by Lemma 1. Moreover, by the definition of Γ, we have that $I^\gamma_t = I^\gamma_x$ for $t > x$. Hence, $z(G[A^\gamma_t \setminus I^\gamma_t]) = z^*(G)$ for all $t > x$ and the first claim of the proposition follows.

To prove the second claim, we consider two possible cases. Specifically, one has that $G[A^\gamma_t]$ is either I^γ_t-spare for some $t \leq |A|$ or not. In the former case, the arguments above imply that $\tau^\gamma(G, \mathcal{C}_0) \leq |A|$ and the result follows. In the latter case, because $\ell(P^\gamma_t) = z(G[A \setminus I^\gamma_t])$ for all t, by equation (A.4) we have that

$$z(G[A^\gamma_t \setminus I^\gamma_t]) > z(G[A \setminus I^\gamma_t]) \quad t \leq |A|\,. \tag{A.5}$$

The above implies that $P_t^\gamma \not\subseteq A_t^\gamma$ for all $t \leq |A|$. Indeed, suppose that this is not the case. Since $P_t^\gamma \cap I_t^\gamma = \emptyset$ and I_t^γ is a set of k-most vital arcs of $G[A_t^\gamma]$, then $P_t^\gamma \subseteq A_t^\gamma$ implies that $\ell(P_t^\gamma) \geq z(G[A_t^\gamma \setminus I_t^\gamma])$. Thus, from equation (A.5) we have that

$$\ell(P_t^\gamma) > z(G[A \setminus I_t^\gamma]), \tag{A.6}$$

which contradicts the fact that P_t^γ is a shortest path in $G[A \setminus I_t^\gamma]$. We conclude that $A_t^\gamma \subset A_{t+1}^\gamma$, which implies the required result. ∎

Lemma 4. *Suppose that $t \in \mathcal{T}$ is such that $z(G_t^\lambda[A_t^\lambda \setminus I_t^\lambda]) = z(G[A \setminus I_t^\lambda])$, then $(G[A_t^\lambda], I_t^\lambda)$ is k-complete (with respect to G). Moreover, I_t^λ is a set of k-most vital arcs of $G[U]$ for all U such that $A_t^\lambda \subseteq U \subseteq A$.*

Proof. Let L_A and L be sets of k-most vital arcs of G and $G[A_t^\lambda]$, respectively. We have that

$$z(G[A \setminus I_t^\lambda]) \leq z(G[A \setminus L_A]) \overset{(a)}{\leq} z(G[A_t^\lambda \setminus L]) \overset{(b)}{\leq} z(G_t^\lambda[A_t^\lambda \setminus I_t^\lambda]), \tag{A.7}$$

where (a) and (b) follow from Propositions 12 and 13, respectively (see Appendix A.1). This, and the condition that $z(G[A \setminus I_t^\lambda]) = z(G_t^\lambda[A_t^\lambda \setminus I_t^\lambda])$ implies that

$$z(G[A \setminus I_t^\lambda]) = z(G[A \setminus L_A]) = z(G[A_t^\lambda \setminus L]) = z(G_t^\lambda[A_t^\lambda \setminus I_t^\lambda]). \tag{A.8}$$

On the other hand, because $A_t^\lambda \subseteq A$ the following inequalities hold

$$z(G[A \setminus I_t^\lambda]) \leq z(G[A_t^\lambda \setminus I_t^\lambda]) \leq z(G[A_t^\lambda \setminus L]), \tag{A.9}$$

thus, implying that $z(G[A \setminus I_t^\lambda]) = z(G[A_t^\lambda \setminus I_t^\lambda]) = z(G[A_t^\lambda \setminus L])$. Hence, $G[A_t^\lambda]$ is I_t^λ-spare. Moreover, because L is a set of k-most vital arcs of $G[A_t^\lambda]$, the fact that $z(G[A_t^\lambda \setminus I_t^\lambda]) = z(G[A_t^\lambda \setminus L])$ implies that I_t^λ is also a set of k-most vital arcs of $G[A_t^\lambda]$, and the first statement of the proposition follows. The second statement follows directly from Lemma 1. ∎

Lemma 5. *Let $\lambda \in \Lambda$. Then for any $\mathcal{C}_0$ and $G \in \mathbb{G}(\mathcal{C}_0)$:*

1. *$\tau^\lambda(G, \mathcal{C}_0) \leq \widehat{x}^\lambda(G, \mathcal{C}_0)$;*
2. *if $T \geq |A|$, then $\tau^\lambda(G, \mathcal{C}_0) \leq |A|$.*

Proof. To simplify the notation let $x = \widehat{x}^{\lambda}$. Because $z(G_x^{\lambda}[A_x^{\lambda} \setminus I_x^{\lambda}]) = z(G[A \setminus I_x^{\lambda}])$, Lemma 4 implies that I_x^{λ} is a set of k-most vital arcs of G. Thus, $z(G[A \setminus I_x^{\lambda}]) = z^*(G)$. Moreover, from the definition of Λ,we have that $I_t^{\lambda} = I_x^{\lambda}$ for all $t > x$, and the first claim of the proposition follows. The proof of the second statement follows from the arguments in proof of the second statement in Lemma 2, with (A.7) playing the role of (A.4).

■

Lemma 6. Λ *is a homogeneous set both with respect to cumulative regret and with respect to time-stability.*

Proof. We slightly modify the proof of Lemma 3. Consider again $\mathcal{C}_0$, G and G' as given by Figure 5. However, assume that the cost of arc $(1, n)$ in $\mathcal{C}_0$ is not known exactly, but is known to be within range $[1, M]$, where $M \geq 2$, i.e., $(1, n) \notin \widetilde{A}_0$, but $(1, n) \in \widehat{A}_0$. At the first time period the evader travels along the arc $(1, n)$ and its cost becomes known to the interdictor. The result then follows by mimicking the proof of Lemma 3. ■

Proposition 2. *There exists* $\mathcal{C}_0$, $G \in \mathbb{G}(\mathcal{C}_0)$ *and* $\zeta > 0$ *such that if* $T \geq |A|$, *then* $\tau^{\lambda}(G, \mathcal{C}_0) \geq \zeta|A|$. *Moreover, the value of* $R_T^{\lambda}(G, \mathcal{C}_0)$ *can be made arbitrarily large.*

Proof. Consider the same network as in Proposition 1, but with $\widetilde{A}_0^{\lambda} = \emptyset$, and $\widehat{A}_0^{\lambda} = A$. Let $c_{ij} = 1$ for all $(i, j) \in A$, except for $(1, 2)$ with $c_{12} = M \geq 2$, and assume that $\ell_{ij} = 1$ and $u_{ij} = M + 1$ for all $(i, j) \in A$. Proceeding in a similar fashion as in Proposition 1, arc $(1, 2)$ is revealed after $\zeta|A|$ time periods, where $\zeta \leq k/(2k + 2)$. ■

Proposition 3. *Algorithm 2 correctly solves* $LB(G, \mathcal{C}_\prime, \mathcal{T})$.

Proof. Note that only the case when $T_0 < T$ is relevant. Upon convergence one has that either $T' = T$ or $T' < T$. In the first case, $LB(G, \mathcal{C}_\prime, \mathcal{T})$ is solved using formulation (2.11) and the result follows. For the second case suppose that $\{(r^t, p^t, y^t) : \ t \in \mathcal{T}\}$ is not optimal, and consider a solution $\{(\bar{r}^t, \bar{p}^t, \bar{y}^t) : \ t \in \mathcal{T}\}$ feasible for $LB(G, \mathcal{C}_\prime, \mathcal{T})$ such that

$$\sum_{t=0}^{T} \bar{y}_1^t - \bar{y}_n^t > \sum_{t=0}^{T} y_1^t - y_n^t. \tag{A.10}$$

Because $T' < T$ it is necessarily the case that $z^* = y_s^t - y_t^t$ for all $t \geq T'$. Therefore, (A.10) implies that

$$\sum_{t=0}^{T'} \bar{y}_1^t - \bar{y}_n^t > \sum_{t=0}^{T'} y_1^t - y_n^t.$$

However, $\{(\bar{r}^t, \bar{p}^t, \bar{y}^t) : t \leq T'\}$ is feasible for $LB(G, \mathcal{C}_\prime, \mathcal{T}')$. Thus, the equation above contradicts the fact that $\{(r^t, p^t, y^t) : t \leq T'\}$ is an optimal solution of $LB(G, \mathcal{C}_\prime, \mathcal{T}')$. This proves the result. ■

A.3 ADDITIONAL GRAPHS

We provide additional results for the computational experiments described in Section 2.5. Figure 18 corresponds to the discussion in Section 2.5.4 for the left-skewed distributed costs. Figures 19 and 20 are the complementary figures for the second experiment in Section 2.5.5, setting $p_c = 0$ and $p_c = 2/3$ respectively.

