A Study on Multilevel Optimization Subject to Uncertainty

by

Liang Xu

B.S. in Applied Mathematics, Dalian University of Technology, 2006

M.Eng. in Software Engineering, Dalian University of Technology, 2008

M.S. in Finance, University of South Florida, 2014

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

2023

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Liang Xu

It was defended on

December 13 2022

and approved by

Daniel R. Jiang, Ph.D., Assistant Professor, Department of Industrial Engineering

Hoda Bidkhori, Ph.D., Assistant Professor, Department of Industrial Engineering

Zhi-Hong Mao, Ph.D., Professor, Department of Electrical and Computer Engineering

Dissertation Director: Bo Zeng, Ph.D., Associate Professor, Department of Industrial Engineering

A Study on Multilevel Optimization Subject to Uncertainty

Liang Xu, PhD

University of Pittsburgh, 2023

Multilevel, including bilevel, optimization has been widely applied in real world hierarchical systems, such as power systems and transportation systems. However, this optimization scheme is complex and its extension to handle uncertainty is very limited. In this dissertation, we explore the mathematical structure and develop efficient solution methods for two types of such optimization problem: bilevel mixed-integer nonlinear programming and robust bilevel optimization. Two applications, wind farm capacity expansion problem and optimal decision tree problem, are investigated using the proposed methods.

In our first study, we consider general bilevel mixed-integer nonlinear programming problems. By analyzing the structure of the problem, we provide optimality conditions based reformulation and computing scheme for both optimistic and pessimistic cases.

Our second study focuses on bilevel optimization with uncertainty and develops robust bilevel optimization (RBO) models along with solution methods. We first study single-stage RBO problems and provide solution methods to deal with different types of uncertainties. For single-stage RBO with discrete uncertainty set, we develop a novel cut-and-branch algorithm. We then study two-stage RBO problems, which involve wait-and-see decisions. We provide two basic models and their variations, as well as column-and-constraint generation algorithms to exactly handle uncertainties.

Finally, we apply our proposed methods to wind farm capacity expansion problem and optimal decision tree problem. In the first application, we formulate the wind farm investment problem into a two-stage RBO model and solve it by a proposed column-and-constraint generation algorithm. In the second application, we develop a new mixed-integer programming (MIP) based formulation to construct an optimal classification tree. We improve the generalizability of the model through a data-driven hyperparameter tuning approach in the bilevel optimization framework.

Table of Contents

List of Tables

List of Figures

Preface

First, I would like to give special thanks to my advisor, Dr. Bo Zeng, for his consistent support during my Ph.D. study. He dedicates so much time to work with me on my research, helping me not only overcome technical challenges but also grow up to be an independent researcher. His passion toward scientific work impresses and encourages me a lot, especially during my hard time. Without his guidance and encouragement, it is impossible to complete the research in this dissertation. It is my great honor to work with Dr. Zeng, and I believe such experience will benefit my whole career.

I would also like to thank Dr. Daniel R. Jiang, Dr. Hoda Bidkhori, and Dr. Zhi-Hong Mao for serving as my Ph.D. dissertation committee member. They help me to improve my dissertation from more perspectives with their valuable comments. Furthermore, I would like to thank the faculty in the industrial engineering department for their high quality inspiring courses and the staff team for their great support.

Finally, I want to give thanks to my family, especially to my wife and my parents, for their unconditional love and support. Thanks to all my beloved friends for accompanying me through this unforgettable journey. I am so grateful to have you in my life!

1.0 Introduction

Bilevel optimization [43, 49] is a powerful framework to model and investigate hierarchical systems such as power systems and transportation systems. However, solving bilevel optimization problems is very challenging, and the simplest bilevel linear programming problems have been proven NP-hard [53]. Moreover, uncertainty such as missing data and estimation error is almost always involved in a decision making process. In this dissertation, we study two types of bilevel optimization problems, namely, bilevel mixed integer nonlinear programming problems and robust bilevel optimization problems. Two applications that take the advantage of our proposed model and solution method are also included in this dissertation.

1.1 Bilevel Optimization

Many practical systems are organized and operated in a hierarchical structure, where two types of decision makers (DMs) with different interests act in two levels. An upper level DM (referred to as she) first makes her decision, which is then passed to DM(s) in the lower level (referred to as he for a single DM) and affects his reasoning. After the lower level DM makes his decision, he discloses it to the upper level DM. The performance of the upper level DM will be evaluated by the aggregated decisions, i.e., the decision from her and the one(s) from DM(s) in the lower level. For example, in a highway system, the transportation authority, acting as the upper level DM, sets tolls on arcs of a transportation network to generate revenue, and vehicle drivers, acting as the lower level DMs, select their routes considering travelling costs. It is clear that the revenue received by the authority is jointly determined by the overall toll scheme and traffic flows on arcs.

Similar situations can be found in a deregulated electricity market consisting of the market administrator and market participants regarding power generation, market bidding, and capacity expansion issues [114, 72, 77], a waste management system consisting of a regional planning agency and private firms to determine effective pollution control policies

[6], an emergency evacuation system consisting of a central planner and evacuees to determine shelter locations and support facilities [69]. Another typical application can be found in a security system where one defender allocates protective resources with an assumption that attacker(s) respond her decisions subsequently. Also, an interesting application arises in biological engineering where the hierarchical structure includes the human DM in the upper level to make biochemical or genetic changes and the biological system as the lower level DM to respond according to biological mechanisms [110, 34]. Indeed, in many studies of economics, such sequential hierarchical decision making has long been recognized as Stackelberg leader-follower game [123], where the leader and the follower correspond to the aforementioned upper and lower DMs respectively. So, as in most existing literature, the leader (the follower, respectively) and the upper level DM (the lower level DM, respectively) are used interchangeably in this dissertation for ease of exposition.

To quantitatively analyze that hierarchical decision making process, especially to support the upper level DM, the regular monolithic optimization formulation is extended in a nested fashion to the following bilevel optimization (BO) formulation. Let $\mathbf{x}$ and $\mathbf{y}$ denote decision variables, and F and f be objective functions of DMs in the upper and the lower levels, respectively. The BO formulation is

$$\mathbf{BO}: \quad \Theta^* = \min_{(\mathbf{x},\tilde{\mathbf{y}})} F(\mathbf{x},\tilde{\mathbf{y}}) \tag{1.1}$$

$$\text{s.t.} \quad G(\mathbf{x}) \leq \mathbf{0}, \mathbf{x} \in \mathbb{X} \subseteq \mathbb{R}_+^{m_c} \times \mathbb{Z}_+^{m_d}, \tag{1.2}$$

$$\tilde{\mathbf{y}} \in \phi(\mathbf{x}) \equiv \arg\min\{f(\mathbf{x},\mathbf{y}) : g(\mathbf{x},\mathbf{y}) \leq 0, \mathbf{y} \in \mathbb{Y} \subseteq \mathbb{R}_+^{n_c} \times \mathbb{Z}_+^{n_d}\} \tag{1.3}$$

We mention that the embedded optimization model in (1.3), which takes the upper level decision $\mathbf{x}$ as input parameters, represents the decision making problem of the lower level DM. The logic of the complete decision making goes as follows. The upper level DM selects a particular $\mathbf{x}$ considering the upper level constraints, and sends it to the follower. After receiving $\mathbf{x}$, the follower computes and feeds back an optimal solution $\tilde{\mathbf{y}}$ to the upper level. Hence, we, on behalf of the leader, should search for a feasible $\mathbf{x}$ that can generate, in an indirect way, a desired $\tilde{\mathbf{y}}$ to jointly minimize her cost function represented by $F(\mathbf{x},\tilde{\mathbf{y}})$. This formulation was initially presented in [29, 30] and referred to two-level or bilevel optimization model in [36] in 1970s. Since then, because of its strong modeling capacity and practical significance, bilevel optimization has received enormous research interests on its mathematical structures, computational algorithms, and applications to solve real problems [43, 49].

It has been well recognized in the literature of the regular monolithic (i.e., single level) optimization that assuming the deterministic information, i.e., no random or unknown parameters, is rather restrictive. It often fails to yield a solution that is feasible or with an acceptable performance in practice. To address this challenge, countless research has been devoted to modeling, analyzing or mitigating random factors or uncertainties within an optimization model. Up now, many variants that extend and generalize a deterministic single level model to consider uncertainties have been developed, and abundant theoretical and computational studies have been appeared in the literature. Examples include stochastic programming, robust optimization and distributionally robust optimization schemes and related studies. Nevertheless, we observe that the study of bilevel optimization subject to uncertainty is rather limited and many critical issues remain open. As described previously, the challenge of uncertainty is generally fundamental and unavoidable in such a hierarchical decision making problem. Compared with that of regular single level mathematical programs, the uncertainty issue in bilevel optimization is more prevalent, demonstrates a richer variety of forms, and generates more complex impacts directly or indirectly on both DMs.

1.2 Bilevel Optimization With Uncertainty

In the context of regular optimization built for a single decision maker, many random factors presented in the real world have been considered and studied. As they often directly impact the decision making process and solutions' performance, those factors generally can be easily recognized, modeled and then incorporated into an optimization formulation. Nevertheless, for bilevel optimization, the situation could be very different. One reason behind is that some new types of random factors unique to this hierarchical system arise, which might be hidden in this system or harder to be described. Another reason is that the participation of random factors in the decision making process could be much more involved. Indeed, given that two DMs exchange information and take actions in a sequential way, this whole process might expose itself to various random factors and become vulnerable to them. To help us understand random factors' sources and their roles in this hierarchical system, we introduce in Figure 1 a scheme that classifies them into three categories.

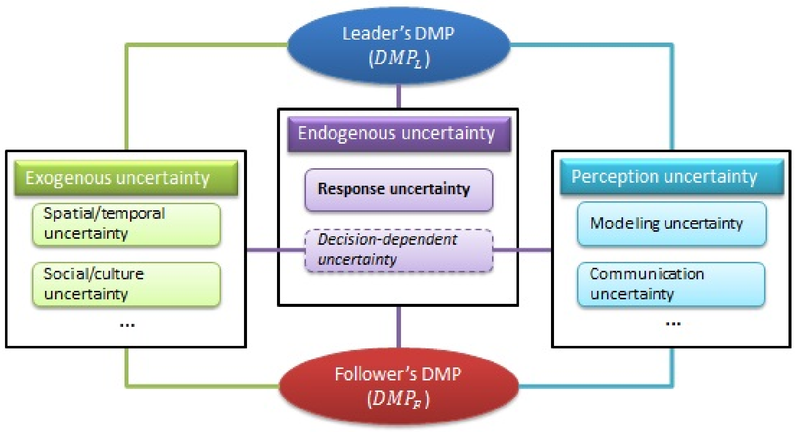

Figure 1: Different Types of Uncertainties

The left category in Figure 1, i.e., the exogenous uncertainties, includes random factors originated outside of this hierarchical system. Basically all random factors occurred in the external environment of this system, including those in the natural or the social/culture environment, can be classified into this group. Typical examples include the wind speed, the prices of products, and the customers' choices, as long as that the nature, the price maker or customers are not DMs in this system. Note that those random factors generally can be observed and modeled easily, and their participation in the decision making process is often straightforward. Hence, up to now, the majority of uncertainty studies on regular monolithic optimization (i.e., optimization with single DM) and almost all of them on bilevel optimization are concerned with this type of randomness [107, 27, 25, 17, 60]. Different from this classical group, the middle and the right categorise of uncertainties are either unique to the fundamental structure of this hierarchical system or much less common, to which we refer as the endogenous uncertainties and the uncertainties in perception, respectively.

A bilevel optimization model is built on half of the leader. It is often the case that the leader might not have perfect information of the follower. As a result, the lower level decision making problem (DMP) she adopts may only be a surrogate of the actual one. In another case, the follower may be a dynamic system switching between two or more working modes over time, and such a system is not transparent to the leader. In both cases, the leader will only have imperfect information regarding the follower's DMP. Note also that two-way communications are involved to transmit information between DMs. Since communications are subject to intentional or unintentional modifications or errors, DMs could receive inaccurate or erroneous information in their decision making process. Since all those uncertainties, as well as other similar uncertainties occurred within the system, are reflected in one DM's understanding of the other one, we classify them as the uncertainties in perception, as shown in the right category in Figure 1. Actually, it is often observed that the leader uses a single lower level DMP or a simple one to approximate a set of lower level DMPs or a complex one (e.g., a piecewise linear function to replace a nonlinear one) to simplify her reasoning. Such approximation inevitably carries uncertainties in her perception.

The last category, i.e., the endogenous uncertainties in Figure 1, is introduced to represent the following two types of random factors that are influenced by decisions.

1. One is the response uncertainty, which reflects an essential challenge underlying bilevel optimization. It happens when the optimal solution set of the lower level DMP is not guaranteed to be a singleton. If the follower is not fully cooperative with the leader and is neutral to any optimal solution, his response towards the leader's decision is subject to implicit uncertainty hidden between DMs. This observation has been well recognized in the bilevel optimization literature for a long time [98, 89, 94]. Actually, researchers do not consider it from the perspective of uncertainty at the beginning. Nevertheless, different treatments on this issue has led to a couple of fundamental bilevel optimization formulations that coincide to several popular strategies on handling uncertainties. It is worth mentioning that, if the follower is tolerable to take an ϵ-optimal decision (which has been interpreted as the bounded rationality), such response uncertainty is almost unavoidable when continuous decisions are involved.
2. Another type is the decision-dependent uncertainty, a concept originally introduced for the regular optimization [64, 101]. It indeed describes a situation when some exogenous random factor's sampling space and/or distribution is changing with respect to

the particular values of decision variables. In bilevel optimization, this concept is clearly applicable when the upper or lower level DM has to handle similar exogenous random factor. Moreover, we mention that some uncertainties in perception could be evolved into decision-dependent ones, noting that the scale of communication error or noise could depend on the genuine signals. As a decision-dependent uncertainty stems from some uncertainty in other two categories, we enclose the general concept by a dashed box with links to two other types of uncertainties in Figure 1 to highlight such connections.

Table 1: Comparison of Different Categories of Uncertainties

Uncertainties	Source	Attributes	Controllability
Exogenous	external factors	independent of the system	not affected by DMs
Endogenous	DM's decisions (upon external factors)	within the system	directly affected by DMs
Perception	insufficient understandings between DMs	within the system	might be reduced or refined

A comparison among the three types of uncertainties is summarized in Table 1. Besides the aforementioned classification scheme, we note that it is common to have multiple random factors of different types co-exist and jointly affect a hierarchical system. For example, the leader needs to handle an exogenous random factor, but she just has incomplete information regarding the follower's DMP. The situation could be more complicated if the follower is tolerable to ϵ-optimal decisions, yielding a pool of choices for him to select. Under such a situation, the leader clearly needs to address those three types of random factors in a holistic approach if she expects a sound decision. Indeed, different combinations of various random factors could be found in practice, which indicates modeling tools should be flexible to capture them within the associated bilevel optimization problem.

The hierarchical structure of bilevel optimization naturally provides a stage-wise decision making interpretation. The consideration of random factors within a regular optimization model generally renders it a stage-wise decision making structure. In particular, if there exist recourse opportunities after (some) random factors are fixed, i.e., some decisions can be flexibly made according to the realized scenario, that regular optimization model will expand into either a two-stage or a multi-stage model, depending on the number of recourse opportunities [105, 13]. Otherwise, that optimization model remains a single-stage formulation. In the context of bilevel optimization, we present in Figure 2 a schematic diagram showing the basic interactions between random factors (collectively denoted by u) and the associated hierarchical decision making system. Note that the upper and lower level variables $\mathbf{x}$ and $\mathbf{y}$ are separated into $(\mathbf{x}_1, \mathbf{x}_2)$ and $(\mathbf{y}_1, \mathbf{y}_2)$, respectively, to indicate the decisions made before and after the randomness is fixed. Additionally, $\mathbf{x}^*$ and $\mathbf{y}^*$ are decisions transmitted between the leader and the follower, which are subjected to the aforementioned communication errors.

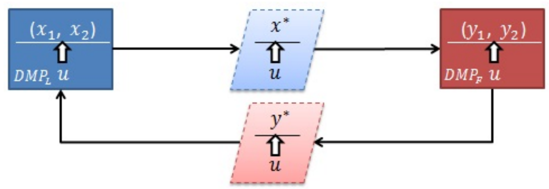

Figure 2: Uncertainty in the Decision Making Process

Regardless of the fact that those two DMs make decisions sequentially, we would like to mention that the existence of scenario-specific recourse decisions, which can be from the leader or the follower, still differentiates between single-stage and two- or multi-stage bilevel optimization. For example, only considering the response uncertainty in a bilevel formulation (i.e., the pessimistic bilevel formulation) does not convert it into a two-stage decision making model, given that there is no scenario-specific adjustments. Also, if both DMs make a single decision across all scenarios and pass it to the other DM, the bilevel formulation remains a single-stage one. The reason behind is that neither DM's decision is scenario-specific. Certainly, if one or both DMPs have scenario-specific decisions, the overall bilevel formulation will naturally evolve into a two-stage or multi-stage one.

1.3 Contribution and Outline

In this dissertation, we investigate mathematical properties of and develop solution methods for bilevel mixed integer nonlinear programming problems and robust bilevel optimization problems. The proposed methods are applied to two practical problems: wind farm capacity expansion problem and data-driven optimal decision tree problem.

In Chapter 2, we study bilevel mixed integer nonlinear programming problems, where the lower level problem is generally non-convex. In this case, the widely adopted KKT conditions based reformulation approach is no long applicable. By exploiting the structure of the problem, we provide optimality conditions based reformulation as well as a decomposition based computing scheme for both optimistic and pessimistic cases.

Chapter 3 and Chapter 4 study robust bilevel optimization (RBO). In particular, Chapter 3 considers single-stage RBO problems, where both the leader and the follower make their decisions before any realization of random variables. Several RBO models along with solution methods are provided to deal with different types of uncertainties. For single-stage RBO with discrete uncertainty set, a novel cut-and-branch algorithm is also developed. In Chapter 4, we further consider two-stage RBO problems, where some of the leader's decisions are made after the realization of random variables. We provide two basic models and their variations to take different types of uncertainties into consideration. Mathematical properties and computational methods of those developed models are also explored.

In Chapter 5, we study wind farm investment problem, taking wind power uncertainty into consideration. In a decentralized electricity market, investment decisions are made before the randomness of wind reveals, and market operates after the wind generators are built and wind intensity is determined. Thus, the wind farm capacity expansion problem is indeed a multistage decision making process. We formulate this decision making process as a two-stage RBO model. Our computational study on IEEE test sets demonstrates the superiority of the proposed model and solution method.

Finally in Chapter 6, an optimal decision tree problem is studied. We develop a new data-driven mixed-integer programming (MIP) based formulation that takes local information into consideration We then apply the bilevel optimization framework to perform hyperparameter tuning such that the generalizability of the model is enhanced. Numerical experiments are performed on benchmark datasets, and the experimental results demonstrate the outstanding performance of the proposed model.

2.0 On Solving Bilevel Mixed Integer Nonlinear Programming Problems

2.1 Motivation

In this chapter, we consider the following bilevel optimization problem

$$
\begin{aligned}
\text{OBO}: \min_{\mathbf{x},\mathbf{y}} & F(\mathbf{x},\mathbf{y}) \\
\text{s.t.} & G(\mathbf{x},\mathbf{y}) \leq \mathbf{0}, \mathbf{x} \in \mathbf{X} \\
& \mathbf{y} \in \arg\min_{\mathbf{y}} \{ f(\mathbf{x},\mathbf{y}) : g(\mathbf{x},\mathbf{y}) \leq \mathbf{0}, \mathbf{y} \in R^{n_\mathbf{y}} \},
\end{aligned}
$$

where $F, f : R^{n_\mathbf{x}} \times R^{n_\mathbf{y}} \to R$, $G : R^{n_\mathbf{x}} \times R^{n_\mathbf{y}} \to R^p$, $g : R^{n_\mathbf{x}} \times R^{n_\mathbf{y}} \to R^q$, and $\mathbf{X} \subseteq R^{n_\mathbf{x}}$. $\mathbf{x}$ is the upper level decision variable and $\mathbf{y}$ is the lower level decision variable. If the lower level problem has multiple optimal solutions for some fixed $\mathbf{x}$, then (2.1) actually represents the optimistic formulation, where the two DMs are cooperative, i.e. among all optimal solutions, the follower picks the one that is favorable to the leader. If the two DMs are not cooperative, pessimistic formulations are needed to reflect this conflict [94]. We use BO and optimistic bilevel optimization (OBO) problem interchangeably in the remainder of this chapter, and explicitly mention pessimistic bilevel optimization (PBO) formulation if needed.

As mentioned in the introduction section, solving BO is computationally challenging. Even in the simplest case, where both the upper level problem and the lower level problem are linear programs (LP), BO is still NP-hard [53]. Linear BO problems are often solved through KKT conditions based single level reformulation. For BO with nonlinear convex lower level problem, Edmunds and Bard [55] propose a branch-and-bound (B&B) algorithm to deal with BO with convex quadratic lower level problem. Recently, Dempe and Franke [50] propose a local algorithm for BO with fully convex lower level problem.

For BO whose lower level problem has integer variables, the KKT conditions based single level reformulation cannot be applied directly, and only limited algorithms have been developed. In particular, special cases of bilevel mixed integer nonlinear programming problems are studied by Gümüş and Floudas in [66] via a reformulation approach, where the lower level integer variables are first relaxed to continuous ones through a constructed convex hull and then replaced by its corresponding KKT conditions, and the resulting single level mixed integer nonlinear problem is solved by B&B based algorithms [3]. For more general BO problems, obtaining an exact solution may not be achievable. As a result, an ϵ-optimal solution is often considered being practically acceptable [96], and algorithms aiming at ϵ-optimal solution have been developed. To the best of our knowledge, Mitsos [97] introduces the first global algorithm that deals with general BO problems by sequentially generating tighter bound to approximate the optimal value function. Very recently, a novel branch-and-sandwich algorithm, which branches the upper level and the lower level variables simultaneously while maintaining the bilevel feasibility, is employed by Kleniati and Adijiman [81, 79, 80]. Due to the non-convex nature of bilevel programs, most of the aforementioned algorithms are developed within the B&B framework. Nevertheless, those B&B based algorithms often result in heavy computational burden, and only small size instances are solved in the literature for demonstration purpose.

In addition to the B&B approach, scholars also attempt to address the challenging problem from other different perspectives, such as simulated annealing approach in [115] and parametric programming approach for bilevel quadratic and bilevel mixed integer linear problems in [56]. We notice that the former one does not guarantee global optimal, and that the later one does not have numerical results.

In [133], a decomposition algorithm is developed and shows very strong capacity in solving bilevel mixed integer linear programming (BiMILP) problems. In this chapter, we further develop this framework to solve more general BO problems with mixed integer nonlinear lower level problem, in both optimistic and pessimistic settings.

The remaining of this chapter is organized as follows. Section 2 provides preliminaries and introduces a generalized pessimistic BO model. Several optimality based reformulations for bilevel mixed integer nonlinear programming (BiMINLP) problems are derived in Section 3. Decomposition algorithms are provided in Section 4. Section 5 presents a systematic computational study on various types of randomly generated BiMINLP instances as well as a case study on pessimistic bilevel gene knockout model. Section 6 concludes this chapter.

2.2 Single Level Reformulation of BO Problems With Convex Lower Level Problem

2.2.1 Preliminaries

In this section, we review some important definitions and assumptions.

1. The constraint region of BO is denoted by

$$\Omega_{\text{BO}} = \{(\mathbf{x}, \mathbf{y}) \in R^{n_{\mathbf{x}}} \times R^{n_{\mathbf{y}}} : G(\mathbf{x}, \mathbf{y}) \leq 0, \mathbf{x} \in \mathbf{X}, g(\mathbf{x}, \mathbf{y}) \leq 0\}.$$

The projection of Ω_{BO} on the upper level variable $\mathbf{x}$ is denoted by

$$\text{Proj}_{\mathbf{x}}(\Omega_{\text{BO}}) = \{\mathbf{x} \in \mathbf{X} : \exists\ \mathbf{y} \in R^{n_{\mathbf{y}}} \text{ such that } G(\mathbf{x}, \mathbf{y}) \leq 0, g(\mathbf{x}, \mathbf{y}) \leq 0\}.$$

For a fixed $\mathbf{x} \in \text{Proj}_{\mathbf{x}}(\Omega_{\text{BO}})$, we denote the lower level feasible region by

$$L(\mathbf{x}) = \{\mathbf{y} \in R^{n_{\mathbf{y}}} : g(\mathbf{x}, \mathbf{y}) \leq 0\},$$

and assume that $L(\mathbf{x})$ is bounded. For a fixed $\mathbf{x} \in \text{Proj}_{\mathbf{x}}(\Omega_{\text{BO}})$, we denote the lower level rational reaction set by

$$R(\mathbf{x}) = \arg\min_{\mathbf{y}}\{f(\mathbf{x}, \mathbf{y}) : \mathbf{y} \in L(\mathbf{x})\}.$$

The inducible region (IR) of BO is denoted by

$$IR_{\text{BO}} = \{(\mathbf{x}, \mathbf{y}) \in R^{n_{\mathbf{x}}} \times R^{n_{\mathbf{y}}} : G(\mathbf{x}, \mathbf{y}) \leq 0, \mathbf{x} \in \mathbf{X}, \mathbf{y} \in R(\mathbf{x})\}.$$

With the concept of IR, we can rewrite BO as

$$\min_{\mathbf{x}, \mathbf{y}}\{F(\mathbf{x}, \mathbf{y}) : (\mathbf{x}, \mathbf{y}) \in IR_{\text{BO}}\}.$$

We assume that $IR_{\text{BO}} \neq \emptyset$, that F, G, f, g are continuous over their domains, and that f and g are convex in $\mathbf{y}$ for fixed $\mathbf{x}$. Denote the lower level of BO by

$$\theta(\mathbf{x}) : \min_{\mathbf{y}}\{f(\mathbf{x}, \mathbf{y}) : g(\mathbf{x}, \mathbf{y}) \leq 0, \mathbf{y} \in R^{n_{\mathbf{y}}}\},$$

then it is easy to verify that $\theta(\mathbf{x})$ is a convex optimization problem for fixed $\mathbf{x}$.

2. The Lagrangian $\mathcal{L}(\mathbf{x}, \mathbf{y}, \mu) : R^{n_\mathbf{y}} \times R^q \rightarrow R$ of $\theta(\mathbf{x})$ for a fixed $\mathbf{x} \in \text{Proj}_\mathbf{x}(\Omega_{\text{BO}})$ is defined as

$$\mathcal{L}(\mathbf{x}, \mathbf{y}, \mu) = f(\mathbf{x}, \mathbf{y}) + \sum_{i=1}^{q} \mu_i g_i(\mathbf{x}, \mathbf{y}).$$

We denote $\mu = (\mu_1, \mu_2, ..., \mu_q)^T$ and $g = (g_1, g_2, ..., g_q)^T$. The Lagrange dual function $h(\mathbf{x}, \mu) : R^q \rightarrow R$ of $\theta(\mathbf{x})$ for a fixed $\mathbf{x}$ is defined as the minimum of $L(\mathbf{x}, \mathbf{y}, \mu)$ over $\mathbf{y}$, i.e.

$$h(\mathbf{x}, \mu) = \inf_{\mathbf{y} \in R^{n_\mathbf{y}}} \mathcal{L}(\mathbf{x}, \mathbf{y}, \mu) = \inf_{\mathbf{y} \in R^{n_\mathbf{y}}} \{f(\mathbf{x}, \mathbf{y}) + \sum_{i=1}^{q} \mu_i g_i(\mathbf{x}, \mathbf{y})\}.$$

The dual problem of $\theta(\mathbf{x})$ is denoted by

$$\lambda(\mathbf{x}) : \max_{\mu} \{h(\mathbf{x}, \mu) : \mu \in R^q_+\},$$

which is also a convex optimization problem. For $\theta(\mathbf{x})$ and its dual problem $\lambda(\mathbf{x})$, we have the following results. (1) Weak duality: for a fixed $\mathbf{x} \in \text{Proj}_\mathbf{x}(\Omega_{\text{BO}})$, if $\mathbf{y}$ is feasible to $\theta(\mathbf{x})$ and μ is feasible to $\lambda(\mathbf{x})$, then $f(\mathbf{x}, \mathbf{y}) \geq h(\mathbf{x}, \mu)$. (2) Strong duality: for a fixed $\mathbf{x} \in \text{Proj}_\mathbf{x}(\Omega_{\text{BO}})$, let $\theta^*(\mathbf{x})$ and $\lambda^*(\mathbf{x})$ be the optimal value of $\theta(\mathbf{x})$ and $\lambda(\mathbf{x})$ respectively, we say the strong duality holds if $\theta^*(\mathbf{x}) = \lambda^*(\mathbf{x})$. The weak duality always holds while the strong duality does not. For a convex optimization problem, the strong duality holds if some constraint qualifications are satisfied. There are various constraint qualifications, and Slater's condition is often used in convex optimization [28].

3. For a fixed $\mathbf{x} \in \text{Proj}_\mathbf{x}(\Omega_{\text{BO}})$, $\theta(\mathbf{x})$ satisfies Slater's condition (also called Slater's constraint qualification) if there exists a $\mathbf{y} \in R^{n_\mathbf{y}}$ such that $g(\mathbf{x}, \mathbf{y}) < 0$. If some of the constraints in $\theta(\mathbf{x})$ are affine, those constraints do not have to hold as strict inequalities. Hence, if $\theta(\mathbf{x})$ is an LP, it satisfies Slater's condition as long as it is feasible.

2.2.2 Bilevel Optimization With Convex Lower Level Problem

In convex optimization, the strong duality provides a condition to verify optimality for a pair of primal-dual problems. Inspired by this observation, we have the following result.

Lemma 2.1. *If $\theta(\mathbf{x})$ satisfies Slater's condition for any $\mathbf{x} \in Proj_{\mathbf{x}}(\Omega_{BO})$, then BO is equivalent to*

1. *The strong duality based single level formulation*

$$\min_{\mathbf{x},\mathbf{y},\mu}\{F(\mathbf{x},\mathbf{y}) : G(\mathbf{x},\mathbf{y}) \leq 0, g(\mathbf{x},\mathbf{y}) \leq 0, f(\mathbf{x},\mathbf{y}) \leq h(\mathbf{x},\mu), \mathbf{x} \in \mathbf{X}, \mathbf{y} \in R^{n_{\mathbf{y}}}, \mu \in R^q_+\}. \tag{2.1}$$

2. *The KKT conditions based single level formulation*

$$\begin{aligned}\min_{\mathbf{x},\mathbf{y},\mu}\{F(\mathbf{x},\mathbf{y}) :&G(\mathbf{x},\mathbf{y}) \leq 0, \nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}) + \mu^T \nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y}) = 0, g(\mathbf{x},\mathbf{y}) \leq 0, \\ &\mu^T g(\mathbf{x},\mathbf{y}) = 0, \mathbf{x} \in \mathbf{X}, \mathbf{y} \in R^{n_{\mathbf{y}}}, \mu \in R^q_+\}.\end{aligned} \tag{2.2}$$

if f and g are continuously differentiable with respect to $\mathbf{y}$ for any $\mathbf{x} \in Proj_{\mathbf{x}}(\Omega_{BO})$.

Proof. 1. To show the equivalence between BO and (2.1), it is sufficient to show that a pair of primal-dual variable $(\mathbf{y}^*, \mu^*)$ is optimal to $\theta(\mathbf{x})$ and $\lambda(\mathbf{x})$ if it satisfies

$$g(\mathbf{x},\mathbf{y}^*) \leq 0, f(\mathbf{x},\mathbf{y}^*) \leq h(\mathbf{x},\mu^*), \mathbf{y}^* \in R^{n_{\mathbf{y}}}, \mu^* \in R^q_+. \tag{2.3}$$

It is obvious that $\mathbf{y}^*$ is feasible to $\theta(\mathbf{x})$ and that μ^* is feasible to $\lambda(\mathbf{x})$. From weak duality, we have $f(\mathbf{x},\mathbf{y}^*) \geq h(\mathbf{x},\mu^*)$, which together with the second constraint of (2.3) implies $f(\mathbf{x},\mathbf{y}^*) = h(\mathbf{x},\mu^*)$. As $\theta(\mathbf{x})$ is a convex problem with Slater's condition satisfied, the strong duality holds. Therefore, the optimality of $\theta(\mathbf{x})$ is guaranteed.

2. As $\theta(\mathbf{x})$ is a convex optimization problem with Slater's condition satisfied for any $\mathbf{x} \in \mathrm{Proj}_{\mathbf{x}}(\Omega_{\mathrm{BO}})$, the KKT conditions are necessary and sufficient condition for its optimality [28]. By replacing $\theta(\mathbf{x})$ with its KKT conditions, we have that (2.2) is equivalent to BO.

□

Both of the two single level equivalent formulations depend on the fact that $\theta(\mathbf{x})$ satisfies Slater's condition for any $\mathbf{x} \in \text{Proj}_{\mathbf{x}}(\Omega_{\text{BO}})$. However, $\theta(\mathbf{x})$ does not necessarily have this property. To address this issue, we introduce an extended formulation for $\theta(\mathbf{x})$ as

$$\hat{\theta}(\mathbf{x}) : \min_{\mathbf{y},\hat{\mathbf{y}}}\{f(\mathbf{x},\mathbf{y}) + M\mathbf{e}^T\hat{\mathbf{y}} : g(\mathbf{x},\mathbf{y}) \leq \hat{\mathbf{y}}, \mathbf{y} \in R^{n_{\mathbf{y}}}, \hat{\mathbf{y}} \in R^q_+\},$$

where $\mathbf{e} \in R^q$ is a vector with all elements being 1 and M is a sufficiently large positive number. Similarly, the dual problem of $\hat{\theta}(\mathbf{x})$ is

$$\hat{\lambda}(\mathbf{x}) : \max_{\mu}\{\hat{h}(\mathbf{x},\mu) : \mu \leq M\mathbf{e}, \mu \in R^{n_q}_+\},$$

where $\hat{h}(\mathbf{x},\mu) : R^q_+ \to R$ is defined as

$$\hat{h}(\mathbf{x},\mu) = \inf_{\mathbf{y}\in R^{n_{\mathbf{y}}},\hat{\mathbf{y}}\in R^q_+}\{f(\mathbf{x},\mathbf{y}) + M\mathbf{e}^T\hat{\mathbf{y}} + \sum_{i=1}^{q}\mu_i(g_i(\mathbf{x},\mathbf{y}) - \hat{y}_i)\},$$

and $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_q)^T$.

Remark 2.1. *$\hat{\theta}(\mathbf{x})$ satisfies Slater's condition as $\hat{\mathbf{y}}$ can be arbitrarily large. Hence, the strong duality holds for $\hat{\theta}(\mathbf{x})$ and $\hat{\lambda}(\mathbf{x})$. Moreover, if f and g are continuously differentiable with respect to $\mathbf{y}$, the KKT conditions are necessary and sufficient optimality condition for $\hat{\theta}(\mathbf{x})$.*

Lemma 2.2. *1. The extended formulation $\hat{\theta}(\mathbf{x})$ is a relaxation of $\theta(\mathbf{x})$.*

2. If $\theta(\mathbf{x})$ has an optimal solution and M is sufficiently large, then $\hat{\theta}(\mathbf{x})$ has an optimal solution that is also optimal to $\theta(\mathbf{x})$.

Proof. 1. For a fixed $\mathbf{x} \in \text{Proj}_{\mathbf{x}}(\Omega_{\text{BO}})$, if $\mathbf{y}$ is feasible to $\theta(\mathbf{x})$, then $(\mathbf{y}, \hat{\mathbf{y}})$ with $\hat{\mathbf{y}} = \mathbf{0}$ is feasible to $\hat{\theta}(\mathbf{x})$.

2. For a fixed $\mathbf{x} \in \text{Proj}_{\mathbf{x}}(\Omega_{\text{BO}})$, $\hat{\theta}(\mathbf{x})$ has an optimal solution since $L(\mathbf{x})$ is nonempty and bounded. Let $(\mathbf{y}^*, \hat{\mathbf{y}}^*)$ be an optimal solution to $\hat{\theta}(\mathbf{x})$, and let $\mathbf{y}^{**}$ be an optimal solution to $\theta(\mathbf{x})$. For a sufficiently large M, we have $f(\mathbf{x}, \mathbf{y}^{**}) + M\sum_i^q 0 < f(\mathbf{x}, \mathbf{y}^*) + M\mathbf{e}^T\hat{\mathbf{y}}^*$ if $\hat{\mathbf{y}}^* > \mathbf{0}$. This contradicts the fact that $(\mathbf{y}^*, \hat{\mathbf{y}}^*)$ is an optimal solution to $\hat{\theta}(\mathbf{x})$. Thus, we have $\hat{\mathbf{y}}^* = \mathbf{0}$ for a sufficiently large M. Since $(\mathbf{y}^{**}, \mathbf{0})$ is feasible to $\hat{\theta}(\mathbf{x})$, it is also optimal to $\hat{\theta}(\mathbf{x})$, i.e. the relaxation is tight.

□

In practice, checking whether $\theta(\mathbf{x})$ satisfies Slater's condition for any $\mathbf{x} \in \text{Proj}_{\mathbf{x}}(\Omega_{\text{BO}})$ could be hard, and we can take the advantage of the extended formulation $\hat{\theta}(\mathbf{x})$ to have single level formulations of BO.

Lemma 2.3. *The strong duality based single level extended formulation*

$$\min_{\mathbf{x},\mathbf{y},\hat{\mathbf{y}},\mu} \{F(\mathbf{x},\mathbf{y}) : G(\mathbf{x},\mathbf{y}) \leq 0, g(\mathbf{x},\mathbf{y}) \leq \hat{\mathbf{y}}, \mu \leq M\mathbf{e}, \mathbf{x} \in \mathbf{X},$$

is equivalent to BO.

Proof. From Remark 2.1 and Lemma 2.2, we know that $\hat{\theta}(\mathbf{x})$ satisfies Slater's condition and is a tight relaxation of $\theta(\mathbf{x})$. Therefore, replacing $\theta(\mathbf{x})$ with $\hat{\theta}(\mathbf{x})$ and applying Lemma 2.1 lead to the result. □

Remark 2.2. *Similar as Lemma 2.1, if f and g are continuously differentiable with respect to $\mathbf{y}$ for any $\mathbf{x} \in Proj_{\mathbf{x}}(\Omega_{BO})$, we can obtain the KKT conditions based single level extended formulation, which is also equivalent to BO.*

Remark 2.3. *With minor modifications, Lemma 2.1 (as well as Lemma 2.3) can be applied to BO with multiple followers. By replacing each lower level problem with its optimality conditions, such a BO problem is converted to a single level problem with multiple sets of optimality conditions based constraints.*

2.2.3 Generalized Pessimistic Bilevel Optimization and Reformulation

As mentioned previously, pessimistic formulations are employed if the follower does not cooperate with the leader, or if the leader wants to hedge against risks by considering the worst case scenario.

If there is no coupled constraints, i.e., $\mathbf{y}$ does not appear in the upper level constraints, the pessimistic counterpart of BO is given by

$$\min_{\mathbf{x}} \max_{\mathbf{y} \in R(\mathbf{x})} \{F(\mathbf{x},\mathbf{y}) : G(\mathbf{x}) \leq 0, \mathbf{x} \in \mathbf{X}\},$$

where $R(\mathbf{x})$ is the lower level rational reaction set. Recent studies on pessimistic bilevel optimization can be found in [40, 126, 52, 136, 93, 11, 86, 132]. For BO problems with pessimistic coupled constraints, $G(\mathbf{x},\mathbf{y}) \leq 0$ needs to be satisfied for all $\mathbf{y} \in R(\mathbf{x})$, and a reformulation method is introduced in [132].

Since BO with coupled constraints can be rewritten as

$$\min_{\mathbf{x},t,\mathbf{y}}\{t : F(\mathbf{x},\mathbf{y}) - t \leq 0, G(\mathbf{x},\mathbf{y}) \leq 0, \mathbf{x} \in \mathbf{X}, t \in R, \mathbf{y} \in R(\mathbf{x})\},$$

we can without loss of generality consider BO in a reduced form as

$$\text{BO-O} : \min_{\mathbf{x},\mathbf{y}}\{F(\mathbf{x}) : \mathbf{x} \in \mathbf{X}, G(\mathbf{x},\mathbf{y}) \leq 0, \mathbf{y} \in R(\mathbf{x})\}.$$

In the optimistic setting, the follower picks a point in $R(\mathbf{x})$ to satisfy the coupled constraints, and the resulting problem is BO-O. However, if the follower does not cooperate with the leader, he can pick a point in $R(\mathbf{x})$ such that some of the coupled constraints are violated. To ensure feasibility, the leader needs to choose an $\mathbf{x} \in \mathbf{X}$ such that $G(\mathbf{x},\mathbf{y}) \leq 0$ holds for all $\mathbf{y}$ in the rational reaction set by considering the following problem

$$\text{BO-P} : \min_{\mathbf{x},\mathbf{y}}\{F(\mathbf{x}) : \mathbf{x} \in \mathbf{X}, G(\mathbf{x},\mathbf{y}) \leq 0, \forall \mathbf{y} \in R(\mathbf{x})\}.$$

Denoting the IR of BO-O by

$$IR_{\text{BO-O}} = \{(\mathbf{x},\mathbf{y}) \in R^{n_{\mathbf{x}}} \times R^{n_{\mathbf{y}}} : \mathbf{x} \in \mathbf{X}, G(\mathbf{x},\mathbf{y}) \leq 0, \mathbf{y} \in R(\mathbf{x})\}$$

and the IR of BO-P by

$$IR_{\text{BO-P}} = \{(\mathbf{x},\mathbf{y}) \in R^{n_{\mathbf{x}}} \times R^{n_{\mathbf{y}}} : \mathbf{x} \in \mathbf{X}, G(\mathbf{x},\mathbf{y}) \leq 0, \forall \mathbf{y} \in R(\mathbf{x})\},$$

we have $IR_{\text{BO-P}} \subseteq IR_{\text{BO-O}}$, which implies BO-O is a relaxation of BO-P.

In fact, BOP-O and BOP-P are two extreme cases. In BO-O, all the coupled constraints are optimistic, i.e., they are satisfied if there exist $\mathbf{y} \in R(\mathbf{x})$ and $\mathbf{x} \in \mathbf{X}$ such that $G(\mathbf{x},\mathbf{y}) \leq 0$. In contrast, all the coupled constraints are pessimistic in BO-P, and thus $G(\mathbf{x},\mathbf{y}) \leq 0$ must hold for all $\mathbf{y} \in R(\mathbf{x})$. We now consider a more general case where some coupled constraints are optimistic while others are pessimistic and introduce a new bilevel model.

Let $\mathcal{CC} = \{1, 2, ..., p\}$ be the index set of the coupled constraints, $\mathcal{PCC} \subseteq \mathcal{CC}$ be the index set of the pessimistic coupled constraints, and G_i be the ith coupled constraint, then a generalized pessimistic bilevel optimization (GPBO) problem is given by

$$\begin{aligned}\text{GPBO}: \min_{\mathbf{x},\mathbf{y}}\{F(\mathbf{x}):& G_i(\mathbf{x},\mathbf{y}) \leq 0, \forall i \in \mathcal{CC}\backslash\mathcal{PCC}, \\ & G_i(\mathbf{x},\mathbf{y}) \leq 0, \forall \mathbf{y} \in R(\mathbf{x}), i \in \mathcal{PCC}, \mathbf{x} \in \mathbf{X}, \mathbf{y} \in R(\mathbf{x})\},\end{aligned}$$

where is the lower level rational reaction set.

Remark 2.4. *GPBO reduces to BO-O if $\mathcal{PCC} = \emptyset$, and reduces to BO-P if $\mathcal{PCC} = \mathcal{CC}$. Moreover, if $\mathcal{PCC}_1 \subseteq \mathcal{PCC}_2 \subseteq \mathcal{CC}$, then GPBO with $\mathcal{PCC} = \mathcal{PCC}_1$ is a relaxation of the one with $\mathcal{PCC} = \mathcal{PCC}_2$.*

If we simply drop off the pessimistic coupled constraints, GPBO reduces to

$$\text{GPBO-O}: \min_{\mathbf{x},\mathbf{y}}\{F(\mathbf{x}): G_i(\mathbf{x},\mathbf{y}) \leq 0, \forall i \in \mathcal{CC}\backslash\mathcal{PCC}, \mathbf{x} \in \mathbf{X}, \mathbf{y} \in R(\mathbf{x})\},$$

which is a instance of BO-O and a relaxation of GPBO. By applying Lemma 2.3, we can obtain a single level reformulation of GPBO-O, and obtain a lower bound of GPBO by solving this single level problem.

The projection of the IR of GPBO-O on $\mathbf{x}$ is denoted by

$$\text{Proj}_{\mathbf{x}}(IR_{\text{GPBO-O}}) = \{\mathbf{x} \in \mathbf{X} : \exists\ \mathbf{y} \in R(\mathbf{x}) \text{ with } G_i(\mathbf{x},\mathbf{y}) \leq 0, \forall i \in \mathcal{CC}\backslash\mathcal{PCC}\}.$$

For a fixed $\mathbf{x}^* \in \text{Proj}_{\mathbf{x}}(IR_{\text{GPBO-O}})$ and $j \in \mathcal{PCC}$, denote the optimal value of

$$\max_{\mathbf{y}}\{G_j(\mathbf{x}^*,\mathbf{y}) : \mathbf{y} \in R(\mathbf{x}^*)\} \tag{2.4}$$

by $v_j(\mathbf{x}^*)$, then it is not hard to verify that

$$G_j(\mathbf{x}^*,\mathbf{y}) \leq 0, \forall \mathbf{y} \in R(\mathbf{x}^*) \Longleftrightarrow v_j(\mathbf{x}^*) \leq 0.$$

Hence, we can rewrite GPBO as

$$\min_{\mathbf{x},\mathbf{y}}\{F(\mathbf{x}): G_i(\mathbf{x},\mathbf{y}) \leq 0, \forall i \in \mathcal{CC}\backslash\mathcal{PCC}, v_i(\mathbf{x}) \leq 0, \forall i \in \mathcal{PCC}, \mathbf{x} \in \mathbf{X}, \mathbf{y} \in R(\mathbf{x})\}. \tag{2.5}$$

We see that (2.5) is actually a tri-level problem since each constraint in the form of $v_i(\mathbf{x}) \leq 0$ requires solving a BO problem defined in (2.4). To solve such a complicated problem, we introduce a reformulation method that converts GPBO to a standard BO problem with multiple followers.

