Ann Oper Res (2007) 153: 235–256
DOI 10.1007/s10479-007-0176-2

An overview of bilevel optimization

Benoît Colson · Patrice Marcotte · Gilles Savard

Published online: 20 April 2007

Abstract This paper is devoted to bilevel optimization, a branch of mathematical programming of both practical and theoretical interest. Starting with a simple example, we proceed towards a general formulation. We then present fields of application, focus on solution approaches, and make the connection with MPECs (Mathematical Programs with Equilibrium Constraints).

Keywords Bilevel programming · Mathematical programs with equilibrium constraints · Nonlinear programming · Optimal pricing

1 Introduction

This paper is devoted to bilevel optimization. Its purpose is to provide the reader with the key concepts, applications and solution methods associated with this class of hierarchical mathematical programs. It is an updated version of the survey of Colson et al. (2005b) that originally appeared in *4OR*. Apart from minor modifications to the text, this version includes

B. Colson now at SAMTECH s.a., Liège, Belgium.

This article is an updated version of a paper that appeared in *4OR* 3, 87–107, 2005.

B. Colson (✉)
Department of Mathematics, The University of Namur, Namur, Belgium
e-mail: benoit.colson@samcef.com

P. Marcotte
Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Montréal, QC, Canada
e-mail: marcotte@iro.umontreal.ca

G. Savard
Département de Mathématiques et de Génie Industriel, Ecole Polytechnique de Montréal, Montréal, QC, Canada
e-mail: gilles.savard@polymtl.ca

now a section on applications and puts more emphasis on the combinatorial structure of bilevel programs.

Let us start with a simple example concerned with a *toll-setting problem* that consists in maximizing the revenue raised from tolls set on some links of a transportation network. Taking into account that network users minimize their travel costs, an optimal toll schedule will be such that toll levels are not too high—otherwise the users are deterred from using the toll arcs—though still generating "large" revenues. Once the network managers have set tolls, travelers react to these values and select their itinerary such that total travel cost, i.e. standard costs (time, distance, etc.) plus tolls, is minimized. An important feature of this problem—and more generally of bilevel programs—is the hierarchical relationship between two autonomous, and possibly conflictual, decision makers. In this sense, it is related to Stackelberg (leader-follower) games in economics.

Let us now introduce some notation. We denote by $\mathcal{A}$ the set of links of the network and by $\bar{\mathcal{A}}$ the subset of toll links. Since the network manager's goal is to maximize revenues, it faces the mathematical program

$$\max_{T,x} \quad \sum_{a\in\bar{\mathcal{A}}} T_a x_a \tag{1.1a}$$

$$\text{s.t.} \quad l_a \le T_a \le u_a, \quad \forall a \in \bar{\mathcal{A}}, \tag{1.1b}$$

where T_a and x_a denote the toll and the flow on link a respectively, and l_a (respectively u_a) is a lower (respectively upper) bound on the toll.[1]

The selfish behaviour of network users results in an equilibrium where all users are assigned to paths of minimum cost with respect to the current congestion levels (see e.g. Patriksson 1994 for more details). In the simplest situation, e.g., in a congestion-free environment, such *user equilibrium* coincides with a flow assignment that minimizes total system cost. It follows that the path-flow vector f, together with the link-flow vector x, is solution of the linear program:

$$\min_{f,x} \quad \sum_{a\in\mathcal{A}} c_a x_a + \sum_{a\in\bar{\mathcal{A}}} T_a x_a \tag{1.2a}$$

$$\text{s.t.} \quad \sum_{p\in\mathcal{P}_{rs}} f_p^{rs} = d_{rs}, \quad \forall (r,s) \in \Theta, \tag{1.2b}$$

$$x_a = \sum_{(r,s)\in\Theta} \sum_{p\in\mathcal{P}_{rs}} \delta_{a,p}^{rs} f_p^{rs}, \quad \forall a \in \mathcal{A}, \tag{1.2c}$$

$$f_p^{rs} \ge 0, \quad \forall p \in \mathcal{P}_{rs}, \ \forall (r,s) \in \Theta. \tag{1.2d}$$