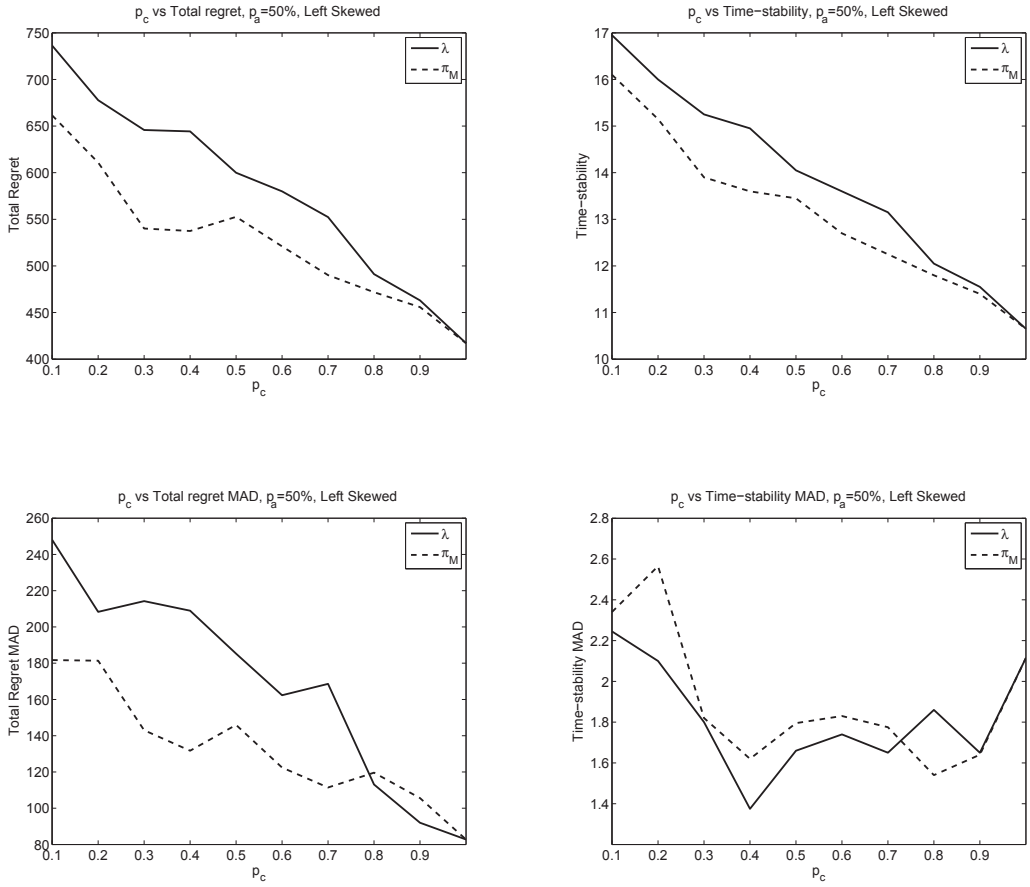


Figure 18: Behavior of the average time-stability, average total regret, time-stability MAD and total regret MAD as p_c increases. The cost distribution is left-skewed and $p_a = 1/2$.

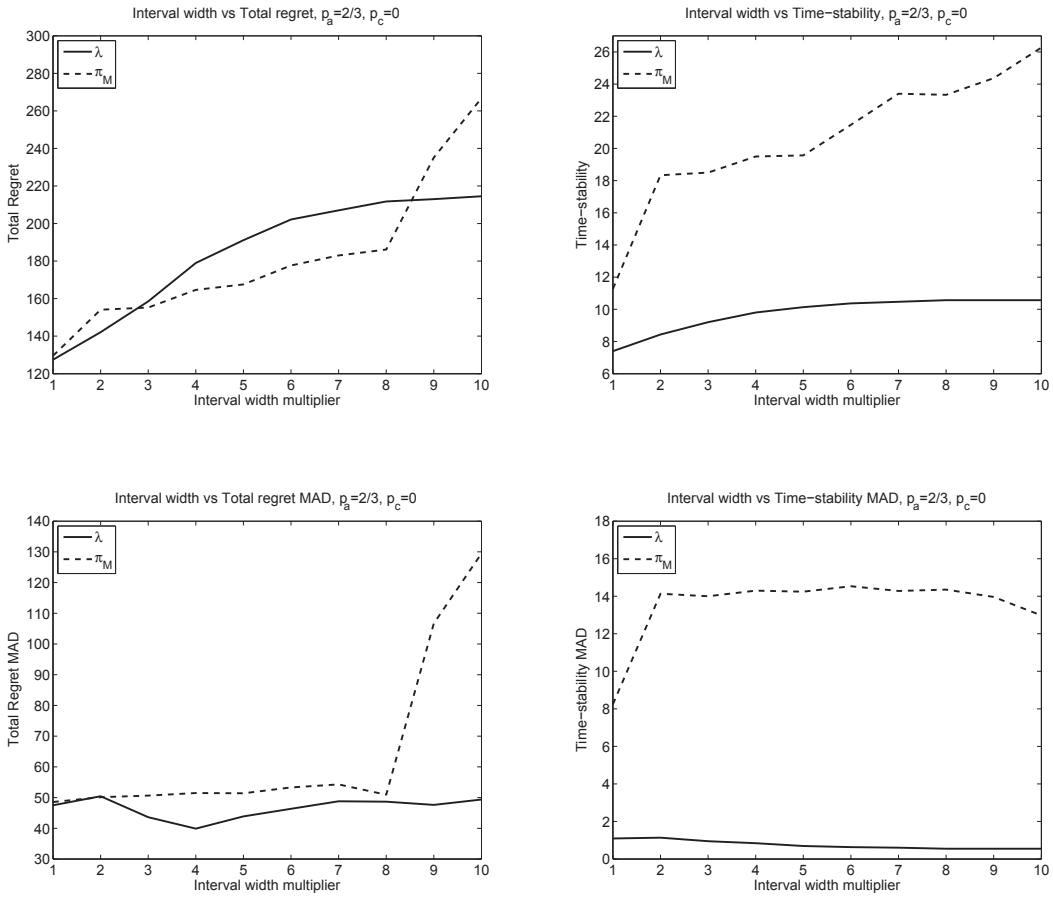


Figure 19: Behavior of the average time-stability, average total regret, time-stability MAD and total regret MAD as the cost intervals widen for the case of $p_a = 2/3$ and $p_c = 0$. Given the interval-width multiplier m, the lower and upper bounds of the arc cost in $\widehat{A}_0$ are $l_a = c_a - mx_a$ and $u_a = c_a + my_a$, respectively.