Lemma 2.4. *For a fixed* $\mathbf{x}^* \in Proj_{\mathbf{x}}(IR_{GPBO\text{-}O})$ *and* $j \in \mathcal{PCC}$, $v_j(\mathbf{x}^*) \leq 0$ *if and only if there exist* $\bar{\mathbf{y}}_j^* \in L(\mathbf{x}^*)$ *and* $\mathbf{y}_j^* \in S_j(\mathbf{x}^*, \bar{\mathbf{y}}_j^*) \equiv \arg\max_{\mathbf{y}}\{G_j(\mathbf{x}^*, \mathbf{y}) : g(\mathbf{x}^*, \mathbf{y}) \leq 0, f(\mathbf{x}^*, \mathbf{y}) \leq f(\mathbf{x}^*, \bar{\mathbf{y}}_j^*), \mathbf{y} \in R^{n_{\mathbf{y}}}\}$ *such that* $G_j(\mathbf{x}^*, \mathbf{y}_j^*) \leq 0$

Proof. We first show the "⇒" part. Sicne $\mathbf{x}^* \in \text{Proj}_{\mathbf{x}}(IR_{\text{GPBO-O}})$, we have $R(\mathbf{x}^*) \neq \emptyset$. As f, g are continuous and $L(\mathbf{x}^*)$ is bounded, $R(\mathbf{x}^*)$ is compact. By the continuity of G_j, the problem

$$\max_{\mathbf{y}}\{G_j(\mathbf{x}^*, \mathbf{y}) : \mathbf{y} \in R(\mathbf{x}^*)\} \tag{2.6}$$

has an optimal solution. Let $\mathbf{y}_j^*$ be an optimal solution to (2.6) and let $\bar{\mathbf{y}}_j^* = \mathbf{y}_j^*$, then we have $G_j(\mathbf{x}^*, \mathbf{y}_j^*) = v_j(\mathbf{x}^*) \leq 0$, $\bar{\mathbf{y}}_j^* \in L(\mathbf{x}^*)$, and $R(\mathbf{x}^*) = \{\mathbf{y} \in R^{n_{\mathbf{y}}} : g(\mathbf{x}^*, \mathbf{y}) \leq 0, f(\mathbf{x}^*, \mathbf{y}) \leq f(\mathbf{x}^*, \bar{\mathbf{y}}_j^*)\}$. Hence, $\mathbf{y}_j^* \in S_j(\mathbf{x}^*, \bar{\mathbf{y}}_j^*)$.

We next show the "⇐" part. For fixed $\mathbf{x}^* \in \text{Proj}_{\mathbf{x}}(IR_{\text{GPBO-O}})$ and $\bar{\mathbf{y}}_j^* \in L(\mathbf{x}^*)$, let $\mathbf{y}_j^* \in S_j(\mathbf{x}^*, \bar{\mathbf{y}}_j^*)$, i.e., $\mathbf{y}_j^*$ is an optimal solution to the problem

$$z_j(\mathbf{x}^*, \bar{\mathbf{y}}_j^*) = \max_{\mathbf{y}}\{G_j(\mathbf{x}^*, \mathbf{y}) : g(\mathbf{x}^*, \mathbf{y}) \leq 0, f(\mathbf{x}^*, \mathbf{y}) \leq f(\mathbf{x}^*, \bar{\mathbf{y}}_j^*), \mathbf{y} \in R^{n_{\mathbf{y}}}\}, \tag{2.7}$$

then we have

$$v_j(\mathbf{x}^*) \leq z_j(\mathbf{x}^*, \bar{\mathbf{y}}_j^*) \leq G_j(\mathbf{x}^*, \mathbf{y}_j^*) \leq 0.$$

The first inequality follows the fact $R(\mathbf{x}^*) \subseteq \{\mathbf{y} \in R^{n_{\mathbf{y}}} : g(\mathbf{x}^*, \mathbf{y}) \leq 0, f(\mathbf{x}^*, \mathbf{y}) \leq f(\mathbf{x}^*, \bar{\mathbf{y}}_j^*)\}$, and the second inequality is implied by the optimality of $\mathbf{y}_j^*$ to (2.7). □

Theorem 2.1. *Let* $S_i(\mathbf{x}, \bar{\mathbf{y}})$ *be defined as in Lemma 2.4, then GPBO is equivalent to the following BO problem*

$$\begin{aligned} \textit{GPBO-R}: \min_{\boldsymbol{x}, \bar{\mathbf{y}}, \mathbf{y}, \mathbf{y}_i} \{F(\mathbf{x}) : & G_i(\mathbf{x}, \mathbf{y}) \leq 0, \forall i \in \mathcal{CC} \backslash \mathcal{PCC}, G_i(\mathbf{x}, \mathbf{y}_i) \leq 0, \forall i \in \mathcal{PCC} \\ & \mathbf{y}_i \in S_i(\mathbf{x}, \bar{\mathbf{y}}), \forall i \in \mathcal{PCC}, \mathbf{x} \in \mathbf{X}, \bar{\mathbf{y}} \in L(\mathbf{x}), \mathbf{y} \in R(\mathbf{x})\}. \end{aligned}$$

Proof. According to Lemma 2.4, it is easy to verify that GPBO is equivalent to

$$\begin{aligned} \min_{\mathbf{x}, \bar{\mathbf{y}}_i, \mathbf{y}, \mathbf{y}_i} \{F(\mathbf{x}) : & G_i(\mathbf{x}, \mathbf{y}) \leq 0, \forall i \in \mathcal{CC} \backslash \mathcal{PCC}, G_i(\mathbf{x}, \mathbf{y}_i) \leq 0, \forall i \in \mathcal{PCC} \\ & \mathbf{y}_i \in S_i(\mathbf{x}, \bar{\mathbf{y}}_i), \forall i \in \mathcal{PCC}, \mathbf{x} \in \mathbf{X}, \bar{\mathbf{y}}_i \in L(\mathbf{x}), \mathbf{y} \in R(\mathbf{x})\}. \end{aligned}$$

Let $\bar{\mathbf{y}} = \min_i\{\bar{\mathbf{y}}_i\}$, then $\bar{\mathbf{y}} \in L(\mathbf{x})$. For a fixed $\mathbf{x}^* \in \text{Proj}_{\mathbf{x}}(IR_{\text{GPBO-O}})$ and an arbitrary $i \in \mathcal{PCC}$, let $\mathbf{y}_i^* \in S_i(\mathbf{x}^*, \bar{\mathbf{y}}_i)$ and $\mathbf{y}_i^0 \in S_i(\mathbf{x}^*, \bar{\mathbf{y}})$, then $\mathbf{y}_i^0$ is an optimal solution to

$$\max_{\mathbf{y}_i}\{G_i(\mathbf{x}^*, \mathbf{y}_i) : g(\mathbf{x}^*, \mathbf{y}_i) \leq 0, f(\mathbf{x}^*, \mathbf{y}_i) \leq f(\mathbf{x}^*, \bar{\mathbf{y}}), \mathbf{y}_i \in R^{n_{\mathbf{y}}}\}.$$

If $v_i(\mathbf{x}^*) \leq 0$, then we have

$$G_i(\mathbf{x}^*, \mathbf{y}_i^0) \leq \max_{\mathbf{y}_i \in R^{n_{\mathbf{y}}}}\{G_i(\mathbf{x}^*, \mathbf{y}_i) : g(\mathbf{x}^*, \mathbf{y}_i) \leq 0, f(\mathbf{x}^*, \mathbf{y}_i) \leq f(\mathbf{x}^*, \bar{\mathbf{y}}_i)\} \leq G_i(\mathbf{x}^*, \mathbf{y}_i^*) \leq 0.$$

The first inequality holds as $\bar{\mathbf{y}} \leq \bar{\mathbf{y}}_i$, and the second inequality follows as $\mathbf{y}_i^* \in S_i(\mathbf{x}^*, \bar{\mathbf{y}}_i)$. Conversely, for $i \in \mathcal{PCC}$, if $\bar{\mathbf{y}} \in L(\mathbf{x}^*)$, $\mathbf{y}_i^* \in S_i(\mathbf{x}^*, \bar{\mathbf{y}})$ and $G_i(\mathbf{x}^*, \mathbf{y}_i^*) \leq 0$, then we have

$$\begin{aligned} v_i(\mathbf{x}^*) &= \max_{\mathbf{y}_i}\{G_i(\mathbf{x}^*, \mathbf{y}_i) : \mathbf{y}_i \in R(\mathbf{x}^*)\} \\ &\leq \max_{\mathbf{y}_i}\{G_i(\mathbf{x}^*, \mathbf{y}_i) : g(\mathbf{x}^*, \mathbf{y}_i) \leq 0, f(\mathbf{x}^*, \mathbf{y}_i) \leq f(\mathbf{x}^*, \bar{\mathbf{y}}_i), \mathbf{y}_i \in R^{n_{\mathbf{y}}}\} \leq G_i(\mathbf{x}^*, \mathbf{y}_i^*) \leq 0. \end{aligned}$$

Since $i \in \mathcal{PCC}$ is arbitrary, the result follows. □

It is worthy to mention that we cannot replace $\bar{\mathbf{y}}$ with $\mathbf{y}$ unless $\mathcal{PCC} = \mathcal{CC}$. Theorem 2.1 shows that GPBO is equivalent to a standard BO problem with $|\mathcal{PCC}| + 1$ followers. Furthermore, if $G_i(\mathbf{x}, \mathbf{y})$ is concave in $\mathbf{y}$ for fixed $\mathbf{x} \in \text{Proj}_{\mathbf{x}}(IR_{\text{GPBO-O}})$, then the maximization problem $\max_{\mathbf{y}}\{G_i(\mathbf{x}, \mathbf{y}) : g(\mathbf{x}, \mathbf{y}) \leq 0, f(\mathbf{x}, \mathbf{y}) \leq f(\mathbf{x}, \bar{\mathbf{y}}_j), \mathbf{y} \in R^{n_{\mathbf{y}}}\}$ is a convex optimization problem. Therefore, if $G_i(\mathbf{x}, \mathbf{y})$ is concave in $\mathbf{y}$ for all $i \in \mathcal{PCC}$, we can apply Lemma 2.3 to obtain a single level problem that is equivalent to GPBO.

From Lemma 2.1, if $\theta(\mathbf{x})$ satisfies Slater's condition, then $\mathbf{y} \in R(\mathbf{x})$ implies there exists $\mu \in R_+^q$ such that $g(\mathbf{x}, \mathbf{y}) \leq 0$ and $f(\mathbf{x}, \mathbf{y}) \leq h(\mathbf{x}, \mu)$, where $h(\mathbf{x}, \mu)$ is the dual function. Let $\hat{R}(\mathbf{x}) = \{(\mathbf{y}, \mu) \in R^{n_{\mathbf{y}}} \times R_+^q : g(\mathbf{x}, \mathbf{y}) \leq 0, f(\mathbf{x}, \mathbf{y}) \leq h(\mathbf{x}, \mu)\}$, then for a fixed $\mathbf{x}^* \in \text{Proj}_{\mathbf{x}}(IR_{\text{GPBO-O}})$, we can rewrite $\max_{\mathbf{y}}\{G_j(\mathbf{x}^*, \mathbf{y}) : \mathbf{y} \in R(\mathbf{x}^*)\}$ as

$$\max_{\mathbf{y},\mu}\{G_j(\mathbf{x}^*, \mathbf{y}) : (\mathbf{y}, \mu) \in \hat{R}(\mathbf{x}^*)\}. \tag{2.8}$$

Denote the optimal solution set of (2.8) by $\hat{S}_j(\mathbf{x}^*)$, we can have an alternative reformulation of GPBO as

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y},\mu,\mathbf{y}_i,\mu_i}\{F(\mathbf{x}) : & G_i(\mathbf{x}, \mathbf{y}) \leq 0, \forall i \in \mathcal{CC}\backslash\mathcal{PCC}, G_i(\mathbf{x}, \mathbf{y}_i) \leq 0, \forall i \in \mathcal{PCC}, \\ & (\mathbf{y}_i, \mu_i) \in \hat{S}_i(\mathbf{x}), \forall i \in \mathcal{PCC}, \mathbf{x} \in \mathbf{X}, (\mathbf{y}, \mu) \in \hat{R}(\mathbf{x})\}. \end{aligned}$$

Similar as GPBO-R, we can also apply Lemma 2.3 to obtain a single level equivalent problem if $G_i(\mathbf{x}, \mathbf{y})$ is concave for all $i \in \mathcal{PCC}$. We indicate that applying the extended formula is essential as both $\max_{\mathbf{y}}\{G_j(\mathbf{x}^*, \mathbf{y}) : \mathbf{y} \in R(\mathbf{x}^*)\}$ and (2.8) may not satisfy Slater's condition, which guarantees the strong duality.

2.2.4 Bilevel Optimization With Bounded Rationality and Suboptimality

In bilevel optimization, the follower is traditionally assumed to be fully rational, i.e., he always selects a point in the rational reaction set $R(\mathbf{x})$ as a response to the leader. However, in practice, the follower's rationality is often limited due to imperfect information.

For a fixed $\mathbf{x} \in \{\mathbf{x} \in \mathbf{X} : L(\mathbf{x}) \neq \emptyset\}$ and $\epsilon \geq 0$, we denote the optimal value of the lower level problem by $\theta^*(\mathbf{x})$, and denote the bounded rational reaction set of the follower by

$$R_\epsilon(\mathbf{x}) = \{\mathbf{y} \in R^{n_{\mathbf{y}}} : g(\mathbf{x}, \mathbf{y}) \leq 0, f(\mathbf{x}, \mathbf{y}) \leq \theta^*(\mathbf{x}) + \epsilon\}. \tag{2.9}$$

The parameter ϵ measures the willingness and capacity of the follower to achieve optimality, and a small ϵ corresponds to strong willingness and capacity. It is easy to see that $R_\epsilon(\mathbf{x})$ reduces to $R(\mathbf{x})$ if $\epsilon = 0$, and reduces to $L(\mathbf{x})$ if ϵ is sufficiently large. In general, we have $R(\mathbf{x}) \subseteq R_\epsilon(\mathbf{x}) \subseteq L(\mathbf{x})$ for $\epsilon \geq 0$.

Before incorporating $R_\epsilon(\mathbf{x})$ into GPBO, we revisit BO-O and BO-P to gain some insights. Consider BO-O with a bounded rational reaction set

$$\text{BO-O-BR} : p(\epsilon) = \min_{\mathbf{x},\mathbf{y}}\{F(\mathbf{x}) : \mathbf{x} \in \mathbf{X}, G(\mathbf{x}, \mathbf{y}) \leq 0, \mathbf{y} \in R_\epsilon(\mathbf{x})\},$$

then we have that $IR_{\text{BO-O}} \subseteq IR_{\text{BO-O-BR}}$, and that $p(\epsilon)$ is decreasing in ϵ. By introducing a new variable $\hat{\mathbf{y}}$, we can rewrite BO-O-BR as

$$\min_{\mathbf{x},\mathbf{y},\hat{\mathbf{y}}}\{F(\mathbf{x}) : G(\mathbf{x}, \mathbf{y}) \leq 0, \mathbf{x} \in \mathbf{X}, f(\mathbf{x}, \mathbf{y}) \leq f(\mathbf{x}, \hat{\mathbf{y}}) + \epsilon, \mathbf{y} \in L(\mathbf{x}), \hat{\mathbf{y}} \in R(\mathbf{x})\}, \tag{2.10}$$

which, according to Lemma 2.3, can be further converted to a single level problem.

In fact, BO-O-BR has a very meaningful economic interpretation and can be used for sensitivity analysis. From the follower's perspective, ϵ can be interpreted as the maximum loss that he would like to incur in order to benefit the leader. From the leader's perspective, the additional benefit gained from the follower can be quantified by $p(0)-p(\epsilon)$. If $p(0)-p(\epsilon) \geq \epsilon$, the leader has incentive to compensate the follower up to ϵ such that both she and the follower are better off.

For BO-P, we can also introduce a bounded rational reaction set, and have the pessimistic counterpart of BO-O-BR as

$$\text{BO-P-BR} : q(\epsilon) = \min_{\mathbf{x}}\{F(\mathbf{x}) : \mathbf{x} \in \mathbf{X}, G(\mathbf{x}, \mathbf{y}) \leq 0, \forall \mathbf{y} \in R_{\epsilon}(\mathbf{x})\}.$$

Similarly, we have $IR_{\text{BO-P-BR}} \subseteq IR_{\text{BO-P}}$, and $q(\epsilon)$ is increasing in ϵ.

We point out that the follower in BO-P-BR is implicitly assumed to act consistently, i.e., to have a homogeneous ϵ toward all different pessimistic coupled constraints. However, as we mentioned previously, the follower may not act consistently due to his bounded rationality, and such inconsistency results in heterogeneous ϵ. Suppose there are $|\mathcal{PCC}|$ pessimistic coupled constraints in BO-P, then a generalized model is given by

$$\text{GBO-P-BR} : q(\epsilon_1, \epsilon_2, ..., \epsilon_{|\mathcal{PCC}|}) = \min_{\mathbf{x}}\{F(\mathbf{x}) : \mathbf{x} \in \mathbf{X}, G_i(\mathbf{x}, \mathbf{y}) \leq 0, \forall \mathbf{y} \in R_{\epsilon_i}(\mathbf{x}), i \in \mathcal{PCC}\},$$

which has multiple bounded rational reaction sets. GBO-P-BR reduces to BO-P-BR if $\epsilon_i = \epsilon$ for all i, and reduces to BO-P if $\epsilon_i = 0$ for all i. Without loss of generality, we can set $\epsilon = \max\{\epsilon_1, \epsilon_2, ..., \epsilon_{|\mathcal{PCC}|}\}$, and have $IR_{\text{BO-P-BR}} \subseteq IR_{\text{GBO-P-BR}} \subseteq IR_{\text{BO-P}}$.

In practice, if the leader does not know the safety margin, i.e., ϵ_i, for each constraint, she can employ BO-P-BR with $\epsilon = \max\{\epsilon_1, \epsilon_2, ..., \epsilon_{|\mathcal{PCC}|}\}$ as a conservative estimation. However, the leader can be significantly benefited if she has an accurate estimation of each ϵ_i. Consider an illustrative example, where the lower level feasible region is set to be independent of x for simplicity.

Example 2.1.

$$\begin{aligned} q(\epsilon_1, \epsilon_2) = \min_{x}\{-x :& 0 \leq x \leq 5, x + y \leq 2, \forall y \in R_{\epsilon_1}(x) \\ & x + 2y \leq 4, \forall y \in R_{\epsilon_2}(x), R_0(x) = \arg\min_{y}\{y : 1 \leq y \leq 2\}\}. \end{aligned}$$

The rational reaction set $R_0(x) = \{1\}$. For $\epsilon_1 = \epsilon_2 = 0$, the unique optimal solution is $x^* = 1$, and $q(0,0) = -1$. If the leader has little information about the follower and thus conservatively estimates $\epsilon_1 = \epsilon_2 = 0.5$, the bounded rational reaction set will be $R_{0.5}(x) = \{y : 1 \leq y \leq 1.5\}$. The optimal solution will be $x^* = 0.5$, and $q(0.5, 0.5) = -0.5$, resulting in 50% loss of the optimal value. If the leader has more accurate information and is able to specify $(\epsilon_1, \epsilon_2) = (0.2, 0.5)$, then the optimal solution will be $x^* = 0.8$, and $q(0.2, 0.5) = -0.8$, which is much better than -0.5.

Now we can take bounded rationality into consideration by incorporating multiple bounded rational reaction sets into the general GPBO model. Let $R_\epsilon(\mathbf{x})$ be defined as in (2.9), we can extend GPBO to

$$\begin{aligned}\text{GPBO-BR} : \min_{\mathbf{x},\mathbf{y}} \{F(\mathbf{x}) : & G_i(\mathbf{x},\mathbf{y}) \leq 0, \forall i \in \mathcal{CC} \backslash \mathcal{PCC} \\ & G_i(\mathbf{x},\mathbf{y}) \leq 0, \forall \mathbf{y} \in R_{\epsilon_i}(\mathbf{x}), i \in \mathcal{PCC}, \mathbf{x} \in \mathbf{X}, \mathbf{y} \in R_{\epsilon_0}(\mathbf{x})\},\end{aligned}$$

which has $|\mathcal{PCC}| + 1$ bounded rational reaction sets. Let $S_{\epsilon_i}(\mathbf{x}, \bar{\mathbf{y}}) = \arg\max\{G_i(\mathbf{x},\mathbf{y}) : g(\mathbf{x},\mathbf{y}) \leq 0, f(\mathbf{x},\mathbf{y}) \leq f(\mathbf{x},\bar{\mathbf{y}}) + \epsilon_i, \mathbf{y} \in R^{n_\mathbf{y}}\}$, we can apply Theorem 2.1 and (2.10) to have a reformulation of GPBO-BR as

$$\begin{aligned}\text{GPBO-BR-R} : \min_{\mathbf{x},\bar{\mathbf{y}},\mathbf{y},\mathbf{y}_i} \{F(\mathbf{x}) : & G_i(\mathbf{x},\mathbf{y}) \leq 0, \forall i \in \mathcal{CC} \backslash \mathcal{PCC}, f(\mathbf{x},\mathbf{y}) \leq f(\mathbf{x},\bar{\mathbf{y}}) + \epsilon_0, G_i(\mathbf{x},\mathbf{y}_i) \leq 0, \\ & \forall i \in \mathcal{PCC}, \mathbf{y}_i \in S_{\epsilon_i}(\mathbf{x},\bar{\mathbf{y}}), \forall i \in \mathcal{PCC}, \mathbf{x} \in \mathbf{X}, \mathbf{y} \in L(\mathbf{x}), \bar{\mathbf{y}} \in R(\mathbf{x})\}.\end{aligned}$$

Similar as for GPBO, we can also apply Lemma 2.3 to obtain a single level equivalent problem to GPBO-BR if $G_i(\mathbf{x},\mathbf{y})$ is concave in $\mathbf{y}$ for $i \in \mathcal{PCC}$.

2.3 Optimality Conditions Based Reformulation of Bilevel Mixed Integer Nonlinear Programming Problems

2.3.1 Reformulation of Optimistic Bilevel Mixed Integer Programming Problems

In this section, we consider general bilevel mixed integer nonlinear programming (BiMINLP) problem

$$\begin{aligned}\text{BiMINLP}: &\min_{\mathbf{x},\mathbf{y},\mathbf{z}} F(\mathbf{x},\mathbf{y},\mathbf{z})\\ &\text{s.t.} G(\mathbf{x},\ \mathbf{y},\ \mathbf{z}) \leq 0, \mathbf{x} \in \mathbf{X}\\ &\quad(\mathbf{y},\ \mathbf{z}) \in \arg\min_{\mathbf{y},\mathbf{z}}\{f(\mathbf{x},\mathbf{y},\mathbf{z}) : g(\mathbf{x},\mathbf{y},\mathbf{z}) \leq 0, \mathbf{y} \in R^{n_\mathbf{y}}, \mathbf{z} \in \mathbf{Z}\},\end{aligned}$$

where $F, f : R^{n_\mathbf{x}} \times R^{n_\mathbf{y}} \times Z^{n_\mathbf{z}} \to R$, $G : R^{n_\mathbf{x}} \times R^{n_\mathbf{y}} \times Z^{n_\mathbf{z}} \to R^p$, $g : R^{n_\mathbf{x}} \times R^{n_\mathbf{y}} \times Z^{n_\mathbf{z}} \to R^q$, $\mathbf{X} \subseteq R^{n_\mathbf{x}}$, $\mathbf{Z} \subseteq Z^{n_\mathbf{z}}$, and $|\mathbf{Z}| = K < +\infty$. We assume that f and g are convex in $\mathbf{y}$ in the remainder of this chapter. It is easy to see that BiMINLP is an extension of BO, and thus the concepts reviewed in the preliminaries section can be naturally extended. For example, the projection of Ω_{BiMINLP} on the upper level variable $\mathbf{x}$ is given by

$$\text{Proj}_\mathbf{x}(\Omega_{\text{BiMINLP}}) = \{\mathbf{x} \in \mathbf{X} : \exists\ (\mathbf{y},\mathbf{z}) \in R^{n_\mathbf{y}} \times \mathbf{Z} \text{ with } G(\mathbf{x},\mathbf{y},\mathbf{z}) \leq 0, g(\mathbf{x},\mathbf{y},\mathbf{z}) \leq 0\}.$$

Denote the projection of Ω_{BiMINLP} on the lower level integer variable by

$$\text{Proj}_\mathbf{z}(\Omega_{\text{BiMINLP}}) = \{\mathbf{z} \in \mathbf{Z} : \exists\ (\mathbf{x},\mathbf{y}) \in \mathbf{X} \times R^{n_\mathbf{y}} \text{ with } G(\mathbf{x},\mathbf{y},\mathbf{z}) \leq 0, g(\mathbf{x},\mathbf{y},\mathbf{z}) \leq 0\},$$

then BiMINLP is feasible only if $\text{Proj}_\mathbf{z}(\Omega_{\text{BiMINLP}}) \neq \emptyset$. We without loss of generality assume that $\mathbf{Z} = \text{Proj}_\mathbf{z}(\Omega_{\text{BiMINLP}})$ as otherwise we can re-define $\mathbf{Z}$ by $\text{Proj}_\mathbf{z}(\Omega_{\text{BiMINLP}})$ as in [131].

Due to the presence of the integer variable $\mathbf{z}$, we cannot apply Lemma 2.1 or Lemma 2.3 to solve BiMINLP since neither the strong duality nor the KKT conditions holds for mixed integer programming problems. However, we notice that for fixed $(\mathbf{x},\mathbf{z}) \in \text{Proj}_\mathbf{x}(\Omega_{\text{BiMINLP}}) \times \mathbf{Z}$, the remaining lower level problem

$$\xi(\mathbf{x},\mathbf{z}) : \min_\mathbf{y}\{f(\mathbf{x},\mathbf{y},\mathbf{z}), g(\mathbf{x},\mathbf{y},\mathbf{z}) \leq 0, \mathbf{y} \in R^{n_\mathbf{y}}\}$$

and its dual problem

$$\rho(\mathbf{x}, \mathbf{z}) : \max_{\mu}\{s(\mathbf{x}, \mathbf{z}, \mu) : \mu \in R^q_+\}$$

are convex optimization problems, where $s(\mathbf{x}, \mathbf{z}, \mu) = \inf_{\mathbf{y} \in R^{n_\mathbf{y}}}\{f(\mathbf{x}, \mathbf{y}, \mathbf{z}) + \sum_{i=1}^q \mu_i g_i(\mathbf{x}, \mathbf{y}, \mathbf{z})\}$.

Denote the feasible region of $\xi(\mathbf{x}, \mathbf{z})$ by $P(\mathbf{x}, \mathbf{z}) = \{\mathbf{y} \in R^{n_\mathbf{y}} : g(\mathbf{x}, \mathbf{y}, \mathbf{z}) \leq 0\}$, and that of $\rho(\mathbf{x}, \mathbf{z})$ by $Q(\mathbf{x}, \mathbf{z})$, then we can also rewrite $\xi(\mathbf{x}, \mathbf{z})$ and $\rho(\mathbf{x}, \mathbf{z})$ as $\min_\mathbf{y}\{f(\mathbf{x}, \mathbf{y}, \mathbf{z}) : \mathbf{y} \in P(\mathbf{x}, \mathbf{z})\}$ and $\max_\mu\{s(\mathbf{x}, \mathbf{z}, \mu) : \mu \in Q(\mathbf{x}, \mathbf{z})\}$, respectively.

If $P(\mathbf{x}, \mathbf{z})$ is nonempty and bounded, then $\xi(\mathbf{x}, \mathbf{z})$ has an optimal solution. We say the problem $\xi(\mathbf{x}, \mathbf{z})$ has the relatively complete response property if it has a finite optimal value for any fixed $(\mathbf{x}, \mathbf{z}) \in \mathrm{Proj}_\mathbf{x}(\Omega_{\mathrm{BiMINLP}}) \times \mathbf{Z}$ [133].

Remark 2.5. *If for any fixed $(\mathbf{x}, \mathbf{z}) \in Proj_\mathbf{x}(\Omega_{BiMINLP}) \times \mathbf{Z}$, $P(\mathbf{x}, \mathbf{z})$ is bounded, and $\xi(\mathbf{x}, \mathbf{z})$ satisfies Slater's condition, then $\xi(\mathbf{x}, \mathbf{z})$ has the relatively complete response property. Moreover, if the lower level problem of BiMINLP is a mixed integer linear programming problem that has the relatively complete response property, then $\xi(\mathbf{x}, \mathbf{z})$ satisfies Slater's condition for any fixed $(\mathbf{x}, \mathbf{z}) \in Proj_\mathbf{x}(\Omega_{BiMINLP}) \times \mathbf{Z}$.*

Since the cardinality of $\mathbf{Z}$ is finite, we can obtain a single level reformulation of BiMINLP by enumerating all the possible values of the integer variable $\mathbf{z}$.

Theorem 2.2. *If $\xi(\mathbf{x}, \mathbf{z})$ has the relatively complete response property and satisfies Slater's condition for any $(\mathbf{x}, \mathbf{z}) \in Proj_\mathbf{x}(\Omega_{BiMINLP}) \times \mathbf{Z}$, then BiMINLP is equivalent to*

1. *The strong duality based single level formulation*

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{y}^0, \mathbf{z}^0, \mu^k} & F(\mathbf{x}, \mathbf{y}^0, \mathbf{z}^0) \\ s.t. & G(\mathbf{x}, \mathbf{y}^0, \mathbf{z}^0) \leq 0 \\ & g(\mathbf{x}, \mathbf{y}^0, \mathbf{z}^0) \leq 0 \\ & f(\mathbf{x}, \mathbf{y}^0, \mathbf{z}^0) \leq s(\mathbf{x}, \mu^k, \mathbf{z}^k), k = 1, 2, ..., K \\ & \mathbf{x} \in \mathbf{X}, \mathbf{y}^0 \in R^{n_\mathbf{y}}, \mathbf{z}^0 \in \mathbf{Z}, \mu^k \in R^q_+, k = 1, 2, ..., K, \end{aligned} \tag{2.11}$$

where $\mathbf{z}^k \in \mathbf{Z}$ for $k = 1, 2, ..., K$, and μ^k are dual variables corresponding to $\mathbf{z}^k$.

2. The KKT conditions based single level formulation

$$
\begin{aligned}
\min_{\mathbf{x},\mathbf{y}^0,\mathbf{z}^0,\mathbf{y}^k,\mu^k} & F(\mathbf{x},\mathbf{y}^0,\mathbf{z}^0) \\
& s.t. G(\mathbf{x},\mathbf{y}^0,\mathbf{z}^0) \leq 0 \\
& g(\mathbf{x},\mathbf{y}^0,\mathbf{z}^0) \leq 0 \\
& f(\mathbf{x},\mathbf{y}^0,\mathbf{z}^0) \leq f(\mathbf{x},\mathbf{y}^k,\mathbf{z}^k),\ k=1,2,...,K \\
& \nabla_{\mathbf{y}^k} f(\mathbf{x},\mathbf{y}^k,\mathbf{z}^k) + (\mu^k)^T \nabla_{\mathbf{y}^k} g(\mathbf{x},\mathbf{y}^k,\mathbf{z}^k) = 0,\ k=1,2...,K \\
& g(\mathbf{x},\mathbf{y}^k,\mathbf{z}^k) \leq 0,\ k=1,2...,K \\
& (\mu^k)^T g(\mathbf{x},\mathbf{y}^k,\mathbf{z}^k) = 0,\ k=1,2...,K \\
& \mathbf{x}\in\mathbf{X}, \mathbf{y}^0 \in R^{n_\mathbf{y}}, \mathbf{z}^0\in\mathbf{Z}, \mathbf{y}^k\in R^{n_\mathbf{y}}, \mu^k\in R^q_+,\ k=1,2...,K,
\end{aligned}
\tag{2.12}
$$

where $\mathbf{z}^k \in \mathbf{Z}$ *for* $k = 1, 2, ..., K,$ *and* $\mathbf{y}^k$, μ^k *are corresponding to* $\mathbf{z}^k$ *if* f *and* g *are continuously differentiable with respect to* $\mathbf{y}$ *for any fixed* $(\mathbf{x},\mathbf{z}) \in Proj_{\mathbf{x}}(\Omega_{BiMINLP}) \times \mathbf{Z}$.

Proof. 1. It is obvious that BiMINLP can be equivalently rewritten as

$$
\begin{aligned}
\min_{\mathbf{x},\mathbf{y}^0,\mathbf{z}^0} \{F(\mathbf{x},\mathbf{y}^0,\mathbf{z}^0) :& G(\mathbf{x},\mathbf{y}^0,\mathbf{z}^0) \leq 0, g(\mathbf{x},\mathbf{y}^0,\mathbf{z}^0) \leq 0, \mathbf{x}\in\mathbf{X}, \mathbf{y}^0\in R^{n_\mathbf{y}}, \mathbf{z}^0\in\mathbf{Z} \\
& f(\mathbf{x},\mathbf{y}^0,\mathbf{z}^0) \leq \min_{\mathbf{y},\mathbf{z}}\{f(\mathbf{x},\mathbf{y},\mathbf{z}) : g(\mathbf{x},\mathbf{y},\mathbf{z}) \leq 0, \mathbf{y}\in R^{n_\mathbf{y}}, \mathbf{z}\in\mathbf{Z}\}\}.
\end{aligned}
\tag{2.13}
$$

Since $\xi(\mathbf{x},\mathbf{z})$ has the relatively complete response property, it has a finite optimal value for any fixed $(\mathbf{x},\mathbf{z}) \in \mathrm{Proj}_{\mathbf{x}}(\Omega_{\mathrm{BiMINLP}}) \times \mathbf{Z}$. Thus, by enumerating all the possible values of $\mathbf{z}$ and introducing corresponding variable $\mathbf{y}$ for each $\mathbf{z}$, the last constraint in (2.13) can be equivalently rewritten as

$$
f(\mathbf{x},\mathbf{y}^0,\mathbf{z}^0) \leq \min_{\mathbf{y}^k}\{f(\mathbf{x},\mathbf{y}^k,\mathbf{z}^k) : g(\mathbf{x},\mathbf{y}^k,\mathbf{z}^k) \leq 0, \mathbf{y}^k\in R^{n_\mathbf{y}}\}, k=1,2,...,K, \tag{2.14}
$$

where $\mathbf{z}^k \in \mathbf{Z}$ are fixed for $k = 1, 2, ..., K$. Since $\xi(\mathbf{x},\mathbf{z})$ satisfies Slater's condition for each k, the strong duality holds for the right hand side minimization problem of (2.14). Replacing the minimization problem by its dual problem for each k, we have

$$
f(\mathbf{x},\mathbf{y}^0,\mathbf{z}^0) \leq \max_{\mu^k}\{s(\mathbf{x},\mu^k,\mathbf{z}^k) : \mu^k \in R^q_+\}, k=1,2,...,K. \tag{2.15}
$$

Since the right hand side of (2.15) is a maximization problem, the "max" operator can be dropped. Hence, (2.15) is equivalent to

$$f(\mathbf{x}, \mathbf{y}^0, \mathbf{z}^0) \leq s(\mathbf{x}, \mu^k, \mathbf{z}^k), \mu^k \in R^q_+, k = 1, 2, ..., K. \tag{2.16}$$

Replacing the last constraint in (2.13) by (2.16) leads to the result.

2. Since the right hand side problem of (2.14) is a convex problem with Slater's condition satisfied, and f and g are continuously differentiable with respect to $\mathbf{y}$, we can replace it with its KKT conditions and rewrite (2.14) as

$$\begin{aligned}
&f(\mathbf{x}, \mathbf{y}^0, \mathbf{z}^0) \leq f(\mathbf{x}, \mathbf{y}^k, \mathbf{z}^k), \ k = 1, 2, ..., K \\
&\nabla_{\mathbf{y}^k} f(\mathbf{x}, \mathbf{y}^k, \mathbf{z}^k) + (\mu^k)^T \nabla_{\mathbf{y}^k} g(\mathbf{x}, \mathbf{y}^k, \mathbf{z}^k) = 0, \ k = 1, 2..., K \\
&g(\mathbf{x}, y^k, \mathbf{z}^k) \leq 0, \ k = 1, 2..., K \\
&(\mu^k)^T g(\mathbf{x}, \mathbf{y}^k, \mathbf{z}^k) = 0, \ k = 1, 2..., K \\
&\mathbf{x} \in \mathbf{X}, \mathbf{y}^0 \in R^{n_\mathbf{y}}, \mathbf{z}^0 \in \mathbf{Z}, \mathbf{y}^k \in R^{n_\mathbf{y}}, \mu^k \in R^q_+, \ k = 1, 2..., K,
\end{aligned} \tag{2.17}$$

where $\mathbf{z}^k \in \mathbf{Z}$ are fixed, and $\mathbf{y}^k$, μ^k are primal and dual variables corresponding to $\mathbf{z}^k$. Replacing the last constraint in (2.13) by (2.17) leads to the result.

□

Remark 2.6. *For any $\bar{\mathbf{Z}} \subseteq \mathbf{Z}$, problem (2.11) (same for (2.12)) with $\mathbf{Z}$ replaced by $\bar{\mathbf{Z}}$ is a relaxation of BiMINLP.*

We indicate that the strong duality based formulation is indeed more general than the KKT conditions based one. To see this, we consider the following BiMINLP instance whose lower level problem is a mixed integer quadratic programming (MIQP) problem.

Example 2.2.

$$\min_{x,y,z}\{-3x + y + z : 0 \leq x \leq 1, (y, z) \in \arg\min_{y,z}\{\frac{1}{2}y^2 : y - 2z \geq x, 0 \leq y \leq 2, z \in \{0, 1\}\}\}.$$

Example 2.2 has an unique optimal solution $(x^*, y^*, z^*) = (1, 1, 0)$ with the optimal value of -2. The parameterized lower level problem of is

$$\xi(x, z) : \min_{y}\{\frac{1}{2}y^2 : y \geq x + 2z, 0 \leq y \leq 2\},$$

and $P(x, z) = \{y \in R : y \geq x + 2z, 0 \leq y \leq 2\}$ is the feasible region of $\xi(x, z)$. For $(x, z) = (1, 1)$, $P(x, z) = \emptyset$, and thus $\xi(x, z)$ does not have the relatively complete response property. Applying the strong duality based formulation to Example 2.2, we have

$$\begin{aligned}
&\min -3x + y^0 + z^0 \\
&\text{s.t.} 0 \leq x \leq 1, x - y^0 + 2z^0 \leq 0, 0 \leq y^0 \leq 2 \\
&\quad \frac{1}{2}(y^0)^2 \leq -\frac{1}{2}(\mu_1^1 + \mu_2^1 - \mu_3^1)^2 + x\mu_1^1 - 2\mu_3^1 \\
&\quad \frac{1}{2}(y^0)^2 \leq -\frac{1}{2}(\mu_1^2 + \mu_2^2 - \mu_3^2)^2 + (x+2)\mu_1^2 - 2\mu_3^2 \\
&\quad z^0 \in \{0, 1\}, \mu_1^1, \mu_2^1, \mu_3^1, \mu_1^2, \mu_2^2, \mu_3^2 \geq 0.
\end{aligned} \tag{2.18}$$

It is easy to verify that $(x^*, y^{0*}, z^{0*}) = (1, 1, 0)$ is optimal to (2.18), and that the optimal value is -2. The KKT conditions based formulation of Example 2.2 is given by

$$\begin{aligned}
&\min -3x + y^0 + z^0 \\
&\text{s.t.} 0 \leq x \leq 1, x - y^0 + 2z^0 \leq 0, 0 \leq y^0 \leq 2 \\
&\quad \frac{1}{2}(y^0)^2 \leq \frac{1}{2}(y^1)^2 \\
&\quad x - y^1 \leq 0, 0 \leq y^1 \leq 2 \\
&\quad y^1 - \mu_1^1 - \mu_2^1 + \mu_3^1 = 0 \\
&\quad \mu_1^1(y^1 - x) = 0, \mu_2^1 y^1 = 0, \mu_3^1(2 - y^1) = 0 \\
&\quad \frac{1}{2}(y^0)^2 \leq \frac{1}{2}(y^2)^2 \\
&\quad x - y^2 + 2 \leq 0, 0 \leq y^2 \leq 2 \\
&\quad y^2 - \mu_1^2 - \mu_2^2 + \mu_3^2 = 0 \\
&\quad \mu_1^2(y^2 - x - 2) = 0, \mu_2^2 y^2 = 0, \mu_3^2(2 - y^2) = 0 \\
&\quad z^0 \in \{0, 1\}, \mu_1^1, \mu_2^1, \mu_3^1, \mu_1^2, \mu_2^2, \mu_3^2 \geq 0.
\end{aligned} \tag{2.19}$$

The optimal solution to (2.19) is $(x^*, y^{0*}, z^{0*}) = (0, 0, 0)$ with the optimal value of 0. However, the optimal solution is infeasible to (2.19) since $x - y^2 + 2 \leq 0$ is violated when $x = 1$.

For the aforementioned example, the strong duality based formulation is equivalent to the original bilevel problem while the KKT conditions based one is not. This is because the lower level primal constraints that may be violated are not in the strong duality based formulation, and thus do not cut off an optimal solution. Indeed, the strong duality based formulation is applicable even if $\xi(\mathbf{x}, \mathbf{z})$ does not have the relatively complete response property.

Theorem 2.3. *If $Q(\mathbf{x}, \mathbf{z})$ is independent of $\mathbf{z}$, and $\xi(\mathbf{x}, \mathbf{z})$ satisfies Slater's condition for fixed $(\mathbf{x}, \mathbf{z}) \in Proj_{\mathbf{x}}(\Omega_{BiMINLP}) \times \mathbf{Z}$ such that $P(\mathbf{x}, \mathbf{z}) \neq \emptyset$, then the strong duality based formulation (2.11) is equivalent to BiMINLP, i.e. if $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*)$ is optimal to BiMINLP then there exists $(\mu_1^*, \mu_2^*, ..., \mu_K^*)$ such that $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*, \mu_1^*, \mu_2^*, ..., \mu_K^*)$ is optimal to (2.11); and if $(\boldsymbol{x}^*, \boldsymbol{y}^{0*}, \boldsymbol{z}^{0*}, \mu_1^*, \mu_2^*, ..., \mu_K^*)$ is optimal to (2.11), $(\mathbf{x}^*, \mathbf{y}^{0*}, \mathbf{z}^{0*})$ is optimal to BiMINLP.*

Proof. We first show the existence of $(\mu_1^*, \mu_2^*, ..., \mu_K^*)$ and then show that the optimal value of (2.11) is $F(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*)$. As $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*)$ is optimal to BiMINLP, we have $G(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*) \leq 0$ and $g(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*) \leq 0$. Moreover, as $Q(\mathbf{x}, \mathbf{z})$ is independent of $\mathbf{z}$, we have $Q(\mathbf{x}^*, \mathbf{z}^k) \neq \emptyset$ for all k. Otherwise we would have $Q(\mathbf{x}^*, \mathbf{z}^k) = \emptyset$ for all k, and thus $\xi(\mathbf{x}, \mathbf{z})$ is either infeasible or unbounded for all k, which contradicts the fact that BiMINLP has an optimal solution. For a fixed k, if $P(\mathbf{x}^*, \mathbf{z}^k) = \emptyset$, the problem

$$\max_{\mu^k} \{s(\mathbf{x}^*, \mu^k, \mathbf{z}^k) : \mu^k \in Q(\mathbf{x}^*, \mathbf{z}^k)\}$$

is unbounded, and the constraint $f(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*) \leq s(\mathbf{x}^*, \mu^k, \mathbf{z}^k)$ naturally holds. If $P(\mathbf{x}^*, \mathbf{z}^k) \neq \emptyset$, we have

$$\begin{aligned} f(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*) &\leq \min_{\mathbf{y}, \mathbf{z}} \{f(\mathbf{x}^*, \mathbf{y}, \mathbf{z}) : g(\mathbf{x}^*, \mathbf{y}, \mathbf{z}) \leq 0, \mathbf{y} \in R^{n_\mathbf{y}}, \mathbf{z} \in \mathbf{Z}\} \\ &\leq \min_{\mathbf{y}^k} \{f(\mathbf{x}^*, \mathbf{y}^k, \mathbf{z}) : g(\mathbf{x}^*, \mathbf{y}^k, \mathbf{z}) \leq 0, \mathbf{y}^k \in R^{n_\mathbf{y}}, \mathbf{z} = \mathbf{z}^k\} \\ &= \max_{\mu^k} \{s(\mathbf{x}^*, \mu^k, \mathbf{z}^k) : \mu^k \in Q(\mathbf{x}^*, \mathbf{z}^k)\}. \end{aligned}$$

The first inequality holds due to the optimality of BiMINLP, and the second one holds because $(\mathbf{y}^k, \mathbf{z}^k)$ is feasible to BiMINLP. The last equality holds by the strong duality and implies that there exits a μ^k such that $f(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*) \leq s(\mathbf{x}^*, \mu^k, \mathbf{z}^k)$. Since such a dual variable μ^k can be found for each k, the existence of $(\mu_1^*, \mu_2^*, ..., \mu_K^*)$ follows.

We next show that the optimal value of (2.11) is $F(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*)$ by contradiction. Suppose there exists a feasible solution $(\bar{\mathbf{x}}, \bar{\mathbf{y}}^0, \bar{\mathbf{z}}^0, \bar{\mu}_1, \bar{\mu}_2, ..., \bar{\mu}_K)$ such that $F(\bar{\mathbf{x}}, \bar{\mathbf{y}}^0, \bar{\mathbf{z}}^0) < F(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*)$. By the feasibility of $(\bar{\mathbf{x}}, \bar{\mathbf{y}}^0, \bar{\mathbf{z}}^0, \bar{\mu}_1, \bar{\mu}_2, ..., \bar{\mu}_K)$, we have

$$f(\bar{\mathbf{x}}, \bar{\mathbf{y}}^0, \bar{\mathbf{z}}^0) \leq s(\bar{\mathbf{x}}, \bar{\mu}^k, \bar{\mathbf{z}}^k), \bar{\mu}^k \in Q(\bar{\mathbf{x}}, \bar{\mathbf{z}}^k), k = 1, 2, ..., K,$$

which is the same as

$$f(\bar{\mathbf{x}}, \bar{\mathbf{y}}^0, \bar{\mathbf{z}}^0) \leq \max_{\mu^k}\{s(\bar{\mathbf{x}}, \mu^k, \bar{\mathbf{z}}^k) : \mu^k \in Q(\bar{\mathbf{x}}, \bar{\mathbf{z}}^k)\}, k = 1, 2, ..., K.$$

For each k, if $P(\bar{\mathbf{x}}, \bar{\mathbf{z}}^k) = \emptyset$ then the problem

$$\min_{\mathbf{y}^k}\{f(\bar{\mathbf{x}}, \mathbf{y}^k, \bar{\mathbf{z}}^k) : g(\bar{\mathbf{x}}, \mathbf{y}^k, \bar{\mathbf{z}}^k) \leq 0, \mathbf{y}^k \in R^{n_{\mathbf{y}}}\}$$

is infeasible, and the constraint

$$f(\bar{\mathbf{x}}, \bar{\mathbf{y}}^0, \bar{\mathbf{z}}^0) \leq \min_{\mathbf{y}^k}\{f(\bar{\mathbf{x}}, \mathbf{y}^k, \bar{\mathbf{z}}^k) : g(\bar{\mathbf{x}}, \mathbf{y}^k, \bar{\mathbf{z}}^k) \leq 0, \mathbf{y}^k \in R^{n_{\mathbf{y}}}\}$$

holds. If $P(\bar{\mathbf{x}}, \bar{\mathbf{z}}^k) \neq \emptyset$, by the weak duality, the above inequality also holds, and we have

$$f(\bar{\mathbf{x}}, \bar{\mathbf{y}}^0, \bar{\mathbf{z}}^0) \leq \min_{\mathbf{y}^k}\{f(\bar{\mathbf{x}}, \mathbf{y}^k, \bar{\mathbf{z}}^k) : g(\bar{\mathbf{x}}, \mathbf{y}^k, \bar{\mathbf{z}}^k) \leq 0, \mathbf{y}^k \in R^{n_{\mathbf{y}}}\}, k = 1, 2, ..., K.$$

Thus $(\bar{\mathbf{x}}, \bar{\mathbf{y}}^0, \bar{\mathbf{z}}^0)$ is feasible to BiMINLP, and $F(\bar{\mathbf{x}}, \bar{\mathbf{y}}^0, \bar{\mathbf{z}}^0) < F(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*)$ contradicts the fact that $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*)$ is optimal to BiMINLP.