The objective (1.2a) is the sum of tolls T_a ($a \in \bar{\mathcal{A}}$) and other costs (duration, length, etc.), aggregated in a measure c_a for each link. Constraint (1.2b) expresses demand satisfaction in the sense that, for a given origin-destination pair (r,s) (the set of all such pairs is denoted by Θ), the sum of the flows f_p^{rs} on all paths p connecting r to s (these paths being regrouped in $\mathcal{P}_{rs}$) equals the travel demand, d_{rs}. Constraint (1.2c) links path flows f_p^{rs} and link flows

[1] While it seems natural to have $l_a = 0$, it is sometimes advantageous to set tolls to negative values. This corresponds to subsidies.

x_a, with

$$\delta_{a,p}^{rs} = \begin{cases} 1 & \text{if path } p \in \mathcal{P}_{rs} \text{ uses link } a, \\ 0 & \text{otherwise.} \end{cases}$$

Mathematical programs (1.1) and (1.2) are connected through the use of common variables, namely tolls T_a ($a \in \bar{\mathcal{A}}$) and flows x_a ($a \in \mathcal{A}$). Also, the profit of the network manager (see (1.1a)) cannot be computed until flows are known. These flows are not in the direct control of the manager, but the solution of a mathematical program parameterized in the toll vector T. This yields the bilevel formulation[2]

$$\begin{aligned}
&\max_{T,f,x} \quad \sum_{a\in\bar{\mathcal{A}}} T_a x_a \\
&\text{s.t.} \quad l_a \le T_a \le u_a, \quad \forall a \in \bar{\mathcal{A}}, \\
&\qquad (f,x) \in \arg\min_{f',x'} \quad \sum_{a\in\mathcal{A}} c_a x'_a + \sum_{a\in\bar{\mathcal{A}}} T_a x'_a \\
&\qquad \text{s.t.} \quad \sum_{p\in\mathcal{P}_{rs}} f'^{rs}_p = d_{rs}, \quad \forall (r,s) \in \Theta, \\
&\qquad\qquad x'_a = \sum_{(r,s)\in\Theta} \sum_{p\in\mathcal{P}_{rs}} \delta_{a,p}^{rs} f'^{rs}_p, \quad \forall a \in \mathcal{A}, \\
&\qquad\qquad f'^{rs}_p \ge 0, \quad \forall p \in \mathcal{P}_{rs},\ \forall (r,s) \in \Theta.
\end{aligned}$$

The hierarchical relationship results from the fact that the mathematical program related to the users' behaviour is part of the manager's constraints. This is the major feature of bilevel programs: they include two mathematical programs within a single instance, one of these problems being part of the constraints of the other one. In view of this hierarchical relationship, the program (1.1) is called the *upper-level problem* while (1.2) corresponds to the *lower-level problem*.

We now leave the framework of toll-setting problems. The interested reader is referred to e.g. Labbé et al. (1998) and Brotcorne et al. (2001) for further details. The next section describes bilevel programs from a more general point of view, and includes a list of fields of application. This will be followed by a survey of existing methods for solving various types of bilevel programs (Sect. 3). Mathematical programs with equilibrium constraints, which are very similar to bilevel programs, will be the subject of Sect. 4. We will conclude this paper with a review of perspectives and challenges for future research.