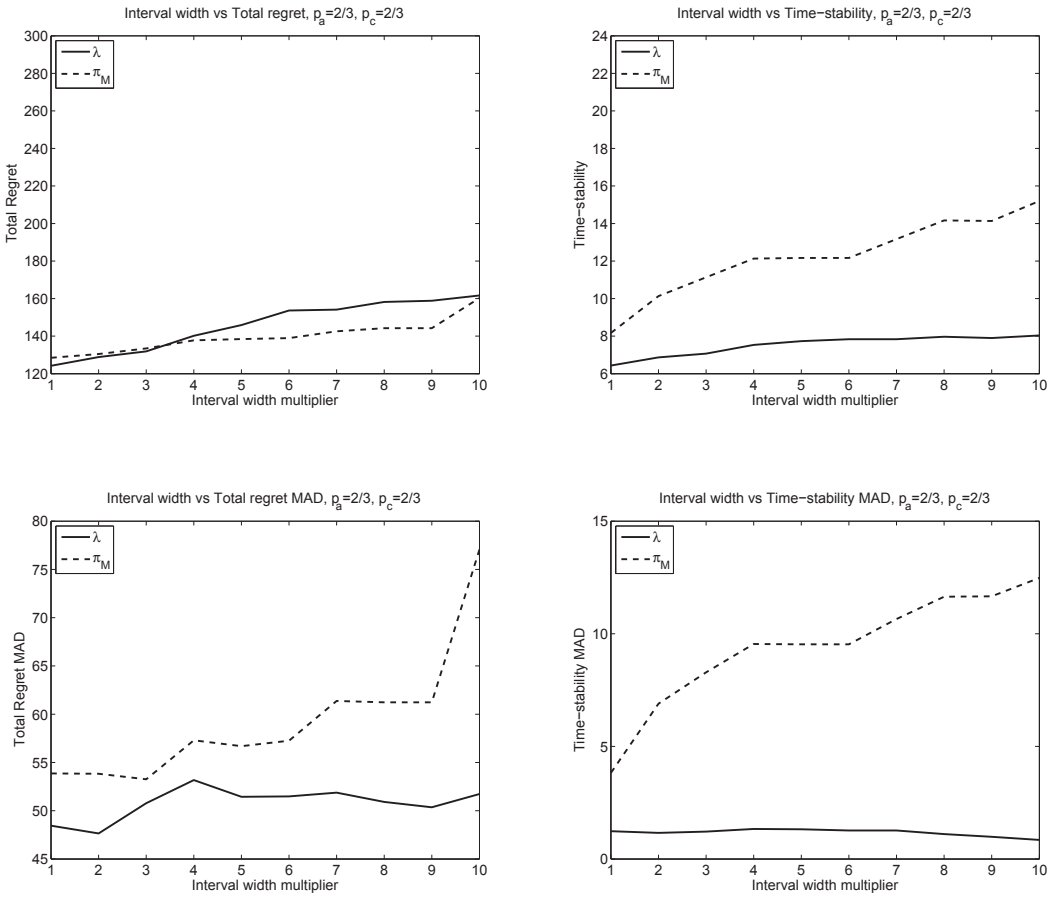


Figure 20: Behavior of the average time-stability, average total regret, time-stability MAD and total regret MAD as the cost intervals widen for the case of $p_a = 2/3$ and $p_c = 2/3$. Given the interval-width multiplier m, the lower and upper bounds of the arc costs in $\widehat{A}_0$ are $l_a = c_a - mx_a$ and $u_a = c_a + my_a$, respectively.

APPENDIX B

SUPPLEMENT FOR CHAPTER 3

This appendix contains most of the proofs and additional example for Chapter 3.

B.1 PROOFS OF THE RESULTS FOR THE BASIC COST MODEL

We first introduce some auxiliary results. We have the following basic observation:

Lemma 17. *For any $t \in \mathcal{T}$ and $x \in X^t$,*

$$z_R^t(x) = \min_{y'} \{ \left(\boldsymbol{d}^t\right)^\top y' \ : \ y' \in Y_R^t(x) \} \tag{B.1}$$

where $\boldsymbol{d}^t = (1, 0, \ldots, 0)^\top$ and

$$Y_R^t(x) := \left\{ (y_0, y) \in \mathbb{R} \times \mathbb{R}_+^{|A^t|} \ : \ \ -y_0 + \left(\boldsymbol{c}^t\right)^\top y \leq 0 \ \forall \hat{\boldsymbol{c}}^t \in \mathcal{U}^t, y \in Y^t(x) \right\}.$$

Proof. Note that $z_R^t(x)$ can be equivalently posed as

$$\begin{aligned} \min_{y_0, y} \ & y_0 \\ \text{s.t.} \ & y_0 \geq \max_{\hat{\boldsymbol{c}}^t \in \mathcal{U}^t} \left(\hat{\boldsymbol{c}}^t\right)^\top y \\ & \boldsymbol{F}^t y + \boldsymbol{L}^t x \leq \boldsymbol{f}^t \\ & y \geq 0. \end{aligned}$$

The result follows after noting that (y_0, y) satisfies the first of the above constraints if and only if $y_0 \geq (\boldsymbol{c}^t)^\top y$ for all $\hat{\boldsymbol{c}}^t \in \mathcal{U}^t$. ∎

In all the proofs that follow we use the representation of $z_R^t(x)$ given by equation (B.1) instead of the representation given by the original definition.

Proof of Theorem 3. (i) We first prove the statement that $z^{t,\lambda} \leq z^* \leq z_R^{t,*}$. For the leftmost inequality, the result follows from the definition of both z^* and $z^{t,\lambda}$ (see equations (3.1) and (4.7)) because $\bar{x}^{t,\lambda} \in X$, 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 (3.4). 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

$$(y_0, \hat{y}) \in \arg\min\{\left(\boldsymbol{d}^t\right)^\top y' \,:\, y' \in Y_R^t(\hat{x})\},$$

hence $(\boldsymbol{c}^t)^\top \hat{y} = z_R^t(\hat{x})$. By the definition of $z_R^{t,*}$ we have that

$$\left(\boldsymbol{c}^t\right)^\top \hat{y} \leq z_R^{t,*}. \tag{B.2}$$

Define $\bar{y}$ as $\bar{y}_a := \hat{y}_a$ if $a \in A^t$, and $\bar{y}_a := 0$ if $a \in A \setminus A^t$. Since $\mathcal{F}$ is standard, Assumption **(A4)** holds, and $\hat{y} \in Y^t(\hat{x})$, it follows that $\bar{y} \in Y(x^*)$. Therefore,

$$z^* \leq \boldsymbol{c}^\top \bar{y}. \tag{B.3}$$

As $\boldsymbol{c}^\top \bar{y} = (\boldsymbol{c}^t)^\top \hat{y}$, equations (B.2) and (B.3) yield the desired result.

(ii) Next, we show that $\tau^\lambda \leq \xi^\lambda$. Recall the definition of $\bar{x}^{t,\lambda}$ (see equations (3.2) and (4.7)), i.e., $\bar{x}_i^{t,\lambda} = x_i^{t,\lambda}$ if $i \in I^t$ and $\bar{x}_i^{t,\lambda} = 0$ if $i \notin I^t$. 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 (3.4)) implies that $\bar{x}^{\xi,\lambda} \in X$. Since by definition of ξ we have that $z^{\xi,\lambda} = z_R^{\xi,*}$, part (i) implies that (recall that by definition we have $z^{t,\lambda} = z(\bar{x}^{t,\lambda})$ for any period t)

$$z(\bar{x}^{\xi,\lambda}) = z^*,$$

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 $z^{\xi,\lambda} = z^*$. The desired claim follows. ∎

Proof of Proposition 4. As $y^{t,\lambda} \in Y(x^{t,\lambda})$ and $y^t_a = 0$ for all $a \notin A^t$, it follows that

$$\sum_{a \in A^t} F_{da} y^{t,\lambda}_a + \sum_{i \in I^t} L_{di} x^{t,\lambda}_i \leq f_d \qquad \forall d \in C^t_F,$$

which implies that $(y^{t,\lambda}_a)_{a \in A^t} \in Y^t(x^{t,\lambda})$. On the other hand, as $\mathcal{U}^t$ has dimension zero, the set $Y^t_R(x^{t,\lambda})$ becomes

$$Y^t_R(x^{t,\lambda}) = \{(y_0, y) \in \mathbb{R}^{|A^t|}_+ : \ -y_0 + \left(\boldsymbol{c}^t\right)^\top y \leq 0, y \in Y(x^{t,\lambda})\},$$

and hence, $z^t_R(x^{t,\lambda}) \leq \left(\boldsymbol{c}^t\right)^\top (y^{t,\lambda})_{a \in A^t}$. Therefore, from the first set of inequalities of Theorem 3 (see part (i) in its proof above) and as $z^t_R(x^{t,\lambda}) = z^{t,*}_R$ by definition of $x^{t,\lambda}$, we have that

$$z^{t,\lambda} \leq z^{t,*}_R \leq \left(\boldsymbol{c}^t\right)^\top (y^{t,\lambda})_{a \in A^t} \tag{B.4}$$

but on the other hand, from the definition of $y^{t,\lambda}$ we have that

$$z^{t,\lambda} = \left(\boldsymbol{c}^t\right)^\top (y^{t,\lambda})_{a \in A^t}. \tag{B.5}$$

Equations (B.4) and (B.5) imply that $z^{t,\lambda} = z^{t,*}_R$, 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 3 (see part (ii) in its proof above). ∎

Proof of Lemma 7. First, note that if $y^{t,\lambda}_a > 0$ for some $a \notin A^t$, then the result follows from the assumptions of Value–Perfect feedback. Therefore, suppose that $y^{t,\lambda}_a = 0$ for all $a \notin A^t$. We claim that there exists an activity $a \in A^t \setminus \widetilde{A}^t$ such that $y^{t,\lambda}_a > 0$; the existence of such an activity implies the desired result from the assumptions of Value–Perfect feedback. Indeed, to proceed by contradiction, suppose that this is not the case, i.e., $y^{t,\lambda}_a = 0$ for all $a \in A^t \setminus \widetilde{A}^t$.