For the converse part, let $(\mathbf{x}^*, \mathbf{y}^{0*}, \mathbf{z}^{0*}, \mu_1^*, \mu_2^*, ..., \mu_K^*)$ be optimal to (2.11) but $(\mathbf{x}^*, \mathbf{y}^{0*}, \mathbf{z}^{0*})$ not optimal to BiMINLP. Then there exists an optimal solution $(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}})$ to BiMINLP such that $F(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}) < F(\mathbf{x}^*, \mathbf{y}^{0*}, \mathbf{z}^{0*})$. By the first part of Theorem 2.3, there exists $(\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_K)$ such that $(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}, \hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_K)$ is optimal to (2.11). As $F(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}) < F(\mathbf{x}^*\mathbf{y}^{0*}, \mathbf{z}^{0*})$, it contradicts the fact $(\mathbf{x}^*, \mathbf{y}^{0*}, \mathbf{z}^{0*}, \mu_1^*, \mu_2^*, ..., \mu_K^*)$ is optimal to (2.11) and the result follows. □

In practice, verifying whether $\xi(\mathbf{x},\mathbf{z})$ satisfies Slater's condition for any fixed $(\mathbf{x},\mathbf{z}) \in \text{Proj}_{\mathbf{x}}(\Omega_{\text{BiMINLP}}) \times \mathbf{Z}$ can be challenging. In this case, we can adopt the method introduced in Section 2.2.2 to get an extended formulation of $\xi(\mathbf{x},\mathbf{z})$ as $\hat{\xi}(\mathbf{x},\mathbf{z}) : \min_{\mathbf{y},\hat{\mathbf{y}}}\{f(\mathbf{x},\mathbf{y},\mathbf{z}) + M\mathbf{e}^T\hat{\mathbf{y}} : g(\mathbf{x},\mathbf{y},\mathbf{z}) \leq \hat{\mathbf{y}}, \mathbf{y} \in R^{n_{\mathbf{y}}}, \hat{\mathbf{y}} \in R^q_+\}$, where $\mathbf{e} \in R^q$ with all elements being 1, and $M \geq 0$. Same as $\hat{\theta}(\mathbf{x})$, $\hat{\xi}(\mathbf{x},\mathbf{z})$ satisfies Slater's condition for any $(\mathbf{x},\mathbf{z}) \in \text{Proj}_{\mathbf{x}}(\Omega_{\text{BiMINLP}}) \times \mathbf{Z}$, thus we can replace the lower level problem of BiMINLP by $\min_{\mathbf{y},\hat{\mathbf{y}},\mathbf{z}}\{f(\mathbf{x},\mathbf{y},\mathbf{z}) + M\mathbf{e}^T\hat{\mathbf{y}} : g(\mathbf{x},\mathbf{y},\mathbf{z}) \leq \hat{\mathbf{y}}, \mathbf{y} \in R^{n_{\mathbf{y}}}, \hat{\mathbf{y}} \in R^q_+, \mathbf{z} \in \mathbf{Z}\}$ and apply Theorem 2.2 or Theorem 2.3 to obtain a single level formulation.

2.3.2 Reformulation of Generalized Pessimistic Bilevel Mixed Integer Programming Problems

In this subsection, we consider the extension of GPBO with integer variables, which we refer as generalized pessimistic bilevel mixed integer programming problem (GPBiMINLP)

$$\begin{aligned}\text{GPBiMINLP} : \min_{\mathbf{x},\mathbf{y},\mathbf{z}}\{F(\mathbf{x}) :& G_i(\mathbf{x},\mathbf{y},\mathbf{z}) \leq 0, \forall i \in \mathcal{CC}\backslash\mathcal{PCC}, \\ & G_i(\mathbf{x},\mathbf{y},\mathbf{z}) \leq 0, \forall (\mathbf{y},\mathbf{z}) \in \tilde{R}(\mathbf{x}), i \in \mathcal{PCC}, \mathbf{x} \in \mathbf{X}, (\mathbf{y},\mathbf{z}) \in \tilde{R}(\mathbf{x})\},\end{aligned}$$

where $\tilde{R}(\mathbf{x}) = \arg\min_{\mathbf{y},\mathbf{z}}\{f(\mathbf{x},\mathbf{y},\mathbf{z}) : g(\mathbf{x},\mathbf{y},\mathbf{z}) \leq 0, \mathbf{y} \in R^{n_{\mathbf{y}}}, \mathbf{z} \in \mathbf{Z}\}$ is the lower level rational reaction set. For GPBiMINLP to have an optimal solution, it is necessary to assume that

$$\begin{aligned}\text{Proj}_{\mathbf{x}}(\Omega_{\text{GPBiMINLP}}) = \{\mathbf{x} \in \mathbf{X} :& \exists\ (\mathbf{y},\mathbf{z}) \in R^{n_{\mathbf{y}}} \times \mathbf{Z} \text{ with } g(\mathbf{x},\mathbf{y},\mathbf{z}) \leq 0, \\ & G_i(\mathbf{x},\mathbf{y},\mathbf{z}) \leq 0, \forall i \in \mathcal{CC}\backslash\mathcal{PCC}\} \neq \emptyset.\end{aligned}$$

Furthermore, since Theorem 2.1 makes no convexity assumption, we can apply it to convert GPBiMINLP to a standard BiMINLP problem with multiple followers.

Corollary 2.1. *Denote the lower level feasible set of GPBiMINLP by* $\tilde{L}(\mathbf{x}) = \{(\mathbf{y},\mathbf{z}) \in R^{n_{\mathbf{y}}} \times \mathbf{Z} : g(\mathbf{x},\mathbf{y},\mathbf{z}) \leq 0\}$*, and let* $\tilde{S}_i(\bar{\mathbf{y}}_i,\bar{\mathbf{z}}_i) = \arg\max_{\mathbf{y},\mathbf{z}}\{G_i(\mathbf{x},\mathbf{y},\mathbf{z}) : g(\mathbf{x},\mathbf{y},\mathbf{z}) \leq 0, f(\mathbf{x},\mathbf{y},\mathbf{z}) \leq f(\mathbf{x},\bar{\mathbf{y}}_i,\bar{\mathbf{z}}_i), \mathbf{y} \in R^{n_{\mathbf{y}}}, \mathbf{z} \in \mathbf{Z}\}$*, then GPBiMINLP is equivalent to*

$$\begin{aligned}\min_{\mathbf{x},\bar{\mathbf{y}},\bar{\mathbf{z}},\mathbf{y},\mathbf{y}_i,\mathbf{z},\mathbf{z}_i}\{F(\mathbf{x}) :& G_i(\mathbf{x},\mathbf{y},\mathbf{z}) \leq 0, \forall i \in \mathcal{CC}\backslash\mathcal{PCC}, G_i(\mathbf{x},\mathbf{y}_i,\mathbf{z}_i) \leq 0, \forall i \in \mathcal{PCC} \\ & (\mathbf{y}_i,\mathbf{z}_i) \in S_i(\mathbf{x},\bar{\mathbf{y}},\bar{\mathbf{z}}), \forall i \in \mathcal{PCC}, \mathbf{x} \in \mathbf{X}, (\bar{\mathbf{y}},\bar{\mathbf{z}}) \in \tilde{L}(\mathbf{x}), (\mathbf{y},\mathbf{z}) \in \tilde{R}(\mathbf{x})\},\end{aligned}$$

which is denoted by GPBiMINLP-R.

Proof. We omit the proof as it is almost identical to that of Theorem 2.1. □

2.4 Solution Method for BiMINLP Problems

2.4.1 Decomposition Algorithm and Computational Complexity

Based on the reformulation method introduced in Section 2.3, a single level equivalent formulation can be obtained for BiMINLP under certain conditions. However, as the cardinality of $\mathbf{Z}$ can be extremely large, solving (2.11) or (2.12) directly is not practically possible for medium or large size problems. According to Remark 2.6, if $\mathbf{Z}$ is replaced by its subset, a relaxation problem can be obtained, and the optimal value of the relaxation problem provides a lower bound. Moreover, by enlarging the subset of $\mathbf{Z}$, the low bound can be tightened.

Following this observation, we propose a decomposition algorithm to solve BiMINLP. In particular, in each iteration, we solve a relaxation problem of BiMINLP as the master problem (MP) to get an optimal upper level solution $\mathbf{x}^*$, and then solve sub-problems for $\mathbf{x}^*$. If the optimality gap is not within a predetermined tolerance, a new $\mathbf{z}^*$ from the sub-problem is added to the subset of $\mathbf{Z}$ so that a tighter lower bound can be obtained. Furthermore, by solving the sub-problems, we may get a feasible solution that leads to an upper bound of BiMINLP, and high quality upper bounds can significantly accelerate the convergence.

Let LB and UB be the lower bound and upper bound respectively, ϵ be the optimality tolerance, and l be the iteration index. We present the detailed solution procedure based on the single level formulation (2.11), and it can also be used for (2.12).

Theorem 2.4. *For a BiMINLP instance that satisfies the conditions in Theorem 2.2 or Theorem 2.3, Algorithm 1 returns an optimal solution of the instance in finite number of iterations, and the optimality gap is bounded by ϵ.*

Proof. Since the cardinality of $\mathbf{Z}$ is finite, it is sufficient to show that a repeated $\mathbf{z}^*$ obtained from SP1 or SP2 leads to the termination of the algorithm. Suppose in iteration l, we obtain $(\mathbf{x}^*, \mathbf{y}^{0*}, \mathbf{z}^{0*})$ from MP, and $\mathbf{z}^*$ from SP1 or SP2, but $\mathbf{z}^*$ has been obtained in a previous iteration. If $LB - UB \leq \epsilon$, then the algorithm terminates; otherwise, new variables and constraints are added to MP with $\mathbf{z}^{l+1} = \mathbf{z}^*$. In iteration $l+1$, LB does not change since $\mathbf{z}^*$

Algorithm 1 Decomposition algorithm for solving BiMINLP

Initialize $LB = -\infty$, $UB = +\infty$, $\bar{\mathbf{Z}} = \emptyset$, and $l = 0$

while $UB - LB \geq \epsilon$ **do**

    Solve MP:

$$\eta^* = \min\{F(\mathbf{x}, \mathbf{y}^0, \mathbf{z}^0) : G(\mathbf{x}, \mathbf{y}^0, \mathbf{z}^0) \leq 0, g(\mathbf{x}, \mathbf{y}^0, \mathbf{z}^0) \leq 0, \mathbf{x} \in \mathbf{X}, \mathbf{y}^0 \in R^{n_\mathbf{y}}, \mathbf{z}^0 \in \mathbf{Z}, f(\mathbf{x}, \mathbf{y}^0, \mathbf{z}^0) \leq s(\mathbf{x}, \mu^k, \mathbf{z}^k), \mu^k \in R^q_+, \forall \mathbf{z}^k \in \bar{\mathbf{Z}}\},$$

    obtain an optimal solution $(\mathbf{x}^*, \mathbf{y}^{0*}, \mathbf{z}^{0*}, \mu^{1*}, ..., \mu^{|\bar{\mathbf{Z}}|*})$, and update $LB = \eta^*$

    For $\mathbf{x}^*$ obtained from MP, solve the first sub-problem (SP1)

$$P(\mathbf{x}^*) = \min_{\mathbf{y},\mathbf{z}}\{f(\mathbf{x}^*, \mathbf{y}, \mathbf{z}) : g(\mathbf{x}^*, \mathbf{y}, \mathbf{z}) \leq 0, \mathbf{y} \in R^{n_\mathbf{y}}, \mathbf{z} \in \mathbf{Z}\},$$

    and get an optimal solution $(\mathbf{y}^*_F, \mathbf{z}^*_F)$.

    Solve the second sub-problem (SP2)

$$\min_{\mathbf{y},\mathbf{z}}\{F(\mathbf{x}^*, \mathbf{y}, \mathbf{z}) : G(\mathbf{x}^*, \mathbf{y}, \mathbf{z}) \leq 0, g(\mathbf{x}^*, \mathbf{y}, \mathbf{z}) \leq 0, f(\mathbf{x}^*, \mathbf{y}, \mathbf{z}) \leq P(\mathbf{x}^*), \mathbf{y} \in R^{n_\mathbf{y}}, \mathbf{z} \in \mathbf{Z}\}.$$

    if SP2 has an optimal solution $(\mathbf{y}^*_S, \mathbf{z}^*_S)$ **then**

        Set $\mathbf{z}^* = \mathbf{z}^*_S$ and update $UB = \min\{UB, F(\mathbf{x}^*, \mathbf{y}^*_S, \mathbf{z}^*_S)\}$.

    else

        Set $\mathbf{z}^* = \mathbf{z}^*_F$

    end if

    Update $\bar{\mathbf{Z}} = \bar{\mathbf{Z}} \cup \{\mathbf{z}^*\}$

    Set $l = l + 1$

end while

Return $(\mathbf{x}^*, \mathbf{y}^{0*}, \mathbf{z}^{0*})$ as an optimal solution.

is repeated and the constraints added are identical to those added previously, thus we have

$$
\begin{aligned}
LB &= F(\mathbf{x}^*, \mathbf{y}^{0*}, \mathbf{z}^{0*}) \\
&= \min_{\mathbf{y}^0, \mathbf{z}^0, \mu^k} \{F(\mathbf{x}^*, \mathbf{y}^0, \mathbf{z}^0) : G(\mathbf{x}^*, \mathbf{y}^0, \mathbf{z}^0) \leq 0, g(\mathbf{x}^*, \mathbf{y}^0, \mathbf{z}^0) \leq 0, \\
&\qquad f(\mathbf{x}^*, \mathbf{y}^0, \mathbf{z}^0) \leq s(\mathbf{x}^*, \mu^k, \mathbf{z}^k),\ k = 1, 2, ..., l+1, \\
&\qquad \mathbf{y}^0 \in R^{n_\mathbf{y}}, \mathbf{z}^0 \in \mathbf{Z}, \mu^k \in R^q_+, k = 1, 2, ..., l+1\} \\
&\geq \min_{\mathbf{y}^0, \mathbf{z}^0, \mu^{l+1}} \{F(\mathbf{x}^*, \mathbf{y}^0, \mathbf{z}^0) : G(\mathbf{x}^*, \mathbf{y}^0, \mathbf{z}^0) \leq 0, g(\mathbf{x}^*, \mathbf{y}^0, \mathbf{z}^0) \leq 0, \\
&\qquad f(\mathbf{x}^*, \mathbf{y}^0, \mathbf{z}^0) \leq s(\mathbf{x}^*, \mu^{l+1}, \mathbf{z}^{l+1}), \\
&\qquad \mathbf{y}^0 \in R^{n_\mathbf{y}}, \mathbf{z}^0 \in \mathbf{Z}, \mu^{l+1} \in R^q_+\} \\
&\geq \min_{\mathbf{y}^0, \mathbf{z}^0} \{F(\mathbf{x}^*, \mathbf{y}^0, \mathbf{z}^0) : G(\mathbf{x}^*, \mathbf{y}^0, \mathbf{z}^0) \leq 0, g(\mathbf{x}^*, \mathbf{y}^0, \mathbf{z}^0) \leq 0, \\
&\qquad f(\mathbf{x}^*, \mathbf{y}^0, \mathbf{z}^0) \leq P(\mathbf{x}^*), \mathbf{y}^0 \in R^{n_\mathbf{y}}, \mathbf{z}^0 \in \mathbf{Z}\} \\
&\geq UB.
\end{aligned}
$$

The first inequality holds as removing l set of constraints results in a relaxation of MP. For the second inequality, since $\mathbf{z}^{l+1}$ is optimal to SP1, by weak duality, $s(\mathbf{x}^*, \mu^{l+1}, \mathbf{z}^{l+1}) \leq P(\mathbf{x}^*)$. Thus, replacing $s(\mathbf{x}^*, \mu^{l+1}, \mathbf{z}^{l+1})$ with $P(\mathbf{x}^*)$ leads to a further relaxation of MP. □

The number of iterations is bounded by $|\mathbf{Z}| + 1$. Although $|\mathbf{Z}|$ can be very large, later numerical study shows that the actual number of iterations is much smaller than $|\mathbf{Z}|$ as the linking constraints $f(\mathbf{x}, \mathbf{y}^0, \mathbf{z}^0) \leq s(\mathbf{x}, \mu^k, \mathbf{z}^k)$ along with the dual constraints make the MP a much tighter relaxation than a high point problem [99]. In addition to tighter relaxation, the algorithm can be further improved if a high quality feasible solution can be found through some heuristics. It is important to mention that finding a feasible solution to BiMINLP may not be easy as the lower level variable $\mathbf{y}$ and $\mathbf{z}$ are involved in the upper level constraint. As a result, solving the lower level problem for a fixed $\mathbf{x}$ does not necessarily lead to a feasible solution. For fixed $\mathbf{x}^*$, the lower level problem may have multiple optimal solutions, and the second sub-problem finds those satisfying the upper level coupled constraint and minimizing the upper level objective function.

2.4.2 Illustration Examples

In this section, two instances are solved by the proposed decomposition algorithm. In particular, we first solve Example 2.2 using the strong duality based formulation (2.11) and then solve another BiMINLP instance using the KKT conditions based formulation (2.12).

In the previous section, we see that 2.2 can be reformulated into a single level problem through (2.11). Now we solve it through Algorithm 1. Specifically, the initial MP

$$\min_{x,y^0,z^0}\{-3x+y^0+z^0 : 0\le x\le 1, x-y^0+2z^0\le 0, 0\le y^0\le 2, z^0\in\{0,1\}\}$$

is solved. We have $(x^*, y^{0*}, z^{0*}) = (1,1,0)$ and $LB = -2$. We then solve SP1

$$P(x^*) = \min_{y,z}\{\frac{1}{2}y^2 : y-2z\ge 0, 0\le y\le 2, z\in\{0,1\}\}$$

is solved for $x^* = 1$. We have $(y_F^*, z_F^*) = (1,0)$ and $P(x^*) = 0.5$. Next, SP2

$$\min_{y,z}\{y+z : y-2z-1\ge 0, \frac{1}{2}y^2\le 0.5, 0\le y\le 2, z\in\{0,1\}\}$$

is solved for $x^* = 1$ and $P(x^*) = 0.5$. As SP2 has an optimal solution of $(y_S^*, z_S^*) = (1,0)$, we have $UB = \min\{+\infty, F(x^*, y_S^*, z_S^*)\} = -2 = LB$. Therefore, the algorithm stops and returns the unique optimal solution, which is $(1,1,0)$.

The second example is adopted from [97], where we replace $x_1\in[-1,1]$ with $x_1\in[-1,0]$ so that the lower level problem is convex in y for any fixed (x_1, x_2, z). It is easy to verify that Example 2.3 satisfies the conditions in Theorem 2.2, and thus can be solved by Algorithm 1.

Example 2.3.

$$\min_{x_1,x_2,y,z}\{-x_2-x_1-z+x_1y+10y^2 : x_1-0.5x_2\le 0, -1\le x_1\le 0, x_2\in\{0,1\},$$
$$(y,z)\in\arg\min_{y,z}\{z-x_1y^2+0.5y^4 : y-0.2z\le 0, -1\le y\le 1, z\in\{0,1\}\}\}$$

In the first iteration, we obtain an optimal solution $(x_1^*, x_2^*, y^{0*}, z^{0*}) = (0, 1, 0, 1)$ from the following MP

$$\min_{x_1,x_2,y^0,z^0} \{-x_2 - x_1 - z^0 + x_1 y^0 + 10(y^0)^2 : x_1 - 0.5x_2 \le 0, y^0 - 0.2z^0 \le 0$$
$$- 1 \le x_1 \le 0, -1 \le y^0 \le 1, x_2, z^0 \in \{0, 1\}\},$$

and we have $LB = F(x_1^*, x_2^*, y^{0*}, z^{0*}) = F(0, 1, 0, 1) = -2$. Then, SP1

$$P(x_1^*, x_2^*) = \min_{y,z}\{z + 0.5y^4 : y - 0.2z \le 0, -1 \le y \le 1, z \in \{0, 1\}\}$$

and SP2

$$\min_{y,z}\{-1 - z + 10y^2 : y - 0.2z \le 0, -1 \le y \le 1, z \in \{0, 1\}, z + 0.5y^4 \le P(x_1^*, x_2^*)\}$$

are solved for $(x_1^*, x_2^*) = (0, 1)$. An optimal solution $(y_S^*, z_S^*) = (0, 0)$ is obtained form the SP2, and $UB = \min\{+\infty, F(0, 1, 0, 0)\} = -1$. Since $UB > LB$, the following variables and constraints

$$z^0 - x_1(y^0)^2 + 0.5(y^0)^4 \le -x_1(y^1)^2 + 0.5(y^1)^4, 2x_1y^1 - 2(y^1)^3 + \mu^1 \ge 0$$
$$y^1 + 1 \ge 0, y^1(2x_1y^1 - 2(y^1)^3 + \mu^1) = 0, \mu^1(y^1 + 1) = 0, y^1 \le 0, \mu^1 \ge 0$$

are added to MP with $z^1 = z_S^* = 0$. In the next iteration, we obtain an optimal solution $(x_1^*, x_2^*, y^{0*}, z^{0*}, y^{1*}, \mu^{1*}) = (0, 1, 0, 0, 0, 0)$ and $LB = -1$ from the augmented MP. Then the two sub-problems are solved for $(x_1^*, x_2^*) = (0, 1)$. The optimal solution to SP2 is $(y_S^*, z_S^*) = (0, 0)$, and $UB = \min\{UB, F(0, 0)\} = -1$. Since $LB = UB$, the algorithm terminates.

The optimal solution $(x_1^*, x_2^*, y^*, z^*) = (0, 1, 0, 0)$ and the optimal value -1 are the same as those reported in [97]. As the original problem is a relaxation of Example 2.3 and its optimal solution is feasible to Example 2.3, the two problems have the same optimal solution.

2.5 Numerical Study

In this section, the proposed solution method is applied to BiMINLP instances. The experiments are implemented in Julia [26], and the commercial solver Gurobi [68] and Mosek [9] are used to solve single level formulations such as MP, SP1, and SP2. Some instances are solved with the help of the Julia package BilevelJuMP [62].

Our numerical study includes two parts. In the first part, we apply the proposed method to solve a metabolic network [117, 140] optimization problem in the pessimistic settings. In the second part, we test the method on randomly generated instances of general BiMINLP problems. In particular, we study bilevel mixed integer quadratic programming (BiMIQP) problems, bilevel mixed integer second-order cone programming (BiMISOCP) problems, and bilevel mixed integer bilinear programming (BiMIBLP) problems.

2.5.1 Metabolic Network Optimization

Bilevel optimization has been applied in metabolic engineering [117, 140]. The upper level problem is to achieve an engineering target such as to maximize the production of certain chemicals through gene knockouts. The lower level problem is a metabolic networks problem. Under the MOMA assumption [117], such a problem is formulated as a bilevel quadratic programming (BiQP) problem [110]. We consider the pessimistic counterpart of this problem, which is given by

$$\begin{aligned}
\text{P-MOMA:} \max_{\mathbf{y}} \min_{\mathbf{v}\in R(\mathbf{y})} \ & v_{chemical} \\
\text{s.t.} \ & \sum_{j\in J}(1-y_j) \le K,\ y_j \in \{0,1\},\ \forall j \in J \\
& R(\mathbf{y}) = \arg\min_{\mathbf{v}} \{ \sum_{j\neq chemical} (v_j - w_j)^2 : \\
& \sum_{j\in J} S_{ij} v_j = 0,\ \forall i \in I \\
& v_{glc} = v_{glc_uptake} \\
& v_{biom} \ge v_{biom}^{target} \\
& v_j^{min} y_j \le v_j \le v_j^{max} y_j,\ \forall \in J\}.
\end{aligned}$$

The upper level objective is to maximize the desired biochemical production. Binary variables $y_j \in \{0,1\}$ for $j \in J$ indicates if reaction j is knocked out ($y_j = 0$), we allow up to K genes or reactions to be knocked out. The lower level problem is to minimize the metabolic adjustment process subject to flux balance and bound restrictions. Since our purpose is to evaluate the proposed algorithm rather than to study the metabolic network problem itself, we do not provide the derivation of the bilevel MOMA model, which can be found in [110].

We also consider bounded rationality in P-MOMA. We denote the optimal value of the lower level problem by $\theta(\mathbf{y})$, denote the lower level feasible set by $L(\mathbf{y})$, and denote the enlarged lower level optimal solution set by

$$R_\epsilon(\mathbf{y}) = \{\mathbf{v} : \mathbf{v} \in L(\mathbf{y}), \sum_{j \neq chemical} (v_j - w_j)^2 \leq (1+\epsilon)\theta(\mathbf{y})\}.$$

We replace $R(\mathbf{y})$ with $R_\epsilon(\mathbf{y})$ in P-MOMA, and then apply Corollary 2.1 and Lemma 2.1 to P-MOMA so that it can be solved by Gurobi.

The test data are based on [8] with modifications. The metabolic network has 77 nodes with three configurations, i.e. with 80, 100, and 120 arcs (i.e., removable reactions). We evaluate the solution of both the optimistic and pessimistic MOMA model for different ϵ, and the objective value and computing time are reported in Table 2 and Table 3, respectively.

Table 2: Computational Result of Metabolic Network Problem

Arcs	ϵ	0.1	0.2	0.3	0.4
80	O-MOMA	59.06	43.78	32.08	22.23
	P-MOMA	58.55	43.98	32.86	25.98
100	O-MOMA	81.97	68.47	58.12	49.47
	P-MOMA	106.17	106.17	106.17	106.17
120	O-MOMA	86.93	73.05	62.35	53.36
	P-MOMA	107.77	107.77	107.77	107.77

Table 3: Computing Time for Solving P-MOMA

Arcs	0.1	0.2	0.3	0.4	Avg.(s)
80	739.85	675.40	365.56	312.95	523.44
100	1707.74	736.57	424.38	984.27	963.24
120	3147.29	1695.46	728.68	695.20	1566.66

We observe that P-MOMA is better able to provide reliable solutions against uncertainty caused by bounded rationality. Almost in all cases, P-MOMA has better evaluation result. Moreover, in the 100-arc and 120-arc cases, P-MOMA gives unique solution regardless of ϵ, demonstrated its strong robustness.

From Table 3, we see that most instances are solved within half an hour. The computing time increases as the number of arcs increases. We also notice that it takes much time to solve P-MOMA than to solve O-MOMA, which typically only takes less than 5 minutes.

2.5.2 Numerical Study on General BiMINLP Problems

In this section, we further evaluate the proposed solution method on general BiMINLP problems. In particular, we solve three types of BiMINLP instances: namely, BiMIQP, BiMISOCP, and BiMIBLP. The formulations are provided in the appendix.

The computational results are reported in Table 4, Table 5, and Table 6. "NC" and "NV" denote the number of constraints and that of variables, respectively. Moreover, we use "NV(U)" to denote the number of upper level variables, and use "NV(C/D)" to indicate the number of continuous/discrete variables in the lower level problem. For each setting, we compute 10 randomly generated instances and report the average time and number of iterations. The computing time is limited to one hour, and "Gap (%)" is reported if an instance was not solved to optimality in one hour.

Table 4: Experiment Results on Randomly Generated BiMIQP Instances

Type	NC(U+L)	NV(U)	NV(C/D)	# of iterations	Time (s)	Gap (%)
BiMIQP	5	5	5/5	4.8	12.99	
		10	10/10	3.7	317.23	
	10	5	5/5	2.4	3.11	
		10	10/10	4.7	495.93	
	20	10	10/10	5.2	392.14	
		15	15/15	3.0	1644.35	4.88

We observe that the proposed method can solve small to moderate size BiMINLP problems efficiently in just few iterations. For BiMIOP problem, instances with up to 20 constraints and 45 variables are solved in half an hour on average. For BiMISOCP, the computing time is even much faster, and all the instances are solved in just few seconds. Furthermore,

Table 5: Experiment Results on Randomly Generated BiMISOCP Instances

Type	NC(U+L)	NV(U)	NV(C/D)	# of iterations	Time (s)	Gap (%)
MISOCP	5	5	5/5	3.1	1.21	
		10	10/10	3.5	2.07	
	10	5	5/5	2.1	0.50	
		10	10/10	2.1	0.96	
	20	10	10/10	1.8	1.08	
		15	15/15	1.9	1.49	

Table 6: Experiment Results on Randomly Generated BiMIBLP Instances

Type	NC(U+L)	NV(U)	NV(C/D)	# of iterations	Time (s)	Gap (%)
BiMIBLP	10	10	20/10	1.6	1.17	
		20	40/20	1.8	5.03	
	20	10	20/10	1.6	10.63	
		20	40/20	1.8	358.69	
	30	10	20/10	2.4	46.45	
		20	40/20	1.83	1627.55	5.34

our method is able to solve BiMIBLP instances with up to 30 constraints and 80 variables within 2 iterations averagely. To the best of our to knowledge, no systematic computational study on instances with similar size has been found in the existing literature.

2.6 Conclusion

In this chapter, we investigate general BiMINLP problems. Several optimality conditions based reformulations are developed for both BO with convex lower level problem and that with lower level integer variables. A generalized pessimistic BO framework is introduce that includes optimistic and pessimistic optimization into special case. Based on the single level reformulation, a computing scheme is developed. The numerical studies on real world metabolic network problem and on general BiMINLP problems demonstrate the computational efficiency of the proposed solution method. Better reformulation and fast approximation algorithms could be a future research topic for large scale BiMINLP problems.

3.0 Robust Bilevel Optimization

3.1 Motivation and Preliminaries

As mentioned in the introduction part, the current research on bilevel optimization considering uncertainties is rather limited. Indeed, existing methodologies in both modeling and computational aspects are insufficient to address practical challenges. Up to now, we observe that stochastic programming based approaches are the dominating strategy to handle random factors, primarily exogenous ones. Nevertheless, the uncertain response issue has been long recognized and studied, although it is not treated as a uncertainty challenge. In this chapter, we use BO to refer the problem defined by (1.1) - (1.3).

3.1.1 Current Status on Bilevel Optimization With Uncertainties

Stochastic Bilevel Optimization: Stochastic programming (SP) [27] is a methodology initially developed for the regular monolithic optimization to handle probabilistic uncertainties. The basic idea is to consider the expected performance (or similar risk measure) across all possible scenarios, especially for a finite set of scenarios that might be obtained from sampling. When recourse decisions exist, the deterministic model will be extended to build a two-stage (or multi-stage) SP model by introducing recurse variables and constraints for every scenario. This simple and effective strategy can easily be applied to a bilevel optimization model to develop a stochastic bilevel one. We note that if only the follower needs to handle randomness, it is also related to stochastic programs with equilibrium constraints, noting that the lower level DMP can be replaced by its optimality conditions [103, 91]. A particular advantage of stochastic bilevel optimization is that its deterministic equivalent derived from enumerating the finite number of scenarios (i.e., the associated recurse variables and constraints), although large-scale, demonstrates a block structure friendly for developing decomposition algorithms, e.g. [70] for the linear case. Hence, stochastic bilevel optimization has been often adopted to study practical hierarchical systems. When the underlying distribution is continuous, similar to SP in the regular optimization, sampling or discretizing methods are commonly used to approximate it by generating a finite set of scenarios [119]. In addition to capturing uncertainties occurring in external environment or in perception, this SP based approach can model a decision-dependent uncertainty by allowing scenarios' realization probabilities change according to some decision variables. For stochastic bilevel linear program directly dealing with a general distribution, theoretical analysis regarding the continuity, differentiability, and stability issues is reported in [35].

Applications of stochastic bilevel optimization can be found in addressing many practical problems, such as pricing and operational problems of electricity market (e.g., [39, 83]), capacity expansion problems in power systems (e.g., [127, 133]), and network design problems in logistics, supply chain and transportation systems (e.g., [111, 130, 5]). It is noted that the majority of applications consider exogenous random factors, e.g., demands, traffic flows or wind power generations, which appear in the right-hand-sides of the lower level DMPs. Some studies also recognize the unsureness regarding the follower's DMP in practice, e.g., [41, 90, 75]. Also, the concerned randomness is often modeled by a finite set of scenarios with fixed realization probabilities. One exception appears in [135], which studies a generation capacity expansion problem subject to wind, demand and price uncertainties with decision-dependent probabilities. So, it considers both exogenous and decision-dependent uncertainties.

Robust Bilevel Optimization: Robust optimization (RO) is a relatively new methodology that assumes no probabilistic information on uncertainties [17]. Rather, it simply assumes that DMP's random parameters belong to a set, referred to as an *uncertainty set*. Hence, it is very suitable to model uncertainties whose distributional information is not available or less reliable. Then, instead of considering probability-based risk measures, RO seeks to derive solutions of the best performance in any worst situation within the uncertainty set. Indeed, the existence of multiple optimal solutions in the lower level DMP has been recognized long ago, and the philosophy behind RO has already been applied in bilevel optimization to hedge against the associated response uncertainty [98, 89, 94]. Different from BO, the resulting model imposes an additional maximization operation over $\tilde{\mathbf{y}}$ in the upper level objective function as the following.

$$\textbf{PBL}: \quad \Phi^* = \min_{\mathbf{x}} \max_{\tilde{\mathbf{y}}} \{F(\mathbf{x}, \tilde{\mathbf{y}}) : (1.2) - (1.3)\} \tag{3.1}$$

Clearly, this formulation reflects the leader's pessimistic belief that the follower is not cooperative and always feed back an optimal solution against her interest. Hence, it is referred to as the *pessimistic bilevel optimization(PBL)* model, also known as the weak formulation [94]. On the contrary, BO demonstrates that the follower is fully cooperative to the leader and thereafter is referred to as the *optimistic* model, also known as the strong formulation.

Compared to BO, PBL, a complex tri-level formulation, is much less investigated or utilized. Existing studies on optimality conditions and solution methods include [94, 52, 126, 132] and references therein. Research on robust bilevel optimization considering other types of uncertainties only appear in a few papers. One study in [42] considers a robust bilevel polynomial optimization model with linear upper level constraints and a linear lower level DMP subject to interval uncertainty sets in constraints of both levels. A solution procedure based on computing a sequence of semidefinite programming relaxations is developed. Note that both the leader and the follower are robust optimizers when handle their own (exogenous) uncertainties, and there is no scenario-specific recourse decision. Hence, it is a single stage robust bilevel model. A couple of recent studies in [116, 32] analyze robust bilevel linear programs with uncertain coefficients in the lower level objective functions from the pessimistic perspective, therefore considering uncertainties both in perception and in response. Because those models are of simple or particular structures, special algorithms have been designed. Noting that the follower makes his decision after the exact information is revealed, those models are actually two-stage robust bilevel formulations.

Hybrid Bilevel Optimization: Given the strength of SP and the very limited development of RO in the context of bilevel optimization, we note that hybrid strategies making use of both SP and RO have been developed. One is the strong-weak bilevel optimization formulation proposed and studied in [1, 38] that computes a weighted sum of BO and PBL. Since weights can be interpreted as the probabilities of the follower being cooperative and non-cooperative, it employs SP and RO to jointly consider uncertainties in perception and in response. Another one is a two-stage stochastic bilevel optimization model studied in [129].

It adopts the pessimistic view to handel the response uncertainty and develops a decision rule based approximation method to derive lower and upper bounds. Also, it is noted in [35] that several structural properties developed for stochastic (optimistic) bilevel linear program naturally extend to its pessimistic counterpart.

Overall, we mention that the existing research on bilevel optimization with uncertainties is at its early stage and we do not have a full set of strong and general methodologies to address practical challenges yet, regardless the ubiquitous existence of random factors within a real hierarchical system. This is particularly the case for robust optimization based approaches, which are believed to be more appropriate when a system cares more on reliability or the distributional information is not available or accurate. To change such a situation and to provide practical tools, we present a systematic study on RO based modeling and solution methodologies to address the impacts of uncertainties in bilevel optimization. Specifically, our provide a set of models that are able to capture all types of uncertainties and their interactions, except the decision-dependent uncertainty (which requires an application-specific function to describe the change of uncertainty set with regard to decisions). Also, we consider decision making models within a single stage and over two different stages explicitly. Then, a set of effective solution methods, along with their convergence analysis, are developed to accurately compute all the proposed robust bilevel models. We expect that the presented results provide a substantial support for bilevel optimization in practice, and pave the way to address uncertainty challenges in hierarchical decision making systems.

3.1.2 Basic Concepts and Properties of Bilevel Optimization

In the remainder of this chapter, we assume that all functions involved in bilevel optimization, i.e., F, G, f, and g, are continuously differentiable on their respective domains, and $\mathbb{X}$ and $\mathbb{Y}$ are non-empty and compact with no discrete variable in $\mathbf{y}$. Note that the concepts and properties listed in this subsection are mainly developed for the optimistic formulation BO, as it has been the mainstream in the literature.

The *constraint set* of a bilevel optimization model is defined by all constraints from both the upper and lower level problems, i.e.,

$$\Omega = \{(\mathbf{x}, \mathbf{y}) : G(\mathbf{x}) \leq \mathbf{0}, g(\mathbf{x}, \mathbf{y}) \leq \mathbf{0}, \mathbf{x} \in \mathbb{X}, \mathbf{y} \in \mathbb{Y}\}.$$

A regular monolithic optimization model defined on Ω with the upper level objective function F is called the *high point problem* [99], denoted by $\mathbf{H}(\boldsymbol{\Omega})$.

For a fixed upper level variable $\mathbf{x}$, the lower level feasible set is

$$\psi(\mathbf{x}) = \{\mathbf{y} : g(\mathbf{x}, \mathbf{y}) \leq \mathbf{0}, \mathbf{y} \in \mathbb{Y}\},$$

and the lower level optimal solution set (also known as *rational reaction set*) is

$$\phi(\mathbf{x}) = \arg\min_{\mathbf{y}} \{f(\mathbf{x}, \mathbf{y}) : \mathbf{y} \in \psi(\mathbf{x})\}.$$

The *inducible region* of a bilevel optimization model is defined as

$$\mathcal{IR} = \{(\mathbf{x}, \tilde{\mathbf{y}}) : G(\mathbf{x}) \leq \mathbf{0}, \mathbf{x} \in \mathbb{X}, \tilde{\mathbf{y}} \in \phi(\mathbf{x})\},$$

which is actually the feasible region of **BO**. Therefore, we can rewrite **BO** as

$$\min_{\mathbf{x}, \tilde{\mathbf{y}}} \{F(\mathbf{x}, \tilde{\mathbf{y}}) : (\mathbf{x}, \tilde{\mathbf{y}}) \in \mathcal{IR}\}.$$

The *lower level optimal value function* is

$$v(\mathbf{x}) = \min_{\mathbf{y}} \{f(\mathbf{x}, \mathbf{y}) : \mathbf{y} \in \psi(\mathbf{x})\},$$

and **BO** can also be rewritten as

$$\min_{\mathbf{x}, \tilde{\mathbf{y}}} \{F(\mathbf{x}, \tilde{\mathbf{y}}) : G(\mathbf{x}) \leq \mathbf{0}, \mathbf{x} \in \mathbb{X}, g(\mathbf{x}, \tilde{\mathbf{y}}) \leq \mathbf{0}, \tilde{\mathbf{y}} \in \mathbb{Y}, f(\mathbf{x}, \tilde{\mathbf{y}}) \leq v(\mathbf{x})\}.$$

Given that $\mathcal{IR} \subseteq \Omega$ and they share the same objective function, the high point problem $\mathbf{H}(\boldsymbol{\Omega})$ is a relaxation of **BO**. This property and $\mathbf{H}(\boldsymbol{\Omega})$ have often been employed as a basis to make use of various methods of regular optimization to attack complex **BO**.

Remark 3.1. *When f is continuous and $\psi(\mathbf{x})$ is compact for fixed $\mathbf{x}$, the lower level problem has an optimal solution as long as $\psi(\mathbf{x})$ is not empty. Therefore, it is clear that $\mathcal{IR} \neq \emptyset$ if and only if $\Omega \neq \emptyset$.*

Different from regular optimization problems, the existence of optimal solutions of BO might not be straightforward. In the following, we present a set of sufficient conditions that ensure the existence of an optimal solution to them.

Theorem 3.1. *Suppose that* $\Omega \neq \emptyset$*, then* **BO** *has an optimal solution if (i) [Adapted from [49]] the Mangasarian-Fromovitz constraint qualification (with respect to* $\mathbf{y}$*) is satisfied for all* $(\mathbf{x}, \mathbf{y}) \in \Omega$*, or (ii)* $\mathbf{x}$ *is discrete, i.e.,* $\mathbb{X} \subseteq \mathbb{Z}_+^{m_d}$*.*

Corollary 3.1 (Adapted from [49])**.** *If* $\Omega \neq \emptyset$ *and all functions involved in* **BO** *are linear, then it has an optimal solution.*

Note that this result holds if $\mathbf{x}$ also involves discrete variables, given that we can enumerate discrete variables.

Remark 3.2. *(i) Bilevel optimization problems are generally difficult to solve. Even for the simplest linear BO with both upper and lower DMPs being linear programs, it is strongly NP-hard [53]. Since all robust bilevel optimization models presented in this chapter reduce to standard BO if no uncertainty exists, they are strongly NP-hard too.*
(ii) Regarding the computational issue of BO, many analytical and heuristic methods have been developed [14, 49, 43], including vertex enumeration, penalty and descent methods. We note that the most popular strategy adopted is to compute a regular single level reformulation. Specifically, if BO has a convex lower level DMP satisfying some constraint qualification, then the lower level DMP can be replaced by its optimality conditions, e.g., those based on Karush–Kuhn–Tucker (KKT) conditions and based on the strong duality. The resulting single level formulation can be either computed by customizing nonlinear programming algorithms according to its particular structure, or converted into a mixed integer program (MIP) [10] and computed by general-purpose MIP solvers.

3.2 Robust Bilevel Model With Exogenous Uncertainty

In this section, we study robust bilevel optimization without scenario-specific recourse decisions, i.e., both the leader and the follower make a single decision across all possible scenarios. We start with the simple case where only exogenous uncertainty is involved. Then we proceed to consider uncertainties in perception and response. In fact, if the response uncertainty is ignored, it implies the optimistic settings are considered, where the follower is cooperative towards the leader.

3.2.1 Model Development and Basic Properties

Let variables $\mathbf{u} \in \mathbb{U}$ (and $\mathbf{w} \in \mathbb{W}$) respectively) represent the exogenous random factors considered by the leader (and the follower, respectively). Without loss of generality, we assume in the remainder of this chapter that uncertainty sets $\mathbb{U}$ and $\mathbb{W}$ are compact and non-empty. Also, objective functions of both DMs, i.e., F and f, are independent of uncertainties, noting that we can always introduce dummy variables to represent objective functions while treat those objective functions as constraints. The single-stage robust bilevel optimization model subject to exogenous uncertainties can be formulated next.

$$\mathbf{R1-BO}: \quad \Theta^*_{R1} = \min F(\mathbf{x}, \tilde{\mathbf{y}}) \tag{3.2}$$

$$\text{s.t.} \quad G(\mathbf{x}, \mathbf{u}) \leq \mathbf{0}, \forall \mathbf{u} \in \mathbb{U}, \mathbf{x} \in \mathbb{X} \tag{3.3}$$

$$\tilde{\mathbf{y}} \in \phi_{\mathbb{W}}(\mathbf{x}) \equiv \arg\min\{f(\mathbf{x}, \mathbf{y}) : g(\mathbf{x}, \mathbf{y}, \mathbf{w}) \leq \mathbf{0}, \forall \mathbf{w} \in \mathbb{W}, \mathbf{y} \in \mathbb{Y}\}. \tag{3.4}$$

Following the convention in bilevel optimization literature, we define the constraint set of $\mathbf{R1-BO}$ as

$$\Omega_{\mathbf{R1-BO}} = \{(\mathbf{x}, \mathbf{y}) : G(\mathbf{x}, \mathbf{u}) \leq \mathbf{0}, \forall \mathbf{u} \in \mathbb{U}, g(\mathbf{x}, \mathbf{y}, \mathbf{w}) \leq \mathbf{0}, \forall \mathbf{w} \in \mathbb{W}, \mathbf{x} \in \mathbb{X}, \mathbf{y} \in \mathbb{Y}\},$$

and denote the corresponding high point problem by $H(\Omega_{\mathbf{1R-BO}})$. Again, we have that $H(\Omega_{\mathbf{R1-BO}})$ is a relaxation of $\mathbf{R1-BO}$. Note that $H(\Omega_{\mathbf{R1-BO}})$ can be treated as an RO formulation of $\mathbf{H}(\boldsymbol{\Omega})$. Correspondingly, the inducible region of this robust bilevel optimization is defined as

$$\mathcal{IR}_{\mathbf{R1-BO}} = \{(\mathbf{x}, \tilde{\mathbf{y}}) : G(\mathbf{x}, \mathbf{u}) \leq \mathbf{0}, \forall \mathbf{u} \in \mathbb{U}, \mathbf{x} \in \mathbb{X}, \tilde{\mathbf{y}} \in \phi_{\mathbb{W}}(\mathbf{x})\}.$$

It is easy to verify that $\mathcal{IR}_{\mathbf{R1-BO}} \neq \emptyset$ if and only if $\Omega_{\mathbf{R1-BO}} \neq \emptyset$. Moreover, by using maximization operation to replace the constraint satisfaction with respect to all $\mathbf{u}$ and $\mathbf{w}$ in

$\Omega_{\mathbf{R1-BO}}$, it is clear that computing the following bilevel optimization model helps check the feasibility of $\mathbf{R1-BO}$.