2 General formulation and basic concepts

The general formulation of a bilevel programming problem (BLPP) is

$$\min_{x\in X, y} \quad F(x,y) \tag{2.1a}$$

$$\text{s.t.} \quad G(x,y) \le 0, \tag{2.1b}$$

[2] In the sequel, we will simply write down the upper and lower level problems, dispensing with the "prime" and "arg min" notation. The resulting "vertical" format is indeed less heavy and more transparent.

$$\min_{y} \quad f(x, y) \tag{2.1c}$$

$$\text{s.t.} \quad g(x, y) \leq 0, \tag{2.1d}$$

where $x \in \mathbb{R}^{n_1}$ and $y \in \mathbb{R}^{n_2}$. The variables of problem (2.1) are divided into two classes, namely the *upper-level variables* $x \in \mathbb{R}^{n_1}$ and the *lower-level variables* $y \in \mathbb{R}^{n_2}$. Similarly, the functions $F : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}$ and $f : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}$ are the *upper-level* and *lower-level objective functions* respectively, while the vector-valued functions $G : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}^{m_1}$ and $g : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}^{m_2}$ are called the *upper-level* and *lower-level constraints* respectively. Upper-level constraints involve variables from both levels (in contrast with the constraints specified by the set X) and play a very specific role. Indeed, they must be enforced indirectly, as they do not bind the lower-level decision-maker.

From a historical point of view, multilevel optimization is closely related to the economic problem of Stackelberg (1952) in the field of game theory, which we briefly describe now. To this end, we consider an economic planning process involving interacting agents at two distinct levels: some of the individuals—collectively called the *leader*—issue directives to the remaining agents—called the *followers*. In the particular framework of Stackelberg games, the leader is assumed to anticipate the reactions of the followers; this allows him to choose his best—or *optimal*—strategy accordingly. More precisely, the leader chooses a strategy x in a set $X \subseteq \mathbb{R}^n$, and every follower i has a strategy set $Y_i(x) \subseteq \mathbb{R}^{m_i}$ corresponding to each $x \in X$. The sets $Y_i(x)$ are assumed to be closed and convex. Any follower i also has a cost function depending on both the leader's and *all* followers' strategies and which may be expressed as

$$\theta_i(x, \cdot) : \prod_{j=1}^{M} \mathbb{R}^{m_j} \to \mathbb{R},$$

where M is the number of followers. It is further assumed that for fixed values of $x \in X$ and y_j $(j \neq i)$ the function θ_i is convex and continuously differentiable in $y_i \in Y_i(x)$. The followers behave collectively according to the *noncooperative principle* of Nash (1951) which means that, for each $x \in X$, they will choose a joint response vector

$$y^{\text{opt}} \equiv (y_i^{\text{opt}})_{i=1}^{M} \in C(x),$$

where $C(x) = \prod_{i=1}^{M} Y_i(x)$, such that, for every $i = 1, \ldots, M$, there holds

$$y_i^{\text{opt}} \in \operatorname{argmin}\{\theta_i(x, y_i, y_{j\neq i}^{\text{opt}}) : y_i \in Y_i(x)\}.$$

In the above setting, considered by Sherali et al. (1983) in an oligopolistic situation, Stackelberg problems possess a hierarchical structure similar to that of BLPP, although the lower-level program is an equilibrium rather than an optimization problem. This class of problems will be discussed in more details in Sect. 4.

Bilevel programs were initially considered by Bracken and McGill in a series of papers—see Bracken and McGill (1973, 1974, 1978)—that dealt with applications in the military field as well as in production and marketing decision making. By that time, such problems were called *mathematical programs with optimization problems in the constraints*, which exactly reflects the situation formulated in (2.1), the terms *bilevel* and *multilevel programming* being introduced later by Candler and Norton (1977). Notice however that the problems studied in the latter paper did not involve joint upper-level constraints, that is, constraints

depending on both x and y. To our knowledge, the general formulation with $G(x, y) \leq 0$ as upper-level constraints first appeared in Aiyoshi and Shimizu (1981).