As $y^{t,\lambda} \in Y(\bar{x}^{t,\lambda})$ 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 $\hat{\boldsymbol{c}}^t \in \mathcal{U}^t$

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

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

$$z_R^{t,*} \leq \left(\boldsymbol{c}^t\right)^\top (y_a^{t,\lambda})_{a\in A^t}. \tag{B.6}$$

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 3 along with (B.6), we have that $z^{t,\lambda} = z_R^{t,*}$, yielding the desired contradiction. ■

Proof of Lemma 8. As $z^{t,\lambda} < z_R^{t,*}$ there must exist $\tilde{\boldsymbol{c}}^t \in \mathcal{U}^t$ such that

$$z^{t,\lambda} < \left(\tilde{\boldsymbol{c}}^t\right)^\top (y_a^{t,\lambda})_{a\in A^t}.$$

Because $A^{t+1} = A^t$, we have that

$$\mathcal{U}^{t+1} = \{\hat{\boldsymbol{c}}^t \in \mathbb{R}^{|A^t|} : \ \left(\hat{\boldsymbol{c}}^t\right)^\top (y_a^{t,\lambda})_{a\in A^t} = z^{t,\lambda}, \hat{\boldsymbol{c}}^t \in \mathcal{U}^t\},$$

and therefore $\tilde{\boldsymbol{c}}^t \notin \mathcal{U}^{t+1}$.

Now, in view of equation above, $\boldsymbol{G}^{t+1} = (\boldsymbol{G}^t; (y^{t,\lambda})^\top)$ 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))

$$\dim(\mathcal{U}^t) = |A^t| - \operatorname{rank}(\boldsymbol{G}^{t,=}, \boldsymbol{g}^{t,=}). \tag{B.7}$$

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 (B.7). Indeed, arguing by contradiction, suppose that $\operatorname{rank}(\boldsymbol{G}^{t+1,=}, \boldsymbol{g}^{t+1,=}) =$

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

B.2 PROOFS OF THE RESULTS FOR THE MATRIX MODEL

Proof of Proposition 7. For (i), the leftmost inequality follows from the definition of both z^* and $z^{t,\lambda}$ (see equations (3.1) and (4.7)), the fact that the feedback 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 partition of matrices given by (3.4). Therefore, because $x^* \in X$ and **(A3)** holds, one has that $\hat{x} \in X^t$. Now, suppose that $Y_E^t(\hat{x})$ is non-empty (if it is empty then it must be the case that $z_R^{t,*} = \infty$ and the result holds) and let $\hat{y}$ be such that

$$\hat{y} \in \arg\min\{\left(\boldsymbol{c}^t\right)^\top y \colon y \in Y_E^t(\hat{x})\},$$

hence $(\boldsymbol{c}^t)^\top \hat{y} = z_E^t(\hat{x})$. By definition of $z_E^{t,*}$ we have that

$$\left(\boldsymbol{c}^t\right)^\top \hat{y} \le z_E^{t,*}. \tag{B.1}$$

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)** hold, and $\hat{y} \in Y_E^t(\hat{x})$, it follows that $\bar{y} \in Y(x^*)$. Therefore,

$$z^* \le \boldsymbol{c}^\top \bar{y}. \tag{B.2}$$

As $\boldsymbol{c}^\top \bar{y} = \left(\boldsymbol{c}^t\right)^\top \hat{y}$, equations (B.1) and (B.2) yield the desired result.

The proof of (ii) is the same as the proof of (ii) of Theorem 3. For (iii) note that $y^{t,\lambda} \in Y(x^{t,\lambda})$ and $y_a^t = 0$ for all $a \notin A^t$, hence it follows that

$$\sum_{a \in A^t} F_{da} y_a^{t,\lambda} + \sum_{i \in I^t} L_{di} x_i^{t,\lambda} \leq f_d \qquad \forall d \in C_F^t. \tag{B.3}$$

On the other hand, as $\mathcal{U}^t$ has dimension zero, the set $Y_E^t(x^{t,\lambda})$ becomes

$$Y_E^t(x^{t,\lambda}) = \{y \in \mathbb{R}_+^{|A^t|} \colon \boldsymbol{F}^t y + \boldsymbol{L}^t x^{t,\lambda} \leq \boldsymbol{f}^t\},$$

and hence, from equation (B.3) it follows that $(y^{t,\lambda})_{a \in A^t} \in Y_E^t(x^{t,\lambda})$. Therefore, from part (i) and the definition of $z_E^{t,*}$ we have that

$$z^{t,\lambda} \leq z_E^{t,*} \leq \left(\boldsymbol{c}^t\right)^\top (y^{t,\lambda})_{a \in A^t}, \tag{B.4}$$

but on the other hand, from the definition of $y^{t,\lambda}$ we have that

$$z^{t,\lambda} = \left(\boldsymbol{c}^t\right)^\top (y^{t,\lambda})_{a \in A^t}. \tag{B.5}$$

Equations (B.4) and (B.5) imply that $z^{t,\lambda} = z_E^{t,*}$, and hence $\xi^\lambda \leq t$ as desired. The later part of the proposition is a consequence of the above result and part (ii). ∎

Proof of Proposition 8. For (i), note that if $y_a^{t,\lambda} > 0$ for some $a \notin A^t$, then the result follows from the assumptions 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. 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 $y^{t,\lambda} \in Y(\bar{x}^{t,\lambda})$, this assumption implies that

$$\sum_{a \in \widetilde{A}^t} F_{da} y_a^{t,\lambda} + \sum_{i \in I^t} L_{di} x_i^{t,\lambda} \leq f_d \qquad \forall d \in C_F^t. \tag{B.6}$$

Define $y' \in \mathbb{R}_+^{|A^t|}$ as $y'_a := y_a^{t,\lambda}$ for all $a \in A^t$. For any $\hat{\boldsymbol{F}} \in \mathcal{U}^t$, the vector y' satisfies

$$\sum_{a \in \widetilde{A}^t} \hat{F}_{da} y'_a + \sum_{i \in I^t} L_{di} x_i^{t,\lambda} \leq f_d \qquad \forall d \in C_F^t. \tag{B.7}$$

Now, from the definition of $\widetilde{A}^t$, $\hat{F}_{da} = F_{da}$ for all $a \in \widetilde{A}^t$ and $\hat{\boldsymbol{F}} \in \mathcal{U}^t$, hence equations (B.6) and (B.7) imply that $y' \in Y_E^t(x^{t,\lambda})$. Therefore, from the definition of $z_E^{t,*}$ we have that

$$z_E^{t,*} \leq \left(\boldsymbol{c}^t\right)^\top y', \tag{B.8}$$

but because it is readily checked that $z^{t,\lambda} = \left(\boldsymbol{c}^t\right)^\top y'$, and, moreover, that $z^{t,\lambda} \leq z_E^{t,*}$ by part (i) of Proposition 7, Equation (B.8) implies that $z^{t,\lambda} = z_E^{t,*}$, yielding the desired contradiction.