Proposition 3.1. *Let* $\mathbf{e}_1$ *and* $\mathbf{e}_2$ *be vectors with all elements being* 1*, and*

$$z^* = \min_{\mathbf{x}\in\mathbb{X},\mathbf{y}\in\mathbb{Y},\mathbf{t}_1,\mathbf{t}_2\geq\mathbf{0}} \{\mathbf{e}_1^T\mathbf{t}_1 + \mathbf{e}_2^T\mathbf{t}_2 : G(\mathbf{x},\mathbf{u}) \leq \mathbf{t}_1, \forall\mathbf{u}\in\mathbb{U},\ g(\mathbf{x},\mathbf{y},\mathbf{w}) \leq \mathbf{t}_2, \forall\mathbf{w}\in\mathbb{W}\}. \quad (3.5)$$

We have that $\mathbf{R1-BO}$ *is feasible (equivalently* $\mathcal{IR}_{\mathbf{R1-BO}} \neq \emptyset$*) if and only if* $z^* = 0$*.*

Proof. If $\mathcal{IR}_{\mathbf{R1-BO}} \neq \emptyset$, then $\Omega_{\mathbf{R1-BO}} \neq \emptyset$. Thus, there exists $(\mathbf{x}^*, \mathbf{y}^*)$ such that $G(\mathbf{x}^*, \mathbf{u}) \leq \mathbf{0}, \forall\mathbf{u}\in\mathbb{U}, g(\mathbf{x}^*, \mathbf{y}^*, \mathbf{w}) \leq \mathbf{0}, \forall\mathbf{w}\in\mathbb{W}, \mathbf{x}^*\in\mathbb{X}, \mathbf{y}^*\in\mathbb{Y}$. Let $\mathbf{t}_1^* = \mathbf{t}_2^* = \mathbf{0}$, then $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{t}_1^*, \mathbf{t}_2^*)$ is optimal to (3.5), and $z^* = 0$. Conversely, if $z^* = 0$, then we must have $\mathbf{t}_1^* = \mathbf{t}_2^* = \mathbf{0}$. This implies that $\Omega_{\mathbf{R1-BO}} \neq \emptyset$ and thus $\mathcal{IR}_{\mathbf{R1-BO}} \neq \emptyset$. □

Note that the feasibility check problem (3.5) is a RO formulation of a single-level optimization problem. It can be solved by a couple of matured methods in existing RO literature such as those in [58, 22]. The next result provides a sufficient condition for the existence of optimal solutions to $\mathbf{R1-BO}$. We omit the proof as it is directly implied by Theorem 3.1.

Corollary 3.2. *Let* $\Omega_{\mathbf{R1-BO}} \neq \emptyset$*, then* $\mathbf{R1-BO}$ *has an optimal solution if the Mangasarian-Fromovitz constraint qualification (MFCQ) is satisfied with respect to* $\mathbf{y}$ *for all* $(\mathbf{x},\mathbf{y}) \in \Omega_{\mathbf{R1-BO}}$.

Regarding the instance of $\mathbf{R1-BO}$ studied in [42], its lower level problem is a robust LP with matrix coefficients belonging to a box uncertainty set. As that robust LP can be equivalently reformulated as a larger LP with constraints populated by enumerating extreme points of the box set, MFCQ is satisfied and the problem has an optimal solution as long as its constraint set is non-empty. Note that using the duality based reduction method described in the following, this conclusion holds for any polyhedral uncertainty set.

3.2.2 Reformulations and Relaxations for Computing $\mathbf{R1-BO}$

As both upper and low level constraints need to hold for all possible $\mathbf{u} \in \mathbb{U}$ and $\mathbf{w} \in \mathbb{W}$, the upper and lower level DMPs in $\mathbf{R1-BO}$ could be two complex semi-infinite problems, which makes the complete formulation $\mathbf{R1-BO}$ very challenging to analyze and compute. By using maximization operation to replace such constraint satisfaction, $\mathbf{R1-BO}$ can be readily converted into the following multi-level formulation.

Theorem 3.2. *Let $\mathbb{C}_l = \{1, \ldots, p\}$ and $\mathbb{C}_f = \{1, \ldots, q\}$ denote the index sets of constraints of the upper and lower level DMPs, respectively. The robust bilevel model $\mathbf{R1-BO}$ is equivalent to the following one.*

$$\min \quad F(\mathbf{x}, \tilde{\mathbf{y}}) \tag{3.6}$$

$$\max_{\mathbf{u}_i \in \mathbb{U}} G_i(\mathbf{x}, \mathbf{u}_i) \quad \leq 0, \ \forall i \in \mathbb{C}_l \tag{3.7}$$

$$\mathbf{x} \in \mathbb{X}, \tag{3.8}$$

$$\tilde{\mathbf{y}} \in \arg\min\{f(\mathbf{x}, \mathbf{y}) : \tag{3.9}$$

$$\max_{\mathbf{w}_j \in \mathbb{W}} g_j(\mathbf{x}, \mathbf{y}, \mathbf{w}_j) \quad \leq 0, \ \forall j \in \mathbb{C}_f \tag{3.10}$$

$$\mathbf{y} \in \mathbb{Y}\}. \tag{3.11}$$

Proof. Note that an upper level solution $\mathbf{x}^0$ satisfies all upper level constraints simultaneously for $\mathbf{u} \in \mathbb{U}$ in $\mathbf{R1-BO}$ if and only if

$$G_i(\mathbf{x}^0, \mathbf{u}_i) \leq 0, \forall \mathbf{u}_i \in \mathbb{U}, \ i \in \mathbb{C}_l$$

which is equivalent to

$$\max_{\mathbf{u}_i \in \mathbb{U}} G_i(\mathbf{x}^0, \mathbf{u}_i) \leq 0, \ i \in \mathbb{C}_l$$

Since G_i is continuous and $\mathbb{U}$ is non-empty compact, the maximization problems in (3.7) achieve their optimal values. For a fixed upper level decision $\mathbf{x}^0$, the same argument holds for a feasible lower level solution $\mathbf{y}^0$ with respect to $\mathbb{W}$. Hence, the conclusion follows. □

Compared to the initial formulation $\mathbf{R1-BO}$, this constraint-wise equivalent one allows us to consider the impact of randomness on constraints individually. Note that variable $\mathbf{u}_i$ or $\mathbf{w}_j$ in one constraint is independent of those of other constraints. In particular, if $\mathbb{U}$ or $\mathbb{W}$ is a convex set and G_i or g_j is concave in $\mathbf{u}_i$ or $\mathbf{w}_j$, the KKT conditions or the strong duality for the maximization operation can be utilized to further reduce the model complexity. We illustrate a strong duality based reduction for $\mathbf{R1-BO}$ as follows.

Corollary 3.3. *Suppose that $\mathbb{U}$ and $\mathbb{W}$ are convex sets satisfying Slater's condition, and G_i and g_j are concave in $\mathbf{u}_i$ and $\mathbf{w}_j$ for $i \in \mathbb{C}_l$ and $j \in \mathbb{C}_f$, respectively. Assume the dual problem of* (3.7) *can be analytically represented as* $\min_{\mathbf{v}_{l_i}}\{G_i'(\mathbf{x}, \mathbf{v}_{l_i}) : \mathbf{v}_{l_i} \in \mathbb{V}_{l_i}(\mathbf{x}) \text{ for } i \in \mathbb{C}_l\}$, *and that of* (3.10) *as* $\min_{\mathbf{v}_{f_j}}\{g_j'(\mathbf{x}, \mathbf{y}, \mathbf{v}_{f_j}) : \mathbf{v}_{f_j} \in \mathbb{V}_{f_j}(\mathbf{x}, \mathbf{y}) \text{ for } j \in \mathbb{C}_f\}$. $\mathbf{R1-BO}$ *can be reformulated as*

$$
\begin{aligned}
\mathbf{R1-BO/D}: \min \quad & F(\mathbf{x}, \tilde{\mathbf{y}}) \\
& G_i'(\mathbf{x}, \mathbf{v}_{l_i}) \le 0, \ i = 1, \dots, p \\
& \mathbf{v}_{l_i} \in \mathbb{V}_{l_i}(\mathbf{x}), \ i = 1, \dots, p \\
& \mathbf{x} \in \mathbb{X}, \\
& \tilde{\mathbf{y}} \in \arg\min\{f(\mathbf{x}, \mathbf{y}) : \\
& \qquad g_j'(\mathbf{x}, \mathbf{y}, \mathbf{v}_{f_j}) \le 0, \ j = 1, \dots, q \\
& \qquad \mathbf{v}_{f_j} \in \mathbb{V}_{f_j}(\mathbf{x}, \mathbf{y}), \ j = 1, \dots, q \\
& \qquad \mathbf{y} \in \mathbb{Y}\}.
\end{aligned}
$$

Proof. Because the uncertainty set $\mathbb{U}_l$ is convex with Slater's condition satisfied and G_i is concave in $\mathbf{u}_{l_i}$, the strong duality holds and (3.7) can be replaced by its dual problem to have

$$\min G_i'(\mathbf{x}, \mathbf{v}_{l_i}) \le 0, \ \mathbf{v}_{l_i} \in \mathbb{V}_{l_i}(\mathbf{x})$$

for all i. The minimization operation can be safely removed as it appears in the left-hand-side (LHS) of the "$\le$" sign. By applying the same argument to the robust lower level DMP, the desired result follows. □

By applying Corollary 3.3, the multi-level robust optimization model $\mathbf{R1-BO}$ can be reduced to a deterministic bilevel optimization model $\mathbf{R1-BO/D}$. Note that the KKT conditions can also be used to achieve a similar reduced formulation, which then has complementary slackness constraints.

Remark 3.3. *(i) Many uncertainty sets adopted in the robust optimization literature, including interval, general polyhedral and ellipsoidal sets, readily support this duality based reduction as they are convex sets satisfying Slater's condition. Indeed, such reduction operation might not drastically increase the complexity of the upper or lower level DMP. As shown in [23, 19], the dual problem of a robust linear constraint over a polyhedral uncertainty set is just a linear program of a size proportional to the dimensions of the uncertainty set and the number of uncertain coefficients.*
(ii) The deterministic bilevel reformulation $\mathbf{R1-BO/D}$ *provides us a great convenience to study and compute the original* $\mathbf{R1-BO}$ *using existing methodologies for classical bilevel optimization. For example, if the lower level DMP in* $\mathbf{R1-BO/D}$ *is convex and satisfies some constraint qualification, it can be replaced by its KKT conditions or dual problems to convert* $\mathbf{R1-BO/D}$ *into a single level formulation that could be directly solved by off-the-shelf packages.*
(iii) If $\mathbf{u}_i$ *only appears in the right-hand-side (RHS) of the* $i-th$ *constraint for some* $i \in \mathbb{C}_l$, *i.e., the constraint is in the form of* $G_i(\mathbf{x}) \leq h(\mathbf{u}_i), \forall \mathbf{u}_i \in \mathbb{U}$, *it can be simplified by preprocessing. Specifically, let* $h_i^* = \min\{h(\mathbf{u}_i) : \mathbf{u}_i \in \mathbb{U}\}$. *Then, this constraint can be replaced by* $G_i(\mathbf{x}) \leq h_i^*$. *The same strategy can be applied to a constraint in the lower level DMP if the uncertainty only appears in its RHS.*

Note that the duality or KKT conditions based reduction method is not applicable to a general discrete uncertainty set. We might be interested in solving $\mathbf{R1-BO}$'s less complicated relaxations. For the classical RO formulations of a single level optimization, it is observed that an RO model built upon a subset of the original uncertainty set is a relaxation, regardless of the original uncertainty set's continuity. This observation actually is the foundation to develop the iterative cutting plane (also known as constraint generation) method for exact solutions [58, 22]. Nevertheless, as shown in the next, that relaxation strategy is only valid for the upper level DMP of $\mathbf{R1-BO}$, while it can be extended in a rather counter-intuitive way to derive a correct relaxation of the complete robust formulation.

Theorem 3.3. *Let $\check{\mathbb{U}}$, $\check{\mathbb{W}}$ and $\hat{\mathbb{W}}$ be three uncertainty sets such that $\check{\mathbb{U}} \subseteq \mathbb{U}$ and $\check{\mathbb{W}} \subseteq \mathbb{W} \subseteq \hat{\mathbb{W}}$, respectively. The following robust bilevel optimization problem is a relaxation to* $\mathbf{R1-BO}$

$$
\begin{aligned}
\mathbf{R1-BO/R}(\check{\mathbb{U}}, \check{\mathbb{W}}, \hat{\mathbb{W}}): \quad & \min F(\mathbf{x}, \tilde{\mathbf{y}}) \\
st. \quad & \mathbf{x} \in \mathbb{X},\ G(\mathbf{x}, \mathbf{u}) \leq \mathbf{0}, \forall \mathbf{u} \in \check{\mathbb{U}}, \\
& \tilde{\mathbf{y}} \in \mathbb{Y},\ g(\mathbf{x}, \tilde{\mathbf{y}}, \mathbf{w}) \leq \mathbf{0}, \forall \mathbf{w} \in \check{\mathbb{W}} \\
& f(\mathbf{x}, \tilde{\mathbf{y}}) \leq \eta \\
& \eta = \min\{f(\mathbf{x}, \mathbf{y}) : g(\mathbf{x}, \mathbf{y}, \mathbf{w}) \leq \mathbf{0}, \forall \mathbf{w} \in \hat{\mathbb{W}}, \mathbf{y} \in \mathbb{Y}\}.
\end{aligned}
$$

Denoting its optimal value by $\tilde{\Theta}^*_{R1-e}(\check{\mathbb{U}}, \check{\mathbb{W}}, \hat{\mathbb{W}})$, we have

$$\tilde{\Theta}^*_{R1-e}(\check{\mathbb{U}}, \check{\mathbb{W}}, \hat{\mathbb{W}}) \leq \tilde{\Theta}^*_{R1-e}(\mathbb{U}, \mathbb{W}, \mathbb{W}) = \Theta^*_{R1}.$$

Proof. As η is a dummy variable to enforce the optimality regarding $\tilde{y}$, it is easy to see that $\mathbf{R1-BO}$ and $\mathbf{R1-BO/R}(\mathbb{U}, \mathbb{W}, \mathbb{W})$ are equivalent. Let the optimal value of $\mathbf{R1-BO/R}(\check{\mathbb{U}}, \check{\mathbb{W}}, \hat{\mathbb{W}})$ denoted by $\tilde{\Theta}^*_{R1-e}(\check{\mathbb{U}}, \check{\mathbb{W}}, \hat{\mathbb{W}})$. As $\mathbb{W} \subseteq \hat{\mathbb{W}}$, we have

$$
\begin{aligned}
&\min\{f(\mathbf{x}, \mathbf{y}) : \mathbf{y} \in \mathbb{Y}, g(\mathbf{x}, \mathbf{y}, \mathbf{w}) \leq \mathbf{0}, \forall \mathbf{w} \in \hat{\mathbb{W}}\} \geq \\
&\min\{f(\mathbf{x}, \mathbf{y}) : \mathbf{y} \in \mathbb{Y}, g(\mathbf{x}, \mathbf{y}, \mathbf{w}) \leq \mathbf{0}, \forall \mathbf{w} \in \mathbb{W}\},
\end{aligned}
$$

regardless the value of $\mathbf{x}$. Hence,

$$\tilde{\Theta}^*_{R1-e}(\mathbb{U}, \mathbb{W}, \hat{\mathbb{W}}) \leq \tilde{\Theta}^*_{R1-e}(\mathbb{U}, \mathbb{W}, \mathbb{W}) = \Theta^*_{R1}.$$

Moreover, for any fixed $\hat{\mathbb{W}}$, we have

$$\tilde{\Theta}^*_{R1-e}(\check{\mathbb{U}}, \check{\mathbb{W}}, \hat{\mathbb{W}}) \leq \tilde{\Theta}^*_{R1-e}(\mathbb{U}, \mathbb{W}, \hat{\mathbb{W}}),$$

as the feasible set of $(\mathbf{x}, \tilde{\mathbf{y}})$ in the first problem is at least as large as that in the last problem. Hence, the desired conclusion follows. □

The next result simply allows us to consider the continuous relaxation of discrete $\mathbb{W}$.

Corollary 3.4. *Assume* $\mathbb{W}$ *is a discrete set and* $\tilde{\mathbb{W}}$ *is a continuous set subsuming* $\mathbb{W}$*, and sets* $\check{\mathbb{U}}$ *and* $\check{\mathbb{W}}$ *are defined as in Theorem 3.3. Then,* $\mathbf{R1-BO/R}(\check{\mathbb{U}},\check{\mathbb{W}},\tilde{\mathbb{W}})$ *is a relaxation to* $\mathbf{R1-BO}$*, and we have* $\tilde{\Theta}^*_{R1-e}(\check{\mathbb{U}},\check{\mathbb{W}},\tilde{\mathbb{W}}) \le \tilde{\Theta}^*_{R1-e}(\mathbb{U},\mathbb{W},\mathbb{W}) = \Theta^*_{R1}$.

Remark 3.4. *Let* $Cr(\mathbb{W})$ *denote the continuous relaxation of* $\mathbb{W}$*. If the convexity conditions in Corollary 3.3 hold for* $Cr(\mathbb{W})$ *and* g_j*, the lower level problem of* $\mathbf{R1-BO/R}(\check{\mathbb{U}},\check{\mathbb{W}},Cr(\mathbb{W}))$ *renders itself suitable for the optimality conditions based reduction method, through which we will have a robust bilevel model with a deterministic lower level DMP as that of* $\mathbf{R1-BO/D}$*. By replacing the lower level DMP with its optimality conditions,* $\mathbf{R1-BO/R}(\check{\mathbb{U}},\check{\mathbb{W}},Cr(\mathbb{W}))$ *can be converted into an RO formulation for a single level optimization problem. As mentioned, this type of RO formulations, regardless of the continuity of* $\mathbb{U}$*, can be directly computed by the cutting plane method [58, 22].*

By Corollary 3.4 and the aforementioned discussion, we can derive a more computationally friendly relaxation and solve $\mathbf{R1-BO}$ with discrete uncertainty sets approximately. Moreover, it also provides a basis to develop an exact solution procedure by iteratively strengthening that relaxation.

3.2.3 Cut-and-Branch Algorithm for Discrete Uncertainty Sets

In this section, we develop a novel algorithm to solve $\mathbf{R1-BO}$ with discrete uncertainty set, $\mathbb{W} = \{\mathbf{w} \in \mathbb{Z}^{n_\mathbf{w}} : \mathbf{D}(\mathbf{w}) \le \mathbf{0}\}$. As discussed previously, such a problem is equivalent to $\mathbf{R1-BO/R}(\mathbb{U},\mathbb{W},\mathbb{W})$, and we also have its continuous relaxation problem $\mathbf{R1-BO/R}(\mathbb{U},\mathbb{W},Cr(\mathbb{W}))$, where $Cr(\mathbb{W})$ is the continuous relaxation of $\mathbb{W}$. Moreover, for $\mathbb{W}_1,\mathbb{W}_2 \subseteq Cr(\mathbb{W})$ such that $\mathbb{W}_1 \cap \mathbb{W}_2 = \emptyset$ and $\mathbb{W} \subseteq \mathbb{W}_1 \cup \mathbb{W}_2$, if we denote $\mathbb{W}' = \mathbb{W}_1 \cup \mathbb{W}_2$, then we have $\mathbf{R1-BO/R}(\mathbb{U},\mathbb{W},\mathbb{W}')$ is a stronger relaxation of $\mathbf{R1-BO/R}(\mathbb{U},\mathbb{W},\mathbb{W})$.

To solve $\mathbf{R1-BO}$, we first solve $\mathbf{R1-BO/R}(\mathbb{U},\mathbb{W},Cr(\mathbb{W}))$ and obtain an optimal $\mathbf{x}^*$ as well as a lower bound (LB). Then we solve the lower level problem through cutting plane method for $\mathbf{x}^*$ and obtain an optimal $\mathbf{y}^*$ as well as an upper bound (UB). If the difference between UB and LB is not within a pre-defined tolerance (ϵ), we solve separation problems to find $\mathbf{w} \in Cr(\mathbb{W})\backslash\mathbb{W}$ and then branch on $\mathbf{w}$ to have $\mathbb{W}_1,\mathbb{W}_2 \subseteq Cr(\mathbb{W})$ satisfying the aforementioned conditions to get a tighter LB. We name this solution method Cut-and-Branch Algorithm since we borrow the branching idea from integer programming and integrates it with the classical cutting plane method developed for the regular RO [58, 22].

In addition to UB, LB and ϵ, we also denote by $\mathbb{W}^k$ the relaxation of $\mathbb{W}$ in the kth iteration. Following Remark 3.4, we assume that an oracle based on the cutting plane method (i.e., the cutting plane oracle) is available to solve the RO formulation of a single level optimization model, including $\mathbf{R1-BO/R}(\mathbb{U},\mathbb{W},\mathbb{C}r(\mathbb{W}))$ with the lower level problem satisfying necessary convexity conditions and replaced by its optimality conditions.

Algorithm 2 Cut-and-Branch Algorithm for $\mathbf{R1-BO}$

Initialize $UB=+\infty$, $UB=-\infty$, $k=0$, and $\mathbb{W}^k=\mathbb{C}r(\mathbb{W})$
while $UB-LB\geq\epsilon$ **do**
    Solve $\mathbf{R1-BO/R}(\mathbb{U},\mathbb{W},\mathbb{W}^k)$ by the oracle, obtain an optimal solution $(\mathbf{x}^{k*},\tilde{\mathbf{y}}^{k*},\eta^{k*})$
    Update $LB=F(\mathbf{x}^{k*},\tilde{\mathbf{y}}^{k*})$
    For $\mathbf{x}^{k*}$, solve the lower level problem in (3.9-3.11) and obtain an optimal $\mathbf{y}^{k*}$.
    Update $UB=\min\{UB,F(\mathbf{x}^{k*},\mathbf{y}^{k*})\}$
    for $j\in\mathbb{C}_f$ **do**
        Solve the separation problem

$$\xi_j(\mathbf{x}^{k*},\mathbf{y}^{k*})=\max\{g_j(\mathbf{x}^{k*},\mathbf{y}^{k*},\mathbf{w}_j^k):\mathbf{w}_j^k\in\mathbb{W}^k\}$$

    and obtain an optimal solution $\mathbf{w}_j^{k*}$ for given $(\mathbf{x}^{k*},\mathbf{y}^{k*})$
        if $\xi_j(\mathbf{x}^{k*},\mathbf{y}^{k*})>0$ **then**
            Branch on $\mathbf{w}_j^{k*}$, i.e., choose $l\in\{1,2,...n_{\mathbf{w}}\}$ such that $\mathbf{w}_{jl}^{k*}$ is not an integer, then set $\mathbb{W}_{j_1}^k=\{\mathbf{w}_j^k\in\mathbb{W}^k:\mathbf{w}_{jl}^k\leq\lfloor\mathbf{w}_{jl}^{k*}\rfloor\}$ and $\mathbb{W}_{j_2}^k=\{\mathbf{w}_j^k\in\mathbb{W}^k:\mathbf{w}_{jl}^k\geq\lceil\mathbf{w}_{jl}^{k*}\rceil\}$
            Update $\mathbb{W}^k=\mathbb{W}_{j_1}^k\cup\mathbb{W}_{j_2}^k$, $k=k+1$
            Break
        else
            Return $(\mathbf{x}^{k*},\tilde{\mathbf{y}}^{k*})$ as a solution for $\mathbf{R1-BO}$ and terminate
        end if
    end for
end while
Return $(\mathbf{x}^{k*},\tilde{\mathbf{y}}^{k*})$ as a solution for $\mathbf{R1-BO}$ and terminate

Theorem 3.4. *Suppose* $(\mathbf{x}^{k*}, \tilde{\mathbf{y}}^{k*}, \eta^{k*}, \mathbf{y}^{k*})$ *is an optimal solution to* $\mathbf{R1}-\mathbf{BO}/\mathbf{R}(\mathbb{U},\mathbb{W},\mathbb{W}^k)$ *in the kth iteration, and* $\mathbf{y}^{k*} \in \phi_{\mathbb{W}}(\mathbf{x}^{k*})$*, i.e.,* $\mathbf{y}^{k*}$ *is an optimal solution to the lower level problem of* $\mathbf{R1}-\mathbf{BO}$*, then* $(\mathbf{x}^{k*}, \tilde{\mathbf{y}}^{k*})$ *is optimal to* $\mathbf{R1}-\mathbf{BO}$ *if*

$$\arg\max\{g_j(\mathbf{x}^{k*}, \mathbf{y}^{k*}, \mathbf{w}_j^k) : \mathbf{w}_j^k \in \mathbb{W}^k\} \cap \mathbb{W} \neq \emptyset, \forall j \in \mathbb{C}_f.$$

Proof. Without loss of generality, we assume $|\mathbb{C}_f| = 1$, then $\arg\max\{g_j(\mathbf{x}^{k*}, \mathbf{y}^{k*}, \mathbf{w}_j^k) : \mathbf{w}_j^k \in \mathbb{W}^k\} \cap \mathbb{W} \neq \emptyset$ for all $j \in \mathbb{C}_f$ reduces to $\arg\max\{g(\mathbf{x}^{k*}, \mathbf{y}^{k*}, \mathbf{w}^k) : \mathbf{w}^k \in \mathbb{W}^k\} \cap \mathbb{W} \neq \emptyset$. Let $\mathbf{w}^{k*} \in \arg\max\{g(\mathbf{x}^{k*}, \mathbf{y}^{k*}, \mathbf{w}^k) : \mathbf{w}^k \in \mathbb{W}^k\} \cap \mathbb{W}$, then we have

$$\max\{g(\mathbf{x}^{k*}, \mathbf{y}^{k*}, \mathbf{w}^k) : \mathbf{w}^k \in \mathbb{W}^k\} \leq g(\mathbf{x}^{k*}, \mathbf{y}^{k*}, \mathbf{w}^{k*}) \leq 0,$$

where the second inequality follows the fact that $\mathbf{w}^{k*} \in \mathbb{W}$ and $\mathbf{y}^{k*} \in \phi_{\mathbb{W}}(\mathbf{x}^{k*})$. Hence, $\mathbf{y}^{k*}$ is feasible and thus optimal to the lower level problem of $\mathbf{R1}-\mathbf{BO}/\mathbf{R}(\mathbb{U},\mathbb{W},\mathbb{W}^k)$ since the lower level problem of $\mathbf{R1}-\mathbf{BO}$ is a relaxation of that of $\mathbf{R1}-\mathbf{BO}/\mathbf{R}(\mathbb{U},\mathbb{W},\mathbb{W}^k)$. Thus, the lower level problem of $\mathbf{R1}-\mathbf{BO}/\mathbf{R}(\mathbb{U},\mathbb{W},\mathbb{W}^k)$ and that of $\mathbf{R1}-\mathbf{BO}$ have the same optimal value for $\mathbf{x}^{k*}$, which is $f(\mathbf{x}^{k*}, \mathbf{y}^{k*})$. This implies that $\tilde{\mathbf{y}}^{k*} \in \phi_{\mathbb{W}}(\mathbf{x}^{k*})$. Therefore, $(\mathbf{x}^{k*}, \tilde{\mathbf{y}}^{k*})$ is feasible and thus optimal to $\mathbf{R1}-\mathbf{BO}$. □

Corollary 3.5. *If* $UB - LB \geq \epsilon$*, then there exists* $\mathbf{w}_j^{k*} \in \mathbb{W}^k$ *such that* $g_j(\mathbf{x}^{k*}, \mathbf{y}^{k*}, \mathbf{w}_j^{k*}) > 0$ *for some* $j \in \mathbb{C}_f$*. In this case, we can branch on* $\mathbf{w}_j^{k*}$*. As* $|\mathbb{W}| < +\infty$*, Algorithm 2 converges in finite number of iterations.*

Proof. This result is directly implied by Theorem 3.4. □

Remark 3.5. *(i) If the lower level uncertainty set* $\mathbb{W}$ *is a mixed integer set, we mention that the Cut-and-Branch algorithm remains valid and still converges to an optimal solutions, noting that the branching operation is readily applicable with little change. Also, if the upper level uncertainty set* $\mathbb{U}$ *is a friendly convex set, the duality or KKT conditions based reduction method can be adopted within the cutting plane oracle for the related robust constraints (as showed in Corollary 3.3).*

(ii) In the aforementioned implementation, only one branching operation is performed every iteration on a single uncertainty set (and the associated constraint) in the lower level problem of $\mathbf{R1}-\mathbf{BO}/\mathbf{R}(\mathbb{U},\mathbb{W},\mathbb{W}^k)$*. A straightforward extension is to branch multiple uncertainty*

sets and their associated constraints per iteration. By doing that, the number of iterations certainly should also be reduced significantly, while the size of the lower level DMP also increases quickly. Hence, a study on the number of branching operations can be considered to achieve the desired trade-off with a better performance.

3.3 Robust Bilevel Model Under Uncertainties in Perception

In this subsection, we first consider single-stage robust bilevel optimization with uncertainty in perception. As discussed in introduction part, uncertainty in perception occurs when inaccurate or erroneous information is transmitted within this hierarchical system, or either the leader or the follower has less confidence regarding the other's DMP. We consider a relative easier case here where the leader proactively hedges an inexact feedback from the follower. We will investigate more sophisticated cases in the next chapter.

Specifically, let $\mathbb{U}_y$ be a compact uncertainty set, and $(\tilde{\mathbf{y}}+\mathbf{y}')$ with $\mathbf{y}' \in \mathbb{U}_y$ represent the perceived decisions in place of $\tilde{\mathbf{y}}$, respectively. The robust bilevel optimization is formulated as the next.

$$\mathbf{R1p-BO}: \quad \min_{\mathbf{x},\tilde{\mathbf{y}}} \max_{\mathbf{y}'\in\mathbb{U}_y} F(\mathbf{x},\tilde{\mathbf{y}}+\mathbf{y}') \tag{3.12}$$

$$\text{s.t.} \quad (1.2)-(1.3) \tag{3.13}$$

Since $\mathbf{y}'$ does not appear in any constraint, it will not affect the feasibility of $\mathbf{R1p-BO}$. Hence, it is clear that $\mathbf{R1p-BO}$ is feasible if and only if $\mathbf{BO}$ is feasible. Next, we give some sufficient conditions guaranteeing the existence of an optimal solution.

Theorem 3.5. *Assume that the deterministic counterpart* $\mathbf{BO}$ *satisfies the condition in Theorem 3.1, then* $\mathbf{R1p-BO}$ *has an optimal solution (i) if* $F'(\mathbf{x},\tilde{\mathbf{y}}) \equiv \max_{\mathbf{y}'\in\mathbb{U}_\mathbf{y}}\{F(\mathbf{x},\tilde{\mathbf{y}}+\mathbf{y}')\}$ *has a unique optimal solution for all* $(\mathbf{x},\tilde{\mathbf{y}}) \in \Omega_{\mathbf{BO}}$*. or (ii) if the lower level optimal value function* $v(\mathbf{x})$ *is continuous.*

Proof. (i) Since $\max_{\mathbf{y}'\in\mathbb{U}_\mathbf{y}}\{F(\mathbf{x},\tilde{\mathbf{y}}+\mathbf{y}')\}$ has a unique optimal solution for fixed $(\mathbf{x},\tilde{\mathbf{y}})$, its optimal solution can be denoted by $\mathbf{y}'(\mathbf{x},\tilde{\mathbf{y}})$ for fixed $(\mathbf{x},\mathbf{y})$, and thus $\mathbf{R1p-BO}$ can be rewritten as

$$\min_{\mathbf{x},\tilde{\mathbf{y}}}\{F(\mathbf{x},\tilde{\mathbf{y}},\mathbf{y}'(\mathbf{x},\tilde{\mathbf{y}})):\mathbf{x}\in\mathbb{X},G(\mathbf{x})\leq\mathbf{0},\tilde{\mathbf{y}}\in\phi(\mathbf{x})\}.$$

As F is continuously differentiable with respect to $\mathbf{x}$ and $\tilde{\mathbf{y}}$, by the *implicit function theorem* [44, 84], F' is also continuously differentiable. Noting that $\mathbf{R1p-BO}$ and $\mathbf{BO}$ have identical constraints, so $\mathbf{R1p-BO}$ also satisfies the conditions in Theorem 3.1, and thus has an optimal solution. (ii) It is easy to see that $\mathbf{R1p-BO}$ is equivalent to

$$\min_{\mathbf{x},\tilde{\mathbf{y}}}\{F'(\mathbf{x},\tilde{\mathbf{y}}):\mathbf{x}\in\mathbb{X},G(\mathbf{x})\leq\mathbf{0},\tilde{\mathbf{y}}\in\mathbb{Y},g(\mathbf{x},\tilde{\mathbf{y}})\leq\mathbf{0},f(\mathbf{x},\tilde{\mathbf{y}})\leq v(\mathbf{x})\}. \tag{3.14}$$

Since $v(\mathbf{x})$ is continuous, the feasible region of (3.14) is non-empty and compact. Moreover, as F is continuous, so is F'. Hence, $\mathbf{R1p-BO}$ has an optimal solution by the Weierstrass theorem. □

Remark 3.6. *Although $v(\mathbf{x})$ is not continuous in general, it has this property for several typical lower level problems such as LP, second-order cone programming (SCOP), and semidefinite programming (SDP), where $\mathbf{x}$ appears in the RHS of the lower level problem.*

3.4 Numerical Study

Facility location problem is a well-studied topic, and it has been widely applied in practice. In our numerical study, we consider a robust bilevel facility location problem. By convention, we use a regular lower case letter to represent a scalar and a bolded letter to represent the corresponding vector.

Sets

I	the index set of customers, indexed by i;
J	the index set of potential building locations, indexed by j;

Parameters

f_j	the fixed cost of using location j;
g_j	the unit capacity cost of facility j;
h_j	the operating cost of facility j;
K	budget;
U_j	the upper bound of the capacity of facility j;
c_{ij}	the cost incurred by customer i if going to lacation j;
d_i	the demand of customer i;

Variables

x_j	binary variables: $x_j = 1$ if location j is used for building a facility, and $x_j = 0$ otherwise;
w_j	the capacity of facility j;
y_{ij}	the fraction of demand of customer i provided by location j.

Using the aforementioned notations, a bilevel facility location model is defined as

$$\begin{aligned} \min_{\mathbf{x},\mathbf{w}} & \sum_{j\in J} f_j x_j + \sum_{j\in J} g_j w_j + \min_{\mathbf{y}\in S(\mathbf{x},\mathbf{w})} \sum_{j\in J} h_j \sum_{i\in I} y_{ij} \\ s.t. & \sum_{j\in J} x_j \le K \\ & w_j \le U_j x_j, \ \forall j \in J \\ & x_j \in \{0,1\}, w_j \ge 0, \forall j \in J \end{aligned} \tag{3.15}$$

where $S(\mathbf{x}, \mathbf{w})$ is the set of optimal solutions to the following problem for fixed $\mathbf{x}$ and $\mathbf{w}$.

$$\begin{aligned} \min_{\mathbf{y}} & \sum_{i\in I} \sum_{j\in J} c_{ij} y_{ij} \\ s.t. & \sum_{j\in J} y_{ij} \ge d_i, \ \forall i \in I \\ & \sum_{i\in I} y_{ij} \le w_j, \ \forall j \in J \\ & y_{ij} \ge 0, \ \forall i \in I, j \in J \end{aligned} \tag{3.16}$$

The upper level problem is to minimize the total cost subject to budget constraint and capacity bound constraints, and the lower level problem is to minimize transportation cost while meeting all the demands subject to capacity constraints.

In practice, the system may be subject to various types of uncertainties. For instance, some products may get damaged in transit. Such a scenario can be reflected by a stochastic coefficient matrix in (3.16). Suppose that the coefficient matrix of the constraint $\sum_{j\in J} y_{ij} \geq d_i$ is of the form $A = [\tilde{a}_{ij}]$ and that $\tilde{a}_{ij} \in [a_{ij} - \hat{a}_{ij}, a_{ij}], \forall i \in I, j \in J$, then we have the robust counterpart of (3.16) as

$$\begin{aligned}
\min_{\mathbf{y}} & \sum_{i\in I}\sum_{j\in J} c_{ij} y_{ij} \\
s.t. & \sum_{j\in J} a_{ij} y_{ij} - \max_{S_i\subseteq J, |S_i|\leq \Gamma_i} \{\sum_{j\in S_i} \hat{a}_{ij} y_{ij}\} \geq d_i, \ \forall i \in I \\
& \sum_{i\in I} y_{ij} \leq w_j, \ \forall j \in J \\
& y_{ij} \geq 0, \ \forall i \in I, j \in J
\end{aligned} \tag{3.17}$$

where Γ_i is the maximum number of the uncertain parameters in the ith constraint. We replace (3.16) with (3.17) and refer the resulting problem as robust bilevel facility location problem (RBFLP).

We first solve RBFLP instances approximately by applying Theorem 3.2 and Corollary 3.3, and then solve them by the proposed cut-and-branch algorithm. From Table 7, we see that the total cost increases in Γ. Table 8 reports the performance of the proposed cut-and-branch algorithm. First, we observe that the relaxation is strong as the average gap is only 1.42%. Also, we see that the cut-and-branch algorithm is efficient in solving problems of small to moderate size. All the instances are solved in just few minutes, and only very small number of branching operations were performed.

Table 7: Result of RBFL

I	J	Γ	0	0.3	0.6	0.9	1.2
5	5	Optimal value (App)	3659.41	3775.25	3891.09	4006.93	4813.29
		Optimal value (C&B)	3659.41	3756.90	3720.58	3754.19	4813.29
	10	Optimal value (App)	2218.12	2291.29	2956.18	3111.06	3172.97
		Optimal value (C&B)	2218.12	2291.29	2322.49	2335.58	3172.97

Table 8: Performance of the Cut-and-Branch Algorithm

I	J	Γ	0	0.3	0.6	0.9	1.2	Avg.
5	5	Optimality gap (App)	0.00%	1.16%	2.28%	3.36%	0.31%	1.42%
		Optimality gap (C&B)	0.00%	0.98%	0.62%	0.96%	0.31%	0.57%
		Gap closed via branching	0.00%	0.18%	1.66%	2.41%	0.00%	0.85%
		Time (C&B)	0.04	0.2	1.03	0.58	1.15	0.60
		Number of branching performed	0	2	5	5	0	2.4
	10	Optimality gap (App)	0.00%	0.00%	3.33%	5.18%	0.12%	1.73%
		Optimality gap (C&B)	0.00%	0.00%	0.00%	0.00%	0.12%	0.02%
		Gap closed via branching	0.00%	0.00%	3.33%	5.18%	0.00%	1.70%
		Time (C&B)	0.06	0.06	0.44	1.62	0.33	0.50
		Number of branching performed	0	0	2	4	0	1.2

3.5 Conclusion

In this chapter, we study single-stage robust bilevel optimization problems. We develop a general single-stage robust bilevel optimization model that considers both upper level and lower level uncertainty. We further provide a novel relaxation of this model, based on which a novel cut-and-branch is developed for single-stage RBO problem with discrete uncertainty set. The numerical study shows the effectiveness and the efficiency of our proposed model and solution method in handling with uncertainty.

4.0 Two-Stage Robust Bilevel Optimization

In the previous chapter, we assume that all the decisions are made before the uncertainty is revealed, and thus no scenario-specific decision is involved. In practice, the leader may have an opportunity to make some scenario-specific "wait and see" decisions. Such decision making processes are often described by a two-stage optimization model.

In this chapter, we first introduce robust bilevel optimization formulations with scenario-specific decisions. Then we study their structural properties and develop algorithms to solve this type of problem.

4.1 Bilevel Optimization With Scenario-Specific Decisions Under Exogenous Uncertainty

We first consider two-stage bilevel optimization model with exogenous uncertainty. As discussed in the introduction section, exogenous uncertainty refers to random factors outside of the system. Such factors are often assumed to be in an uncertainty set that is independent of the decisions made by the two DMs in a bilevel optimization problem.

4.1.1 Two-Stage Robust Bilevel Optimization Formulations and Properties

Denote the first and the second stage decision of the leader by $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. Also denote the uncertain parameter by $\mathbf{u}$. Then, the follower makes a decision based on given upper level decisions and a realized scenario. Hence, for fixed $(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$ the lower level feasible region is

$$\psi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) = \{\mathbf{y} : g(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) \leq \mathbf{0}, \mathbf{y} \in \mathbb{Y}\},$$

and the lower level rational reaction set is

$$\phi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) = \arg\min_{\mathbf{y}}\{f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) : \mathbf{y} \in \psi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\}.$$

Then, the aforementioned decision making process is modeled by a two-stage robust bilevel optimization (TSROBLO) formulation

$$\textbf{TSROBLO}: \min_{\{\mathbf{x}_1\in\mathbb{X}_1, G(\mathbf{x}_1)\leq\mathbf{0}\}} \max_{\{\mathbf{u}\in\mathbb{U}\}} \min_{\{\mathbf{x}_2\in\mathbb{X}_2, H(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2)\leq\mathbf{0}, \tilde{\mathbf{y}}\in\phi(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2)\}} F(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2,\tilde{\mathbf{y}}).$$

In addition to the assumptions we made in the previous chapters, we further assume that H is continuous, and that $\mathbb{X}_2 \subseteq \mathbb{R}_+^{m_{2c}} \times \mathbb{Z}_+^{m_{2d}}$ is a non-empty compact set. We refer $\{\mathbf{x}_2 \in \mathbb{X}_2, H(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2) \leq \mathbf{0}\}$ as the second stage feasible region.

Since both the second stage and the lower level feasible region are jointly determined by the first stage decision and the random factors, the leader needs to guarantee the second stage and the lower level feasibility while making the first stage decision. Denote $\mathbb{S}(\mathbf{x}_1,\mathbf{u}) = \{(\mathbf{x}_2,\mathbf{y}) : H(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2) \leq \mathbf{0}, g(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2,\mathbf{y}) \leq \mathbf{0}, \mathbf{x}_2 \in \mathbb{X}_2, \mathbf{y} \in \mathbb{Y}\}$, then $\mathbb{S}(\mathbf{x}_1,\mathbf{u})$ is compact for fixed $(\mathbf{x}_1,\mathbf{u})$. For a fixed $\mathbf{x}_1^0 \in \{\mathbf{x}_1 \in \mathbb{X}_1, G(\mathbf{x}_1) \leq \mathbf{0}\}$, either $\mathbb{S}(\mathbf{x}_1^0,\mathbf{u}) \neq \emptyset, \forall \mathbf{u} \in \mathbb{U}$ or $\exists \mathbf{u}^0 \in \mathbb{U}$ such that $\mathbb{S}(\mathbf{x}_1^0,\mathbf{u}^0) = \emptyset$. Hence, to make the second stage and the lower level problem feasible for all $\mathbf{u} \in \mathbb{U}$, $\mathbf{x}_1$ has to be in

$$\mathbb{X}_{\mathbb{U}} = \{\mathbf{x}_1 : \mathbf{x}_1 \in \mathbb{X}_1, G(\mathbf{x}_1) \leq \mathbf{0}, \mathbb{S}(\mathbf{x}_1,\mathbf{u}) \neq \emptyset, \forall \mathbf{u} \in \mathbb{U}\}.$$

Theorem 4.1. *Let $\mathbf{e}_1$ and $\mathbf{e}_2$ be a vector with all elements being one, and let $z^*(\mathbf{x}_1)$ be the optimal value of the following problem for fixed $\mathbf{x}_1$*

$$\max_{\mathbf{u}\in\mathbb{U}}\{\min_{\mathbf{x}_2,\mathbf{y},\mathbf{t}_1,\mathbf{t}_2\geq\mathbf{0}} \mathbf{e}_1^T\mathbf{t}_1 + \mathbf{e}_2^T\mathbf{t}_2 : H(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2) \leq \mathbf{t}_1, g(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2,\mathbf{y}) \leq \mathbf{t}_2, \mathbf{x}_2 \in \mathbb{X}_2, \mathbf{y} \in \mathbb{Y}\}, \quad (4.1)$$

then $\mathbf{x}_1 \in \mathbb{X}_{\mathbb{U}}$ if and only if $\mathbf{x}_1 \in \mathbb{X}_1$, $G(\mathbf{x}_1) \leq \mathbf{0}$ and $z^*(\mathbf{x}_1) = 0$.

Proof. For fixed $(\mathbf{x}_1,\mathbf{u})$, let $w^*(\mathbf{x}_1,\mathbf{u})$ be the optimal value of the inner problem of (4.1)

$$w^*(\mathbf{x}_1,\mathbf{u}) = \min_{\mathbf{x}_2,\mathbf{y},\mathbf{t}_1,\mathbf{t}_2\geq\mathbf{0}}\{\mathbf{e}_1^T\mathbf{t}_1 + \mathbf{e}_2^T\mathbf{t}_2 : H(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2) \leq \mathbf{t}_1, g(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2,\mathbf{y}) \leq \mathbf{t}_2, \mathbf{x}_2 \in \mathbb{X}_2, \mathbf{y} \in \mathbb{Y}\}, \quad (4.2)$$

then it is obvious that $w^*(\mathbf{x}_1,\mathbf{u}) \geq 0$. If $\mathbb{S}(\mathbf{x}_1,\mathbf{u}) \neq \emptyset$, then there exists $(\mathbf{x}_2,\mathbf{y}) \in \mathbb{X}_2 \times \mathbb{Y}$ such that $H(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2) \leq \mathbf{0}$ and $g(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2,\mathbf{y}) \leq \mathbf{0}$. Let $\mathbf{t}_1 = \mathbf{t}_2 = \mathbf{0}$, then $(\mathbf{x}_2,\mathbf{y},\mathbf{t}_1,\mathbf{t}_2)$ is an optimal solution to (4.2), and $w^*(\mathbf{x}_1,\mathbf{u}) = \mathbf{e}_1^T\mathbf{t}_1 + \mathbf{e}_2^T\mathbf{t}_2 = 0$. Conversely, if $(\mathbf{x}_2^*,\mathbf{y}^*,\mathbf{t}_1^*,\mathbf{t}_2^*)$ is an optimal solution to (4.2), then we have $w^*(\mathbf{x}_1,\mathbf{u}) = \mathbf{e}_1^T\mathbf{t}_1^* + \mathbf{e}_2^T\mathbf{t}_2^* = 0$, which implies

$\mathbf{t}_1^* = \mathbf{t}_2^* = \mathbf{0}$. Hence, we have $\mathbf{x}_2^* \in \mathbb{X}_2$, $\mathbf{y}^* \in \mathbb{Y}$, $H(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2^*) \leq \mathbf{0}$ and $g(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2^*, \mathbf{y}^*) \leq \mathbf{0}$. As $(\mathbf{x}_2^*, \mathbf{y}^*) \in \mathbb{S}(\mathbf{x}_1, \mathbf{u})$, we have $\mathbb{S}(\mathbf{x}_1, \mathbf{u}) \neq \emptyset$.