We now return to problem (2.1) to introduce some further concepts of bilevel programming. The *relaxed* problem associated with (2.1) is

$$\begin{array}{ll} \min\limits_{x \in X, y} & F(x, y) \\ \text{s.t.} & G(x, y) \leq 0, \\ & g(x, y) \leq 0, \end{array} \tag{2.2}$$

and its optimal value is a lower bound for the optimal value of (2.1). The *relaxed feasible region* (or *constraint region*) is

$$\Omega = \{(x, y) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} : x \in X, G(x, y) \leq 0 \text{ and } g(x, y) \leq 0\}.$$

For a given (fixed) vector $\bar{x} \in X$, the *lower-level feasible set* is defined by

$$\Omega(\bar{x}) = \{y \in \mathbb{R}^{n_2} : g(\bar{x}, y) \leq 0\}$$

while the *lower-level reaction set*[3] (or *rational reaction set*) is

$$R(\bar{x}) = \{y \in \mathbb{R}^{n_2} : y \in \text{ argmin } \{f(\bar{x}, \hat{y}) : \hat{y} \in \Omega(\bar{x})\}\}.$$

Every $y \in R(\bar{x})$ is a *rational response*. For a given x, $R(x)$ is an implicitly defined multivalued function of x that may be empty for some values of its argument. Finally, the set

$$\mathcal{IR} = \{(x, y) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} : x \in X, G(x, y) \leq 0, y \in R(x)\},$$

that regroups the feasible points of the BLPP, corresponds to the feasible set of the leader, and is known as the *induced region* (or *inducible region*). This set is usually nonconvex and it can even be disconnected or empty in presence of upper-level constraints.

We conclude this section with a short discussion on two modelling approaches to bilevel programming. In the case of *optimistic bilevel programming*, it is assumed that, whenever the reaction set $R(x)$ is not a singleton, the leader is allowed to select the element in $\Omega(x)$ that suits him best. In this situation, a point $(x^*, y^*) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ is said to be a *local optimistic solution* for problem (2.1) if

$$\begin{aligned} & x^* \in X, \\ & G(x^*, y^*) \leq 0, \\ & y^* \in R(x^*), \\ & F(x^*, y^*) \leq F(x^*, y) \quad \text{for all } y \in R(x^*) \end{aligned}$$

and there exists an open neighbourhood $V(x^*; \delta)$ of x^* (with radius $\delta > 0$) such that

$$\phi_o(x^*) \leq \phi_o(x) \quad \text{for all } x \in V(x^*; \delta) \cap X,$$

[3] According to the definition of a bilevel program, the lower level problem must be solvable for global minima. In practice, this requires that the lower-level program be convex.

where $\phi_o(x) = \min_y\{F(x, y) : y \in R(x)\}$. It is called a *global optimistic solution* if $\delta = \infty$ can be selected, corresponding to $V(x^*) = X$.

When cooperation of the leader and the follower is not allowed, or if the leader is risk-averse and wishes to limit the "damage" resulting from an undesirable selection by the follower, then a point $(x^*, y^*) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ is said to be a *local pessimistic solution* for problem (2.1) if

$$\begin{aligned} &x^* \in X, \\ &G(x^*, y^*) \leq 0, \\ &y^* \in R(x^*), \\ &F(x^*, y^*) \geq F(x^*, y) \quad \text{for all } y \in R(x^*) \end{aligned}$$

and there exists an open neighbourhood $V(x^*; \delta)$ of x^* (with $\delta > 0$) such that

$$\phi_p(x^*) \leq \phi_p(x) \quad \text{for all feasible } x \in V(x^*; \delta),$$

where this time $\phi_p(x) = \max_y\{F(x, y) : y \in R(x)\}$. It is called a *global pessimistic solution* if $\delta = \infty$ can be selected. Note that the difference between the optimistic and pessimistic approaches can also be explained from the follower viewpoint: the optimistic solution results from a friendly or cooperative behaviour while an aggressive follower produces a pessimistic solution. A more complete discussion of these issues may be found in Loridan and Morgan (1996) and Dempe (2002).

2.1 A sample of applications

Section 1 illustrated the usefulness of bilevel programs through a specific example. Several other transportation issues may be modelled by bilevel programs—see Migdalas (1995) for a review—and, more generally, real-world problems involving a hierarchical relationship between two decision levels. These are encountered in fields as diverse as management (facility location, environmental regulation, credit allocation, energy policy, hazardous materials), economic planning (social and agricultural policies, electric power pricing, oil production), engineering (optimal design, structures and shape), chemistry, environmental sciences, optimal control, etc.