The proof of (ii) is the same as the proof of Theorem 4. ∎

Proof of Lemma 9. It is clear that because $z^{t,\lambda} < z_E^{t,*}$, it must be that $y^{t,\lambda} \notin Y_E^t(x^{t,\lambda})$. By definition of $Y_E^t(x^{t,\lambda})$, this means that there exist $\widetilde{\boldsymbol{F}}^t \in \mathcal{U}^t$ and $d \in C_F^t$ such that $\left(\widetilde{\boldsymbol{F}}^t\right)_d^\top y^{t,\lambda} > f_d - \left(\boldsymbol{L}^t\right)_d^\top x^{t,\lambda}$, as desired. ∎

Before proceeding with the proof of Proposition 9, 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 $r_d^t \coloneqq \sum_{a\colon y_a^t > 0} F_{da} y_a^t = f_d - q_d^t - \boldsymbol{L}_d^\top x^t$. Then, by using the information from the feedback, the leader updates $\mathcal{U}^t$ by including the linear constraints

$$\sum_{a \in y_a^t > 0} y_a^t \hat{F}_{da} = r_d^t \qquad \text{for all } d \in C_F^{t+1}, \tag{B.9}$$

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

$$n_d^t \coloneqq |\{a \in A^t \colon d \in C_F(a)\}|.$$

For any given time $t \in \mathcal{T}$ we have that

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

Suppose that $m^t = |C_F^t|$ 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 (B.9) can be rewritten as

$$\boldsymbol{v}_d^\top \hat{\boldsymbol{F}} = r_d^t, \tag{B.10}$$

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 $\mathcal{S}^\pi(\mathcal{D}^0, \mathcal{D}) := \{t \in \mathcal{T} \colon \exists a \notin A^t \text{ s.t. } y_a^t > 0\}$, 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 $N^k := \{a \in A \setminus A^{s_k} \colon y_a^{s_k} > 0\}$, i.e., N^k is the set of activities the leader learns by the end of time period s_k.

Lemma 18. *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,*

$$\dim(\mathcal{U}^{t+1}) - \dim(\mathcal{U}^t) \leq \begin{cases} \sum_{a \in N^k} |C_F(a)| - \left| \bigcup_{a \in N^k} C_F(a) \right|, & \text{if } t = s_k \text{ for some } k \leq p, \\ -1, & \text{otherwise.} \end{cases} \tag{B.11}$$

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 (B.10) 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 (B.10), 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 $\mathcal{U}^{s_k}$. This 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)

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

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

$$\dim(\mathcal{U}^{s_k+1}) \leq \dim(\mathcal{U}^{s_k}) + \sum_{a\in N^k}|C_F(a)| - \Big|\bigcup_{a\in N^k} C_F(a)\Big|. \tag{B.12}$$

On the other hand, let $t < \xi^\lambda$ such that $t \notin \mathcal{S}^\lambda$; i.e., $y^t_a = 0$ for all $a \notin A^t$. Note that because $\xi^\lambda > t$ one has that $\boldsymbol{c}^\top y^{t,\lambda} < z^t_R$ (by part (i) of Proposition 7). We claim that (recall from the proof of Lemma 8 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 9 hold, let $\widetilde{\boldsymbol{F}}^t$ such that

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

Now consider $\mathcal{U}^t$ after adding the equation $\boldsymbol{v}_d^\top \hat{\boldsymbol{F}} = r^t_d$. Because $q^t_d \geq 0$, one has that $\widetilde{\boldsymbol{F}}^T_d y^{t,\lambda} > f_d - (\boldsymbol{L}^t)^\top_d x^t - q^t_d$ 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 8, 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 19. *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 18, $\dim(\mathcal{U}^t) - \dim(\mathcal{U}^{t-1}) \leq -1$ 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 18 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 (B.12) 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 19 we have the following important result.

Lemma 20. *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,*

$$\tau^\lambda \leq \xi^\lambda \leq \dim(\mathcal{U}^0) + p + \sum_{k=1}^{p} \Big(\sum_{a \in N^k} |C_F(a)| - \Big| \bigcup_{a \in N^k} C_F(a) \Big| \Big). \tag{B.13}$$

Proof. By repeated application of Lemma 19 it is verified that

$$s_p + \dim(\mathcal{U}^{s_p}) \leq \dim(\mathcal{U}^0) + (p-1) + \sum_{k=1}^{p-1} \Big(\sum_{a \in N^k} |C_F(a)| - \Big| \bigcup_{a \in N^k} C_F(a) \Big| \Big). \tag{B.14}$$

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 7 implies that $\xi^\lambda \leq s_p + \tilde{t}$, and hence equation (B.14) and the selection of t yield the desired result. ∎

Proof of Proposition 9. 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 20 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 $\Big|\bigcup_{a\in N^k} C_F(a)\Big| \geq 1$. 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 $y^t_a = 0$ for all $a \notin A^t$. In this case, following the same arguments as in Lemma 18, equation (B.11) 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 20, after mimicking the proofs of the previous results, as in this case equation (B.13) becomes

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

■

Proof of Proposition 10. Observe that as $z^t_R - \boldsymbol{c}^\top y^{t,\lambda} > 0$, then $D^t \neq \emptyset$. Pick an arbitrary $d \in D^t$, and let $y^{t,*}$ be

$$y^{t,*} \in \arg\min\{\left(\boldsymbol{c}^t\right)^\top y \colon y \in Y^t_E(x^{t,\lambda})\},$$

hence $y^{t,*}$ is (one of) the solution(s) the leader expects from the follower after deciding $x^{t,\lambda}$. Now, let $\widetilde{\boldsymbol{F}}$ be given by

$$\widetilde{\boldsymbol{F}} \in \arg\max\{(y^{t,*})^\top \hat{\boldsymbol{F}}_d \,:\, \hat{\boldsymbol{F}} \in \mathcal{U}^t\}$$

thus, $\widetilde{\boldsymbol{F}}$ is (one of) the value(s) the leader assigns to the row d of the follower lower-level matrix by deciding robustly (i.e., by using policy λ). Observe that because $d \in D^t$ it follows that

$$\widetilde{\boldsymbol{F}}_d^\top y^{t,\lambda} > f_d - \boldsymbol{L}_d^\top x^{t,\lambda},$$

and, henceforth, $\widetilde{\boldsymbol{F}} \in \Delta^t$. Now, define the hyperplane $\mathcal{P}$ as

$$\mathcal{P} := \{\hat{\boldsymbol{F}} \in \mathbb{R}^{\sum_{d \in C_F^t} n_d^t} \colon \boldsymbol{v}_d^\top \hat{\boldsymbol{F}} = f_d - \boldsymbol{L}_d^\top x^{t,\lambda}\},$$

and observe that $\mathcal{U}^{t+1}$ is at one side of $\mathcal{P}$ while $\widetilde{\boldsymbol{F}}$ is at the other side. That is, $\mathcal{P}$ separates $\mathcal{U}^{t+1}$ and $\widetilde{\boldsymbol{F}}$. Let $\boldsymbol{F}' \in \mathcal{U}^{t+1}$ be the closest point of $\mathcal{U}^{t+1}$ to $\widetilde{\boldsymbol{F}}$. It is clear that $\boldsymbol{F}' \in \mathcal{P}$, hence $\|\widetilde{\boldsymbol{F}} - \boldsymbol{F}'\|$ is the *distance* from $\widetilde{\boldsymbol{F}}$ to $\mathcal{P}$, which by standard linear algebra is given by the projection of $\widetilde{\boldsymbol{F}} - \boldsymbol{F}'$ onto the vector $\boldsymbol{v}_d$; that is

$$\|\widetilde{\boldsymbol{F}} - \boldsymbol{F}'\| = \frac{(\widetilde{\boldsymbol{F}} - \boldsymbol{F}')^\top \boldsymbol{v}_d^T}{\|y^{t,\lambda}\|} = \frac{(\widetilde{\boldsymbol{F}}_d - \boldsymbol{F}'_d)^\top y^{t,\lambda}}{\|y^{t,\lambda}\|}.$$