If $\mathbf{x}_1 \in \mathbb{X}_{\mathbb{U}}$, then for any $\mathbf{u} \in \mathbb{U}$, $\mathbb{S}(\mathbf{x}_1, \mathbf{u}) \neq \emptyset$. This implies that $w^*(\mathbf{x}_1, \mathbf{u}) = 0$ for any $\mathbf{u} \in \mathbb{U}$, and thus $z^*(\mathbf{x}_1) = 0$. Conversely, if $z^*(\mathbf{x}_1) = 0$ but $\mathbf{x}_1 \notin \mathbb{X}_{\mathbb{U}}$, then there exists $\mathbf{u}^* \in \mathbb{U}$ such that $\mathbb{S}(\mathbf{x}_1, \mathbf{u}^*) = \emptyset$. Hence, $w^*(\mathbf{x}_1, \mathbf{u}^*) > 0 = z^*(\mathbf{x}_1)$, which contradicts the fact that the optimal value of (4.1) is 0. Therefore, $z^*(\mathbf{x}_1) = 0$ implies $\mathbf{x}_1 \in \mathbb{X}_{\mathbb{U}}$. □

Remark 4.1. *We say* **TSROBLO** *has the relatively complete recourse property [133] if* $\mathbb{S}(\mathbf{x}_1, \mathbf{u}) \neq \emptyset$ *for any* $(\mathbf{x}_1, \mathbf{u})$*. In this case,* $\mathbb{X}_{\mathbb{U}}$ *simply reduces to the upper level first stage feasible region* $\{\mathbf{x}_1 : \mathbf{x}_1 \in \mathbb{X}_1, G(\mathbf{x}_1) \leq \mathbf{0}\}$.

For fixed first stage decision $\mathbf{x}_1$ and uncertain parameter $\mathbf{u}$, **TSROBLO** reduces to

$$\min\{F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \tilde{\mathbf{y}}) : \mathbf{x}_2 \in \mathbb{X}_2, H(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq \mathbf{0}, \tilde{\mathbf{y}} \in \phi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\}. \tag{4.3}$$

Notice that (4.3) is solved based on a particular realized scenario, and thus $\mathbf{x}_2$ and $\mathbf{y}$ are scenario-specific decisions. Denoting the index set of $\mathbb{U}$ by $\mathbb{K}_{\mathbb{U}} \equiv \{k : \mathbf{u}_k \in \mathbb{U}\}$ and introducing $(\mathbf{x}_{2_k}, \tilde{\mathbf{y}}_k)$ for each k, then we can rewrite **TSROBLO** as

$$\begin{aligned}
& \min \eta \\
\text{s.t.} \quad & G(\mathbf{x}_1) \leq \mathbf{0}, \mathbf{x}_1 \in \mathbb{X}_1 && (4.4)\\
& H(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}) \leq \mathbf{0}, \mathbf{x}_{2_k} \in \mathbb{X}_2, \forall k \in \mathbb{K}_{\mathbb{U}} && (4.5)\\
& \eta \geq F(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}, \tilde{\mathbf{y}}_k), \forall k \in \mathbb{K}_{\mathbb{U}} && (4.6)\\
& \tilde{\mathbf{y}}_k \in \phi(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}), \forall k \in \mathbb{K}_{\mathbb{U}}, && (4.7)
\end{aligned}$$

Indeed, we can have a relaxation of **TSROBLO** by considering a subset of the scenarios in the uncertainty set, i.e., by replacing $\mathbb{U}$ with $\bar{\mathbb{U}}$. Such a relaxation provides a lower bound. By generating scenarios and enlarging $\bar{\mathbb{U}}$, we can tighten the bound to achieve optimality.

4.1.2 Decomposition Algorithm

As mentioned previously, for $\bar{\mathbb{U}} \subseteq \mathbb{U}$, we have a relaxation of **TSROBLO** as

$$\mathbf{MP}_1 : \min\{\eta : (4.4), (4.5) - (4.7), \forall k \in \mathbb{K}_{\bar{\mathbb{U}}}\},$$

whose optimal value provides a lower bound.

An optimal $\mathbf{x}_1^*$ obtained from $\mathbf{MP}_1$ may not be in $\mathbb{X}_{\mathbb{U}}$. In this case, a feasibility cut is generated by the following "$\max - \min$" problem, which we refer as a feasibility cut generation problem (FCGP)

$$\begin{aligned}\mathbf{FCGP} : z^*(\mathbf{x}_1) = \quad & \max_{\mathbf{u}\in\mathbb{U}}\{ \min_{\mathbf{x}_2,\mathbf{y},\mathbf{t}_1,\mathbf{t}_2\geq\mathbf{0}} \mathbf{e}_1^T\mathbf{t}_1 + \mathbf{e}_2^T\mathbf{t}_2 : H(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2) \leq \mathbf{t}_1, \\ & g(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2,\mathbf{y}) \leq \mathbf{t}_2, \mathbf{x}_2 \in \mathbb{X}_2, \mathbf{y} \in \mathbb{Y}\},\end{aligned}$$

where $\mathbf{e}_1$ and $\mathbf{e}_2$ are both a vector with all elements being one. According to Theorem 4.1, if $\mathbf{u}^*$ is an optimal solution to **FCGP** with $z^*(\mathbf{x}_1^*) > 0$, then $\mathbb{S}(\mathbf{x}_1^*, \mathbf{u}^*) = \emptyset$, and thus we can add $H(\mathbf{x}_1, \mathbf{u}^*, \mathbf{x}_2) \leq \mathbf{0}$ and $g(\mathbf{x}_1, \mathbf{u}^*, \mathbf{x}_2, \mathbf{y}) \leq \mathbf{0}$ to $\mathbf{MP}_1$. If $z^*(\mathbf{x}_1^*) = 0$, then $\mathbf{x}_1^* \in \mathbb{X}_{\mathbb{U}}$, and we solve the following sub-problem (SP)

$$\mathbf{SP}_1 : \max_{\{\mathbf{u}\in\mathbb{U}\}} \min_{\{\mathbf{x}_2\in\mathbb{X}_2, H(\mathbf{x}_1^*,\mathbf{u},\mathbf{x}_2)\leq\mathbf{0}, \tilde{\mathbf{y}}\in\phi(\mathbf{x}_1^*,\mathbf{u},\mathbf{x}_2)\}} F(\mathbf{x}_1^*, \mathbf{u}, \mathbf{x}_2, \tilde{\mathbf{y}}).$$

If $(\mathbf{u}^*, \mathbf{x}_2^*, \mathbf{y}^*)$ is an optimal solution to $\mathbf{SP}_1$, then $F(\mathbf{x}_1^*, \mathbf{u}^*, \mathbf{x}_2^*, \mathbf{y}^*)$ is an upper bound of **TSROBLO**. If the gap between the upper bound and the lower bound is not within our predetermined optimality tolerance, we set $\bar{\mathbb{U}} = \bar{\mathbb{U}} \cup \{\mathbf{u}^*\}$ and go back to solve $\mathbf{MP}_1$; otherwise we can terminate the algorithm. Denote the optimal value function of the lower level problem by $v(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$ for fixed $(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$, we can rewrite $\phi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$ as $\{\mathbf{y} : g(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) \leq \mathbf{0}, \mathbf{y} \in \mathbb{Y}, f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) \leq v(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\}$.

Let LB and UB be the lower bound and upper bound respectively, ϵ be the optimality tolerance, and k be the iteration index, we propose a decomposition algorithm for **TSROBLO**.

Theorem 4.2. *Suppose v is continuous, $\mathbb{X}_1$ and $\mathbb{U}$ are compact, then Algorithm 3 either terminates in finite number of iterations returning an ϵ-optimal solution, or generates a sequence $\{\mathbf{x}_1^i\}_{i\in\mathbb{N}}$ containing a limit point $\mathbf{x}_1^*$ that is an ϵ-optimal solution to* **TSROBLO**.

Algorithm 3 Decomposition algorithm for **TSROBLO**

```
Initialize LB = −∞, UB = +∞, Ū = ∅, k = 0
while UB − LB > ϵ do
    Solve MP_1, obtain an optimal solution x_1^*
    Update LB = η^*
    Solve FCGP and obtain an optimal solution u^*
    if z^*(x_1^*) > 0 then
        Add H(x_1, u^*, x_2) ≤ 0 and g(x_1, u^*, x_2, y) ≤ 0 to MP_1, k = k + 1
    else
        Solve SP_1 for x_1^*, obtain an optimal solution (u^*, x_2^*, ỹ^*)
        Update UB = min{UB, F(x_1^*, u^*, x_2^*, ỹ^*)}
    end if
    Set Ū = Ū ∪ {u^*}, add corresponding variables and constraints to MP_1, k = k + 1
end while
Return x_1^* as an optimal solution and terminate
```

Proof. We first assume that **TSROBLO** has the relatively complete recourse property and extend the result to general case later. Let z^* be the optimal value of **TSROBLO**, if Algorithm 3 terminates, then we have $UB - z^* \leq UB - LB \leq \epsilon$, and the best feasible solution found is ϵ-optimal. If Algorithm 3 does not terminate, by the Bolzano-Weierstrass theorem and by taking necessary sub-sequence, we can assume that the sequence $\{(\mathbf{x}_1^i, \eta^i, \mathbf{u}^i)\}_{i \in \mathbb{N}}$ generated by Algorithm 3 converges to a limit point $(\mathbf{x}_1^*, \eta^*, \mathbf{u}^*)$. By the relatively complete recourse property, we have $\mathbb{S}(\mathbf{x}_1^*, \mathbf{u}) \neq \emptyset, \forall \mathbf{u} \in \mathbb{U}$. Let $\mathbb{Q}(\mathbf{x}_1, \mathbf{u}) = \{(\mathbf{x}_2, \mathbf{y}) : H(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq \mathbf{0}, \mathbf{x}_2 \in \mathbb{X}_2, \mathbf{y} \in \phi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\}$, then $\mathbb{Q}(\mathbf{x}_1^*, \mathbf{u}) \neq \emptyset, \forall \mathbf{u} \in \mathbb{U}$, and there are two possible cases.

1) For any $\mathbf{u}^0 \in \mathbb{U}$, there exists $(\mathbf{x}_2^0, \mathbf{y}^0) \in \mathbb{Q}(\mathbf{x}_1^*, \mathbf{u}^0)$ such that $\eta^* \geq F(\mathbf{x}_1^*, \mathbf{u}^0, \mathbf{x}_2^0, \mathbf{y}^0)$. Then we have $LB = \eta^* \geq \max_{\mathbf{u} \in \mathbb{U}} \min_{(\mathbf{x}_2, \mathbf{y}) \in \mathbb{Q}(\mathbf{x}_1^*, \mathbf{u})} F(\mathbf{x}_1^*, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) \geq UB$. As $LB = UB$, $\mathbf{x}_1^*$ is optimal to **TSROBLO**.

2) There exists $\mathbf{u}^0 \in \mathbb{U}$ such that $\eta^* < F(\mathbf{x}_1^*, \mathbf{u}^0, \mathbf{x}_2, \mathbf{y})$, $\forall (\mathbf{x}_2, \mathbf{y}) \in \mathbb{Q}(\mathbf{x}_1^*, \mathbf{u}^0)$. Hence, we

have

$$\begin{aligned}\eta^* &< \min_{(\mathbf{x}_2,\mathbf{y})\in\mathbb{Q}(\mathbf{x}_1^*,\mathbf{u}^0)} F(\mathbf{x}_1^*,\mathbf{u}^0,\mathbf{x}_2,\mathbf{y}) \\ &\leq \max_{\mathbf{u}\in\mathbb{U}} \min_{(\mathbf{x}_2,\mathbf{y})\in\mathbb{Q}(\mathbf{x}_1^*,\mathbf{u})} F(\mathbf{x}_1^*,\mathbf{u},\mathbf{x}_2,\mathbf{y}) \\ &= \min_{(\mathbf{x}_2,\mathbf{y})\in\mathbb{Q}(\mathbf{x}_1^*,\mathbf{u}^*)} F(\mathbf{x}_1^*,\mathbf{u}^*,\mathbf{x}_2,\mathbf{y}),\end{aligned} \tag{4.8}$$

where the equality follows the fact that $\mathbf{u}^*$ is an optimal solution to $\mathbf{SP}_1$ for $\mathbf{x}_1^*$. By the continuity of F, f, and v, and the compactness of $\mathbb{S}$, for sufficiently large i, we have

$$\begin{aligned}\eta^i &< \min_{\mathbf{x}_2,\mathbf{y}}\{F(\mathbf{x}_1^i,\mathbf{u}^0,\mathbf{x}_2,\mathbf{y}) : (\mathbf{x}_2,\mathbf{y})\in\mathbb{S}(\mathbf{x}_1^i,\mathbf{u}^0), f(\mathbf{x}_1^i,\mathbf{u}^0,\mathbf{x}_2,\mathbf{y})\leq v(\mathbf{x}_1^i,\mathbf{u}^0,\mathbf{x}_2)\} \\ &\leq \max_{\mathbf{u}\in\mathbb{U}}\min_{\mathbf{x}_2,\mathbf{y}}\{F(\mathbf{x}_1^i,\mathbf{u},\mathbf{x}_2,\mathbf{y}) : (\mathbf{x}_2,\mathbf{y})\in\mathbb{S}(\mathbf{x}_1^i,\mathbf{u}), f(\mathbf{x}_1^i,\mathbf{u},\mathbf{x}_2,\mathbf{y})\leq v(\mathbf{x}_1^i,\mathbf{u},\mathbf{x}_2)\} \\ &= \min_{\mathbf{x}_2,\mathbf{y}}\{F(\mathbf{x}_1^i,\mathbf{u}^i,\mathbf{x}_2,\mathbf{y}) : (\mathbf{x}_2,\mathbf{y})\in\mathbb{S}(\mathbf{x}_1^i,\mathbf{u}^i), f(\mathbf{x}_1^i,\mathbf{u}^i,\mathbf{x}_2,\mathbf{y})\leq v(\mathbf{x}_1^i,\mathbf{u}^i,\mathbf{x}_2)\} \\ &\geq UB.\end{aligned}$$

If $UB - LB \leq \epsilon$, the algorithm terminates; otherwise the following constraints

$$\eta \geq F(\mathbf{x}_1,\mathbf{u}^i,\mathbf{x}_2^i,\mathbf{y}^i), (\mathbf{x}_2^i,\mathbf{y}^i)\in\mathbb{S}(\mathbf{x}_1,\mathbf{u}^i), f(\mathbf{x}_1,\mathbf{u}^i,\mathbf{x}_2^i,\mathbf{y}^i)\leq v(\mathbf{x}_1,\mathbf{u}^i,\mathbf{x}_2^i)$$

are added to $\mathbf{MP}_1$. Let $(\mathbf{x}_1^{i+1},\eta^{i+1})$ be an optimal solution to $\mathbf{SP}_1$ in the next iteration, we have

$$\eta^{i+1} \geq \min_{\mathbf{x}_2,\mathbf{y}}\{F(\mathbf{x}_1^{i+1},\mathbf{u}^i,\mathbf{x}_2,\mathbf{y}) : (\mathbf{x}_2,\mathbf{y})\in\mathbb{S}(\mathbf{x}_1^{i+1},\mathbf{u}^i), f(\mathbf{x}_1^{i+1},\mathbf{u}^i,\mathbf{x}_2,\mathbf{y})\leq v(\mathbf{x}_1^{i+1},\mathbf{u}^i,\mathbf{x}_2)\}.$$

By the continuity of F, f, and v, and the compactness of $\mathbb{S}$, as $i\to+\infty$, we have

$$\begin{aligned}\eta^* &\geq \min_{\mathbf{x}_2,\mathbf{y}}\{F(\mathbf{x}_1^*,\mathbf{u}^*,\mathbf{x}_2,\mathbf{y}) : (\mathbf{x}_2,\mathbf{y})\in\mathbb{S}(\mathbf{x}_1^*,\mathbf{u}^*), f(\mathbf{x}_1^*,\mathbf{u}^*,\mathbf{x}_2,\mathbf{y})\leq v(\mathbf{x}_1^*,\mathbf{u}^*,\mathbf{x}_2)\} \\ &= \min_{(\mathbf{x}_2,\mathbf{y})\in\mathbb{Q}(\mathbf{x}_1^*,\mathbf{u}^*)} F(\mathbf{x}_1^*,\mathbf{u}^*,\mathbf{x}_2,\mathbf{y}),\end{aligned}$$

which leads to a contradiction with (4.8), and thus completes the proof.

For a **TSROBLO** instance without the relatively complete recourse property, the feasibility cuts generated by **FCGP** lead to an ϵ-feasible solution [73, 82]. □

Remark 4.2. *As mentioned in Remark 3.6, v is continuous for several typical problems that are often used in practice. Furthermore, if the lower level problem is an LP, and the uncertainty set $\mathbb{U}$ is a nonempty polytope, then there always exists an $\mathbf{u}$ that is optimal to* **FCGP** *and is an extreme point of $\mathbb{U}$. In this case, the number of iterations to generate an $\mathbf{x}_1$ in $\mathbb{X}_{\mathbb{U}}$ is bounded by the number of extreme points of $\mathbb{U}$, which is finite.*

4.2 Bilevel Optimization With Scenario-Specific Decisions Under Endogenous Uncertainty

In this section, we consider endogenous uncertainty, i.e., the follower may not cooperate with the leader. To hedge against such uncertainty, we have the following formulation

$$\textbf{TSRPBLO}: \min_{\{\mathbf{x}_1\in\mathbb{X}_1, G(\mathbf{x}_1)\leq\mathbf{0}\}} \max_{\{\mathbf{u}\in\mathbb{U}\}} \min_{\{\mathbf{x}_2\in\mathbb{X}_2, H(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2)\leq\mathbf{0}\}} \max_{\{\tilde{\mathbf{y}}\in\phi(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2)\}} F(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2,\tilde{\mathbf{y}}),$$

which is the counterpart of **TSROBLO**. Since the second stage and the lower level feasible region of **TSRPBLO** are identical to those of **TSROBLO**, Theorem 4.1 and Remark 4.1 can be directly applied to **TSRPBLO**. Same as for **TSROBLO**, we can have a relaxation of **TSRPBLO** by only considering a subset of the uncertainty set.

We can also obtain a lower bound by replacing $\phi(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2)$ with $\psi(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2)$, i.e., by ignoring the lower level objective function. The resulting problem is

$$\textbf{TSRBLO}_\mathbf{L}: \min_{\{\mathbf{x}_1\in\mathbb{X}_1, G(\mathbf{x}_1)\leq\mathbf{0}\}} \max_{\{\mathbf{u}\in\mathbb{U}\}} \min_{\{\mathbf{x}_2\in\mathbb{X}_2, H(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2)\leq\mathbf{0}, \mathbf{y}\in\psi(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2)\}} F(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2,\mathbf{y}).$$

Furthermore, the following problem

$$\textbf{TSRBLO}_\mathbf{U}: \min_{\{\mathbf{x}_1\in\mathbb{X}_1, G(\mathbf{x}_1)\leq\mathbf{0}\}} \max_{\{\mathbf{u}\in\mathbb{U}\}} \min_{\{\mathbf{x}_2\in\mathbb{X}_2, H(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2)\leq\mathbf{0}\}} \max_{\{\mathbf{y}\in\psi(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2)\}} F(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2,\mathbf{y}),$$

gives an over pessimistic solution and provides an upper bound for **TSROBLO** and **TSRPBLO**.

Theorem 4.3. *Let z_1^*, z_2^*, z_3^*, z_4^* be the optimal value of* $\textbf{TSRBLO}_\mathbf{L}$*,* **TSROBLO***,* **TSRPBLO***, and* $\textbf{TSRBLO}_\mathbf{U}$*, respectively, then*

1) $z_1^* \leq z_2^* \leq z_3^* \leq z_4^*$;

2) If there exist a non-negative number α such that $f(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2,\mathbf{y}) = \alpha F(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2,\mathbf{y})$, then $z_1^* = z_2^*$ and $z_3^* = z_4^*$;

3) If $\phi(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2)$ is a singleton for any $(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2)$ such that $\phi(\mathbf{x}_1,\mathbf{u},\mathbf{x}_2)\neq\emptyset$, then $z_2^* = z_3^*$.

Proof. We provide the proof in the appendix. □

In practice, uncertainty may occur after all decisions are made by the leader. In this case, there is no scenario-specific decisions for the leader, then **TSROBLO** reduces to

$$\mathbf{TSROBLO_2}: \min_{\{\mathbf{x}\in\mathbb{X},G(\mathbf{x})\leq\mathbf{0}\}} \max_{\{\mathbf{u}\in\mathbb{U}\}} \min_{\{\tilde{\mathbf{y}}\in\phi(\mathbf{x},\mathbf{u})\}} F(\mathbf{x},\mathbf{u},\tilde{\mathbf{y}}),$$

and **TSRPBLO** reduces to

$$\mathbf{TSRPBLO_2}: \min_{\{\mathbf{x}\in\mathbb{X},G(\mathbf{x})\leq\mathbf{0}\}} \max_{\{\mathbf{u}\in\mathbb{U},\tilde{\mathbf{y}}\in\phi(\mathbf{x},\mathbf{u})\}} F(\mathbf{x},\mathbf{u},\tilde{\mathbf{y}}),$$

where $\phi(\mathbf{x},\mathbf{u}) = \arg\min_{\mathbf{y}}\{f(\mathbf{x},\mathbf{u},\mathbf{y}) : g(\mathbf{x},\mathbf{u},\mathbf{y}) \leq \mathbf{0}, \mathbf{y} \in \mathbb{Y}\}$. Similarly, we have

$$\mathbf{TSROBLO_{2L}} \min_{\{\mathbf{x}\in\mathbb{X},G(\mathbf{x})\leq\mathbf{0}\}} \max_{\{\mathbf{u}\in\mathbb{U}\}} \min_{\{\mathbf{y}\in\psi(\mathbf{x},\mathbf{u})\}} F(\mathbf{x},\mathbf{u},\mathbf{y}),$$

and

$$\mathbf{TSROBLO_{2U}}: \min_{\{\mathbf{x}\in\mathbb{X},G(\mathbf{x})\leq\mathbf{0}\}} \max_{\{\mathbf{u}\in\mathbb{U},\mathbf{y}\in\psi(\mathbf{x},\mathbf{u})\}} F(\mathbf{x},\mathbf{u},\mathbf{y}),$$

where $\psi(\mathbf{x},\mathbf{u}) = \{\mathbf{y} : g(\mathbf{x},\mathbf{u},\mathbf{y}) \leq \mathbf{0}, \mathbf{y} \in \mathbb{Y}\}$. As $\mathbf{TSROBLO_{2U}}$ is a deterministic BO problem, it is relatively easy to compute, and its optimal value provides an upper bound for $\mathbf{TSROBLO_2}$ and $\mathbf{TSRPBLO_2}$. On the contrary, $\mathbf{TSROBLO_{2L}}$ is a two-stage robust optimization problem, which is still not easy to solve. The following result shows that we can obtain a lower bound of $\mathbf{TSRPBLO_2}$ by solving a standard BO problem, and that the bound is tight under some easily verifiable conditions.

Theorem 4.4. *1) Let* $\zeta(\mathbf{x},\mathbf{u},\bar{\mathbf{y}}) = \{\mathbf{y} : g(\mathbf{x},\mathbf{u},\mathbf{y}) \leq \mathbf{0}, f(\mathbf{x},\mathbf{u},\mathbf{y}) \leq f(\mathbf{x},\mathbf{u},\bar{\mathbf{y}}), \mathbf{y} \in \mathbb{Y}\}$*, then the BO problem*

$$\mathbf{TSROBLO_{2L}^R}: \min_{\{\mathbf{x}\in\mathbb{X},G(\mathbf{x})\leq\mathbf{0},g(\mathbf{x},\mathbf{u},\bar{\mathbf{y}})\leq\mathbf{0},\bar{\mathbf{y}}\in\mathbb{Y}\}} \max_{\{\mathbf{u}\in\mathbb{U},\mathbf{y}\in\zeta(\mathbf{x},\mathbf{u},\bar{\mathbf{y}})\}} F(\mathbf{x},\mathbf{u},\mathbf{y})$$

is a relaxation of $\mathbf{TSRPBLO_2}$.

2) If $(\mathbf{x}^*, \bar{\mathbf{y}}_R^*, \mathbf{u}_R^*, \mathbf{y}_R^*)$ *is an optimal solution to* $\mathbf{TSROBLO_{2L}^R}$, $(\mathbf{u}^*, \mathbf{y}^*)$ *is an optimal solution to* $\max_{\mathbf{u},\mathbf{y}}\{F(\mathbf{x}^*,\mathbf{u},\mathbf{y}) : \mathbf{u} \in \mathbb{U}, \mathbf{y} \in \phi(\mathbf{x}^*,\mathbf{u})\}$ *for fixed* $\mathbf{x}^*$*, and* $g(\mathbf{x}^*,\mathbf{u}^*,\bar{\mathbf{y}}_R^*) \leq \mathbf{0}$*, then* $(\mathbf{x}^*,\mathbf{u}^*,\mathbf{y}^*)$ *is an optimal solution to* $\mathbf{TSRPBLO_2}$.

Proof. 1) It is sufficient to show that a feasible solution to $\mathbf{TSRPBLO_2}$ is also feasible to $\mathbf{TSROBLO^R_{2L}}$. Suppose $(\mathbf{x}^0, \mathbf{u}^0, \mathbf{y}^0)$ is a feasible solution to $\mathbf{TSRPBLO_2}$, then we have $\mathbf{x}^0 \in \mathbb{X}, G(\mathbf{x}^0) \leq \mathbf{0}$, $(\mathbf{u}^0, \mathbf{y}^0) \in \arg\max_{\mathbf{u},\mathbf{y}}\{F(\mathbf{x}^0, \mathbf{u}, \mathbf{y}) : \mathbf{u} \in \mathbb{U}, \mathbf{y} \in \phi(\mathbf{x}^0, \mathbf{u})\}$, where $\phi(\mathbf{x}^0, \mathbf{u}) = \arg\min_{\mathbf{y}}\{f(\mathbf{x}^0, \mathbf{u}, \mathbf{y}) : g(\mathbf{x}^0, \mathbf{u}, \mathbf{y}) \leq \mathbf{0}, \mathbf{y} \in \mathbb{Y}\}$. Let $\bar{\mathbf{y}}^0 = \mathbf{y}^0$, then for $\mathbf{TSROBLO^R_{2L}}$, we have $\mathbf{x}^0 \in \mathbb{X}, G(\mathbf{x}^0) \leq \mathbf{0}, \bar{\mathbf{y}}^0 \in \mathbb{Y}, g(\mathbf{x}^0, \mathbf{u}, \bar{\mathbf{y}}^0) \leq \mathbf{0}$. Furthermore, as $\bar{\mathbf{y}}^0 = \mathbf{y}^0$ and $\mathbf{y}^0 \in \phi(\mathbf{x}^0, \mathbf{u}^0)$, we have $\phi(\mathbf{x}^0, \mathbf{u}^0) = \zeta(\mathbf{x}^0, \mathbf{u}^0, \bar{\mathbf{y}}^0)$, and thus $(\mathbf{u}^0, \mathbf{y}^0) \in \arg\max_{\mathbf{u},\mathbf{y}}\{F(\mathbf{x}^0, \mathbf{u}, \mathbf{y}) : \mathbf{u} \in \mathbb{U}, \mathbf{y} \in \zeta(\mathbf{x}^0, \mathbf{u}, \bar{\mathbf{y}}^0)\}$. Therefore, $(\mathbf{x}^0, \bar{\mathbf{y}}^0, \mathbf{u}^0, \mathbf{y}^0)$ is feasible to $\mathbf{TSROBLO^R_{2L}}$, and the result follows.
2) It is easy to verify that $(\mathbf{x}^*, \mathbf{u}^*, \mathbf{y}^*)$ is feasible to $\mathbf{TSRPBLO_2}$, and thus $F(\mathbf{x}^*, \mathbf{u}^*, \mathbf{y}^*) \geq F(\mathbf{x}^*, \mathbf{u}^*_R, \mathbf{y}^*_R)$. Moreover, we have

$$F(\mathbf{x}^*, \mathbf{u}^*, \mathbf{y}^*) = \max_{\mathbf{u},\mathbf{y}}\{F(\mathbf{x}^*, \mathbf{u}, \mathbf{y}) : \mathbf{u} \in \mathbb{U}, \mathbf{y} \in \phi(\mathbf{x}^*, \mathbf{u})\}, \tag{4.9}$$

and

$$F(\mathbf{x}^*, \mathbf{u}^*_R, \mathbf{y}^*_R) = \max_{\mathbf{u},\mathbf{y}}\{F(\mathbf{x}^*, \mathbf{u}, \mathbf{y}) : \mathbf{u} \in \mathbb{U}, g(\mathbf{x}^*, \mathbf{u}, \bar{\mathbf{y}}^*_R) \leq \mathbf{0}, \mathbf{y} \in \zeta(\mathbf{x}^*, \mathbf{u}, \bar{\mathbf{y}}^*_R)\}. \tag{4.10}$$

Let z^* be the optimal value of the following problem

$$\max_{\mathbf{u},\mathbf{y}}\{F(\mathbf{x}^*, \mathbf{u}, \mathbf{y}) : \mathbf{u} \in \mathbb{U}, g(\mathbf{x}^*, \mathbf{u}, \bar{\mathbf{y}}^*_R) \leq \mathbf{0}, \mathbf{y} \in \phi(\mathbf{x}^*, \mathbf{u})\}, \tag{4.11}$$

then (4.9) is a relaxation of (4.11), and thus $F(\mathbf{x}^*, \mathbf{u}^*, \mathbf{y}^*) \geq z^*$. If $g(\mathbf{x}^*, \mathbf{u}^*, \bar{\mathbf{y}}^*_R) \leq \mathbf{0}$, then $(\mathbf{x}^*, \mathbf{u}^*, \mathbf{y}^*)$ is optimal to (4.11), and $F(\mathbf{x}^*, \mathbf{u}^*, \mathbf{y}^*) = z^*$. As $\phi(\mathbf{x}^*, \mathbf{u}) \subseteq \zeta(\mathbf{x}^*, \mathbf{u}, \bar{\mathbf{y}}^*_R)$, (4.10) is a relaxation of (4.11), and thus $F(\mathbf{x}^*, \mathbf{u}^*_R, \mathbf{y}^*_R) \geq z^*$, which implies $F(\mathbf{x}^*, \mathbf{u}^*, \mathbf{y}^*) \leq F(\mathbf{x}^*, \mathbf{u}^*_R, \mathbf{y}^*_R)$. Therefore, $(\mathbf{x}^*, \mathbf{u}^*, \mathbf{y}^*)$ is optimal to $\mathbf{TSRPBLO_2}$, and $F(\mathbf{x}^*, \mathbf{u}^*, \mathbf{y}^*) = F(\mathbf{x}^*, \mathbf{u}^*_R, \mathbf{y}^*_R)$, i.e. the relaxation is tight. □

Remark 4.3. *For a special case of* $\mathbf{TSRPBLO_2}$*, where* $\mathbf{u}$ *only appears in the lower level objective function [116, 32], the relaxation problem* $\mathbf{TSROBLO^R_{2L}}$ *is tight since the constraint* $g(\mathbf{x}^*, \mathbf{u}^*, \bar{\mathbf{y}}^*_R) \leq \mathbf{0}$ *always holds in the absence of* $\mathbf{u}$.

Algorithm 4 Decomposition algorithm for **TSRPBLO**

```
Initialize LB = −∞, UB = +∞, Ū = ∅, k = 0
while UB − LB > ϵ do
    Solve MP_2, obtain an optimal solution x_1^*
    Update LB = η^*
    Solve FCGP and obtain an optimal solution u^*
    if z^*(x_1^*) > 0 then
        Add H(x_1, u^*, x_2) ≤ 0 and g(x_1, u^*, x_2, y) ≤ 0 to MP_2, k = k + 1
    else
        Solve SP_2 for x_1^*, obtain an optimal solution (u^*, x_2^*, ỹ^*)
        Update UB = min{UB, F(x_1^*, u^*, x_2^*, ỹ^*)}
    end if
    Set Ū = Ū ∪ {u^*}, add corresponding variables and constraints to MP_2, k = k + 1
end while
Return x_1^* as an optimal solution and terminate
```

To solve **TSRPBLO**, we can slightly modify Algorithm 3. In particular, we have

$$\mathbf{MP}_2 : \min\{\eta : (4.4), (4.5) - (4.6), \hat{\mathbf{y}}_k \in \psi(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}), \tilde{\mathbf{y}}_k \in \zeta(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}, \hat{\mathbf{y}}_k), \forall k \in \mathbb{K}_{\bar{\mathbb{U}}}\},$$

where $\zeta(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \hat{\mathbf{y}}) = \{\mathbf{y} : g(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) \leq \mathbf{0}, f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) \leq f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \hat{\mathbf{y}}), \mathbf{y} \in \mathbb{Y}\}$; and

$$\mathbf{SP}_2 : \max_{\{\mathbf{u} \in \mathbb{U}\}} \min_{\{\mathbf{x}_2 \in \mathbb{X}_2, H(\mathbf{x}_1^*, \mathbf{u}, \mathbf{x}_2) \leq \mathbf{0}\}} \max_{\{\tilde{\mathbf{y}} \in \phi(\mathbf{x}_1^*, \mathbf{u}, \mathbf{x}_2)\}} F(\mathbf{x}_1^*, \mathbf{u}, \mathbf{x}_2, \tilde{\mathbf{y}})$$

for **TSRPBLO**, respectively.

Theorem 4.2 holds for Algorithm 4 if $\bar{w}(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) = \max\{F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) : g(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) \leq \mathbf{0}, \mathbf{y} \in \mathbb{Y}, f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) \leq v(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\}$ is continuous, where $v(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$ is the optimal value function of the lower level problem. We further denote $\underline{w}(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) = \min_{\mathbf{y}}\{F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) : g(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) \leq \mathbf{0}, \mathbf{y} \in \mathbb{Y}, f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) \leq v(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\}$. In practice, verifying the existence of optimal solutions of **TSROBLO** and **TSRPBLO** analytically can be very hard, and the decomposition algorithm indeed provides us a method to investigate such problems.

Theorem 4.5. *Denote the lower level optimal value function of* **TSROBLO** *and* **TSRPBLO** *by* v *and let* $\underline{w}$ *and* $\bar{w}$ *be defined as previously. Suppose that* $\mathbb{X}_{\mathbb{U}} \neq \emptyset$*,* $\mathbb{X}_1, \mathbb{U}$*, and* $\mathbb{X}_2$ *are compact, and that* v*,* $\underline{w}$ *and* $\bar{w}$ *are continuous in* $\mathbf{y}$*, then we have the following sufficient conditions for the existence of optimal solutions of* **TSROBLO** *and* **TSRPBLO***.*

1) If the cardinality of $\mathbb{U}$ *or that of* $\mathbb{X}_1$ *is finite, then both* **TSROBLO** *and* **TSRPBLO** *have an optimal* $\mathbf{x}_1$*.*

2) If $\mathbf{u}$ *does not affect* $\mathbf{x}_2$*, i.e. the constraint* $H(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq \mathbf{0}$ *reduces to* $H(\mathbf{x}_1, \mathbf{x}_2) \leq \mathbf{0}$*, let* $\phi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) = \arg\min_{\mathbf{y}}\{f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) : g(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) \leq \mathbf{0}\}$ *as defined previously, then* **TSROBLO** *has an optimal* $\mathbf{x}_1$ *if the bilevel optimization problem*

$$\min_{\mathbf{x}_2, \mathbf{y}}\{F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) : H(\mathbf{x}_1, \mathbf{x}_2) \leq \mathbf{0}, \mathbf{x}_2 \in \mathbb{X}_2, \mathbf{y} \in \phi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\} \tag{4.12}$$

has an optimal solution for any fixed $(\mathbf{x}_1, \mathbf{u}) \in \mathbb{X}_{\mathbb{U}} \times \mathbb{U}$*; and* **TSRPBLO** *has an optimal* $\mathbf{x}_1$ *if the pessimistic bilevel optimization problem*

$$\min_{\mathbf{x}_2} \max_{\mathbf{y}}\{F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) : H(\mathbf{x}_1, \mathbf{x}_2) \leq \mathbf{0}, \mathbf{x}_2 \in \mathbb{X}_2, \mathbf{y} \in \phi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\} \tag{4.13}$$

has an optimal solution for any fixed $(\mathbf{x}_1, \mathbf{u}) \in \mathbb{X}_{\mathbb{U}} \times \mathbb{U}$*.*

3) If f*,* g *and* H *are convex and continuously differentiable,* f *is separable,* $\mathbf{x}_2$ *and* $\mathbf{y}$ *are continuous variables, and the lower level problem satisfies Slater's condition for any fixed* $(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$*, then* **TSROBLO** *has an optimal* $\mathbf{x}_1$ *if the problem*

$$\min_{\mathbf{x}_2, \mathbf{y}}\{F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) : H(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq \mathbf{0}, \mathbf{x}_2 \in \mathbb{X}_2, \mathbf{y} \in \phi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\} \tag{4.14}$$

has an optimal solution for any fixed $(\mathbf{x}_1, \mathbf{u}) \in \mathbb{X}_{\mathbb{U}} \times \mathbb{U}$*.*

Proof. We provide the proof in the appendix. □

Remark 4.4. *Bounded rationality can be readily incorporated into the two-stage robust bilevel models. For example, we can introduce bounded rationality to* **TSROBLO** *and* **TSRPBLO***. The resulting problems are*

$$\min_{\{\mathbf{x}_1 \in \mathbb{X}_1, G(\mathbf{x}_1) \leq \mathbf{0}\}} \max_{\{\mathbf{u} \in \mathbb{U}\}} \min_{\{\mathbf{x}_2 \in \mathbf{X}_2, H(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq \mathbf{0}, \mathbf{y} \in \phi_\epsilon(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\}} F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) \tag{4.15}$$

and

$$\min_{\{\mathbf{x}_1 \in \mathbb{X}_1, G(\mathbf{x}_1) \leq \mathbf{0}\}} \max_{\{\mathbf{u} \in \mathbb{U}\}} \min_{\{\mathbf{x}_2 \in \mathbb{X}_2, H(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq \mathbf{0}\}} \max_{\{\mathbf{y} \in \phi_\epsilon(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\}} F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}). \tag{4.16}$$

With slight modifications, (4.15) and (4.16) can be solved by the proposed algorithms.

4.3 Extensions of Robust Bilevel Optimization Models

In this section, we extend our previously proposed robust bilevel models to deal with different types of uncertainties and their combinations.

4.3.1 Robust Bilevel Optimization With Multiple Objectives

As mentioned previously, imperfect information may cause modeling errors. One typical case is that the follower may have multiple objectives, i.e., for a fixed decision made by the leader, the follower's DMP is given by

$$\min_{\mathbf{y}}\{\sum_{i=1}^{|I|} w_i f_i(\mathbf{x},\mathbf{y}) : g(\mathbf{x},\mathbf{y}) \leq \mathbf{0}, \mathbf{y} \in \mathbb{Y}\}, \tag{4.17}$$

where w_i is the weight of the objective function f_i for $i = 1, 2, ..., |I|$.

As the leader may not know the exact value of the weight, she can consider the worse case scenario through a robust bilevel model, where the weight is considered in an uncertainty set $\mathbb{W}$. Let $\phi(\mathbf{x}, \mathbf{w})$ be the optimal solution set of (4.17) for fixed $\mathbf{x}$ and $\mathbf{w}$, then one has

$$\min_{\{\mathbf{x}\in\mathbb{X}, G(\mathbf{x})\leq\mathbf{0}\}} \max_{\{\mathbf{w}\in\mathbb{W}\}} \min_{\{\mathbf{y}\in\phi(\mathbf{x},\mathbf{w})\}} F(\mathbf{x},\mathbf{y}) \tag{4.18}$$

if the follower is expected to cooperate, and

$$\min_{\{\mathbf{x}\in\mathbb{X}, G(\mathbf{x})\leq\mathbf{0}\}} \max_{\{\mathbf{w}\in\mathbb{W}, \mathbf{y}\in\phi(\mathbf{x},\mathbf{w})\}} F(\mathbf{x},\mathbf{y}) \tag{4.19}$$

if the follower is not expected to do so. From Remark 4.3, we know that (4.19) can be solved through the tight relaxation introduced in Theorem 4.4.

In addition to the uncertainty of the weight, other random factors may also be involved. For this more general case, we can incorporate the modeling errors into the basic robust bilevel models. For fixed $(\mathbf{x}_1, \mathbf{u}, \mathbf{w}, \mathbf{x}_2)$, the lower level problem is extended to

$$\min_{\mathbf{y}}\{\sum_{i=1}^{|I|} w_i f_i(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) : g(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) \leq \mathbf{0}, \mathbf{y} \in \mathbb{Y}\}, \tag{4.20}$$

whose optimal solution set is denoted by $\phi(\mathbf{x}_1, \mathbf{u}, \mathbf{w}, \mathbf{x}_2)$. Thus, we can modify **TSROBLO** as

$$\min_{\{\mathbf{x}_1 \in \mathbb{X}_1, G(\mathbf{x}_1) \leq \mathbf{0}\}} \max_{\{\mathbf{u} \in \mathbb{U}, \mathbf{w} \in \mathbb{W}\}} \min_{\{\mathbf{x}_2 \in \mathbb{X}_2, H(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq \mathbf{0}, \mathbf{y} \in \phi(\mathbf{x}_1, \mathbf{u}, \mathbf{w}, \mathbf{x}_2)\}} F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}), \tag{4.21}$$

and **TSRPBLO** as

$$\min_{\{\mathbf{x}_1 \in \mathbb{X}_1, G(\mathbf{x}_1) \leq \mathbf{0}\}} \max_{\{\mathbf{u} \in \mathbb{U}, \mathbf{w} \in \mathbb{W}\}} \min_{\{\mathbf{x}_2 \in \mathbf{X}_2, H(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq \mathbf{0}\}} \max_{\{\mathbf{y} \in \phi(\mathbf{x}_1, \mathbf{u}, \mathbf{w}, \mathbf{x}_2)\}} F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}), \tag{4.22}$$

such that both the exogenous uncertainty (represented by $\mathbb{U}$) and the perception uncertainty (represented by $\mathbb{W}$) are taken into consideration.

We point out that all the four models (i.e., (4.18), (4.19), (4.21), and (4.22)) can be readily solved by Algorithm 3 or Algorithm 4. In particular, for fixed $\mathbf{x}_1$, the worst case $(\mathbf{u}, \mathbf{w})$ is obtained from the sub-problem and its corresponding variables and constraints are added to the master problem.

4.3.2 Robust Bilevel Optimization With Objective Function Uncertainty

The follower may partially cooperate with the leader [38]. This leads to a strong-weak formulation

$$\begin{aligned}
\min_{\mathbf{x}, \mathbf{y}_s, \mathbf{y}_w} \; & \alpha F_s + (1-\alpha) F_w \\
\text{s.t.} \; & G(\mathbf{x}) \leq \mathbf{0}, \mathbf{x} \in \mathbb{X} \\
& F_s = \min_{\mathbf{y}_s \in \phi(\mathbf{x})} F(\mathbf{x}, \mathbf{y}_s) \\
& F_w = \max_{\mathbf{y}_w \in \phi(\mathbf{x})} F(\mathbf{x}, \mathbf{y}_w) \\
& \phi(\mathbf{x}) = \arg\min_{\mathbf{y}} \{ f(\mathbf{x}, \mathbf{y}) : g(\mathbf{x}, \mathbf{y}) \leq \mathbf{0}, \mathbf{y} \in \mathbb{Y} \}.
\end{aligned} \tag{4.23}$$

The parameter $0 \leq \alpha \leq 1$ represents the cooperation degree of the follower, and the optimal value of (4.23) is non-increasing in α.

It is shown in [38, 137] that the follower can be better off by partially cooperating, i.e. by setting $0 < \alpha < 1$. In this case, it is better for the leader to consider α being in an uncertain set rather than a constant and to solve the following robust bilevel optimization problem

$$\min_{\{\mathbf{x} \in \mathbb{X}, G(\mathbf{x}) \leq \mathbf{0}\}} \max_{0 \leq \alpha \leq 1} \{ \alpha F_s + (1-\alpha) F_w : F_s = \min_{\mathbf{y}_s \in \phi(\mathbf{x})} F(\mathbf{x}, \mathbf{y}_s), F_w = \max_{\mathbf{y}_w \in \phi(\mathbf{x})} F(\mathbf{x}, \mathbf{y}_w) \}. \tag{4.24}$$

As (4.24) has a special structure, it can be solved more efficiently. Notice that the optimal value of $\min_{\mathbf{y}_s \in \phi(\mathbf{x})} F(\mathbf{x}, \mathbf{y}_s)$ and that of $\max_{\mathbf{y}_w \in \phi(\mathbf{x})} F(\mathbf{x}, \mathbf{y}_w)$ depend only on $\mathbf{x}$, we do not have to solve a complicated multi-level sub-problem. Instead, for fixed $\mathbf{x}$, we can solve those two problems independently to get the value of F_s and F_w. Then the worst case α can be found by solving a very simple LP $\max_{0 \leq \alpha \leq 1} \alpha F_s + (1-\alpha) F_w$, where F_s and F_w are fixed.