For one, it can be argued that most managerial decisions are of a bilevel nature, in the sense that they impact systems with some degree of autonomy and conflicting objectives, few real-life studies have adopted this paradigm. In the following, we provide a small selection of actual or potential fields of application considered in the literature.

2.1.1 Revenue management

Revenue management is a generic term that covers a set of optimization procedures aimed at maximizing the profitability of firms characterized by high investment costs, low operating costs, and perishable inventories. It was initially implemented in the airline industry, under the name "yield management", and involved four issues: ticket pricing, seat allocation, demand forecasting, and overbooking. Notwithstanding the third issue, which is of a tactical nature, the first three lend themselves to a bilevel formulation that extends the toll setting problem described in the introduction of this survey. Such model, that involves the pricing and seat allocation policies, is described in Côté et al. (2003).

2.1.2 Congestion management

In urban areas, marginal tolls can be used to minimize overall congestion. If only a subset of the arcs are subject to tolls, the latter scheme is not applicable, and one faces a "second best" problem of true bilevel nature. See Hearn and Ramana (1998) and Larsson and Patriksson (1998) for more details on this topic.

2.1.3 Origin-destination matrix estimation

A classical problem in urban studies consists in estimating the demand for transportation between nodes of a transportation network, which is usually obtained from statistical surveys, and serves as an input to the equilibrium assignment process. If the latter yields flows that disagree with observed (reliable) flows, then one may set up a bilevel problem where one looks for a demand matrix (upper level) that matches as well as possible the induced equilibrium flows (lower level traffic assignment problem). The reader is referred to Florian and Chen (1995) for further details.

2.1.4 Management of hazardous materials

When designing a sub-network to be used by carriers of hazardous material, it makes sense to take into account the behaviour of carriers, who may favour shortest routes over safest itineraries. A bilevel formulation corresponding to this situation is analyzed in Kara and Verter (2004).

2.1.5 Network design problems

Design problems involving autonomous agents are a rich source of bilevel models. One such example is concerned with capacity improvement of a road network, where one must balance investment costs against congestion reduction, in a network where traffic flows achieve an equilibrium compatible with the design parameters, i.e., optimize their own objective. Introduced by LeBlanc (1973), this model was further analyzed by Marcotte (1986), who provided worst-case bounds on the performance of easily implementable heuristic procedures. Another example is that of the optimization of frequencies in a transit network where users minimize their individual travel time (Constantin and Florian 1995).

2.1.6 Energy sector

The energy sector, in particular the power sector, has been the topic of some interesting bilevel modelizations. Hobbs and Nelson (1992) consider an electric utility that "seeks to minimize costs or maximize benefits while controlling electric rates and subsidizing energy conservation programs". In the model of Haurie et al. (1992), the interaction between a power utility and cogenerators is set within the framework of a leader-follower game, where the demand side is modelled as a large-scale techno-economic model. In a joint energy-agriculture setting, Bard et al. (2000) address the problem faced by a government (the leader) that wishes to induce, through minimal subsidies, the conversion of food to biofuel crops. The model, involving bilinear objectives at both levels of decision-making, is reminiscent of the toll-setting model discussed above.

2.1.7 Engineering problems

Optimization models involving chemical or physical equilibria can be modelled as mathematical programs with equilibrium constraints (see Sect. 4), a class of mathematical programs closely related to bilevel optimization. Recently, some researchers have explicitly addressed these control problems as MPECs. See, among several proposals, the work of Pang and Trinkle (1996) in robotics and that of Outrata and Kocvara (2006) for a truss design problem.