Observe that

$$\begin{aligned}
\|\widetilde{\boldsymbol{F}} - \boldsymbol{F}'\| &= \frac{(\widetilde{\boldsymbol{F}}_d - \boldsymbol{F}'_d)^\top y^{t,\lambda}}{\|y^{t,\lambda}\|} && \text{(B.15a)}\\
&= \frac{(\widetilde{\boldsymbol{F}}_d - \boldsymbol{F}'_d)^\top y^{t,\lambda} + \boldsymbol{F}_d'^T y^{t,*} - \boldsymbol{F}_d'^T y^{t,*}}{\|y^{t,\lambda}\|} && \text{(B.15b)}\\
&= \frac{\boldsymbol{F}_d'^T (y^{t,*} - y^{t,\lambda}) + \widetilde{\boldsymbol{F}}_d^\top y^{t,\lambda} - \boldsymbol{F}_d'^T y^{t,*}}{\|y^{t,\lambda}\|} && \text{(B.15c)}\\
&\geq \frac{\boldsymbol{F}_d'^T (y^{t,*} - y^{t,\lambda}) + \widetilde{\boldsymbol{F}}_d^\top (y^{t,\lambda} - y^{t,*})}{\|y^{t,\lambda}\|} && \text{(B.15d)}\\
&= \frac{(y^{t,*} - y^{t,\lambda})^\top (\boldsymbol{F}'_d - \widetilde{\boldsymbol{F}}_d)}{\|y^{t,\lambda}\|}, && \text{(B.15e)}
\end{aligned}$$

where the inequality follows because $\boldsymbol{F}' \in \mathcal{U}^t$. From the definition of $\widetilde{\boldsymbol{F}}$ we have that

$$\boldsymbol{F}_d'^T y^{t,*} \leq \widetilde{\boldsymbol{F}}_d^T y^{t,*}.$$

Now, let U be given by

$$U := \max_{\{a \in A^t \colon \boldsymbol{F}'_{d,a} - \widetilde{\boldsymbol{F}}_{d,a} \neq 0\}} \left| \frac{y_a^{t,*} - y_a^{t,\lambda}}{\boldsymbol{F}'_{d,a} - \widetilde{\boldsymbol{F}}_{d,a}} \right|,$$

therefore, if $K = (U+1)U$, then the Diaz-Metcalf inequality (see e.g Diaz and Metcalf (1964)) implies that

$$(y^{t,*} - y^{t,\lambda})^\top (\boldsymbol{F}'_d - \widetilde{\boldsymbol{F}}_d) \geq \|y^{t,*} - y^{t,\lambda}\|^2 - K \|\boldsymbol{F}'_d - \widetilde{\boldsymbol{F}}_d\|^2. \tag{B.16}$$

Inequalities in (B.15e) and (B.16) imply that

$$\|\widetilde{\boldsymbol{F}} - \boldsymbol{F}'\| \geq \frac{\|y^{t,*} - y^{t,\lambda}\|^2 - K\|\boldsymbol{F}'_d - \widetilde{\boldsymbol{F}}_d\|^2}{\|y^{t,\lambda}\|},$$

and henceforth

$$K\|\widetilde{\boldsymbol{F}} - \boldsymbol{F}'\|^2 + \|y^{t,\lambda}\|\|\widetilde{\boldsymbol{F}} - \boldsymbol{F}'\| \geq \|y^{t,*} - y^{t,\lambda}\|^2.$$

Now, because $z_R^t - \boldsymbol{c}^\top y^{t,\lambda} = \boldsymbol{c}^\top(y^{t,*} - y^{t,\lambda}) > \epsilon$, the Cauchy-Schwartz inequality implies that

$$\|y^{t,*} - y^{t,\lambda}\| > \frac{\epsilon}{\|\boldsymbol{c}\|},$$

and hence

$$K\|\widetilde{\boldsymbol{F}} - \boldsymbol{F}'\|^2 + \|y^{t,\lambda}\|\|\widetilde{\boldsymbol{F}} - \boldsymbol{F}'\| \geq \epsilon^2\|\boldsymbol{c}\|^{-2}.$$

Because $K \geq 0$, the above inequality implies the desired result. ∎

B.3 ADDITIONAL RESULTS AND COMPLEMENTARY MATERIAL

B.3.1 Semi-Oracle Algorithm

In this section we discuss an algorithm that speeds-up the solution of problem (3.13) 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 (3.13) 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 $I^* := \{i \in I \colon x_i^* > 0\}$ be the set of resources that x^* uses. For any $J \subseteq I^*$ define $x^{*,J}$ as $x_i^{*,J} := x_i^*$ 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 $I(y)$ is the set of resources that interfere with the activities that y performs (i.e., with a slight abuse of notation)

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

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 the first observation is straightforward, the proof of the second is a consequence of the following lemma.

Lemma 21. *Let $J \subseteq I^*$ and suppose that $y^J \in \arg\min\{\boldsymbol{c}^\top y\colon y \in Y(x^{*,J})\}$. 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

$$\begin{aligned}\sum_{a\in A} F_{da}y^J_a + \sum_{i\in I^*} L_{di}x^*_i &= \sum_{a\in A} F_{da}y^J_a + \sum_{i\in J} L_{di}x^*_i + \sum_{i\in I^*\setminus J} L_{di}x^*_i \\ &= \sum_{a\in A} F_{da}y^J_a + \sum_{i\in J} L_{di}x^*_i + \sum_{i\in K_1} L_{di}x^*_i + \sum_{i\in K_2} L_{di}x^*_i, \qquad \text{(B.1)}\end{aligned}$$

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 equation (B.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 $\sum_{a\in A} F_{da}y^J_a = 0$; then (B.1) is at most f_d by Assumption **A4**. Hence, suppose that $d \in C_F$ satisfies that $\sum_{a\in A} F_{da}y^J_a \neq 0$. 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 $\sum_{a\in A} F_{da}y^J_a \neq 0$, 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 $\sum_{a\in A} F_{da}y^J_a \neq 0$ then

$$\sum_{a\in A} F_{da}y^J_a + \sum_{i\in I^*} L_{di}x^*_i = \sum_{a\in A} F_{da}y^J_a + \sum_{i\in J} L_{di}x^*_i \leq f_d,$$

where the inequality in the above expression follows from the assumption that $y^J \in Y(x^{*,J})$. Thus, (B.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 4 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 22.

Algorithm 4 Finding an initial feasible solution to (3.13).

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

Lemma 22. *Let T^0 be as computed by Algorithm 4. Then, T^0 is an upper bound on the optimal value of problem (3.13), 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 4 is feasible in (3.13) 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 21 implies that there exist $i \in I(y^s) \cap I^*$ such that $i \notin J^s$. Henceforth, $|J^{s+1} \setminus J^s| \geq 1$.

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,

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

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 (B.2) implies that $J^t = I^*$, and hence, by observation (i) that $z^t = z^*$; which yields the desired contradiction. ∎

By using Algorithm 4, an optimal solution of (3.13) can be readily computed via Algorithm 5. The correctness of Algorithm 5 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 14. *Algorithm 5 correctly solves program (3.13).*

B.3.2 Numerical Computation of Policies in Λ

The following result establishes that $x^{t,\lambda}$ and $z_R^{t,*}$ can be computed by solving a mixed-integer linear problem.

Lemma 23. *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,*

$$z_R^{t,*} = \max \ \left(\boldsymbol{g}^t\right)^\top p \tag{B.3a}$$

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

$$\left(\boldsymbol{G}^t\right)^\top p - y = \boldsymbol{0} \tag{B.3c}$$

Algorithm 5 Finding an optimal solution to (3.13)

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

Compute $(T^0, z^*, \{(x^t, y^t) \colon t \in \mathcal{T}\})$ by calling **Algorithm 4** using $((\mathcal{D}^0, \mathcal{D}), T)$

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

Solve program (3.13) until time T^0 passing $\{(x^t, y^t) \colon t = 0, \cdots, T^0\}$ as an initial feasible solution, let τ^* be the objective value

else

Solve program (3.13) until time T passing $\{(x^t, y^t) \colon t = 0, \cdots, T\}$ as an initial feasible solution, let τ^* be the objective value

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

$\tau^* = \infty$

end if

end if

return τ^*

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

$$\boldsymbol{G}^t \hat{c}^t \leq \boldsymbol{g}^t \tag{B.3e}$$

$$-\left(\boldsymbol{F}^t\right)^\top q - \hat{c}^t \leq \boldsymbol{0} \tag{B.3f}$$

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

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

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

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

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

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

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* (B.3b)–(B.3f), *then the matrix* $\boldsymbol{M}^q$ *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* $d \in C_F^t$ *(the*

matrices $\boldsymbol{M}^p$ and $\boldsymbol{M}^y$ are defined in an analogous manner). Moreover, the vector $x^{t,\lambda}$ 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 (B.3a)–(B.3l).