In practice, the follower may not know the upper level objective function exactly. In this case, he would make decisions based on an approximation of the true upper level objective function. If we denote the approximation of F by $\tilde{F}$ and the lower level optimal solution set again by $\phi(\mathbf{x})$, then the leader's DMP is

$$\begin{aligned} &\min F(\mathbf{x}, \tilde{\mathbf{y}}) \\ &\text{s.t.} G(\mathbf{x}) \leq \mathbf{0}, \mathbf{x} \in \mathbb{X} \\ &\quad \tilde{\mathbf{y}} \in S_1(\mathbf{x}) = \arg\min\{\tilde{F}(\mathbf{x}, \mathbf{y}) : \mathbf{y} \in \phi(\mathbf{x})\}. \end{aligned} \tag{4.25}$$

Due to the limited information the follower has, it might be better for the leader to also consider the response uncertainty, leading to the pessimistic counterpart of (4.25)

$$\begin{aligned} &\min F(\mathbf{x}, \tilde{\mathbf{y}}) \\ &\text{s.t.} G(\mathbf{x}) \leq \mathbf{0}, \mathbf{x} \in \mathbb{X} \\ &\quad \tilde{\mathbf{y}} \in S_2(\mathbf{x}) = \arg\max\{\tilde{F}(\mathbf{x}, \mathbf{y}) : \mathbf{y} \in \phi(\mathbf{x})\}. \end{aligned} \tag{4.26}$$

In general, $\tilde{F}$ is in an uncertainty set $\mathcal{F}$ rather than deterministic. Hence, we can extend (4.25) to

$$\min_{\{G(\mathbf{x}) \leq \mathbf{0}, \mathbf{x} \in \mathbb{X}\}} \max_{\{\tilde{F} \in \mathcal{F}\}} \min_{\{\tilde{\mathbf{y}} \in S_1(\mathbf{x})\}} \tilde{F}(\mathbf{x}, \tilde{\mathbf{y}}), \tag{4.27}$$

and extend (4.26) to

$$\min_{\{G(\mathbf{x}) \leq \mathbf{0}, \mathbf{x} \in \mathbb{X}\}} \min_{\{\tilde{F} \in \mathcal{F}, \tilde{\mathbf{y}} \in S_2(\mathbf{x})\}} \tilde{F}(\mathbf{x}, \tilde{\mathbf{y}}). \tag{4.28}$$

Remark 4.5. *If the true upper level objective function F is in the uncertainty set $\mathcal{F}$, then* (4.27) *reduces to an instance of optimistic bilevel optimization problem*

$$\min_{\{G(\mathbf{x}) \leq \mathbf{0}, \mathbf{x} \in \mathbb{X}, \tilde{\mathbf{y}} \in \phi(\mathbf{x})\}} F(\mathbf{x}, \tilde{\mathbf{y}}),$$

and (4.28) *reduces to an instance of pessimistic bilevel optimization problem*

$$\min_{\{G(\mathbf{x}) \leq \mathbf{0}, \mathbf{x} \in \mathbb{X}\}} \max_{\{\tilde{\mathbf{y}} \in \phi(\mathbf{x})\}} F(\mathbf{x}, \tilde{\mathbf{y}}).$$

4.3.3 Robust Bilevel Optimization With Communication Uncertainty

In a hierarchical system described by a bilevel model, the two DMs interact with each other by exchanging information. However, the information may get lost or affected by noise while passing from one DM to the other. Such incomplete or inaccurate information causes communication uncertainty, which is a typical instance of perception uncertainty.

We first consider the circumstances that information get affected while passing from the leader to the follower. In this case, an upper level decision (denoted by $\mathbf{x}$) along with noise (denoted by $\mathbf{u_x}$) passes to the follower, who then makes a decision based on his DMP, i.e.,

$$\min_{\mathbf{y}}\{f(\mathbf{x},\mathbf{u_x},\mathbf{y}) : g(\mathbf{x},\mathbf{u_x},\mathbf{y}) \leq \mathbf{0}, \mathbf{y} \in \mathbb{Y}\},$$

whose optimal solution set is denoted by $\phi(\mathbf{x},\mathbf{u_x})$.

To consider the worst case noise, we have

$$\min_{\{\mathbf{x}\in\mathbb{X},G(\mathbf{x})\leq\mathbf{0}\}} \max_{\{\mathbf{u_x}\in\mathbb{U}_\mathbf{x}\}} \min_{\{\mathbf{y}\in\phi(\mathbf{x},\mathbf{u_x})\}} F(\mathbf{x},\mathbf{y}) \tag{4.29}$$

for the cooperative case, and

$$\min_{\{\mathbf{x}\in\mathbb{X},G(\mathbf{x})\leq\mathbf{0}\}} \max_{\{\mathbf{u_x}\in\mathbb{U}_\mathbf{x},\mathbf{y}\in\phi(\mathbf{x},\mathbf{u_x})\}} F(\mathbf{x},\mathbf{y}) \tag{4.30}$$

for the non-cooperative case. Both (4.29) and (4.30) can be solved by Algorithm 3 and Algorithm 4, respectively. Moreover, information can also get affected while passing back to the leader, i.e., a lower level decision $\mathbf{y}$ comes back with noise $\mathbf{u_y}$ as a respond to the leader.

If the follower is cooperative, we can consider him and the leader together as one DM, and the noise as exogenous random factor in an uncertainty set $\mathbb{U}_\mathbf{y}$. Since no decision is made after the uncertainty reveals, we can employ a single-stage robust bilevel model

$$\min_{\{\mathbf{x}\in\mathbb{X},G(\mathbf{x})\leq\mathbf{0},\mathbf{y}\in\phi(\mathbf{x})\}} \max_{\{\mathbf{u_y}\in\mathbb{U}_\mathbf{y}\}} F(\mathbf{x},\mathbf{y},\mathbf{u_y}), \tag{4.31}$$

where $\phi(\mathbf{x}) = \arg\min_\mathbf{y}\{f(\mathbf{x},\mathbf{y}) : g(\mathbf{x},\mathbf{y}) \leq \mathbf{0}, \mathbf{y} \in \mathbb{Y}\}$ is independent of $\mathbf{u_y}$.

For fixed $(\mathbf{x},\mathbf{y})$, the worst case scenario is obtained by solving

$$\max_{\mathbf{u_y}}\{F(\mathbf{x},\mathbf{y},\mathbf{u_y}) : \mathbf{u_y} \in \mathbb{U}_\mathbf{y}\}. \tag{4.32}$$

From the continuity of F and the compactness of $\mathbb{U}_{\mathbf{y}}$, we know that (4.32) has an optimal solution for any fixed $(\mathbf{x}, \mathbf{y})$. Furthermore, we can have the dual problem of (4.32), namely

$$\min_{\mathbf{v}_{\mathbf{y}}} \{\hat{F}(\mathbf{x}, \mathbf{y}, \mathbf{v}_{\mathbf{y}}) : \mathbf{v}_{\mathbf{y}} \in \mathbb{V}_{\mathbf{y}}(\mathbf{x}, \mathbf{y})\}. \tag{4.33}$$

A sufficient condition for the strong duality to hold for (4.32) and (4.33) is that $\mathbb{U}_{\mathbf{y}}$ is a convex set with an interior point and F is concave in $\mathbf{u}_{\mathbf{y}}$. Then, we can rewrite (4.31) as

$$\min_{\mathbf{x}, \mathbf{y}, \mathbf{v}_{\mathbf{y}}} \{\hat{F}(\mathbf{x}, \mathbf{y}, \mathbf{v}_{\mathbf{y}}) : \mathbf{x} \in \mathbb{X}, G(\mathbf{x}) \leq \mathbf{0}, \mathbf{y} \in \phi(\mathbf{x}), \mathbf{v}_{\mathbf{y}} \in \mathbb{V}_{\mathbf{y}}(\mathbf{x}, \mathbf{y})\},$$

which is an instance of optimistic bilevel optimization problem.

If the follower is non-cooperative, we can consider him along with the noise together as one follower, who plays against the leader. Such settings lead to a pessimistic bilevel optimization instance as

$$\min_{\{\mathbf{x} \in \mathbb{X}, G(\mathbf{x}) \leq \mathbf{0}\}} \max_{\{\mathbf{y} \in \phi(\mathbf{x}), \mathbf{u}_{\mathbf{y}} \in \mathbb{U}_{\mathbf{y}}\}} F(\mathbf{x}, \mathbf{y}, \mathbf{u}_{\mathbf{y}}), \tag{4.34}$$

which can be solved by existing methods such as [136, 132].

Information interference may occur while going from and coming back to the leader. For the cooperative case, we can combine (4.29) and (4.33) to have a robust bilevel model as

$$\min_{\{\mathbf{x} \in \mathbb{X}, G(\mathbf{x}) \leq \mathbf{0}\}} \max_{\{\mathbf{u}_{\mathbf{x}} \in \mathbb{U}_{\mathbf{x}}\}} \min_{\{\mathbf{y} \in \phi(\mathbf{x}, \mathbf{u}_{\mathbf{x}}), \mathbf{v}_{\mathbf{y}} \in \mathbb{V}_{\mathbf{y}}(\mathbf{x}, \mathbf{y})\}} \hat{F}(\mathbf{x}, \mathbf{y}, \mathbf{v}_{\mathbf{y}}). \tag{4.35}$$

The existence of optimal solution of (4.32) along with the strong duality implies that (4.33) also has an optimal solution for any fixed $(\mathbf{x}, \mathbf{y})$. Hence, we have

$$\phi(\mathbf{x}, \mathbf{u}_{\mathbf{x}}) = \text{Proj}_{\mathbf{y}}(\arg\min_{\mathbf{y}, \mathbf{v}_{\mathbf{y}}} \{f(\mathbf{x}, \mathbf{u}_{\mathbf{x}}, \mathbf{y}) : g(\mathbf{x}, \mathbf{u}_{\mathbf{x}}, \mathbf{y}) \leq \mathbf{0}, \mathbf{y} \in \mathbb{Y}, \mathbf{v}_{\mathbf{y}} \in \mathbb{V}_{\mathbf{y}}(\mathbf{x}, \mathbf{y})\})$$

since $\mathbb{V}_{\mathbf{y}}(\mathbf{x}, \mathbf{y}) \neq \emptyset$ for any fixed $(\mathbf{x}, \mathbf{y})$. Therefore, $\mathbf{y}$ and $\mathbf{v}_{\mathbf{y}}$ can be considered as one follower, and thus (4.35) is reduced to an instance of **TSROBLO**.

For the non-cooperative case, we can combine (4.30) and (4.34) to have

$$\min_{\{\mathbf{x} \in \mathbb{X}, G(\mathbf{x}) \leq \mathbf{0}\}} \max_{\{\mathbf{u}_{\mathbf{x}} \in \mathbb{U}_{\mathbf{x}}, \mathbf{y} \in \phi(\mathbf{x}, \mathbf{u}_{\mathbf{x}}), \mathbf{u}_{\mathbf{y}} \in \mathbb{U}_{\mathbf{y}}\}} F(\mathbf{x}, \mathbf{y}, \mathbf{u}_{\mathbf{y}}),$$

which is reduced to an instance of **TSRPBLO**.

4.3.4 Robust Bilevel Optimization With Multiple Uncertainty Sets

In some cases, we may need multiple uncertainty sets to better capture the uncertainty arising in practice. In the presence of scenario-specific decisions, we can extend the two stage robust bilevel models to include multiple uncertainty sets, which can better reflect the leader's conservativeness and reduce the impact of unrealistic scenarios [7].

Suppose there are $|I|$ uncertainty sets and ρ_i is the weight of uncertainty set $\mathbb{U}_i$, then **TSROBLO** can be extended as

$$\min_{\{\mathbf{x}_1 \in \mathbb{X}_1, G(\mathbf{x}_1) \leq \mathbf{0}\}} \sum_{i \in I} \rho_i (\max_{\{\mathbf{u}_i \in \mathbb{U}_i\}} \min_{\{\mathbf{x}_{2_i} \in \mathbb{X}_2, H_i(\mathbf{x}_1, \mathbf{u}_i, \mathbf{x}_{2_i}) \leq \mathbf{0}, \mathbf{y}_i \in \phi_i(\mathbf{x}_1, \mathbf{u}_i, \mathbf{x}_{2_i})\}} F(\mathbf{x}_1, \mathbf{u}_i, \mathbf{x}_{2_i}, \mathbf{y}_i)), \tag{4.36}$$

and **TSRPBLO** can be extended as

$$\min_{\{\mathbf{x}_1 \in \mathbb{X}_1, G(\mathbf{x}_1) \leq \mathbf{0}\}} \sum_{i \in I} \rho_i (\max_{\{\mathbf{u}_i \in \mathbb{U}_i\}} \min_{\{\mathbf{x}_{2_i} \in \mathbb{X}_2, H_i(\mathbf{x}_1, \mathbf{u}_i, \mathbf{x}_{2_i}) \leq \mathbf{0}\}} \max_{\{\mathbf{y}_i \in \phi_i(\mathbf{x}_1, \mathbf{u}_i, \mathbf{x}_{2_i})\}} F(\mathbf{x}_1, \mathbf{u}_i, \mathbf{x}_{2_i}, \mathbf{y}_i)), \tag{4.37}$$

where $\phi_i(\mathbf{x}_1, \mathbf{u}_i, \mathbf{x}_{2_i}) = \arg\min_{\mathbf{y}_i} \{ f_i(\mathbf{x}_1, \mathbf{u}_i, \mathbf{x}_{2_i}, \mathbf{y}_i) : g_i(\mathbf{x}_1, \mathbf{u}_i, \mathbf{x}_{2_i}, \mathbf{y}_i) \leq \mathbf{0}, \mathbf{y}_i \in \mathbb{Y} \}$.

With minor modifications, (4.36) and (4.37) can be solved by Algorithm 3 and Algorithm 4, respectively. Specifically, for fixed $\mathbf{x}_1$, $|I|$ independent sub-problems are solved, and $|I|$ worst case scenarios are obtained. Then each scenario with its corresponding variables and constraints are added to the MP. As the problem size of increases in the number of uncertainty sets, it is better to keep $|I|$ small in practice.

4.4 Computational Study

In this section, we apply the proposed model and algorithm to solve two practical problems arising from real world systems, namely, a vehicle sharing system design problem and a plant selection problem, both of which are subject to uncertainty.

4.4.1 Design of Vehicle Sharing System Under Uncertainty

In recent years, vehicle sharing systems (VSSs) have been implemented in many different cities [118], and companies such as Zipcar, Hertz, Enterprise, now have car sharing programs. With the help of VSSs, people have more traveling options, and vehicles are utilized more efficiently. As a result, VSSs are expected to increase in the near future [138]. Studies on design of VSSs can be found in [12, 46, 95, 100] and the references therein.

The model for our computational study is adopted from [100] with some modifications. Most notations are the same as those in [100] for consistency purpose. By convention, we use a regular lower case letter to represent a scalar and a bolded letter to represent the corresponding vector.

Sets

$G(V, A)$	a transportation network
V	the set of sites of $G(V, A)$
A	the set of arcs of $G(V, A)$
$V_s \subseteq V$	the set of candidate sharing sites
$A_s \subseteq A$	the set of sharing arcs
$\underline{A} \subseteq A$	the set of non-frequency based arcs
$A \backslash \underline{A}$	the set of frequency based arcs
K	the set of OD pairs, indexed by k
D	the uncertainty set of the demand, and $d_{ik} \in D$

Parameters

r_{ij}	revenue gained from arc $(i,j) \in A_s$
c_s	costs of building a vehicle sharing station
c_p	costs of adding a parking slot
c_v	unit cost for a vehicle
U	the upper bound of the capacity of a candidate site
y^{ub}	upper bound of a parking slot
c_{ij}	costs of traveling through arc $(i,j) \in A$
d_{ik}	demand at site $i \in V$ for OD pair k
f_{ij}	frequency parameter for arc $(i,j) \in A \backslash \underline{A}$
a	the checkout replacement ratio, $a \geq 1$
Variables	
x_i	binary variable, 1 if a vehicle sharing station is built at site $i \in V_s$; 0 otherwise
y_i	the capacity of station $i \in V_s$
z_i	the number of vehicles at station $i \in V_s$
v_{ijk}	flow decision variable over arc $(i,j) \in A$ for OD pair k
w_{ik}	waiting time at site $i \in V$ for OD pair k

Using the aforementioned notations, a robust vehicle sharing system design problem (RVSSDP) can be modeled as follows

$$\text{RVSSDP:} \max_{(\mathbf{x},\mathbf{y},\mathbf{z})\in\mathbf{M}} -(\sum_{i\in V_s} c_s x_i + c_p y_i + c_v z_i) + \min_{\mathbf{d}\in D,(\mathbf{v},\mathbf{w})\in R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{d})} \sum_{k\in K}\sum_{(i,j)\in A_s} r_{ij} v_{ijk}, \tag{4.38}$$

where

$$\begin{aligned} \mathbf{M} = \{(\mathbf{x},\mathbf{y},\mathbf{z}) : U x_i \geq y_i, \quad & \forall i \in V_s \\ z_i \leq y_i, \quad & \forall i \in V_s \\ y_i \leq y^{ub}, \quad & \forall i \in V_s \\ x_i \in \{0,1\}, y_i, z_i \in Z_+^n, & \forall i \in V_s\}, \end{aligned} \tag{4.39}$$

$$R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{d}) = \arg\min_{\mathbf{v},\mathbf{w}} \{\sum_{k\in K}(\sum_{(i,j)\in A} c_{ij} v_{ijk} + \sum_{i\in V} w_{ik}) : (\mathbf{v},\mathbf{w}) \in L(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{d})\}, \tag{4.40}$$

and

$$L(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{d}) = \{(\mathbf{v},\mathbf{w}) : \begin{aligned}[t] & \sum_{j:(i,j)\in A} v_{ijk} - \sum_{j:(j,i)\in A} v_{jik} = d_{ik}, \quad \forall i \in V, k \in K \\ & v_{ijk} \le f_{ij} w_{ik}, \quad \forall (i,j) \in A \backslash \underline{A}, k \in K \\ & U x_i \ge \sum_{k\in K} v_{ijk}, \quad \forall (i,j) \in A_s \\ & U x_j \ge \sum_{k\in K} v_{ijk}, \quad \forall (i,j) \in A_s \\ & \sum_{k\in K} \sum_{j:(i,j)\in A_s} v_{ijk} \le z_i, \quad \forall i \in V_s \\ & \sum_{k\in K} \sum_{j:(j,i)\in A_s} v_{jik} \le a(y_i - z_i), \quad \forall i \in V_s \\ & w_{ik} \ge 0, \quad \forall i \in V, k \in K \\ & v_{ijk} \ge 0, \quad \forall (i,j) \in A, k \in K\}. \end{aligned} \tag{4.41}$$

In the upper level, the system owner seeks to maximize the profit, which is the difference between the revenue and the cost, by making decisions on the location (represented by $\mathbf{x}$), the capacity (represented by $\mathbf{y}$) and the number of vehicles (represented by $\mathbf{z}$) of each vehicle sharing station in a transportation network, subject to capacity upper bound constraints (represented by U). In the lower level, the customers choose an optimal path to minimize their traveling cost for an existing VSS configuration, subject to flow conservation constraints, waiting time constraints, capacity constraints, and the non-negativity constraints.

We see that RVSSDP considers both exogenous and endogenous uncertainty. In particular, demand is an exogenous factor of the system with uncertainty, and is assumed to minimize the revenue. For a fixed vector $\mathbf{d}$, RVSSDP reduces to a deterministic pessimistic bilevel model, which only has endogenous uncertainty. In this case, if a customer has multiple optimal paths, i.e. different paths that give the same travelling cost, one would choose the path that gives the least revenue to the system owner. In practice, customers may choose a suboptimal path as long as it is not too far from an optimal one. There are two main reasons for this irrational behavior. First, as customers do not have perfect information, they are not able to find an optimal path, and thus choose a path according to their subjective judgment. Second, customers may be insensitive to the difference between a suboptimal path and an optimal one as long as their traveling cost does not increase significantly. To deal with such irrational behaviors, we apply the idea of bounded rationality to RVSSDP.

Let $\theta^*(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{d})$ be the optimal value of the lower level problem for fixed $(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{d})$, then we can modify the lower level optimal solution set as

$$R_\epsilon(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{d}) = \{(\mathbf{v},\mathbf{w}) \in L(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{d}) : \sum_{k\in K}(\sum_{(i,j)\in A} c_{ij}v_{ijk} + \sum_{i\in V} w_{ik}) \leq (1+\epsilon) * \theta^*(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{d})\},$$

and modify (4.38) as

$$\text{RVSSDP}_\epsilon : \max_{(\mathbf{x},\mathbf{y},\mathbf{z})\in\mathbf{M}} -(\sum_{i\in V_s} C_s x_i + C_p y_i + C_v z_i) + \min_{\mathbf{d}\in D,(\mathbf{v},\mathbf{w})\in R_\epsilon(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{d})} \sum_{k\in K}\sum_{(i,j)\in A_s} r_{ij}v_{ijk},$$

where $\epsilon \geq 0$ represents the degree of irrationality of the customers. It is obvious that $R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{d}) \subseteq R_\epsilon(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{d})$ for any fixed $(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{d})$, and that $R_\epsilon(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{d})$ reduces to $R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{d})$ if $\epsilon = 0$.

Remark 4.6. *For fixed upper level decision variables* $(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$*,* $L(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{d}) \neq \emptyset, \forall \boldsymbol{d} \in D$*, i.e., the lower level problem has the relatively complete response property. This property indeed comes from practice. In a real transportation network, customers can always find a path to their destination without using a VSS. For example, they can drive, walk, or take a bus, but dot not have to use a VSS. Mathematically, this means that for any fixed* $(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{d})$*, there exists* $(\boldsymbol{v},\boldsymbol{w})$ *such that all the constraints of the lower level problem are satisfied.*

In our computational study, the demand is assumed to vary within an uncertainty set in the form of $\mathbf{d} = \bar{\mathbf{d}} + \sum_{l=1}^{L} \Delta\mathbf{d}^l u_l$ as in [24, 19, 134], where $\bar{\mathbf{d}}$ is the nominal value, $\Delta\mathbf{d}^l$ is the direction of data perturbation, and u^l are random variables . We consider a polyhedral uncertainty set defined as $D = \{\mathbf{d} \in R_+^{|V|*|K|} : \mathbf{d} = \bar{\mathbf{d}} + \sum_{l=1}^{L} \Delta\mathbf{d}^l u_l, 0 \leq u_l \leq 1, l = 1,2,...,L, \sum_{l=1}^{L} u_l \leq \Omega\}$, where Ω is a parameter reflecting the level of uncertainty. We set $L = 2|K|$, and let $\Delta\mathbf{d}^l \in R_+^{|K|}$ be a vector with all its elements being zero except for the lth element, which is $0.30 * \bar{d}_l$, for $l = 1,2,...,|K|$, and is $-0.30 * \bar{d}_l$, for $l = |K|+1, |K|+2, ..., 2|K|$.

Figure 3 illustrates a typical transportation network with corridor structure, where the VSS station candidate are represented by the round nodes, and the origin and destination of customers are represented by the squared nodes. The bold black lines represent the paths

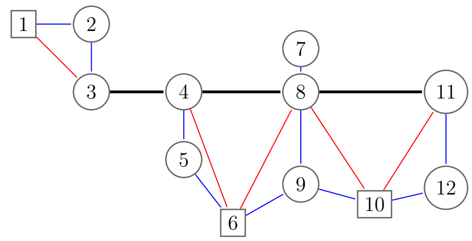

Figure 3: Transportation Network With Corridor Structure

in a public transportation system such as subway; the red lines represent the walking paths; and the blue lines represent the paths through which vehicle sharing service is offered.

We solve the test instances based on the transportation network in Figure 3. The time limit is half an hour, and the optimality gap is 1%. If an instance is not solved to optimality within the time limit, the best feasible solution is reported. Figure 4 shows the worst case profit under different level of irrationality (ϵ) for $\Omega = 0.2$ and $\Omega = 0.4$. The blue line represents the optimistic bilevel model, while the orange line represents the the proposed two-stage robust bilevel model.

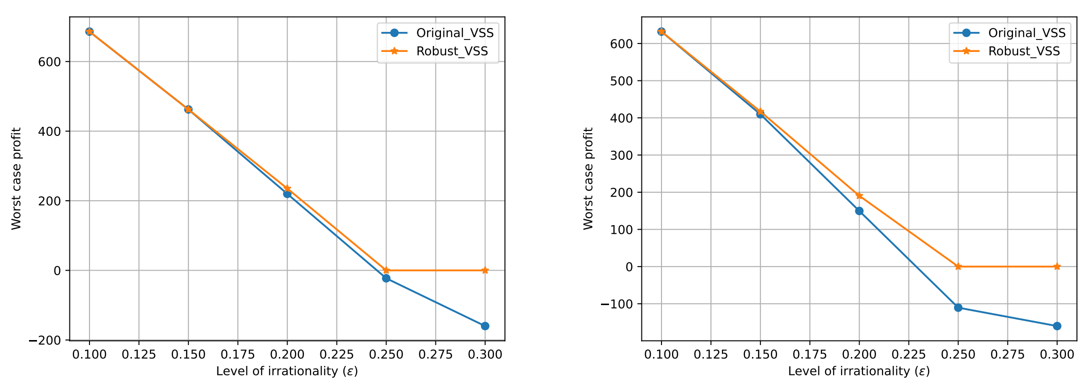

Figure 4: Worst Case Profit for $\Omega = 0.2$ and $\Omega = 0.4$

It is observed that the profit decreases in demand uncertainty (Ω) and irrationality (ϵ). In all cases, RVSSDP provides more profit. Moreover, the difference of the profit from the two models increase in ϵ for fixed Ω and in Ω for fixed ϵ. Such result shows that the robust model is better able to hedge against both the exogenous and endogenous uncertainty. By carefully determining the parameters such as Ω and ϵ, the proposed model can provide strong decision support under various of uncertainties.

4.4.2 Capacitated Plant Selection Problem Under Uncertainty

In this subsection, we apply the two-stage robust belevel optimization model and the proposed decomposition algorithm to solve a robust capacitated plant selection problem (RCPSP), whose deterministic counterpart is originally introduced in [37].

Sets

I	the set of the potential plants, indexed by i
J	the set of the products, indexed by j
IS_i	the group of products that can be produced in plant i, and $IS_i \subseteq J, \forall i \in I$
JS_j	the set of plants that can produce product j, and $JS_j \subseteq I, \forall j \in J$
D	the uncertainty set of demand, and $d_j \in D, \forall j \in J$

Parameters

p_i	opportunity cost for unused production capacity of plant i after it is opened
d_j	customer demand of product j
a_{ij}	capacity consumption ratio for processing product j in plant i
w_i	cost of use production capacity in plant i
Cap_i	available production capacity in plant i
R_{ij}	transportation cost for shipping product j from the principal firm to plant i
f_i	opening cost for plant i

Variables

x_i	binary variables, $x_i = 1$ if plant i is selected and opened; and $x_i = 0$ otherwise
y_{ij}	the number of products j produced in plant i

Using the aforementioned notations, RCPSP is as follows, where we keep most of the notations the same as they are in [37] for consistency purpose.

$$\text{RCPSP:} \min_{\mathbf{x}\in\{0,1\}^{|I|}} \sum_{i\in I} f_i x_i + \max_{\mathbf{d}\in D} \min_{\mathbf{y}\in S(\mathbf{x},\mathbf{d})} \sum_{i\in I} p_i(Cap_i x_i - \sum_{j\in IS_i} a_{ij} y_{ij}), \tag{4.42}$$

where $S(\mathbf{x}, \mathbf{d})$ is the optimal solution set of the following lower level problem

$$\begin{aligned} \min_{\mathbf{y}} & \sum_{i\in I} w_i \sum_{j\in IS_i} a_{ij} y_{ij} + \sum_{i\in I} \sum_{j\in IS_i} R_{ij} y_{ij} \\ \text{s.t.} & \sum_{i\in JS_j} y_{ij} \geq d_j, \ \forall j \in J \\ & \sum_{j\in IS_i} a_{ij} y_{ij} \leq Cap_i x_i, \ \forall i \in I \\ & y_{ij} \geq 0, \forall j \in IS_i, i \in I. \end{aligned} \tag{4.43}$$

In the upper level, the leader decides plant locations to minimize the total cost, which includes opening cost and the opportunity cost of the unused capacity, and the "$\max - \min$" term reflects the worst case consideration. In the lower level, the follower decides the number of product to be made in each opened plant to minimize the total cost, which includes the production cost and the transportation cost. The first lower level constraint ensures that the all the demands are satisfied, the second constraint ensures that the number of products made in a plant is no more than the capacity of that plant, and the last constraint introduces the non-negativity requirement. Same as in the original model, the follower is cooperative with the leader.

In our numerical study, the demand is assumed to be in the form of $\mathbf{d} = \bar{\mathbf{d}} + \sum_{k=1}^{K} \Delta\mathbf{d}^k u_k$, where $\bar{\mathbf{d}}$ is the nominal value, $\Delta\mathbf{d}^k$ is the direction of data perturbation for $k = 1, 2, ..., K$, and $u^k, k = 1, 2, ..., K$ are random variables [24, 19, 134]. We consider a polyhedral uncertainty set defined as $D = \{\mathbf{d} \in R_+^{|J|} : \mathbf{d} = \bar{\mathbf{d}} + \sum_{k=1}^{K} \Delta\mathbf{d}^k u_k, 0 \leq u_k \leq 1, k = 1, 2, ..., K, \sum_{k=1}^{K} u_k \leq \Omega\}$, where Ω is a parameter reflecting the level of uncertainty.

We first demonstrate the impact of uncertainty on the selection of plants using the original deterministic model as a benchmark. The data is obtained from [37]. In particular, we set $\bar{\mathbf{d}}$ equal to the demand in [37], set $K = 2|J|$, and let $\Delta\mathbf{d}^k \in R_+^{|J|}$ be a vector with all its elements being zero except for the kth element, which is $0.3 * \bar{d}_k$, for $k = 1, 2, ..., |J|$, and is $-0.3 * \bar{d}_k$, for $k = |J| + 1, |J| + 2, ..., 2|J|$. We set the optimality tolerance $\epsilon = 0.05\%$, and the time limit is 1800 seconds.

Table 9 shows the result for various Ω. Since the original model is solved for the nominal value of the demand, it may need to purchase products in the market to meet the additional demand. We set the unit price of a product in the marker to 3, and the worst case cost of the original model is calculated as the total of the original cost and the purchase cost. As Ω increases, RCPSP recommends opening more plants to meet potential additional demands, while the original model incurs additional purchase cost. Thus, the leader can make better decision if she can have accurate estimation of the demand and the product price in the market. Such information may be obtained through data mining techniques in practice. With accurately estimated data, the proposed model is able to better support the leader to make better decision under uncertainty.

Table 9: Impact of Uncertainty

Ω	Model	Plants	Purchase	Worst case cost
0	Original	2,3	0	541.00
	RCPSP	2,3	0	541.00
2	Original	2,3	26.00	619.00
	RCPSP	1,6	0	589.33
4	Original	2,3	62.00	727.00
	RCPSP	2,3,6	0	726.43
6	Original	2,3	88.00	805.00
	RCPSP	2,3,5	0	825.40

In addition to solving RCPSP with the original data, we also test the algorithm on randomly generated instances. In particular, we solve RCPSP instances of different sizes over a polyhedral uncertainty set. We set $\Omega = 3$ and solve five instances for each problem size. Table 10 shows the performance of the decomposition algorithm. "$|I|$" and "$|J|$" are the cardinality of I and J, respectively. The time (in seconds) used for solving an instance and the average are reported. We observe that all the instances are solved in half an hour, and that some instances are solved in less than one minute. As expected, the solving time increases as the size of the problem increases. For large size problems, developing advanced approximation algorithm is a good future research direction.

Table 10: Computational Performance on Randomly Generated Instances

$\|I\|$	$\|J\|$	Instance 1	Instance 2	Instance 3	Instance 4	Instance 5	Avg
10	10	19.74	0.27	0.43	1.11	0.38	4.39
	15	11.96	3.03	11.51	12.53	2.64	8.33
15	10	0.38	1.31	1.93	0.64	0.78	1.01
	15	1.01	233.52	65.27	14.4	476.82	158.20
20	10	0.38	0.73	0.44	0.64	0.77	0.59
	15	228.37	37.86	473.07	1044.03	130.19	382.70

4.5 Conclusion

In this chapter, we study two-stage robust bilevel optimization problems under uncertainty. Both theoretical and algorithmic results are derived. The numerical study on two real world applications shows that the proposed algorithm can efficiently solve two-stage robust bilevel optimization instances, and the optimal solutions are robust even under high level of uncertainty.

5.0 Capacity Expansion of Wind Farm in a Market Environment Under Uncertainty

5.1 Motivation

Wind power generation has become a primary clean and sustainable energy in many countries, and investment in wind power facilities is one of the most important decisions in an electricity market as it often involves large amount of capital. In a market environment, investors build wind power generators and gain revenue from the market, which operates to best achieve economic efficiency. Such interactions are often captured through bilevel optimization models [61, 71, 78, 127, 76, 77]. Moreover, there is significant uncertainty involved in wind power generation. On one hand, wind is random and intermittent and thus hard to control and predict; on the other hand, the market desires reliable power to meet the demand. To deal with uncertainty, stochastic bilevel optimization based models are developed in [15, 16, 125]. Nevertheless, due to the limited solution capacity, large number of scenarios can only be considered in small systems since problems of large systems with large number of scenarios are very hard to solve. For example, it takes more than 10 hours for CPLEX to solve an IEEE 118-Bus instance that only has 18 scenarios in [15], and takes about 30 minutes to solve a similar instance that only has 4 scenarios in [125].

To overcome such a big challenge, we propose a novel two-stage robust bilevel optimization model to support wind power investment in an electricity market. The proposed model takes wind power uncertainty into consideration and thus is better able to find a reliable solution. A decomposition algorithm is developed to efficiently solve the two-stage robust bilevel model to global optimal solutions. Numerical experiments show high efficiency of the algorithm and significant benefits by considering wind power uncertainty.

5.2 Bilevel Wind Farm Capacity Expansion Formulation

In an electricity market, the system planner seeks to maximize the profit by investing in wind farms. Such investment decisions are made before the randomness of wind reveals, and market operates after the wind generators are built and wind intensity is determined. Thus, the wind farm capacity expansion problem is indeed a multistage decisions making process. We formulate this decision making process as a two-stage robust bilevel model. Table 11 shows the list of notations, where a regular lower case letter represents a scalar and a bolded letter represents the corresponding vector. A robust wind farm investment problem (RWFIP) is formulated as

$$\text{RWFIP:} \max_{(\mathbf{x},\mathbf{u})\in\mathbf{M}} -\sum_{i\in\Psi}(c_i u_i + h_i x_i) + \min_{\mathbf{k}\in\Omega} \max_{(\mathbf{f},\mathbf{g},\mathbf{q},\mathbf{s},\theta)\in R(\mathbf{x},\mathbf{u},\mathbf{k})} \alpha\beta \sum_{i\in\Psi} q_i, \tag{5.1}$$

where

$$\mathbf{M} = \{(\mathbf{x},\mathbf{u}) : \sum_{i\in\Psi}(C_i u_i + H_i x_i) \leq \hat{C}, u_i \leq \bar{U}_i x_i, \forall i \in \Psi, x_i \in \{0,1\}, u_i \geq 0, \forall i \in \Psi\}, \tag{5.2}$$

and

$$R(\mathbf{x},\mathbf{u},\mathbf{k}) = \arg\min_{\mathbf{f},\mathbf{g},\mathbf{q},\mathbf{s},\theta}\{\sum_{j\in\mathbf{J}}\sum_{b\in\mathbf{B}_j} p_{jb} g_{jb} + \sum_{i\in\mathbf{I}} \rho_i s_i : \tag{5.3}$$

$$\sum_{j\in\mathbf{J}_i}\sum_{b\in\mathbf{B}_j} g_{jb} + q_i + \sum_{l:d(l)=i} f_l + s_i = d_i + \sum_{l:o(l)=i} f_l, \quad \forall i \in I \tag{5.4}$$

$$f_l = S_l(\theta_{o(l)} - \theta_{d(l)}), \quad \forall l \in \mathbf{L} \tag{5.5}$$

$$-\bar{f}_l \leq f_l \leq \bar{f}_l, \quad \forall l \in \mathbf{L} \tag{5.6}$$

$$0 \leq g_{jb} \leq \bar{g}_{jb}, \quad \forall j \in \mathbf{J}, b \in \mathbf{B}_j \tag{5.7}$$

$$0 \leq q_i \leq k_i u_i, \quad \forall i \in \Psi \tag{5.8}$$

$$-\bar{\theta} \leq \theta_i \leq \bar{\theta}, \quad \forall i \in \mathbf{I}\backslash\{r\} \tag{5.9}$$

$$\theta_r = 0 \tag{5.10}$$

$$q_i = 0, \quad \forall i \in I\backslash\Psi \tag{5.11}$$

$$s_i \geq 0, \quad \forall i \in \mathbf{I}\}. \tag{5.12}$$

Table 11: Notation in RWFIP Formulation

Sets	
$\mathbf{B}_j$	set of generator blocks j, $j \in \mathbf{J}$
$\mathbf{I}$	set of buses, indexed by i
$\mathbf{J}$	set of fuel-based generators, indexed by j
$\mathbf{J}_i$	set of fuel-based generators at bus i, $\mathbf{J}_i \subseteq \mathbf{J}$
$\mathbf{L}$	set of lines, indexed by l
Ψ	set of buses eligible for wind farms, $\Psi \subseteq \mathbf{I}$
Ω	the uncertainty set of wind intensity, $k \in \Omega$
$d(l)$	destination bus of transmission line l
$o(l)$	origin bus of transmission line l
r	reference bus
Parameters	
α	hours of the target year
$\hat{C}$	overall budget of wind power investment
d_i	demand at bus i
$\bar{f}_l$	transmission capacity of line l
h_i, c_i	annualized fixed and variable cost of unit wind power generation capacity at i
H_i, C_i	fixed and variable cost of unit wind power generation capacity at i
k_i	wind intensity at bus i
p_{jb}	price offered by generator j in block b_j
$\bar{g}_{jb}$	upper bound of fuel-based generation in b_jth block
S_l	susceptance of line l
$\bar{U}_i$	upper bound of wind farm capacity at bus i
β	weight coefficient for wind power penetration
ρ_i	load shedding penalty cost at bus i, $\rho_i > 0$ for $i \in \mathbf{I}$
$\bar{\theta}$	maximum value of phase angle by generator j
Decision Variables	
x_i	binary variables, 1 if wind power is set at bus i; 0 otherwise
u_i	wind power installation capacity at bus i
f_l	power flow on line l
g_{jb}	power generation in b_jth block by fuel-based generator j
q_i	wind generation at bus i
s_i	load shedding at bus i
θ_i	phase angle at bus i

In the upper level, the investor seeks to maximize the difference between the wind energy absorption and the investment costs by making decisions on the installation and the capacity of wind power generators, subject to capacity constraints and a budget constraint. The lower level problem represents the market clearing conditions by minimizing the total dispatch cost and load shedding, subject to the optimal power flow (OPF) based economic dispatch constraints. In particular, (5.4) and (5.5) are the flow conservation constraints and power balance constraints, respectively. (5.6) - (5.9) introduce bound constraints for power flow, power generation, wind generation, and phase angle, respectively. (5.10) - (5.11) help to set a reference bus, and (5.12) ensures the non-negativity of the load shedding. The wind intensity parameter $\mathbf{k}$ is assumed to be in an uncertainty set Ω. Without loss of generality, we assume Ω is a non-empty compact set.

Remark 5.1. *The lower level problem of RWFIP has an optimal solution for any combination of* $(\mathbf{x}, \mathbf{u}, \mathbf{k})$*. This property is guaranteed by the existence of the load shedding variables.*

5.3 Solution Method

In RWFIP, all the upper level decisions are made before $\mathbf{k}$ is realized, while all the lower level decisions are made after the realization of the parameter. Hence, the lower level decisions are indeed scenario specific decisions. By introducing a set of lower level decision variables for each scenario, we can have an equivalent reformulation of RWFIP.

Let $(\mathbf{f}^{\mathbf{k}}, \mathbf{g}^{\mathbf{k}}, \mathbf{q}^{\mathbf{k}}, \mathbf{s}^{\mathbf{k}}, \theta^{\mathbf{k}})$ denote the lower level decision variables corresponding to $\mathbf{k}$, which represents a particular scenario, we can rewrite RWFIP as

$$\max -\sum_{i\in\Psi}(c_iu_i + h_ix_i) + t \tag{5.13}$$

$$\text{s.t.}(\mathbf{x},\mathbf{u}) \in \mathbf{M} \tag{5.14}$$

$$t \leq \alpha\beta\sum_{i\in\Psi} q_i^{\mathbf{k}}, \quad \forall\mathbf{k}\in\Omega \tag{5.15}$$

$$(\mathbf{f}^{\mathbf{k}}, \mathbf{g}^{\mathbf{k}}, \mathbf{q}^{\mathbf{k}}, \mathbf{s}^{\mathbf{k}}, \theta^{\mathbf{k}}) \in \arg\min\{\sum_{j\in\mathbf{J}}\sum_{b\in\mathbf{B}_j} p_{jb}g_{jb}^{\mathbf{k}} + \sum_{i\in\mathbf{I}}\rho_i s_i^{\mathbf{k}} : \tag{5.16}$$

$$\sum_{j\in\mathbf{J}_i}\sum_{b\in\mathbf{B}_j} g_{jb}^{\mathbf{k}} + q_i^{\mathbf{k}} + \sum_{l:d(l)=i} f_l^{\mathbf{k}} + s_i^{\mathbf{k}} = d_i + \sum_{l:o(l)=i} f_l^{\mathbf{k}}, \quad \forall i\in I \tag{5.17}$$

$$f_l^{\mathbf{k}} = S_l(\theta_{o(l)}^{\mathbf{k}} - \theta_{d(l)}^{\mathbf{k}}), \quad \forall l\in\mathbf{L} \tag{5.18}$$

$$-\bar{f}_l \leq f_l^{\mathbf{k}} \leq \bar{f}_l, \quad \forall l\in\mathbf{L} \tag{5.19}$$

$$0 \leq g_{jb}^{\mathbf{k}} \leq \bar{g}_{jb}, \quad \forall j\in\mathbf{J}, b\in\mathbf{B}_j \tag{5.20}$$

$$0 \leq q_i^{\mathbf{k}} \leq k_iu_i, \quad \forall i\in\Psi \tag{5.21}$$

$$-\bar{\theta} \leq \theta_i^{\mathbf{k}} \leq \bar{\theta}, \quad \forall i\in\mathbf{I}\backslash\{r\} \tag{5.22}$$

$$\theta_r^{\mathbf{k}} = 0 \tag{5.23}$$

$$q_i^{\mathbf{k}} = 0, \quad \forall i\in I\backslash\Psi \tag{5.24}$$

$$s_i^{\mathbf{k}} \geq 0, \quad \forall i\in\mathbf{I}\}, \quad \forall\mathbf{k}\in\Omega. \tag{5.25}$$

For $\bar{\Omega}\subseteq\Omega$,

$$\text{MP:}\eta^* = \max\{-\sum_{i\in\Psi}(c_iu_i + h_ix_i) + t : (5.14), (5.15) - (5.25), \forall\mathbf{k}\in\bar{\Omega}\}$$

is a relaxation of (5.13) - (5.25). We refer this relaxation problem as the master problem (MP), which provides an upper bound of RWFIP. By expanding $\bar{\Omega}$, the upper bound converges to the optimal value of RWFIP. For fixed $(\mathbf{x}^0, \mathbf{u}^0)$, a wind intensity parameter $\mathbf{k}$ as well as an optimal lower level solution can be obtained from the following pessimistic bilevel optimization problem [94]

$$\text{SP:}\zeta(\mathbf{x}^0, \mathbf{u}^0) = \min_{\mathbf{k}\in\Omega}\max\{\alpha\beta\sum_{i\in\Psi} q_i : (\mathbf{f}, \mathbf{g}, \mathbf{q}, \mathbf{s}, \theta) \in R(\mathbf{x}^0, \mathbf{u}^0, \mathbf{k})\},$$

which we refer as the sub-problem (SP), and its optimal solution corresponds to a worst case scenario. The nature of pessimistic bilevel optimization is better able to identify worst case scenarios, and thus to effectively characterize the uncertainty set.

Starting from a subset $\bar{\Omega} \subseteq \Omega$, we can approach to the optimality by adding scenarios identified by SP to $\bar{\Omega}$. Following this idea, we provide the following decomposition algorithm.

Algorithm 5 Decomposition algorithm for solving RWFIP

Initialize: $LB = 0$, $UB = +\infty$, $\bar{\Omega} = \emptyset$, and $it = 0$
while $UB - LB \geq \epsilon$ **do**
  Solve MP for $\bar{\Omega}$ and obtain an optimal solution $(\mathbf{x}^*, \mathbf{u}^*)$
  Update $UB = \eta^*$
  Solve SP for $(\mathbf{x}^*, \mathbf{u}^*)$ and obtain an optimal solution $(\mathbf{k}^*, \mathbf{f}^{\mathbf{k}^*}, \mathbf{g}^{\mathbf{k}^*}, \mathbf{q}^{\mathbf{k}^*}, \mathbf{s}^{\mathbf{k}^*}, \theta^{\mathbf{k}^*})$
  Update $LB = \max\{LB, \zeta(\mathbf{x}^*, \mathbf{u}^*) - \sum_{i \in \Psi}(c_i u_i^* + h_i x_i^*)\}$
  Update $\bar{\Omega} = \bar{\Omega} \cup \{\mathbf{k}^*\}$
  Set $it = it + 1$
end while
Return $(\mathbf{x}^*, \mathbf{u}^*)$ as an optimal solution.

Remark 5.2. *As $R(\mathbf{x}, \mathbf{u}, \mathbf{k}) \neq \emptyset$ for any combination of $(\mathbf{x}, \mathbf{u}, \mathbf{k})$ and the lower level problem of RWFIP is an LP problem, the lower level optimal value function is a piecewise linear function and thus continuous. According to the convergence result in Chapter 4, the continuity of the lower level optimal value function and the compactness of Ω guarantee that the algorithm returns an ϵ-optimal solution of RWFIP in finite number of iterations for any given ϵ.*

Both MP and SP can be further converted to a solver friendly single level problem. Specifically, MP is an optimistic bilevel optimization problem with multiple lower level problems. By replacing each lower level problem with its KKT conditions, we can convert MP to a single level mixed integer programming (MIP) problem, which can be readily solved by commercial solvers. SP is a pessimistic bilevel optimization problem, to which various of solution methods can be applied, such as those in [2, 126, 136, 51, 132]. For example, we can apply the level reduction technique introduced in [132] to obtain a tight relaxation of SP. For fixed

$(\mathbf{x}^0, \mathbf{u}^0)$, the resulting problem is

$$\text{SP'}: \min\{\alpha\beta \sum_{i\in\Psi} q_i : (\bar{\mathbf{f}}, \bar{\mathbf{g}}, \bar{\mathbf{q}}, \bar{\mathbf{s}}, \bar{\theta}) \text{ satisfy } (5.4)-(5.12), \mathbf{k} \in \Omega, (\mathbf{f}, \mathbf{g}, \mathbf{q}, \mathbf{s}, \theta) \in \Phi(\mathbf{x}^0, \mathbf{u}^0)\},$$

where

$$\Phi(\mathbf{x}^0, \mathbf{u}^0) = \arg\max\{\alpha\beta \sum_{i\in\Psi} q_i : (\mathbf{f}, \mathbf{g}, \mathbf{q}, \mathbf{s}, \theta) \text{ satisfy } (5.4)-(5.12),$$
$$\sum_{j\in\mathbf{J}} \sum_{b\in\mathbf{B}_j} p_{jb} g_{jb} + \sum_{i\in\mathbf{I}} \rho_i s_i \leq \sum_{j\in\mathbf{J}} \sum_{b\in\mathbf{B}_j} p_{jb} \bar{g}_{jb} + \sum_{i\in\mathbf{I}} \rho_i \bar{s}_i\}.$$

Same as for MP, we can also convert SP' to a single level MIP problem, which can be efficiently solved by solvers.