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

$$\max_{x \in X^t} \min_{y_0, y} y_0 \tag{B.4a}$$

$$\text{s.t.} \quad \left(\hat{\boldsymbol{c}}^t\right)^\top y \leq y_0 \qquad \forall \hat{\boldsymbol{c}}^t \in \mathcal{U}^t \tag{B.4b}$$

$$-\boldsymbol{F}^t y \geq \boldsymbol{L}^t x - \boldsymbol{f}^t \tag{B.4c}$$

$$y \geq 0 \tag{B.4d}$$

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

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

(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, problem (B.4) is equivalent to

$$\begin{aligned}
&\max_{x \in X^t} \min_{y} \left(\boldsymbol{g}^t\right)^\top p \\
&\text{s.t.} \quad -y + \left(\boldsymbol{G}^t\right)^\top p = \boldsymbol{0} \\
&\qquad -\boldsymbol{F}^t y \geq \boldsymbol{L}^t x - \boldsymbol{f}^t \\
&\qquad y \geq 0
\end{aligned}$$

Because 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

$$\max_{x \in X^t} \left(\boldsymbol{g}^t\right)^\top p \tag{B.5a}$$

$$\text{s.t.} \quad -y + \left(\boldsymbol{G}^t\right)^\top p = \boldsymbol{0} \tag{B.5b}$$

$$-\boldsymbol{F}^t y \geq \boldsymbol{L}^t x - \boldsymbol{f}^t \tag{B.5c}$$

$$-\left(\boldsymbol{F}^t\right)^\top q - \hat{c}^t \leq \boldsymbol{0} \tag{B.5d}$$

$$\boldsymbol{G}^t \hat{c}^t \leq \boldsymbol{g}^t \tag{B.5e}$$

$$(\boldsymbol{f}^t - \boldsymbol{F}^t - \boldsymbol{L}^t x)^\top q = 0 \tag{B.5f}$$

$$(\boldsymbol{g}^t - \boldsymbol{G}^t \hat{c}^t)^\top p = 0 \tag{B.5g}$$

$$\left(\left(\boldsymbol{F}^t\right)^\top q + \hat{c}^t\right)^\top y = 0 \tag{B.5h}$$

$$y \geq 0, q \geq 0, p \geq 0, \hat{c}^t \text{ free.} \tag{B.5i}$$

Observe that problem (B.5) 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))

$$q \leq \boldsymbol{M}^q v^1 \text{ and } \boldsymbol{f} - \boldsymbol{F}^t y - \boldsymbol{L}^t x \leq \boldsymbol{M}^q(\mathbf{1} - v^1).$$

A similar equivalence exists between the other two set of complementary slackness constraints in problem (B.5). The desired result follows. ■

B.3.3 Sequential Assignment Interdiction

We complement the example applications presented in Section 3.2 by modeling an interdiction assignment problem. Consider the problem discussed in Zenklusen (2010). Here the enemy is the follower, who at each time has to assign each agent in a set V to exactly one job in a set W at minimum cost; assigning agent $v \in V$ to job $w \in W$ costs the follower c_{vw}. Define y_{vw} as 1 if v is assigned to w, and zero otherwise. The follower, absent the interventions of the leader, solves the following minimum-weighted matching (assignment) problem on the bipartite graph $G = (V \cup W, E)$, with $E := V \times W$:

$$y^* \in \arg\min_y \{\boldsymbol{c}^\top y \colon \boldsymbol{M}^V y \leq \mathbf{1}, \boldsymbol{M}^W y \leq \mathbf{1}, -\boldsymbol{M}^W y \leq -\mathbf{1}, y \in \{0,1\}^{|E|}\}.$$

In this formulation $\boldsymbol{M}^V$ is a $|V| \times |E|$ (undirected) vertex-edge adjacency matrix, where $M_{v,(v,w)} = 1$ for all $v \in V$, and zero otherwise; similarly, $\boldsymbol{M}^W$ is a $|W| \times |E|$ matrix where $M_{w,(v,w)} = 1$ for all $w \in W$, and $\mathbf{1}$ is a vector of ones. Observe that the constraints enforce $\boldsymbol{M}^W y = \mathbf{1}$, which means that each job must be taken by some agent. Also, note that while

the above program is binary, the binary restrictions can be relaxed as the constraint matrix is totally unimodular, and hence it can be replaced by its linear programming relaxation.

The leader, on the other hand, has the ability to disable agents in V (the settings where she can disable assignments in E or jobs in W follow similar lines). Disabling agent v during each time period costs her b_v and she has a total budget of B at each period. Thus, if we let x_v take the value 1 if the leader disables v and zero otherwise, she faces the constraints $\sum_{v \in V} b_v x_v \leq B$, and $x_v \in \{0, 1\}$ for all $v \in V$ at each time period.

The above problem can be modeled within our framework as follows: the set of follower activities is E, thus $A = E$, and the set of follower constraints C_F consist of the restrictions regarding the assignment at each vertex, so $|C_F| = |V| + 2|W|$. It is readily seen that $\boldsymbol{F} = [\boldsymbol{M}^V; \boldsymbol{M}^W; -\boldsymbol{M}^W]$, $\boldsymbol{f} = [\mathbf{1}; \mathbf{1}; -\mathbf{1}]$, and that the cost vector $\boldsymbol{c}$ is given by the assignment costs, thus $\boldsymbol{c} = (c_e \colon e \in E)$.

The leader resources are given by $I = V$ and C_L is a singleton consisting on the budgetary constraint, hence $\boldsymbol{H} = \boldsymbol{b}^\top$, where $\boldsymbol{b} = (b_v \colon v \in V)$ and $\boldsymbol{h} = B$. Matrix $\boldsymbol{L}$, on the other hand, has the agent-disabling constraints. Thus, $\boldsymbol{L} = (\boldsymbol{I}; \mathbf{0})$, where $\boldsymbol{I}$ is a $|V| \times |V|$ identity matrix and $\mathbf{0}$ is a $2|W| \times |V|$ matrix of zeros.

Initially, the leader has knowledge about all the jobs W, but potentially ignores all possible agents as well as some of the possible assignments and their corresponding cost. For those assignments $A^0 \subseteq E$ she knows, she has interval estimates $\ell_e \leq c_e \leq m_e$, hence $\mathcal{U}^0 = \{\hat{\boldsymbol{c}}^0 \in \mathbb{R}^{A^0} \colon \ell_e \leq \hat{c}^0_e \leq m_e \ \forall e \in A^0\}$, thus $\boldsymbol{G}^0 = [\boldsymbol{I}; -\boldsymbol{I}]$ and $\boldsymbol{g}^0 = (\boldsymbol{m}; \boldsymbol{\ell})$, with $\boldsymbol{m} = (m_e \colon e \in A^0)$ and $\boldsymbol{\ell} = (\ell_e \colon e \in A^0)$.

Note that in this example the follower performs activity $e = (u, w) \in A$ whenever agent u is assigned to job w. The leader uses resource v if she disables agent v. The set $C_F(v, w)$ consist of the three constraints $\sum_{(v,w') \in E} y_{(v,w')} \leq 1$, $\sum_{(v',w) \in E} y_{(v',w)} \leq 1$ and $-\sum_{(v',w) \in E} y_{(v',w)} \leq -1$, while for any $v \in I$ there is only one constraint in $C_L(v)$ which corresponds to the budgetary constraint, thus $C_L(v) = C_L$ for all $v \in I$. For any $(v, w) \in A$, the set $I(v, w)$ corresponds to that agent whose disabling stops assignment of agent v to job w, i.e., $I(v, w) = \{v\}$.

In this example, standard feedback implies that at each time $t \in \mathcal{T}$ the leader always observes the cost incurred by the follower at time t. If the follower makes an assignment

$(v, w) \in A$ that the leader did not observe before, then the leader learns that the assignment between agent v and job w is possible. Moreover, she learns $C_F(v, w)$, and as such, if agent v was never used before by the follower, she also learns about the existence of agent v. Also, the leader learns $I(v, w)$ and L_{di} for all $d \in C_F(a)$ and all $i \in I(a)$, and as such, she learns that by disabling agent v she can disable the assignment (v, w). Finally, she also learns that disabling v costs her b_v.