5.4 Computational Experiments

The proposed two-stage robust optimization model and the decomposition algorithm are applied to IEEE reliability test system (RTS). Test cases are adopted from [65, 102] with some modifications, and we include RTS-96 24-bus, 57-bus, and 118-bus system in our study. The experiments are implemented using Julia programming language [26], and all the instances are solved by Gurobi [68] with the help of the Julia package BilevelJuMP [62].

Similar as in [24, 19], we assume the wind intensity parameter $\mathbf{k}$ in the form of $\mathbf{k} = \bar{\mathbf{k}} + \sum_{i=1}^{|\Psi|} \triangle\mathbf{k}^i \lambda_i$, where $\bar{\mathbf{k}}$ is the nominal value, $\triangle\mathbf{k}$ is the largest amount of data perturbation, and λ_i are random variables. We take a polyhedral uncertainty set defined by $\Omega = \{\mathbf{k} \in R^{|\Psi|} : \mathbf{k} = \bar{\mathbf{k}} + \sum_{i=1}^{|\Psi|} \triangle\mathbf{k}^i \lambda_i, 0 \leq \lambda_i \leq 1, i = 1, 2, ..., |\Psi|, \sum_{i=1}^{|\Psi|} \lambda_i \leq K\}$ in our computational experiments. The parameter K controls the volume of the uncertainty set, and Ω reduces to a singleton $\{\bar{\mathbf{k}}\}$ for $K = 0$.

For each system, we solve both RWFIP and its deterministic counterpart, which only considers the nominal value. We refer the deterministic model as the "original model". To better reflect the reliability requirement in practice, we evaluate the worst case performance for the two models. In particular, an optimal solution of each model is evaluated through

SP, whose optimal value along with the cost term $\sum_{i\in\Psi}(c_i u_i + h_i x_i)$ represents the worst case performance.

The worst case performance evaluation is reported in Table 12, where "Original" refers to the deterministic model. The worst case optimal value of the upper and lower level problem are denoted by "U" and "L" respectively in the table. We first study the upper level performance evaluation, which is our primary interest. It is observed that RWFIP always provides better solutions, and the performance gap between the two models increases as K increases. We see that RWFIP model provide more than 6%, 30%, and 14% profit than the original model does for the 24-bus, 57-bus, and 118-bus system respectively for $K \geq 3$. Such differences can have huge impact in practice, especially when the wind intensity is expected to vary considerably, i.e., when K is large.

In addition to the great performance in terms of the upper level optimal value, we also notice that RWFIP model provides better lower level solutions in 17 out of 18 cases. Our numerical experiments suggest that considering wind intensity uncertainty not only benefits the system planner but also benefits the whole market as the total dispatch cost of the market is reduced.

Table 12: Worst Case Performance Evaluation

	K	0	1	2	3	4	5
RTS-24	Robust (U)	79844.09	67129.96	58893.39	54974.50	54835.69	54835.69
	Original (U)	79844.09	66724.48	57427.15	51859.18	50890.86	50890.86
RTS-57	Robust (U)	26792.75	20171.78	19481.05	19481.05	19481.05	19481.05
	Original (U)	26792.75	15757.27	14881.26	14881.26	14881.26	14881.26
RTS-118	Robust (U)	51239.17	43218.19	35829.13	35829.13	35829.13	35829.13
	Original (U)	51239.17	40644.98	31908.79	31421.71	31391.35	31391.35
RTS-24	Robust (L)	957687.30	992782.14	976637.98	966507.24	973602.01	973602.01
	Original (L)	957687.30	996627.89	1003588.54	1017494.54	1019974.04	1019974.04
RTS-57	Robust (L)	9092410.68	9187952.49	9210976.83	9210976.83	9210976.83	9210976.83
	Original (L)	9092410.68	9454128.34	9473276.54	9473276.54	9473276.54	9473276.54
RTS-118	Robust (L)	5683685.86	5658006.80	5729231.12	5729231.12	5729231.12	5729231.12
	Original (L)	5683685.86	5633353.93	5748120.45	5749744.05	5749845.24	5749845.24

We notice that the proposed solution method has very strong capacity in solving RWFIP instances. From the left part of Figure 5, we see that all the instances are solved in one minute, and approximate 85% of the instances are solved in just 10 seconds. The right part of Figure 5 illustrates the convergence of the decomposition algorithm on a typical instance. It only takes five iterations to find an optimal solution. This implies that the algorithm can effectively identify crucial scenarios and thus efficiently characterize an optimal solution.

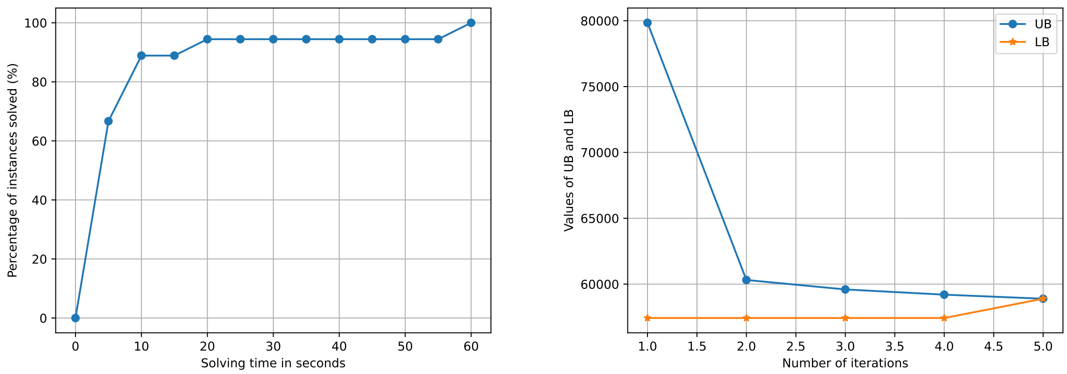

Figure 5: Performance Evaluation of the Solution Method

5.5 Conclusion

In this chapter, we develop a novel two-stage robust bilevel optimization model to investigate the wind farm investment problem in an electricity market under uncertainty. A decomposition algorithm is applied to solve this challenging robust bilevel model. The proposed model and solution method are evaluated on IEEE reliability test systems. The numerical experiments show that both the investors and the electricity market significantly benefit from taking wind intensity uncertainty into consideration. Also, our computational method demonstrates strong solution capacity in solving instances of small to moderate size. We believe it provides a good foundation for future research on more comprehensive large scale problems.

6.0 Data Driven Optimal Decision Trees Considering Local Information

6.1 Motivation and Related Work

Machine learning (ML) models have achieved better-than-human performance in many modern tasks, such as speech recognition, visual object recognition, playing Go [88, 120]. However, most of these techniques are essentially black-box models, i.e., one often has very limited information about how the model makes predictions [112]. The lack of interpretability still limits the potential usages of black-box models in critical domains [113]. ML systems have to be granular in explanation and transparent before trust can be earned [106, 63]. Therefore, there is an urgency in the ML community to develop intrinsic interpretable models [112, 113].

As one of the most classical machine learning models, decision tree enjoys its popularity due to its simplicity, good performance, and especially its interpretability. The traditional decision tree models, e.g., CART [31] and C4.5 [109], generally adopt a greedy and recursive approach to learning a hierarchical model. Different heuristics including different splitting criteria have been proposed, such as Gini Index [31], entropy and information gain [108, 109]. However, as the decision tree presents a hard splitting hierarchy, learning the optimal decision tree is an NP-hard problem in nature [87]. The recursive methods are greedy algorithms, which may lead to sub-optimal solutions. To improve the solution quality, studies on learning optimal decision trees have been proposed in recent years, such as [20, 4, 57]. By formulating the optimization problem using mixed-integer programming (MIP), those methods can solve the optimal classification tree (OCT) problem to optimality, but only for a very limited size (in terms of the number of features and tree depth) problems. The MIP based formulations are good at capturing the non-convex nature of decision trees and thus better able to fit the training data set. However, this advantage of MIP formulations may laed to an OCT that overfits the training data and thus does not perform well on unseen data.

In this chapter, we propose a novel and fully interpretable OCT model as well as an efficient hyperparameter tuning method. The new model learns from both global and local information among samples to improve the generalizability and robustness of an OCT. Our main contributions are summarized as follows.

First, we develop a new mixed-integer programming formulation that takes local information into consideration in addition to the global information. In particular, to improve the robustness of a model, we explicitly consider the local distribution information of each sample in a training set. Such information is then reflected by the sample weight in the loss function through a data-driven approach, reducing the effect of outliers and noises in the training set.

Second, we, for the first time, apply the bilevel optimization framwork to perform hyperparameter tuning for OCT. Such a framework enables us to consider the training set and the validation set simultaneously to maximize the model generalizability.

Third, we propose an efficient decomposition algorithm to solve the resulting bilevel optimization problem for efficient and automatic hyperparameter tuning. Extensive experiments are performed on widely used benchmark datasets in comparison with both conventional recursive heuristics and modern MIP-based OCT algorithms. The experimental results demonstrate the outstanding performance of the proposed OCT model on benchmark datasets.

Learning an optimal decision tree is NP-hard [87]. As a result, this problem has been primarily approached by heuristic methods such as CART [31]. Those greedy methods are simple and efficient, but often lead to sub-optimal solutions, without quality guarantees. In recent years, numerous studies on mathematical programming based decision tree models have been proposed [20, 122, 74, 121, 47, 124, 48, 139, 92, 67].

In this section, we mainly discuss those MIP based OCT learning methods, which are closely related to our work. In [20], the authors proposed an MIP based OCT formulation, which find an optimal branching and labeling decisions in a single formulation rather than making branching decisions at each level like the recursive methods. This formulation leads to higher test accuracy than that from CART for small trees. Later in [122], Verwer and Zhang proposed a binary OCT (BinOCT) formulation, which has fewer variables and constraints compared with the model in [20]. The proposed BinOCT model is solved by MIP solvers, and the numerical results show it runs faster than the one in [20]. In [67], small-size decision trees with categorical data are studied through the integer programming (IP) framework. Similar as in [122], the decision variables are defined based on the features not individual data point. The topology of a tree is an input parameter of the IP formulation. It is selected from a small size candidate pool through a validation set. In [4], the authors propose a big-M free network flow-based MIP formulation to provide strong LP relaxation. By introducing a source node connected to the root node and a sink node connected to all leaf nodes, a classification tree is converted to a directed acyclic graph. We notice that the classical MIP techniques such as cutting planes and dynamic programming are applied in OCT formulation to accelerate the convergence in [92, 47]. Also, hyperplane and soft margin from support vector machines (SVM) are applied to OCT formulation in [139]. The hyperplanes help to achieve higher prediction accuracy at the cost of lower interpretability.

While the majority of the MIP based OCT formulations try to find the best possible classification tree by capturing the global information of a dataset, we propose a novel MIP model that considers both local and global information. In particular, we consider the neighborhood of each sample. Such local information help us to differentiate samples and thus increases the generalizability of the model.

6.2 Problem Formulation

In this section, we present an MIP formulation for OCT. As mentioned previously, an MIP formulation is good at capturing global information to learn ground truth. However, such a powerful feature of MIP formulation may lead to overfitting while being applied to OCT problems. To address this issue, we expand the classical MIP model by taking local information into account. Our basic idea is to assign weight to each sample by considering their neighbourhoods. In particular, we assign a very small weight to an outlier, which has very few or no samples within its neighborhood and thus unlikely to appear in an unseen data set. On the other hand, we assign a large weight to a sample that has many other samples around. Such consideration helps us to focus on representative samples and thus improve the generalizability and robustness of an OCT.

Through this chapter, we use a boldface letter to represent a vector and a regular letter to represent a scalar. For a dataset $(\mathbf{x}_i, y_i)$ for $i \in I$ with the collection of features denoted by F, we assume each feature $f \in F$ as well as the class label y_i are binary, i.e., $(\mathbf{x}_i, y_i) \in \{0, 1\}^{|F|+1}$ for $i \in I$. In fact, categorical and numerical features can be easily binarized via discretization and comparison method [45, 128, 67].

Denote the distance between sample i and sample j by d_{ij}, and assume a distance function $g : F \to R_+$ that maps $(\mathbf{x}_i, \mathbf{x}_j)$ to d_{ij}. For binary valued samples, a typical metric is the Hamming distance, which takes the number of different feature values between two samples as its value. We note that, as the Hamming distance treats all features equally, it does not reflect the reality since some features are more important than others in determining the label of a sample. Hence, we introduce c_f to denote the weight of feature f, with a larger c_f corresponding to a more important feature. For a sample pair $(\mathbf{x}_i, \mathbf{x}_j)$, we can consider d_{ij} as a function of c_f. For example, we can define $d_{ij} = \sum_{f \in F} c_f |x_{if} - x_{jf}|$ as a distance function, which takes the Hamming distance metric as its special case. It is worth to notice that the feature weights are not directly observable and thus need to be learned.

With distance d_{ij} introduced, we further define a function $h(d_{ij}) : R \to R$ to represent the influence of sample j to sample i. To properly reflect the influence, we require the function h to have the following three properties 1) $0 \leq h(d_{ij}) \leq 1$ for i, j; 2) with $h(0) = 1$; 3) h is monotonically decreasing in d. For example, $h(d_{ij}) = \frac{1}{d_{ij}+1}$ and $h(d_{ij}) = 1 - \frac{d_{ij}}{|M|}$ with $M = \max_{i,j \in I}\{d_{ij}\}$ are such a function. We denote the neighborhood of sample i by $B(i, R)$, with parameter R denoting the radius. Let w_i be the weight of sample i in the loss function, then we have $w_i = \sum_{j \in B(i,R)} h(d_{ij}) = \sum_{j \in B(i,R)} h(g(\mathbf{c}))$. Note that once the functions g and h are selected, w_i is determined by the unknown hyperparameters $\mathbf{c}$ and R. We denote the weight of sample i by $w_i(\mathbf{c}, R)$ and assume that both $\mathbf{c}$ and R are given as hyperparameter in this section. We will solve the hyperparameter tuning problem by a data-driven approach in the next section.

A full binary tree of depth K has $2^K - 1$ branching nodes and 2^K leaf nodes. In this chapter, the nodes are ordered according to the breadth-first search method, and the root node is denoted by n_1. We use N_B and N_L to denote the set of branching nodes and that of leaf nodes, respectively. Note that there is an unique path from n_1 to each leaf node, and we define $R(n_l)$ as the set of branching nodes in this unique path to leaf n_l. Moreover, we have parameter a_{if} to indicate if the value of feature f of sample i is 1 and $s_{n_l n_b} \in \{-1, 1\}$ to indicate if node n_l is in the left or the right subtree of node n_b. Specifically, $s_{n_l n_b} = 1$ if n_l is in the left subtree of n_b and -1 otherwise.

For the OCT problem with binary features, the key decision is to configure the structure of the tree, i,e., to assign each branching node a feature and assign each leaf node a label. While most of existing work explicitly makes use of these two sets of variables, we take a slightly different modelling approach.

Specifically, we assign 1 to the leaf node with even number and 0 to those with odd number rather than introducing variables to determine the label of each leaf node. Moreover, we allow the branching nodes that are just one level higher than the leaf nodes in the tree to select features from $\bar{F} = \{\neg f | f \in F\}$ so that all samples can flow to any of the pre-labelled leaf node. We define binary variables $z_{n_b f} \in \{0, 1\}$ to indicate if feature f is selected at node n_b for branching. The full list of notations is shown in Table 13.

For a given $(\mathbf{c}, R)$, the following MIP formulation derives an OCT on the training dataset I_T,

$$L(\mathbf{c}, R) : \min \sum_{i \in I_T} w_i(\mathbf{c}, R) \sum_{n_l \in N_L} u_{n_l i} \tag{6.1}$$

$$\text{s.t.} K - \sum_{n_b \in R(n_l)} \left(\frac{1 - s_{n_l n_b}}{2} + s_{n_l n_b} \sum_{f \in F} a_{if} z_{n_b f}\right)$$

$$+ u_{n_l i} + y_i \geq 1, \forall i \in I_T, n_l \in N_L^1 \tag{6.2}$$

$$K - \sum_{n_b \in R(n_l)} \left(\frac{1 - s_{n_l n_b}}{2} + s_{n_l n_b} \sum_{f \in F} a_{if} z_{n_b f}\right)$$

$$+ u_{n_l i} - y_i \geq 0, \forall i \in I_T, n_l \in N_L^0 \tag{6.3}$$

$$\sum_{f \in F} z_{n_b f} = 1, \quad \forall n_b \in N_B \tag{6.4}$$

$$u_{n_l i}, z_{n_b f} \in \{0, 1\}, \forall n_l \in N_L, n_b \in N_B,$$

$$i \in I_T, f \in F. \tag{6.5}$$

Table 13: Notation in the OCT Formulation

Sets	
I	the index of a dataset, indexed by i
I_V	the index of the validation dataset
I_T	the index of the training dataset
F	the set of features, indexed by f
N_B	the set of branching nodes, indexed by n_b
N_L	the set of leaf nodes, indexed by n_l
N_L^0	the set of leaf nodes that are labelled 0
N_L^1	the set of leaf nodes that are labelled 1
$R(n_l)$	branching nodes in the path from the root node to node n_l
Parameters	
K	the depth of the OCT, $K \geq 0$
a_{if}	indicating if feature f of sample i is 1, $a_{if} \in \{0, 1\}$
$s_{n_l n_b}$	indicating n_l is in the left or right subtree of n_b, $s_{n_l n_b} \in \{-1, 1\}$
c_f	the weight of feature f
R	the radius of the neighborhood
$w_i(\mathbf{c}, R)$	the weight of sample i for given $\mathbf{c}$ and R
Decision Variables	
$z_{n_b f}$	indicating if feature f is selected at node n_b, $z_{n_b f} \in \{0, 1\}$
$u_{n_l i}$	indicating if data point i is correctly classified at n_l, $u_{n_l i} \in \{0, 1\}$

The objective function is to minimize the weighted sum of mis-classified samples. Constraints (6.2) and (6.3) ensure that sample i is correctly classified if and only if it flows into a leaf node whose label is the same as y_i. Specifically, if sample i flows into n_l, then it visits every branching node in $R(n_l)$, we must have $\sum_{n_b \in P(n_l)}(\frac{1-s_{n_l n_b}}{2} + s_{n_l n_b} \sum_{f \in F} a_{if} z_{n_b f}) = K$. To see this, we first consider the case that sample i flows to n_l through the left child of n_b in $R(n_l)$, i.e., $s_{n_l n_b} = 1$. Then there must be at least one feature of sample i with value 1 being selected at node n_b, which implies $\sum_{f \in F} a_{if} z_{n_b f} = 1$. Then we have $\frac{1-s_{n_l n_b}}{2} + s_{n_l n_b} \sum_{f \in F} a_{if} z_{n_b f} = 1$, which in fact holds for all branching node $n_b \in R(n_l)$ with $s_{n_l n_b} = 1$. Similarly, for a branching node n_b with $s_{n_l n_b} = -1$, sample i can flow to n_l only if no feature with value 1 is selected at node n_b. This implies $\sum_{f \in F} a_{if} z_{n_b f} = 0$, which also indicates $\frac{1-s_{n_l n_b}}{2} + s_{n_l n_b} \sum_{f \in F} a_{if} z_{n_b f} = 1$. In other words, $\sum_{n_b \in P(n_l)}(\frac{1-s_{n_l n_b}}{2} + s_{n_l n_b} \sum_{f \in F} a_{if} z_{n_b f}) = K$ if and only if sample i is directed to leaf node n_l. Thus, $u_{n_l i} = 0$ if and only sample i flows into n_l and n_l is labelled with y_i. Finally, Constraints (6.3) ensures that each branching node can only select one feature.

Remark 6.1. *As a special case, if we set $w_i(\mathbf{c}, R) = 1$ for all i, then the proposed OCT model is to minimize the number of mis-classified samples by only considering global information. If, however, $w_i(\mathbf{c}, R)$ is determined by the aforementioned approach, which assigns large weight to representative samples and assigns small weight to outliers, the local information is then taken into consideration.*

6.3 Data-Driven Hyperparameter Tuning

As previously discussed, the hyperparameters $(\mathbf{c}, R)$ play a very crucial role since they provide local information that increases the capacity of the model. In this section, we propose a data-driven hyperparameter tuning method to effectively discover local information.

6.3.1 Bilevel Formulation

In this work, we take a bilevel optimization approach to tune the hyperparameters. The bilevel optimization framework enables us to construct a fully interpretable OCT model, and fully make use of available data.

A bilevel optimization problem defined by (6.6) - (6.9) is indeed an embedded optimization problem. In BiO, (6.6) - (6.8) is the upper level problem with its decision variable $\mathbf{x}$, and (6.9) is the lower level problem with its decision variable $\mathbf{y}$. The decision making process in BiO is in a sequential fashion: the upper level decision maker (often called the leader) determines $\mathbf{x}$ first, the the lower level decision maker (often called the follower) solves (6.9) for a given $\mathbf{x}$. Both the leader and the follower make decision for their own interest. Because of its strong modeling capacity, bilevel optimization has been applied to various problems, including hyperparameter tuning [18, 85, 59].

$$\text{BiO:} \min_{\mathbf{x},\mathbf{y}} \quad F(\mathbf{x}, \mathbf{y}) \tag{6.6}$$

$$\text{s.t.} \quad G(\mathbf{x}) \leq \mathbf{0} \tag{6.7}$$

$$\mathbf{y} \in \phi(\mathbf{x}) \tag{6.8}$$

$$\phi(\mathbf{x}) = \arg\min_{\hat{\mathbf{y}}} \{ f(\mathbf{x}, \hat{\mathbf{y}}) : g(\mathbf{x}, \hat{\mathbf{y}}) \leq \mathbf{0} \} \tag{6.9}$$

For hyperparameter tuning problems, the upper level decision variable $\mathbf{x}$ is the hyperparameter to be tuned, and the lower level decision variable $\mathbf{y}$ is the parameter of a machine learning model to be learned. As discussed in the last section, the hyperparameter we want to tune are the feature weights ($\mathbf{c}$) and radius (R), which then turn to be the decision variables in the upper level problem of the bilevel OCT model. The lower level problem is the OCT model defined by (6.1) - (6.5). Let $I_V \subseteq I$ be a validation dataset and $\phi(\mathbf{c}, R)$ be the optimal solution set of the lower level problem for given $(\mathbf{c}, R)$, then the bilevel OCT formulation is given by (6.10) - (6.15).

$$\text{Bi-OCT:} \min \sum_{i \in I_V} \sum_{n_l \in N_L} \hat{u}_{n_l i} \tag{6.10}$$

$$\text{s.t.} \sum_{f \in F} c_f = 1 \tag{6.11}$$

$$K - \sum_{n_b \in R(n_l)} (\frac{1 - s_{n_l n_b}}{2} + s_{n_l n_b} \sum_{f \in F} a_{if} \hat{z}_{n_b f})$$

$$+ \hat{u}_{n_l i} + y_i \geq 1, \forall i \in I_V, n_l \in N_L^1 \tag{6.12}$$

$$K - \sum_{n_b \in R(n_l)} (\frac{1 - s_{n_l n_b}}{2} + s_{n_l n_b} \sum_{f \in F} a_{if} \hat{z}_{n_b f})$$

$$+ \hat{u}_{n_l i} - y_i \geq 0, \forall i \in I_V, n_l \in N_L^0 \tag{6.13}$$

$$\hat{\mathbf{z}} \in \phi(\mathbf{c}, R) = \arg\min\{\sum_{i \in I_T} w_i(\mathbf{c}, R)$$

$$\sum_{n_l \in N_L} u_{n_l i} : (6.2) - (6.5)\} \tag{6.14}$$

$$R, c_f \geq 0, \hat{u}_{n_l i} \in \{0, 1\}, \forall n_l \in N_L, n_b \in N_B$$

$$i \in I_T, f \in F. \tag{6.15}$$

The objective function of Bi-OCT is to minimize the number of misclassified samples in the validation set. Unlike the objective function in BiO, the samples in the validation set are unweighted. This setting reflects the goal of hyperparameter tuning, i.e., to help the OCT correctly classify unseen data.

Compared with grid search methods, the proposed bilevel model has two main advantages. First, bilevel optimization approach is more computationally friendly as it tunes all parameters at once. As shown in [18], bilevel optimization based hyperparameter tuning method is more favourable than the grid search methods as the number of candidates increases exponentially. Second, it is easier to incorporate domain specific knowledge into the bilevel optimization based hyperparameters tuning process. For example, we add (6.11) in the upper level problem of Bi-OCT. In fact, more sophisticated relationship that may be hard to model in grid search methods could be formulated as constraints in the upper level problem. Next, we present a critical relaxation result.

Proposition 6.1. *Let P be the collection of all possible configurations of an OCT, i.e., $P = \{\mathbf{z} : (6.2) - (6.5)\}$, then*

$$\begin{aligned} &\min \sum_{i \in I_V} \sum_{n_l \in N_L} \hat{u}_{n_l i} \\ &s.t. (6.2) - (6.5), (6.11) - (6.13), (6.15) \\ &\sum_{i \in I_T} w_i(\mathbf{c}, R) \sum_{n_l \in N_L} u_{n_l i} \leq \sum_{i \in I_T} w_i(\mathbf{c}, R) v_i(\mathbf{z}^k), \forall \mathbf{z}^k \in P' \end{aligned} \tag{6.16}$$

is a relaxation of Bi-OCT for any $P' \subseteq P$.

Proof. As $v_i(\mathbf{z}^k) = \sum_{n_l \in N_L} \hat{u}_{n_l i}$ for $i \in I_T$, it indicates whether sample i is correctly classified for a given OCT configuration $\mathbf{z}^k$, and we have $v_i(\mathbf{z}^k) \in \{0, 1\}$ for $i \in I_T$. Moreover, $v_i(\mathbf{z}^k)$ can be calculated once $\mathbf{z}^k$ is fixed. Therefore, for given $(\mathbf{c}, R)$, the optimal value function of the lower level problem can be written as $\Phi(\mathbf{c}, R) = \min_{\mathbf{z}^k \in P}\{\sum_{i \in I_T} w_i(\mathbf{c}, R) v_i(\mathbf{z}^k)\}$ and then we have

$$\begin{aligned} &\min\{\sum_{i \in I_V} \sum_{n_l \in N_L} \hat{u}_{n_l i} : (6.2) - (6.5), (6.11) - (6.13), (6.15), \\ &\qquad \sum_{i \in I_T} w_i(\mathbf{c}, R) \sum_{n_l \in N_L} u_{n_l i} \leq \Phi(\mathbf{c}, R)\} \\ = &\min\{\sum_{i \in I_V} \sum_{n_l \in N_L} \hat{u}_{n_l i} : (6.2) - (6.5), (6.11) - (6.13), (6.15), \\ &\qquad \sum_{i \in I_T} w_i(\mathbf{c}, R) \sum_{n_l \in N_L} u_{n_l i} \leq \sum_{i \in I_T} w_i(\mathbf{c}, R) v_i(\mathbf{z}^k), \forall \mathbf{z}^k \in P\} \\ \leq &\min\{\sum_{i \in I_V} \sum_{n_l \in N_L} \hat{u}_{n_l i} : (6.2) - (6.5), (6.11) - (6.13), (6.15), \\ &\qquad \sum_{i \in I_T} w_i(\mathbf{c}, R) \sum_{n_l \in N_L} u_{n_l i} \leq \sum_{i \in I_T} w_i(\mathbf{c}, R) v_i(\mathbf{z}^k), \forall \mathbf{z}^k \in P'\}. \end{aligned}$$

As the inequality holds for any $P' \subseteq P$, the result follows. □

By making use of this relaxation, we develop an iterative decomposition algorithm to compute Bi-OCT in the following subsection.

6.3.2 Decomposition Algorithm

Bilevel optimization is NP hard even in the linear case [53]. Conventially, it can be reformulated as a single level problem through Karush–Kuhn–Tucker (KKT) conditions if its lower level problem is convex and satisfies some technical conditions. However, this widely used approach is not applicable to Bi-OCT as its lower level problem is an MIP problem, which is highly non-convex. In this section, we develop a decomposition algorithm to tackle this challenge.

According to Proposition 1, for $P' \subseteq P$, we have the following relaxation of Bi-OCT, to which we refer as the master problem (MP).

$$\text{MP:} \quad \eta^* = \min\{\sum_{i\in I_V}\sum_{n_l\in N_L} \hat{u}_{n_l i} : (6.2)-(6.5),(6.11)-(6.13),(6.15)-(6.16)\}$$

MP provides a lower bound (LB) to Bi-OCT, and we can approach the optimality by gradually tightening the LB. This is achieved by expanding P'. After obtaining an optimal solution $(\mathbf{c}^*, R^*)$ from MP, we can solve the lower level problem $L(\mathbf{c}^*, R^*)$ to get an OCT configuration and its optimal value $\Phi(\mathbf{c}^*, R^*)$. To get an upper bound (UB) of Bi-OCT, we solve the following sub-problem (SP), and denote its optimal value by $\zeta(\mathbf{c}^*, R^*)$.

$$\text{SP:} \quad \min\{\sum_{i\in I_V}\sum_{n_l\in N_L} \hat{u}_{n_l i} : (6.2)-(6.5),(6.12)-(6.13), \hat{u}_{n_l i}\in\{0,1\},$$
$$\forall n_l\in N_L, i\in I_T, \sum_{i\in I_T} w_i(\mathbf{c}^*, R^*)\sum_{n_l\in N_L} u_{n_l i} \le \Phi(\mathbf{c}^*, R^*)\}$$

Let $\mathbf{z}^*$ be an optimal solution to SP, we can add $\mathbf{z}^*$ into P' until $UB \le LB$.

Proposition 6.2. *Algorithm 6 converges the global optimum of Bi-OCT in a finite number of iterations.*

Proof. We provide the proof in the appendix. □

Remark 6.2. *For large size datasets, solving Bi-OCT can be very challenging. In this case, we can tune the hyperparameters on a subset of the training dataset, and then train the OCT model for the obtained hyperparameters. Numerical experiments show that this data-driven approach can be efficient and effective.*

Algorithm 6 Decomposition algorithm for solving Bi-OCT

Initialize: $LB = 0$, $UB = |I_V|$, $P' = \emptyset$, and $k = 0$
while $UB > LB$ **do**
  Solve MP for P' and obtain an optimal solution $(\mathbf{c}^*, R^*)$
  Update $LB = \eta^*$
  Solve the lower level problem for $(\mathbf{c}^*, R^*)$ and obtain its optimal value $\Phi(\mathbf{c}^*, R^*)$
  Solve SP for $(\mathbf{c}^*, R^*)$ and obtain an optimal solution $\mathbf{z}^*$
  Update $UB = \min\{UB, \zeta(\mathbf{c}^*, R^*)\}$
  Update $P' = P' \cup \{\mathbf{z}^*\}$
  Set $k = k + 1$
end while
Return $(\mathbf{c}^*, R^*)$ as an optimal solution

6.4 Experiments

In this section, we benchmark the proposed approach with respect to state-of-the-art methods on public datasets.

6.4.1 Experiment Setup

All testing datasets are obtained from the UCI repository [54]. They are first binarized as in [45] and then downsampled so that the number of features is small to moderate. For the comparison purpose, we take a state-of-the-art mathematical programming based package, ”Interpretable AI” (IAI) [20, 21], the random forest (RF) method [104, 33], as well as the popular CART approach as our baseline models. Default settings of IAI, RF, and CART are adopted in our numerical study.

For the proposed method, we practice two implementations: a standard one and a simplified one. In the standard implementation with Bi-OCT, we consider all $(i, j) \in I_T$ to determine radius R. In the simplified version (denoted by Bi-OCT(S)), we define an initial neighborhood for each sample, and only samples in that initial neighborhood are considered in determining R, which helps to reduce the computational burden. Also, we adopt the data driven approach. In particular, we first take 100 samples from the training set as a dataset for solving the bilevel model to get an optimal $(\mathbf{c}, R)$, and then retrain an OCT on the whole training set based on this obtained $(\mathbf{c}, R)$.

As previously mentioned, all the datasets (except for COMPAS-binary and FICO-binary that are already binary featured) are first binarized as in [45]. For each categorical feature $\mathbf{x}_{if} \in \{q_1, q_2, ..., q_f\}$, we create two features for each category. For example, we create $\mathbf{x}_{if_1} = q_1$ and $\mathbf{x}_{if_1} \neq q_1$ for the category q_1. Numerical features are binarized via binning. Note that the number of features may significantly increase after binarization. Hence, we also apply the downsampling technique to each dataset to reduce the number of features. In our experiments, we apply CART with maximum depth 5 to select up to 31 features for each dataset. The data statistics are summarized in Table 14, where #features(O), #features(B), and #features(D) denote the number of features (excluding the label) in the original dataset, after binarization, and after downsampling, respectively.

In MP, the numbers of variables and of constraints increase in the order of $O(|I_T|^2)$. To overcome this computational challenge, we take a data-driven strategy based on a small size subset of the training dataset. Specifically, in our experiments, we first compute w_i by setting $c_f = \frac{1}{|F|}$ for $f \in F$ and $R = 0$. Then, we select 100 samples with the largest weight to form a dataset for hyperparameter tuning, and 25 of them are used as a validation set to populate Bi-OCT. Last, an optimal $(\mathbf{c}, R)$ derived from Bi-OCT is used to solve (6.1) - (6.5) on the original training dataset to learn the final OCT.

In addition to the data-driven approach, we also define an initial neighborhood for each sample to reduce both the number of variables and the number of constraints to build the simplified Bi-OCT(S). The initial neighborhood of sample i is calculated in three steps. First, we let $c_f = \frac{1}{|F|}$ for $f \in F$ and calculate d_{ij} for $i, j \in I_T$. Then we calculate $M = \max_{i,j \in I_T}\{d_{ij}\}$. Finally, we identify $B(i, M * r)$ for each sample as its initial neighborhood, where r is the rate to control the volume of $B(i, M * r)$. It is easy to see that $r = 1$ corresponds to Bi-OCT(S), and we set $r = 0.3$ in our experiments.

6.4.2 Numerical Results

The four methods (including CART(O) that directly handles original features without binarization) are tested on 13 datasets. The depth of a tree is set to be 3 for all methods except for RF, whose maximum depth is set to be 50. Tables 15 and 16 report their training and testing accuracies, as well as standard deviations (in the parentheses), respectively, across 13 datasets. For each dataset, the accuracy is the average over 4-folds tests, and the highest accuracy is highlighted. We include RF as a benchmark baseline, but we do not highlight RF even if it has the highest accuracy since it takes a total different approach. CART was unable to perform classification on the "breast-cancer" dataset, so we report "N/A" in the table. From those numerical results, we have the following observations.

First, both Bi-OCT and Bi-OCT(S) demonstrated strong classification capacities. Bi-OCT(S) ranks 6 times (out of 13) with the highest accuracy, and Bi-OCT ranks 5 times (excluding RF).

Second, Bi-OCT, Bi-OCT(S) and CART(O) have consistent performances over both training and testing stages. In the training stage, both Bi-OCT and Bi-OCT(S) won 5 times first place, CATR(O) 6 times, while IAI and CART fail to won in any datasets. In the testing stage, both Bi-OCT and Bi-OCT(S) maintain their strength, CATR(O) degrades significantly, while IAI and CART actually are not bad if we compare their performances to those in the training stage. From this observation, we believe that the MIP based Bi-OCT and Bi-OCT(S) are better able to learn the ground truths of a dataset.

Third, compared with IAI, another mathematical programming based OCT package, our Bi-OCT and Bi-OCT(S) models are significantly better able to handle challenging datasets. For 5 datasets where IAI's accuracy is lower than 70%, we note that Bi-OCT or Bi-OCT(S) won 4 times. This supports that our Bi-OCT model captures more fundamental structures such as local distribution information.

Fourth, no significant information loss caused by binarization and downsampling is observed from our numerical results. Comparing the result of CART and CART(O) on 10 datasets (excluding breast-cancer, COMPAS-binary, and FICO-binary), we see that CART (CART(O)) won 3 (7) out 10 in terms of the training accuracy and that CART (CART(O)) won 6 (4) out 10 in terms of the test accuracy. This suggests that binarization and down-sampling method could be considered for large dataset in practice.

For the scalability, from Figure 6, we see that our proposed algorithm can generally solve more than 50% (around 50% for Bi-OCT and around 60% for Bi-OCT(S)) instances in 500 seconds. About 80% instances can be solved in half an hour. However, similar as other MIP based OCT models, some cases require much time to solved. While other fast computing methods such as CART can fit a tree in less than a minute for most cases, our methods achieve higher accuracy at the cost of more time spending. Developing strong algorithms that can solve Bi-OCT faster is an important future research direction.

Table 14: Dataset Statistics

Dataset	#samples	#features (O)	#features (B)	#features (D)
breast-cancer	286	9	40	13
heart	303	13	59	23
liver	345	6	53	25
ionosphere	351	34	283	10
WDBC	569	30	270	15
transfusion	748	4	32	21
pima	768	8	67	14
banknote	1372	4	36	17
COMPAS-2016	5020	7	15	14
musk	6598	166	1461	25
COMPAS-binary	6907	12	12	12
FICO	10459	23	156	25
FICO-binary	10459	17	17	15

6.5 Conclusion

In this chapter, we propose a new OCT model that considers local information. Such a task is addressed through a data-driven approach. Specifically, two crucial hyperparameters,

Table 15: Average Training Accuracy

Dataset	Bi-OCT	Bi-OCT (S)	IAI
breast-cancer	**0.8122** (0.0611)	0.8111 (0.0598)	0.7668 (0.0550)
heart	0.8536 (0.0335)	0.8514 (0.0274)	0.8206 (0.0469)
liver	0.7382 (0.0061)	0.7382 (0.0061)	0.7256 (0.0144)
ionosphere	0.9146 (0.0233)	0.9146 (0.0233)	0.9212 (0.0214)
WDBC	0.9151 (0.0063)	0.9156 (0.0064)	0.9098 (0.0094)
transfusion	**0.8039** (0.0251)	0.8026 (0.0258)	0.7843 (0.0262)
pima	0.7513 (0.0236)	0.7526 (0.0215)	0.7522 (0.0228)
banknote	0.9570 (0.0176)	**0.9730** (0.0017)	0.9691 (0.0064)
COMPAS-2016	**0.6708** (0.0035)	**0.6708** (0.0035)	0.6687 (0.0040)
musk	0.9201 (0.0494)	0.9177 (0.0518)	0.9206 (0.0475)
COMPAS-binary	**0.6710** (0.0044)	**0.6710** (0.0044)	0.6699 (0.0045)
FICO	**0.7158** (0.0045)	0.7158 (0.0045)	0.7153 (0.0060)
FICO-binary	**0.7200** (0.0055)	**0.7200** (0.0054)	0.7187 (0.0050)

Dataset	CART	CART(O)	RF
breast-cancer	0.8111 (0.0584)	N/A	0.9160 (0.0173)
heart	0.8492 (0.0265)	**0.8602** (0.0236)	1.0000 (0)
liver	0.7430 (0.0057)	**0.7469** (0.0196)	0.9739 (0.0032)
ionosphere	0.9250 (0.0226)	**0.9288** (0.0134)	0.9582 (0.0103)
WDBC	0.8963 (0.0078)	**0.9731** (0.0053)	0.9438 (0.0043)
transfusion	0.7928 (0.0223)	0.7972 (0.0228)	0.8373 (0.0226)
pima	0.7435 (0.0181)	**0.7669** (0.0092)	0.8477 (0.0084)
banknote	0.9487 (0.0131)	0.9380 (0.0043)	0.9983 (0.0004)
COMPAS-2016	0.6614 (0.0065)	0.6589 (0.0087)	0.6730 (0.0036)
musk	0.9140 (0.0503)	**0.9272** (0.0436)	0.9790 (0.0127)
COMPAS-binary	0.6603 (0.0059)	0.6603 (0.0059)	0.6784 (0.0047)
FICO	0.7109 (0.0011)	0.7104 (0.0013)	0.8303 (0.0079)
FICO-binary	0.7080 (0.0084)	0.7080 (0.0084)	0.7854 (0.0058)

Table 16: Average Test Accuracy

Dataset	Bi-OCT	Bi-OCT (S)	IAI
breast-cancer	0.7543 (0.1536)	0.7437 (0.1663)	0.7610 (0.1826)
heart	**0.7255** (0.1075)	0.6766 (0.0792)	0.6567 (0.0705)
liver	0.6725 (0.0378)	0.6725 (0.0378)	0.6783 (0.0207)
ionosphere	0.8778 (0.0663)	0.8749 (0.0686)	0.8721 (0.0715)
WDBC	0.8911 (0.0256)	0.8981 (0.0226)	0.8893 (0.0267)
transfusion	**0.7901** (0.0727)	**0.7901** (0.0727)	0.7687 (0.1026)
pima	0.6966 (0.0259)	0.6979 (0.0306)	**0.7227** (0.0430)
banknote	0.9570 (0.0174)	**0.9585** (0.0168)	0.9504 (0.0151)
COMPAS-2016	**0.6651** (0.0115)	**0.6651**(0.0115)	0.6612 (0.0140)
musk	0.6940 (0.2175)	0.6925 (0.2154)	0.7057 (0.2569)
COMPAS-binary	**0.6669** (0.0141)	**0.6669** (0.0141)	0.6637 (0.0128)
FICO	**0.6986** (0.0238)	**0.6986** (0.0238)	0.6954 (0.0221)
FICO-binary	0.6998 (0.0253)	0.6998 (0.0253)	**0.7026** (0.0244)
Dataset	CART	CART(O)	RF
breast-cancer	**0.7716** (0.1720)	N/A	0.7402 (0.1551)
heart	0.6393 (0.1789)	0.6627 (0.1228)	0.7058 (0.0866)
liver	**0.6841** (0.0258)	0.6463 (0.0349)	0.7188 (0.0425)
ionosphere	0.8721 (0.0715)	**0.8833** (0.0386)	0.9232 (0.0491)
WDBC	0.8753 (0.0288)	**0.9228** (0.0363)	0.8806 (0.0299)
transfusion	0.7794 (0.0989)	0.7660 (0.1067)	0.7500 (0.0925)
pima	0.7096 (0.0307)	0.7201 (0.0334)	0.7070 (0.0215)
banknote	0.9147 (0.0208)	0.8899 (0.0281)	0.9898 (0.0053)
COMPAS-2016	0.6526 (0.0166)	0.6498 (0.0167)	0.6637 (0.0120)
musk	**0.7204** (0.2266)	0.6866 (0.2588)	0.7164 (0.2508)
COMPAS-binary	0.6576 (0.0116)	0.6576 (0.0116)	0.6673 (0.0114)
FICO	0.6910 (0.0209)	0.6879 (0.0210)	0.6643 (0.0260)
FICO-binary	0.6987 (0.0222)	0.6987 (0.0222)	0.6826 (0.0235)

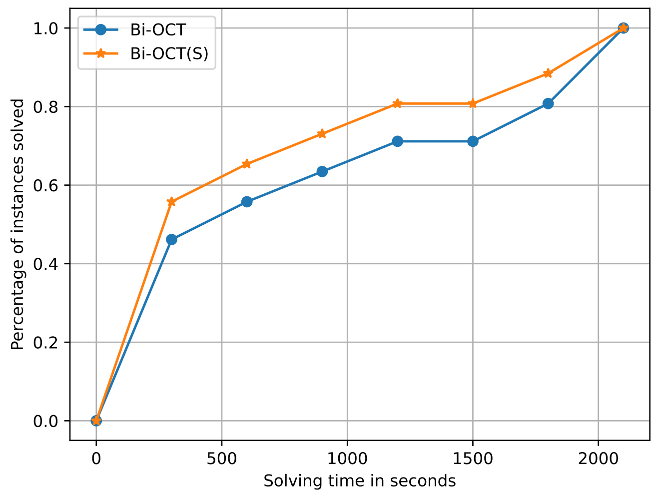

Figure 6: Percentage of Instances Solved Over Time

namely, the feature weight and the radius are introduced to help discover local information of each sample. We then propose a bilevel optimization model to tune the hyperparameters such that the whole model is fully interpretable. A decomposition algorithm is designed to solve the bilevel hyperparameters tuning model globally. The proposed method is evaluated through a comprehensive computational study on various datasets. The superiority of our method over the state-of-the-art is observed. The main limitation of this work is the scalability issue. Hence, a potential future research direction is to develop approximation algorithms for the bilevel hyperparameters tuning problem. Extending this work to other types of decision trees such as regression trees could be another interesting topic.

7.0 Conclusion

In this dissertation, we explore two types of bilevel optimization problems, i.e., bilevel mixed-integer nonlinear programming problem and robust bilevel optimization problem. In addition, we apply the proposed model and solution method to two applications, namely, wind farm capacity expansion problem and optimal decision tree problem.

In Chapter 2, we study general bilevel mixed-integer nonlinear programming problems and provide optimality conditions based reformulation and decomposition algorithm, which is able to solve instances of moderate size efficiently.

Chapter 3 studies single-stage robust bilevel optimization problems with different types of uncertainties. For those with discrete uncertainty set, we provide a strong relaxation, which helps us to develop a novel cut-and-branch algorithm. Numerical study on bilevel facility location problem shows the relaxation provides very strong approximation. However, the cut-and-branch does not work well on large scale instances, and thus more advanced solution methods are a good future research direction.

Chapter 4 studies two-stage robust bilevel optimization problems. Two basic models along with their variations are provided to handle different types of uncertainties. Both theoretical and algorithmic results problems are derived. The numerical study on two real world applications shows that the efficiency of the proposed algorithm .

In Chapter 5, we study a real world wind farm capacity expansion problem. We model such a problem as a two-stage bilevel optimization problem and solve it by a proposed column-and-constraint generation algorithm. The computational study shows that both the investors and the market significantly benefit by considering wind power uncertainty.

In Chapter 6, we study optimal decision tree problem. In particular, we propose a novel mixed-integer programming based formulation that considers both global and local information of a dataset to construct an optimal classification tree. We further take the advantage of the bilevel optimization framework to develop a data-driven hyperparameter tuning approach. Numerical study shows that the proposed model has better generalizability than some state-of-arts method.