Finally, in this setting, as the follower responses are binary, then by assumption **S2**, standard feedback is automatically Response–Perfect. On the other hand, in Value–Perfect feedback, the leader, besides observing the assignments, also observes the costs incurred by the follower when performing each of the assignments made at time $t \in \mathcal{T}$.

BIBLIOGRAPHY

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

Ahuja, R., Magnanti, T. and Orlin, J. (1993), *Network flows: Theory, algorithms, and applications*, Prentice-Hall.

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

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

Auer, P., Cesa-Bianchi, N., Freund, Y. and Schapire, R. E. (1995), Gambling in a rigged casino: The adversarial multi-armed bandit problem, *in* P. Ragbavan, ed., 'Proceedings of 36th Annual Symposium on Foundations of Computer Science', IEEE, pp. 322–331.

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.

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

Awerbuch, B. and Kleinberg, R. D. (2004*a*), Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches, *in* L. Babai, ed., 'Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing', ACM, pp. 45–53.

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

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

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

Beheshti, B., Özaltın, O. Y., Zare, M. H. and Prokopyev, O. A. (2015), 'Exact solution approach for a class of nonlinear bilevel knapsack problems', *Journal of Global Optimization* **61**(2), 291–310.

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.

Birge, J. R. and Louveaux, F. (2011), *Introduction to stochastic programming*, Springer Science & Business Media.

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

Bouhtou, M., Grigoriev, A., Hoesel, S. v., van der Kraaij, A. F., Spieksma, F. C. and Uetz, M. (2007), 'Pricing bridges to cross a river', *Naval Research Logistics* **54**(4), 411–420.

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*

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

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., Freund, Y., Haussler, D., Helmbold, D. P., Schapire, R. E. and Warmuth, M. K. (1997), 'How to use expert advice', *Journal of the ACM* **44**(3), 427–485.

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

Cesa-Bianchi, N. and Lugosi, G. (2006*b*), *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. (2007*a*), 'An overview of bilevel optimization', *Annals of Operations Research* **153**(1), 235–256.

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

Corley, H. and Chang, H. (1974), 'Finding the n most vital nodes in a flow network', *Management Science* **21**(3), 362–364.

Corley, H. and Sha, D. (1982), 'Most vital links and nodes in weighted networks', *Operations Research Letters* **1**(4), 157–160.

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

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

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

Diaz, J. B. and Metcalf, F. T. (1964), 'Complementary inequalities i: Inequalities complementary to cauchy's inequality for sums of real numbers', *Journal of Mathematical Analysis and Applications* **9**(1), 59–74.

Erdös, P. and Rényi, A. (1959), 'On random graphs, I', *Publicationes Mathematicae (Debrecen)* **6**, 290–297.

Freund, Y. and Schapire, R. E. (1997), 'A decision-theoretic generalization of on-line learning and an application to boosting', *Journal of Computer and System Sciences* pp. 119–139.

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

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

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

Granata, D., Steeger, G. and Rebennack, S. (2013), 'Network interdiction via a critical disruption path: Branch-and-price algorithms', *Computers & Operations Research* **40**(11), 2689–2702.

Gyorgy, A., Linder, T., Lugosi, G. and Ottucsak, G. (2007), 'The on-line shortest path problem under partial monitoring', *Journal of Machine Learning Research* **8**(10), 2369–2403.

Hausken, K. and Zhuang, J. (2011), 'Governments' and terrorists' defense and attack in a t-period game', *Decision Analysis* **8**(1), 46–70.

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

Hazan, E., Agarwal, A. and Kale, S. (2007), 'Logarithmic regret algorithms for online convex optimization', *Machine Learning* **69**(2), 169–192.

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.

Held, H. and Woodruff, D. (2005), 'Heuristics for multi-stage interdiction of stochastic networks', *Journal of Heuristics* **11**(5-6), 483–500.

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

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

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

Kalai, A. and Vempala, S. (2005), 'Efficient algorithms for online decision problems', *Journal of Computer and System Sciences* **71**(3), 291–307.

Kleinberg, R., Niculescu-Mizil, A. and Sharma, Y. (2010), 'Regret bounds for sleeping experts and bandits', *Machine Learning* **80**(2-3), 245–272.

Koolen, W. M., Warmuth, M. K. and Kivinen, J. (2010), Hedging structured concepts, *in* A. T. Kalai and M. Mohri, eds, 'Proceedings of the 23rd Annual Conference on Learning Theory (COLT)', Omnipress, pp. 93–105.

Labbé, M., Marcotte, P. and Savard, G. (1998), 'A bilevel model of taxation and its application to optimal highway pricing', *Management Science* **44**(12), 1608–1622.

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

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

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

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

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

McLay, L., Rothschild, C. and Guikema, S. (2012), 'Robust adversarial risk analysis: a level-k approach', *Decision Analysis* **9**(1), 41–54.

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

Migdalas, A., Pardalos, P. M. and Värbrand, P. (2013), *Multilevel optimization: algorithms and applications*, Vol. 20, Springer Science & Business Media.

Modaresi, S., Saure, D. and Vielma, J. (2012), Learning in combinatorial optimization: What and how to explore. Working paper.

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

Neu, G. and Bartók, G. (2013), An efficient algorithm for learning with semi-bandit feedback, *in* S. Jain, R. Munos, F. Stephan and T. Zeugmann, eds, 'Algorithmic Learning Theory', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 234–248.

Pan, F. and Morton, D. (2008), 'Minimizing a stochastic maximum-reliability path', *Networks* **52**(3), 111–119.

Puterman, M. L. (2014), *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons.

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

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

Saharidis, G. K., Conejo, A. J. and Kozanidis, G. (2013), Exact solution methodologies for linear and (mixed) integer bilevel programming, *in* 'Metaheuristics for Bi-level Optimization', Springer, pp. 221–245.

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.

Sen, S. and Sherali, H. D. (2006), 'Decomposition with branch-and-cut approaches for two-stage stochastic mixed-integer programming', *Mathematical Programming* **106**(2), 203–223.

Shen, S. and Smith, J. C. (2012), 'Polynomial-time algorithms for solving a class of critical node problems on trees and series-parallel graphs', *Networks* **60**(2), 103–119.

Shen, S., Smith, J. C. and Goli, R. (2012*a*), 'Exact interdiction models and algorithms for disconnecting networks via node deletions', *Discrete Optimization* **9**(3), 172–188.

Shen, S., Smith, J. and Goli, R. (2012*b*), 'Exact interdiction models and algorithms for disconnecting networks via node deletions', *Discrete Optimization* **9**(3), 172–188.

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.

Van Hoesel, S. (2008), 'An overview of Stackelberg pricing in networks', *European Journal of Operational Research* **189**(3), 1393–1402.

Veremyev, A., Boginski, V. and Pasiliao, E. L. (2014), 'Exact identification of critical nodes in sparse networks via new compact formulations', *Optimization Letters* **8**(4), 1245–1259.

Veremyev, A., Prokopyev, O. A. and Pasiliao, E. L. (2014), 'An integer programming framework for critical elements detection in graphs', *Journal of Combinatorial Optimization* **28**(1), 233–273.

Veremyev, A., Prokopyev, O. A. and Pasiliao, E. L. (2015), 'Critical nodes for distance-based connectivity and related problems in graphs', *Networks* **66**(3), 170–195.

Walteros, J. L. and Pardalos, P. M. (2012), Selected topics in critical element detection, *in* N. J. Daras, ed., 'Applications of Mathematics and Informatics in Military Science', Vol. 71 of *Springer Optimization and Its Applications*, Springer New York, pp. 9–26.

Washburn, A. and Wood, R. (1995), 'Two-person zero-sum games for network interdiction', *Operations Research* **43**(2), 243–251.

Wollmer, R. (1964), 'Removing arcs from a network', *Operations Research* **12**(6), 934–940.

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* .

Xu, J. and Zhuang, J. (2014), 'Modeling costly learning and counter-learning in a defender-attacker game with private defender information', *Annals of Operations Research* . Forthcoming.

Zenklusen, R. (2010), 'Matching interdiction', *Discrete Applied Mathematics* **158**(15), 1676–1690.

Zhuang, J., Bier, V. M. and Alagoz, O. (2010), 'Modeling secrecy and deception in a multiple-period attacker–defender signaling game', *European Journal of Operational Research* **203**(2), 409–418.

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.