Appendix A Computational Study Detail for BiMINLP Problems

In this appendix, we provide the BiMIQP, BiMISOCP, and BiMIBLP formulation used in the numerical study in Chapter 2. We first consider the following BiMIQP problem

$$\begin{aligned}
\text{BiMIQP: } \min\ & \mathbf{c}^T\mathbf{x} + \mathbf{d}^T\mathbf{y} + \mathbf{e}^T\mathbf{z} \\
\text{s.t.} & \mathbf{A}\mathbf{x} \leq \mathbf{b}, \mathbf{x} \in \mathbf{X} \\
& (\mathbf{y}, \mathbf{z}) \in \arg\min\{\frac{1}{2}\mathbf{y}^T\mathbf{Q}_1\mathbf{y} + \frac{1}{2}\mathbf{z}^T\mathbf{Q}_2\mathbf{z} : \\
& \mathbf{G}_1\mathbf{y} + \mathbf{G}_2\mathbf{z} \leq \mathbf{H}\mathbf{x} + \mathbf{f}, \mathbf{y} \in \mathbf{Y}, \mathbf{z} \in \mathbf{Z}\},
\end{aligned}$$

where $\mathbf{G}_1$ and $\mathbf{G}_2$ are positive semidefinite matrix. Similarly, the BiMISOCP problem in the numerical study is

$$\begin{aligned}
\text{BiMISOCP: } \min\ & \mathbf{c}_1^T\mathbf{x} + \mathbf{d}_1^T\mathbf{y} + \mathbf{e}_1^T\mathbf{z} \\
\text{s.t.} & \mathbf{A}\mathbf{x} \leq \mathbf{b}, \mathbf{x} \in \mathbf{X} \\
& (\mathbf{y}, \mathbf{z}) \in \arg\min\{\mathbf{d}_2^T\mathbf{y} + \mathbf{e}_2^T\mathbf{z} : \\
& ||\mathbf{G}_i\mathbf{y} + \mathbf{H}_i\mathbf{z} + \mathbf{f}_i||_2 \leq \mathbf{p}_i^T\mathbf{x} + q_i, i = 1, 2, ..., m, \mathbf{y} \in \mathbf{Y}, \mathbf{z} \in \mathbf{Z}\}.
\end{aligned}$$

By applying Theorem 2.2, we can rewrite BiMIQP and BiMISOCP as a single level problem and solve them by the proposed algorithm. The BiMIBLP problem is

$$\begin{aligned}
\text{BiMIBLP: } \min\ & \mathbf{c}^T\mathbf{x} + \mathbf{d}^T\mathbf{y} + \mathbf{e}_1^T\mathbf{z}_1 + \mathbf{e}_2^T\mathbf{z}_2 \\
\text{s.t.} & \mathbf{A}\mathbf{x} \leq \mathbf{b}, \mathbf{x} \in \mathbf{X} \\
& (\mathbf{y}, \mathbf{z}_1, \mathbf{z}_2) \in \arg\min\{\mathbf{y}^T(\mathbf{z}_1 + \mathbf{z}_2) : \mathbf{G}\mathbf{y} \leq \mathbf{H}\mathbf{x} + \mathbf{f}, \\
& \mathbf{Q}_1\mathbf{z}_1 + \mathbf{Q}_2\mathbf{z}_2 \leq \mathbf{q}, \mathbf{y} \in \mathbf{Y}, \mathbf{z}_1 \in \mathbf{Z}_1, \mathbf{z}_2 \in \mathbf{Z}_2\},
\end{aligned}$$

where $\mathbf{Y} \subseteq \mathbf{R}^{n_1}$, $\mathbf{Z}_1 \subseteq \mathbf{R}^{n_2-m}$, $\mathbf{Z}_2 \subseteq \mathbf{Z}^m$. Although the lower level problem of BiMIBLP is not convex in the continuous variables, the fact that $\mathbf{z}_1$ is independent of $\mathbf{x}$ enables us to apply the proposed solution method. In particular, for fixed $\mathbf{x}^* \in \text{Proj}_{\mathbf{x}}(\Omega_{\text{BiMIBLP}})$, there always exists an optimal solution $(\mathbf{y}^*, \mathbf{z}_1^*, \mathbf{z}_2^*)$ such that $\mathbf{z}_1^*$ is an extreme point of the polyhedron $\{\mathbf{z}_1 \in \mathbf{Z}_1 : \mathbf{Q}_1\mathbf{z}_1 \leq \mathbf{q} - \mathbf{Q}_2\mathbf{z}_2^*\}$. Since $|\mathbf{Z}_2|$ is finite, the number of such polyhedrons are finite, Hence, the solution method is able to solve BiMIBLP in finite number of iterations.

Appendix B Proofs

B.1 Proof of Theorem 4.3

Proof. 1) Using the definition of $\psi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$ and $\phi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$, $\mathbf{TSRBLO_L}$, **TSROBLO**, **TSRPBLO**, and $\mathbf{TSRBLO_U}$ can be respectively rewritten as

$$z_1^* = \min_{\{\mathbf{x}_1 \in \mathbb{X}_1, G(\mathbf{x}_1) \leq \mathbf{0}\}} \max_{\{\mathbf{u} \in \mathbb{U}\}} \min_{\{\mathbf{x}_2 \in \mathbb{X}_2, H(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq \mathbf{0}\}} \min_{\{\mathbf{y} \in \psi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\}} F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}), \tag{B.1}$$

$$z_2^* = \min_{\{\mathbf{x}_1 \in \mathbb{X}_1, G(\mathbf{x}_1) \leq \mathbf{0}\}} \max_{\{\mathbf{u} \in \mathbb{U}\}} \min_{\{\mathbf{x}_2 \in \mathbb{X}_2, H(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq \mathbf{0}\}} \min_{\{\mathbf{y} \in \phi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\}} F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}), \tag{B.2}$$

$$z_3^* = \min_{\{\mathbf{x}_1 \in \mathbb{X}_1, G(\mathbf{x}_1) \leq \mathbf{0}\}} \max_{\{\mathbf{u} \in \mathbb{U}\}} \min_{\{\mathbf{x}_2 \in \mathbb{X}_2, H(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq \mathbf{0}\}} \max_{\{\mathbf{y} \in \phi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\}} F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}), \tag{B.3}$$

$$z_4^* = \min_{\{\mathbf{x}_1 \in \mathbb{X}_1, G(\mathbf{x}_1) \leq \mathbf{0}\}} \max_{\{\mathbf{u} \in \mathbb{U}\}} \min_{\{\mathbf{x}_2 \in \mathbb{X}_2, H(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq \mathbf{0}\}} \max_{\{\mathbf{y} \in \psi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\}} F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}). \tag{B.4}$$

Let $z_1^*(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$, $z_2^*(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$, $z_3^*(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$, and $z_4^*(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$ be the optimal value of (B.1), (B.2), (B.3), and (B.4) respectively for fixed $(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$, then we have

$$z_1^*(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) = \min\{F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) : \mathbf{y} \in \psi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\}, \tag{B.5}$$

$$z_2^*(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) = \min\{F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) : \mathbf{y} \in \phi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\}, \tag{B.6}$$

$$z_3^*(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) = \max\{F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) : \mathbf{y} \in \phi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\}, \tag{B.7}$$

$$z_4^*(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) = \max\{F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) : \mathbf{y} \in \psi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\}. \tag{B.8}$$

It is obvious that $z_2^*(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq z_3^*(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$ for any fixed $(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$, and thus we have $z_2^* \leq z_3^*$. Moreover, since $\phi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \subseteq \psi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$ for any fixed $(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$, (B.5) is a relaxation of (B.6), and (B.8) is a relaxation of (B.7). Hence, $z_1^*(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq z_2^*(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$ and $z_3^*(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq z_4^*(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$ for any fixed $(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$, and we have $z_1^* \leq z_2^*$ and $z_3^* \leq z_4^*$, which imply $z_1^* \leq z_2^* \leq z_3^* \leq z_4^*$.

2) For fixed $(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$, let $P_1(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$, $P_2(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$, $P_3(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$, and $P_4(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$ be the optimal solution set of (B.5), (B.6), (B.7), and (B.8), respectively, then it is sufficient to show that $P_1(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) = P_2(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$ and $P_3(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) = P_4(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$, for any $(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$.

If $\alpha = 0$, then $\phi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) = \psi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$ for any $(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$, and the result holds naturally. Therefore, we focus on the case $\alpha > 0$.

Suppose for a fixed $(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$, $\mathbf{y}_1^* \in P_1(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$, then we have $F(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0, \mathbf{y}_1^*) \leq F(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0, \mathbf{y}), \forall \mathbf{y} \in \psi(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$. To show $\mathbf{y}_1^* \in P_2(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$, it is sufficient to show that $\mathbf{y}_1^* \in \phi(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$ as (B.5) is a relaxation of (B.6). If $\mathbf{y}_1^* \notin \phi(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$, then there exists $\mathbf{y}_2^* \in \psi(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$ such that $f(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0, \mathbf{y}_2^*) < f(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0, \mathbf{y}_1^*)$. As $\alpha > 0$, we have $\frac{1}{\alpha} f(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0, \mathbf{y}_2^*) < \frac{1}{\alpha} f(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0, \mathbf{y}_1^*)$, and thus $F(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0, \mathbf{y}_2^*) < F(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0, \mathbf{y}_1^*)$, which contradicts the fact $\mathbf{y}_1^* \in P_1(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$. Hence, $\mathbf{y}_1^* \in \phi(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$, and thus $\mathbf{y}_1^* \in P_2(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$. As $(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$ is arbitrary, we have $P_1(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \subseteq P_2(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$.

The converse part can be proved through a similar argument. If $\mathbf{y}_2^* \in P_2(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$ for a fixed $(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$, we have $F(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0, \mathbf{y}_2^*) \leq F(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0, \mathbf{y}), \forall \mathbf{y} \in \phi(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$. As $\mathbf{y}_2^* \in \phi(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$, we have $\mathbf{y}_2^* \in \psi(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$. If $\mathbf{y}_2^* \notin P_1(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$, then there exists $\mathbf{y}_1^* \in \psi(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$ such that $F(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0, \mathbf{y}_1^*) < F(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0, \mathbf{y}_2^*)$, which implies $\alpha F(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0, \mathbf{y}_1^*) < \alpha F(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0, \mathbf{y}_2^*)$ and $f(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0, \mathbf{y}_1^*) < f(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0, \mathbf{y}_2^*)$. This contradicts the fact $\mathbf{y}_2^* \in \phi(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$. Therefore, we have $\mathbf{y}_2^* \in P_1(\mathbf{x}_1^0, \mathbf{u}^0, \mathbf{x}_2^0)$ and $P_2(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \subseteq P_1(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$. Since $P_1(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) = P_2(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$ for any fixed $(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$, we have $z_1^* = z_2^*$. Similarly, we have $z_3^* = z_4^*$, and the result follows.

3) The proof of the last statement is simple. In particular, if $\phi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$ is a singleton for any fixed $(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$ whenever it is non-empty, then $P_2(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) = P_3(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$, and thus $z_2^* = z_3^*$. □

B.2 Proof of Theorem 4.5

Proof. 1) If $|\mathbb{U}| < +\infty$, let $|\mathbb{U}| = K_{\mathbb{U}}$, then **TSROBLO** can be rewritten as

$$
\begin{aligned}
&\min \eta \\
&\text{s.t.} G(\mathbf{x}_1) \leq \mathbf{0}, \mathbf{x}_1 \in \mathbb{X}_1 \\
&\quad H(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}) \leq \mathbf{0}, \mathbf{x}_{2_k} \in \mathbb{X}_2, k = 1, 2, ..., K_{\mathbb{U}} \\
&\quad \eta \geq F(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}, \mathbf{y}_k), k = 1, 2, ..., K_{\mathbb{U}} \\
&\quad g(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}, \mathbf{y}_k) \leq \mathbf{0}, \mathbf{y}_k \in \mathbb{Y}, k = 1, 2, ..., K_{\mathbb{U}} \\
&\quad f(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}, \mathbf{y}_k) \leq v(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}), k = 1, 2, ..., K_{\mathbb{U}}.
\end{aligned}
\tag{B.9}
$$

As $\mathbb{X}_{\mathbb{U}} \neq \emptyset$, (B.9) is feasible. Since all the variables are bounded and F, G, H, f, g, and v are continuous, the feasible region of (B.9) is compact. Therefore, by the Weierstrass theorem, **TSROBLO** has an optimal solution. For **TSRPBLO**, a similar argument can be made. In particular, **TSRPBLO** can be rewritten as

$$
\begin{aligned}
&\min \eta \\
&\text{s.t.} G(\mathbf{x}_1) \leq \mathbf{0}, \mathbf{x}_1 \in \mathbb{X}_1 \\
&\quad H(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}) \leq \mathbf{0}, \mathbf{x}_{2_k} \in \mathbb{X}_2, k = 1, 2, ..., K_{\mathbb{U}} \\
&\quad \eta \geq F(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}, \mathbf{y}_k), k = 1, 2, ..., K_{\mathbb{U}} \\
&\quad g(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}, \mathbf{y}_k) \leq \mathbf{0}, \mathbf{y}_k \in \mathbb{Y}, k = 1, 2, ..., K_{\mathbb{U}} \\
&\quad f(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}, \mathbf{y}_k) \leq v(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}), k = 1, 2, ..., K_{\mathbb{U}} \\
&\quad F(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}, \mathbf{y}_k) \geq w_1(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}), k = 1, 2, ..., K_{\mathbb{U}}.
\end{aligned}
\tag{B.10}
$$

Since w_1 is continuous and $\mathbb{X}_{\mathbb{U}} \neq \emptyset$, the feasible region of (B.10) is nonempty and compact, and thus **TSRPBLO** has an optimal solution. We also indicate that for fixed $\mathbf{x}_1 \in \mathbb{X}_{\mathbb{U}}$, by enumerating all possible value of $\mathbf{u} \in \mathbb{U}$, a worst case $\mathbf{u}$ can be obtained. If the cardinality of $\mathbb{X}_1$ is finite, by enumerating all possible value of $\mathbf{x}_1$, an optimal solution can be obtained.

2) As $\mathbb{X}_{\mathbb{U}} \neq \emptyset$, it is sufficient to show that an optimal $\mathbf{x}_1^i$ and $\mathbf{u}^i$ can be obtained from $\mathbf{MP}_1$ and $\mathbf{SP}_1$, respectively. According to 1), we know that an optimal $\mathbf{x}_1^i$ can be obtained

from $\mathbf{MP}_1$ of in each iteration. Moreover, as $\mathbf{u}$ does not affect $\mathbf{x}_2$, we can apply Algorithm 3 to solve $\mathbf{SP}_1$ for fixed $\mathbf{x}_1$. $\mathbf{MP}_1$ in the Kth iteration in this subroutine is

$$\begin{aligned}
&\max_{\mathbf{u}\in\mathbb{U}} \eta \\
&\text{s.t.}\eta \le F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_{2_k}, \mathbf{y}_k), k = 1, 2, ..., K \\
&\quad g(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_{2_k}, \mathbf{y}_k) \le \mathbf{0}, \mathbf{y}_k \in \mathbb{Y}, k = 1, 2, ..., K \\
&\quad f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_{2_k}, \mathbf{y}_k) \le v(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}), k = 1, 2, ..., K \\
&\quad F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_{2_k}, \mathbf{y}_k) \le w_2(\mathbf{x}_1, \mathbf{u}_k, \mathbf{x}_{2_k}), k = 1, 2, ..., K,
\end{aligned}$$

which has an optimal solution due to the compactness of its feasible region and the continuity of the objective function. The sub-problem in this subroutine is (4.12), which has an optimal solution by our assumption. Therefore, we can apply Theorem 4.2 to conclude that $\mathbf{u}^i$ can be obtained from $\mathbf{SP}_1$, and thus complete the proof. For **TSRPBLO**, a very similar argument can be made, and thus we just give a simplified proof. In particular, the master problem of **TSRPBLO** is a special case of (B.10), and thus has an optimal solution. Additionally, the master problem in the subroutine for solving $\mathbf{SP}_2$ is

$$\begin{aligned}
&\max_{\mathbf{u}\in\mathbb{U}} \eta \\
&\text{s.t.}\eta \le F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_{2_k}, \mathbf{y}_k), k = 1, 2, ..., K \\
&\quad g(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_{2_k}, \mathbf{y}_k) \le \mathbf{0}, \mathbf{y}_k \in \mathbb{Y}, k = 1, 2, ..., K \\
&\quad f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_{2_k}, \mathbf{y}_k) \le v(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_{2_k}), k = 1, 2, ..., K,
\end{aligned}$$

which has an optimal solution due to the same reason as for **TSROBLO**, and the sub-problem in the subroutine is (4.13), which has an optimal solution by our assumption. Therefore, $\mathbf{u}^i$ can be obtained from $\mathbf{SP}_2$, and the result follows.

3) Without loss of generality, we assume $\mathbf{y} \in \mathbb{R}^{n_c}$ and $\mathbf{y}_2 \in \mathbb{R}^{m_{2c}}$. Then same as for 2), it is sufficient to show that an optimal $\mathbf{x}_1^i$ and $\mathbf{u}^i$ can be obtained from $\mathbf{MP}_1$ and $\mathbf{SP}_1$, respectively. It is easy to verify that the master problem has an optimal $\mathbf{x}_1^i$ in each iteration as the it is identical to (B.9). For the sub-prboem, the lower level problem is a convex optimization problem with Slater's condition satisfied, and thus the KKT conditions are

necessary and sufficient for optimality. Hence, $\mathbf{y} \in \phi(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$ if and only if there exists $(\mathbf{y}, \mathbf{z})$ that satisfies the KKT conditions, i.e. $(\mathbf{y}, \mathbf{z}) \in M(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)$, where

$$
\begin{aligned}
M(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) = \{(\mathbf{y}, \mathbf{z}) : &\nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) + \sum_{i=1}^{q} z_i \nabla_{\mathbf{y}} g_i(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) = 0 \\
& g_i(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) \leq \mathbf{0}, i = 1, 2, ..., q \\
& z_i g_i(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) = 0, i = 1, 2, ..., q\}.
\end{aligned}
$$

Then the SP can be rewritten as

$$
\begin{aligned}
\max_{\mathbf{u} \in \mathbb{U}} \min_{\mathbf{x}_2, \bar{\mathbf{y}}} \{ & F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \bar{\mathbf{y}}) : H(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq \mathbf{0}, g(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \bar{\mathbf{y}}) \leq \mathbf{0}, \\
& f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \bar{\mathbf{y}}) \leq f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}), (\mathbf{y}, \mathbf{z}) \in M(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2)\}
\end{aligned}
$$

As f is separable, $f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \bar{\mathbf{y}}) - f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y})$ is convex in $\bar{\mathbf{y}}$. Assuming the problem $\min_{\mathbf{x}_2, \bar{\mathbf{y}}} \{F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \bar{\mathbf{y}}) : H(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq \mathbf{0}, g(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \bar{\mathbf{y}}) \leq \mathbf{0}, f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \bar{\mathbf{y}}) \leq f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y})\}$ satisfies Slater's condition for any fixed $(\mathbf{x}_1, \mathbf{u}, \mathbf{y})$ and denoting the ith constraint of $H(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq \mathbf{0}$ by $H_i(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq 0$ and the jth constraint of $g(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq \mathbf{0}$ by $g_j(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq 0$, respectively, we can further rewrite the sub-problem as

$$
\begin{aligned}
& \max_{\mathbf{u} \in \mathbb{U}} F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \bar{\mathbf{y}}) \\
& \text{ s.t.} \nabla_{\mathbf{x}_2} F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \bar{\mathbf{y}}) + \sum_i \lambda_i \nabla_{\mathbf{x}_2} H_i(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) + \sum_j \pi_j \nabla_{\mathbf{x}_2} g_j(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \bar{\mathbf{y}}) \\
& \quad + \mu(\nabla_{\mathbf{x}_2} f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \bar{\mathbf{y}}) - \nabla_{\mathbf{x}_2} f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y})) = 0 \\
& \quad \nabla_{\bar{\mathbf{y}}} F(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \bar{\mathbf{y}}) + \sum_j \pi_j \nabla_{\bar{\mathbf{y}}} g_j(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \bar{\mathbf{y}}) + \mu \nabla_{\bar{\mathbf{y}}} f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \bar{\mathbf{y}}) = 0 \\
& \quad H_i(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \leq 0, \forall i \\
& \quad g_j(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \bar{\mathbf{y}}) \leq 0, \forall j \\
& \quad f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \bar{\mathbf{y}}) \leq f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y}) \\
& \quad \lambda_i H_i(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) = 0, \forall i \\
& \quad \pi_j g_j(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \bar{\mathbf{y}}) = 0, \forall j \\
& \quad \mu(f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \bar{\mathbf{y}}) - f(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2, \mathbf{y})) = 0 \\
& \quad (\mathbf{y}, \mathbf{z}) \in M(\mathbf{x}_1, \mathbf{u}, \mathbf{x}_2) \\
& \quad \lambda_i \geq 0, \forall i, \pi_j \geq 0, \forall j, \mu \geq 0.
\end{aligned}
\tag{B.11}
$$

By enumerating all possible combination of the complementary slackness constraints, an optimal solution of (B.11) can be obtained since (4.14) has an optimal solution, and we can apply Theorem 4.2 to complete the proof. In fact, it is easy to verify that a linear TSROBLO formulation, which is often used in practice, satisfies all the conditions in 3), and thus has an optimal $\mathbf{x}_1$. □

B.3 Proof of Proposition 6.2

Proof. Since the cardinality of P is bounded by $(2^K - 1) * |F|$, it is sufficient to show that a repeated $\mathbf{z}^k$ obtained from SP leads to $UB = LB$. Suppose that we get $(\mathbf{c}^*, R^*)$, η^*, $\Phi(\mathbf{c}^*, R^*)$, $\mathbf{z}^*$, $\zeta(\mathbf{c}^*, R^*)$, and P^k from the respective step in the k-th iteration and that $\mathbf{z}^*$ has been obtained in a previously iteration, then it is easy to verify that P^k does not change. Hence, in the $(k+1)$-th iteration, LB does not change either. Therefore, we have

$$
\begin{aligned}
LB = & \min\{\sum_{i \in I_V} \sum_{n_l \in N_L} \hat{u}_{n_l i} : (6.2)-(6.5), (6.12)-(6.13), \hat{u}_{n_l i} \in \{0,1\}, \forall n_l \in N_L, i \in I_T, \\
& \qquad \sum_{i \in I_T} w_i(\mathbf{c}^k, R^*) \sum_{n_l \in N_L} u_{n_l i} \leq \sum_{i \in I_T} w_i(\mathbf{c}^*, R^*) v_i(\mathbf{z}^k), \forall \mathbf{z}^k \in P^k\} \\
& \geq \min\{\sum_{i \in I_V} \sum_{n_l \in N_L} \hat{u}_{n_l i} : (6.2)-(6.5), (6.12)-(6.13), \hat{u}_{n_l i} \in \{0,1\}, \forall n_l \in N_L, i \in I_T \\
& \qquad \sum_{i \in I_T} w_i(\mathbf{c}^*, R^*) \sum_{n_l \in N_L} u_{n_l i} \sum_{i \in I_T} w_i(\mathbf{c}^*, R^*) v_i(\mathbf{z}^*)\} \\
& = \min\{\sum_{i \in I_V} \sum_{n_l \in N_L} \hat{u}_{n_l i} : (6.2)-(6.5), (6.12)-(6.13), \hat{u}_{n_l i} \in \{0,1\}, \forall n_l \in N_L, i \in I_T \\
& \qquad \sum_{i \in I_T} w_i(\mathbf{c}^*, R^*) \sum_{n_l \in N_L} u_{n_l i} \leq \Phi(\mathbf{c}^*, R^*)\} \\
& \geq UB.
\end{aligned}
$$

The first inequality holds since $k - 1$ constraints are dropped, resulting in a relaxation problem. The second equality holds due to the fact that $\mathbf{z}^*$ is optimal to SP for $(\mathbf{c}^*, R^*)$. Since $LB \geq UB$ in the $(k+1)$-th iteration, the algorithm terminates and returns $(\mathbf{c}^*, R^*)$ as optimal hyperparameters. □

Bibliography

[1] Abdelmalek Aboussoror and Pierre Loridan. Strong-weak stackelberg problems in finite dimensional spaces. *Serdica Mathematical Journal*, 21(2):151–170, 1995.

[2] Abdelmalek Aboussoror and Abdelatif Mansouri. Weak linear bilevel programming problems: existence of solutions via a penalty method. *Journal of Mathematical Analysis and Applications*, 304(1):399–408, 2005.

[3] Claire S Adjiman, Ioannis P Androulakis, and Christodoulos A Floudas. Global optimization of mixed-integer nonlinear problems. *AIChE Journal*, 46(9):1769–1797, 2000.

[4] Sina Aghaei, Andrés Gómez, and Phebe Vayanos. Strong optimal classification trees. *arXiv preprint arXiv:2103.15965*, 2021.

[5] SM Alizadeh, P Marcotte, and G Savard. Two-stage stochastic bilevel programming over a transportation network. *Transportation Research Part B: Methodological*, 58:92–105, 2013.

[6] Mahyar A Amouzegar and Khosrow Moshirvaziri. Determining optimal pollution control policies: An application of bilevel programming. *European Journal of Operational Research*, 119(1):100–120, 1999.

[7] Yu An and Bo Zeng. Exploring the modeling capacity of two-stage robust optimization: Variants of robust unit commitment model. *IEEE transactions on Power Systems*, 30(1):109–122, 2015.

[8] Maciek R Antoniewicz, David F Kraynie, Lisa A Laffend, Joanna González-Lergier, Joanne K Kelleher, and Gregory Stephanopoulos. Metabolic flux analysis in a nonstationary system: fed-batch fermentation of a high yielding strain of e. coli producing 1, 3-propanediol. *Metabolic engineering*, 9(3):277–292, 2007.

[9] MOSEK ApS. *MOSEK Modeling Cookbook*, 2023.

[10] Charles Audet, Pierre Hansen, Brigitte Jaumard, and Gilles Savard. Links between linear bilevel and mixed 0–1 programming problems. *Journal of optimization theory and applications*, 93(2):273–300, 1997.

[11] Didier Aussel and Anton Svensson. Is pessimistic bilevel programming a special case of a mathematical program with complementarity constraints? *Journal of Optimization Theory and Applications*, 181(2):504–520, 2019.

[12] Anjali Awasthi, Dominique Breuil, Satyaveer Singh Chauhan, Michel Parent, and Thierry Reveillere. A multicriteria decision making approach for carsharing stations selection. *Journal of decision systems*, 16(1):57–78, 2007.

[13] Hannah Bakker, Fabian Dunke, and Stefan Nickel. A structuring review on multi-stage optimization under uncertainty: Aligning concepts from theory and practice. *Omega*, 96:102080, 2020.

[14] Jonathan F Bard. *Practical bilevel optimization: algorithms and applications*, volume 30. Springer Science & Business Media, 2013.

[15] Luis Baringo and Antonio J Conejo. Wind power investment: A benders decomposition approach. *IEEE Transactions on Power Systems*, 27(1):433–441, 2011.

[16] Luis Baringo and Antonio J Conejo. Transmission and wind power investment. *IEEE transactions on power systems*, 27(2):885–893, 2012.

[17] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton University Press, 2009.

[18] Kristin P Bennett, Jing Hu, Xiaoyun Ji, Gautam Kunapuli, and Jong-Shi Pang. Model selection via bilevel optimization. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 1922–1929. IEEE, 2006.

[19] Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

[20] Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Machine Learning*, 106(7):1039–1082, 2017.

[21] Dimitris Bertsimas, Jack Dunn, and Nishanth Mundru. Optimal prescriptive trees. *INFORMS Journal on Optimization*, 1(2):164–183, 2019.

[22] Dimitris Bertsimas, Iain Dunning, and Miles Lubin. Reformulation versus cutting-planes for robust optimization. *Computational Management Science*, 13(2):195–217, 2016.

[23] Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations research*, 52(1):35–53, 2004.

[24] Dimitris Bertsimas and Melvyn Sim. Tractable approximations to robust conic optimization problems. *Mathematical programming*, 107(1-2):5–36, 2006.

[25] Hans-Georg Beyer and Bernhard Sendhoff. Robust optimization–a comprehensive survey. *Computer methods in applied mechanics and engineering*, 196(33-34):3190–3218, 2007.

[26] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

[27] John R Birge and Francois Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

[28] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[29] Jerome Bracken and James T McGill. Mathematical programs with optimization problems in the constraints. *Operations Research*, 21(1):37–44, 1973.

[30] Jerome Bracken and James T McGill. Defense applications of mathematical programs with optimization problems in the constraints. *Operations Research*, 22(5):1086–1096, 1974.

[31] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[32] Christoph Buchheim and Dorothee Henke. The bilevel continuous knapsack problem with uncertain follower's objective.

[33] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API

design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[34] Anthony P Burgard, Priti Pharkya, and Costas D Maranas. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering*, 84(6):647–657, 2003.

[35] Johanna Burtscheidt, Matthias Claus, and Stephan Dempe. Risk-averse models in bilevel stochastic linear programming. *SIAM Journal on Optimization*, 30(1):377–406, 2020.

[36] Wilfred Candler and Roger Norton. *Multi-level programming and development policy.* The World Bank, 1977.

[37] Dong Cao and Mingyuan Chen. Capacitated plant selection in a decentralized manufacturing environment: a bilevel optimization approach. *European Journal of Operational Research*, 169(1):97–110, 2006.

[38] Dong Cao and Lawrence C Leung. A partial cooperation model for non-unique linear two-level decision problems. *European Journal of Operational Research*, 140(1):134–141, 2002.

[39] Miguel Carrión, José M Arroyo, and Antonio J Conejo. A bilevel stochastic programming approach for retailer futures market trading. *IEEE Transactions on Power Systems*, 24(3):1446–1456, 2009.

[40] M Červinka, C Matonoha, and Jirí V Outrata. On the computation of relaxed pessimistic solutions to mpecs. *Optimization Methods and Software*, 28(1):186–206, 2013.

[41] Snorre Christiansen, Michael Patriksson, and Laura Wynter. Stochastic bilevel programming in structural optimization. *Structural and multidisciplinary optimization*, 21(5):361–371, 2001.

[42] Thai Doan Chuong and Vaithilingam Jeyakumar. Finding robust global optimal values of bilevel polynomial programs with uncertain linear constraints. *Journal of Optimization Theory and Applications*, 173(2):683–703, 2017.

[43] Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Annals of operations research*, 153(1):235–256, 2007.

[44] John M Danskin. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664, 1966.

[45] Sanjeeb Dash, Oktay Gunluk, and Dennis Wei. Boolean decision rules via column generation. *Advances in neural information processing systems*, 31, 2018.

[46] Gonçalo Homem de Almeida Correia and António Pais Antunes. Optimization approach to depot location and trip selection in one-way carsharing systems. *Transportation Research Part E: Logistics and Transportation Review*, 48(1):233–247, 2012.

[47] Emir Demirović, Anna Lukina, Emmanuel Hebrard, Jeffrey Chan, James Bailey, Christopher Leckie, Kotagiri Ramamohanarao, and Peter J Stuckey. Murtree: Optimal decision trees via dynamic programming and search. *Journal of Machine Learning Research*, 23(26):1–47, 2022.

[48] Emir Demirovic and Peter J Stuckey. Optimal decision trees for nonlinear metrics. *CoRR, abs/2009.06921*, 2020.

[49] Stephan Dempe. *Foundations of bilevel programming*. Springer Science & Business Media, 2002.

[50] Stephan Dempe and Susanne Franke. On the solution of convex bilevel optimization problems. *Computational Optimization and Applications*, 63(3):685–703, 2016.

[51] Stephan Dempe, G Luo, and Susanne Franke. Pessimistic bilevel linear optimization. *Journal of Nepal Mathematical Society*, 1(1):1–10, 2018.

[52] Stephan Dempe, Boris S Mordukhovich, and Alain Bertrand Zemkoho. Necessary optimality conditions in pessimistic bilevel programming. *Optimization*, 63(4):505–533, 2014.

[53] Xiaotie Deng. Complexity issues in bilevel linear programming. In *Multilevel Optimization: Algorithms and Applications*, pages 149–164. Springer, 1998.

[54] Dheeru Dua, Casey Graff, et al. Uci machine learning repository. 2017.

[55] Thomas A Edmunds and Jonathan F Bard. An algorithm for the mixed-integer nonlinear bilevel programming problem. *Annals of Operations Research*, 34(1):149–162, 1992.

[56] Nuno P Faísca, Vivek Dua, Berç Rustem, Pedro M Saraiva, and Efstratios N Pistikopoulos. Parametric global optimisation for bilevel programming. *Journal of Global Optimization*, 38(4):609–623, 2007.

[57] Murat Firat, Guillaume Crognier, Adriana F Gabor, Cor AJ Hurkens, and Yingqian Zhang. Column generation based heuristic for learning classification trees. *Computers & Operations Research*, 116:104866, 2020.

[58] Matteo Fischetti and Michele Monaci. Cutting plane versus compact formulations for uncertain (integer) linear programs. *Mathematical Programming Computation*, 4(3):239–273, 2012.

[59] Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pages 1568–1577. PMLR, 2018.

[60] Virginie Gabrel, Cécile Murat, and Aurélie Thiele. Recent advances in robust optimization: An overview. *European journal of operational research*, 235(3):471–483, 2014.

[61] Lina P Garcés, Antonio J Conejo, Raquel García-Bertrand, and Rubén Romero. A bilevel approach to transmission expansion planning within a market environment. *Power Systems, IEEE Transactions on*, 24(3):1513–1522, 2009.

[62] Joaquim Dias Garcia, Guilherme Bodin, and Alexandre Street. Bileveljump. jl: Modeling and solving bilevel optimization in julia. *arXiv preprint arXiv:2205.02307*, 2022.

[63] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 227–236, 2008.

[64] Vikas Goel and Ignacio E Grossmann. A class of stochastic programs with decision dependent uncertainty. *Mathematical programming*, 108(2-3):355–394, 2006.

[65] Cliff Grigg, Peter Wong, Paul Albrecht, Ron Allan, Murty Bhavaraju, Roy Billinton, Quan Chen, Clement Fong, Suheil Haddad, Sastry Kuruganty, et al. The ieee reliability test system-1996. a report prepared by the reliability test system task force of the application of probability methods subcommittee. *IEEE Transactions on power systems*, 14(3):1010–1020, 1999.

[66] Zeynep H Gümüş and Christodoulos A Floudas. Global optimization of mixed-integer bilevel programming problems. *Computational Management Science*, 2(3):181–212, 2005.

[67] Oktay Günlük, Jayant Kalagnanam, Minhan Li, Matt Menickelly, and Katya Scheinberg. Optimal decision trees for categorical data via integer programming. *Journal of global optimization*, 81(1):233–260, 2021.

[68] Inc. Gurobi Optimization. Gurobi optimizer reference manual, 2016.

[69] Walter J Gutjahr and Nada Dzubur. Bi-objective bilevel optimization of distribution center locations considering user equilibria. *Transportation Research Part E: Logistics and Transportation Review*, 85:1–22, 2016.

[70] Charlotte Henkel. *An algorithm for the global resolution of linear stochastic bilevel programs*. PhD thesis, 2014.

[71] MR Hesamzadeh, Nasser Hosseinzadeh, and Peter J Wolfs. A bi-level formulation of transmission planning problem in liberalised electricity markets. *Australian Journal of Electrical and Electronics Engineering*, 8(2):119–128, 2011.

[72] Benjamin F Hobbs, Carolyn B Metzler, and J-S Pang. Strategic gaming analysis for electric power systems: An mpec approach. *IEEE transactions on power systems*, 15(2):638–645, 2000.

[73] Hui Hu. A one-phase algorithm for semi-infinite linear programming. *Mathematical programming*, 46(1):85–103, 1990.

[74] Xiyang Hu, Cynthia Rudin, and Margo Seltzer. Optimal sparse decision trees. *Advances in Neural Information Processing Systems*, 32, 2019.

[75] Sergey Valerievich Ivanov. A bilevel stochastic programming problem with random parameters in the follower's objective function. *Journal of Applied and Industrial Mathematics*, 12(4):658–667, 2018.

[76] Masoud Jenabi, Seyyed Mohammad Taghi Fatemi Ghomi, and Yves Smeers. Bi-level game approaches for coordination of generation and transmission expansion planning within a market environment. *IEEE Transactions on Power systems*, 28(3):2639–2650, 2013.

[77] Shan Jin and Sarah M Ryan. Capacity expansion in the integrated supply network for an electricity market. *IEEE Transactions on Power Systems*, 26(4):2275–2284, 2011.

[78] S Jalal Kazempour, Antonio J Conejo, and Carlos Ruiz. Strategic generation investment using a complementarity approach. *IEEE Transactions on Power Systems*, 26(2):940–948, 2010.

[79] Polyxeni-M Kleniati and Claire S Adjiman. Branch-and-sandwich: a deterministic global optimization algorithm for optimistic bilevel programming problems. part ii: Convergence analysis and numerical results. *Journal of Global Optimization*, 60(3):459–481, 2014.

[80] Polyxeni-M Kleniati and Claire S Adjiman. A generalization of the branch-and-sandwich algorithm: From continuous to mixed-integer nonlinear bilevel problems. *Computers & Chemical Engineering*, 72:373–386, 2015.

[81] Polyxeni-Margarita Kleniati and Claire S Adjiman. Branch-and-sandwich: a deterministic global optimization algorithm for optimistic bilevel programming problems. part i: Theoretical development. *Journal of Global Optimization*, 60(3):425–458, 2014.

[82] Kenneth O Kortanek and Hoon No. A central cutting plane algorithm for convex semi-infinite programming problems. *SIAM Journal on optimization*, 3(4):901–918, 1993.

[83] Raimund M Kovacevic and Georg Ch Pflug. Electricity swing option pricing by stochastic bilevel optimization: A survey and new approaches. *European Journal of Operational Research*, 237(2):389–403, 2014.

[84] Steven G Krantz and Harold R Parks. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media, 2012.

[85] Gautam Kunapuli, Kristin P Bennett, Jing Hu, and Jong-Shi Pang. Bilevel model selection for support vector machines. *Data mining and mathematical programming*, 45:129–158, 2008.

[86] Lorenzo Lampariello, Simone Sagratella, and Oliver Stein. The standard pessimistic bilevel problem. *SIAM Journal on Optimization*, 29(2):1634–1656, 2019.

[87] Hyafil Laurent and Ronald L Rivest. Constructing optimal binary decision trees is np-complete. *Information processing letters*, 5(1):15–17, 1976.

[88] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[89] George Leitmann. On generalized stackelberg strategies. *Journal of optimization theory and applications*, 26(4):637–643, 1978.

[90] Anna CY Li, Ningxiong Xu, Linda Nozick, and Rachel Davidson. Bilevel optimization for integrated shelter location analysis and transportation planning for hurricane events. *Journal of Infrastructure Systems*, 17(4):184–192, 2011.

[91] Gui-Hua Lin and Masao Fukushima. Stochastic equilibrium problems and stochastic mathematical programs with equilibrium constraints: A survey. *Pacific Journal of Optimization*, 6(3):455–482, 2010.

[92] Jimmy Lin, Chudi Zhong, Diane Hu, Cynthia Rudin, and Margo Seltzer. Generalized and scalable optimal sparse decision trees. In *International Conference on Machine Learning*, pages 6150–6160. PMLR, 2020.

[93] June Liu, Yuxin Fan, Zhong Chen, and Yue Zheng. Pessimistic bilevel optimization: a survey. *International Journal of Computational Intelligence Systems*, 11(1):725–736, 2018.

[94] Pierre Loridan and Jacqueline Morgan. Weak via strong stackelberg problem: new results. *Journal of global Optimization*, 8(3):263–287, 1996.

[95] Luis M Martinez, Luís Caetano, Tomás Eiró, and Francisco Cruz. An optimisation algorithm to establish the location of stations of a mixed fleet biking system: an application to the city of lisbon. *Procedia-Social and Behavioral Sciences*, 54:513–524, 2012.

[96] Alexander Mitsos. *Man-portable power generation devices: product design and supporting algorithms*. PhD thesis, Massachusetts Institute of Technology, 2006.

[97] Alexander Mitsos. Global solution of nonlinear mixed-integer bilevel programs. *Journal of Global Optimization*, 47(4):557–582, 2010.

[98] DA Molodtsov. The solution of a class of non-antagonistic games. *USSR Computational Mathematics and Mathematical Physics*, 16(6):67–72, 1976.

[99] James T Moore and Jonathan F Bard. The mixed integer linear bilevel programming problem. *Operations research*, 38(5):911–921, 1990.

[100] Rahul Nair and Elise Miller-Hooks. Equilibrium network design of shared-vehicle systems. *European Journal of Operational Research*, 235(1):47–61, 2014.

[101] Omid Nohadani and Kartikey Sharma. Optimization under decision-dependent uncertainty. *SIAM Journal on Optimization*, 28(2):1773–1795, 2018.

[102] Christos Ordoudis, Pierre Pinson, Juan Miguel Morales González, and Marco Zugno. *An Updated Version of the IEEE RTS 24-Bus System for Electricity Market and Power System Operation Studies.* Technical University of Denmark, 2016.

[103] Michael Patriksson and Laura Wynter. Stochastic mathematical programs with equilibrium constraints. *Operations research letters*, 25(4):159–167, 1999.

[104] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[105] Georg Ch Pflug and Alois Pichler. *Multistage stochastic optimization*, volume 1104. Springer, 2014.

[106] Wolter Pieters. Explanation and trust: what to tell the user in security and ai? *Ethics and information technology*, 13(1):53–64, 2011.

[107] András Prékopa. *Stochastic programming*, volume 324. Springer Science & Business Media, 2013.

[108] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[109] J Ross Quinlan. *C4. 5: programs for machine learning.* Elsevier, 2014.

[110] Shaogang Ren, Bo Zeng, and Xiaoning Qian. Adaptive bi-level programming for optimal gene knockouts for targeted overproduction under phenotypic constraints. *BMC bioinformatics*, 14(2):S17, 2013.

[111] E Roghanian, Seyed Jafar Sadjadi, and Mir-Bahador Aryanezhad. A probabilistic bi-level linear multi-objective programming problem to supply chain planning. *Applied Mathematics and computation*, 188(1):786–800, 2007.

[112] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[113] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85, 2022.

[114] Carlos Ruiz and Antonio J Conejo. Pool strategy of a producer with endogenous formation of locational marginal prices. *IEEE Transactions on Power Systems*, 24(4):1855–1866, 2009.

[115] Kemal H Sahin and Amy R Ciric. A dual temperature simulated annealing approach for solving bilevel programming problems. *Computers & chemical engineering*, 23(1):11–25, 1998.

[116] Puchit Sariddichainunta and Masahiro Inuiguchi. Global optimality test for maximin solution of bilevel linear programming with ambiguous lower-level objective function. *Annals of Operations Research*, 256(2):285–304, 2017.

[117] Daniel Segre, Dennis Vitkup, and George M Church. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, 99(23):15112–15117, 2002.

[118] Susan A Shaheen, Elliot W Martin, Adam P Cohen, Nelson D Chan, and Mike Pogodzinski. Public bikesharing in north america during a period of rapid expansion: Understanding business models, industry trends & user impacts, mti report 12-29. 2014.

[119] Alexander Shapiro and Huifu Xu. Stochastic mathematical programs with equilibrium constraints, modelling and sample average approximation. *Optimization*, 57(3):395–418, 2008.

[120] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc

Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[121] Hélene Verhaeghe, Siegfried Nijssen, Gilles Pesant, Claude-Guy Quimper, and Pierre Schaus. Learning optimal decision trees using constraint programming. *Constraints*, 25(3):226–250, 2020.

[122] Sicco Verwer and Yingqian Zhang. Learning optimal classification trees using a binary linear program formulation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 1625–1632, 2019.

[123] Heinrich Von Stackelberg and Stackelberg Heinrich Von. *The theory of the market economy*. Oxford University Press, 1952.

[124] Daniël Vos and Sicco Verwer. Efficient training of robust decision trees against adversarial examples. In *International Conference on Machine Learning*, pages 10586–10595. PMLR, 2021.

[125] Yifan Wang, Shixin Liu, Jianhui Wang, and Bo Zeng. Capacity expansion of wind power in a market environment with topology control. *IEEE Transactions on Sustainable Energy*, 2018.

[126] Wolfram Wiesemann, Angelos Tsoukalas, Polyxeni-Margarita Kleniati, and Berç Rustem. Pessimistic bilevel optimization. *SIAM Journal on Optimization*, 23(1):353–380, 2013.

[127] Sonja Wogrin, Efraim Centeno, and Julián Barquín. Generation capacity expansion in liberalized electricity markets: A stochastic mpec approach. *IEEE Transactions on Power Systems*, 26(4):2526–2532, 2011.

[128] Fan Yang, Kai He, Linxiao Yang, Hongxia Du, Jingbang Yang, Bo Yang, and Liang Sun. Learning interpretable decision rule sets: A submodular optimization approach. *Advances in Neural Information Processing Systems*, 34, 2021.

[129] İhsan Yanikoğlu and Daniel Kuhn. Decision rule bounds for two-stage stochastic bilevel programs. *SIAM Journal on Optimization*, 28(1):198–222, 2018.

[130] Kevin Yeh, Craig Whittaker, Matthew J Realff, and Jay H Lee. Two stage stochastic bilevel programming model of a pre-established timberlands supply chain with biorefinery investment interests. *Computers & Chemical Engineering*, 73:141–153, 2015.

[131] Dajun Yue, Jiyao Gao, Bo Zeng, and Fengqi You. A projection-based reformulation and decomposition algorithm for global optimization of a class of mixed integer bilevel linear programs. *Journal of Global Optimization*, 73(1):27–57, 2019.

[132] Bo Zeng. A practical scheme to compute the pessimistic bilevel optimization problem. *INFORMS Journal on Computing*, 32(4):1128–1142, 2020.

[133] Bo Zeng and Yu An. Solving bilevel mixed integer program by reformulations and decomposition. 2014.

[134] Bo Zeng and Long Zhao. Solving two-stage robust optimization problems using a column-and-constraint generation method. *Operations Research Letters*, 41(5):457–461, 2013.

[135] Yiduo Zhan and Qipeng P Zheng. A multistage decision-dependent stochastic bilevel programming approach for power generation investment expansion planning. *IISE Transactions*, 50(8):720–734, 2018.

[136] Yue Zheng, Debin Fang, and Zhongping Wan. A solution approach to the weak linear bilevel programming problems. *Optimization*, 65(7):1437–1449, 2016.

[137] Yue Zheng, Zhongping Wan, Shihui Jia, and Guangmin Wang. A new method for strong-weak linear bilevel programming problem. *Journal of Industrial & Management Optimization*, 11(2):529–547, 2015.

[138] Fan Zhou, Zuduo Zheng, Jake Whitehead, Robert Perrons, Lionel Page, and Simon Washington. Projected prevalence of car-sharing in four asian-pacific countries in 2030: What the experts think. *Transportation Research Part C: Emerging Technologies*, 84:158–177, 2017.

[139] Haoran Zhu, Pavankumar Murali, Dzung Phan, Lam Nguyen, and Jayant Kalagnanam. A scalable mip-based method for learning optimal multivariate decision trees. *Advances in Neural Information Processing Systems*, 33:1771–1781, 2020.

[140] Ali R Zomorrodi and Costas D Maranas. Optcom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS computational biology*, 8(2):e1002363, 2012.