2.1.8 Principal-agent problem

Bilevel programming subsumes the principal-agent paradigm, a classical problem of economics, whereby the leader (principal) sub-contracts a job to an agent (follower). The agent is rewarded by the principal according to the quality of some random outcome that determines the leader's revenue. At the lower level, the agent maximizes an objective that is a function of its reward and its effort level. Of course, the larger the effort level, the larger the expected revenue associated with the outcome. Whenever the set of outcome is continuous, the resulting lower level problem is infinite-dimensional. See Van Ackere (1993) for an overview of the topic.

2.1.9 A Stackelberg-Nash game

Sherali et al. (1983) analyze an oligopoly where one firm acts as the leader, and the remaining ones achieve a Cournot-Nash equilibrium parameterized by the production level of the leader firm. This line of attack can of course be extended to include several leaders, or even several layers of hierarchical players, but that would take us a bit far.

3 A survey of existing methods

Although early work on bilevel programming dates back to the nineteen seventies, it was not until the early nineteen eighties that the usefulness of these mathematical programs in modelling hierarchical decision processes and engineering design problems prompted researchers to pay close attention to bilevel programs. A first bibliographical survey on the subject was written by Kolstad (1985). Bilevel programming problems being intrinsically difficult (see Sect. 3.1 below), it is not surprising that most algorithmic research to date has focused on the simplest cases of bilevel programs, that is problems having nice properties such as linear, quadratic or convex objective and/or constraint functions. In particular, the most studied instance of bilevel programming problems has been for a long time the *linear* BLPP—in which all functions are linear—and therefore this subclass is the subject of several dedicated surveys, such as those by Hsu and Wen (1989), Wen and Hsu (1991) and Ben-Ayed (1993). Over the years, more complex bilevel programs were studied and even those including discrete variables received some attention, as in Vicente et al. (1996). Hence more general surveys appeared, such as those by Savard (1989), Anandalingam and Friesz (1992) and Vicente and Calamai (1994). Colson (1999) deals with both nonlinear bilevel programming problems and mathematical programs with equilibrium constraints and recently Dempe (2003) wrote an annotated bibliography on these same topics. The combinatorial nature of bilevel programming has been reviewed in Marcotte and Savard (2005).

Following the proliferation of research devoted to bilevel programming, a number of dedicated textbooks have also been published in the late nineteen nineties. Among them, those by Shimizu et al. (1997) and Bard (1998) are authored by some of the early protagonists in the field. Another monograph on the subject is that of Migdalas et al. (1997), who consider the more general case of multilevel programming. The most recent book on bilevel programming, as of May 2005, is that of Dempe (2002).

3.1 Properties

Being generically non-convex and non-differentiable, bilevel programs are intrinsically hard. Even the "simplest" instance, the linear-linear BLPP, was shown to be $\mathcal{NP}$-hard by Jeroslow (1985), while Hansen et al. (1992) proved strong $\mathcal{NP}$-hardness, using a reduction from KERNEL (see Garey and Johnson 1979). Vicente et al. (1994) strengthened these results and proved that merely checking strict or local optimality is also $\mathcal{NP}$-hard, based on reductions from 3-SAT. Indeed, many combinatorial optimization problems can be reduced to bilevel programs. As an illustration, consider the mixed binary program

$$\begin{aligned} \max_{x,u} \quad & cx + eu \\ \text{s.t.} \quad & Ax + Eu \leq b, \\ & x \geq 0, \quad u \text{ binary valued}, \end{aligned}$$

where $c \in \mathbb{R}^{n_x}$, $e \in \mathbb{R}^{n_u}$, $A \in \mathbb{R}^{m \times n_x}$, $E \in \mathbb{R}^{m \times n_u}$, $b \in \mathbb{R}^m$.

Note that the binary requirement can be alternatively expressed as

$$0 = \min\{u, \mathbf{1} - u\},$$

where $\mathbf{1}$ denotes the vector of "all ones". We now obtain the linear bilevel program

$$\begin{aligned} \max_{x,y,u} \quad & cx + eu \\ \text{s.t.} \quad & Ax + Eu \leq b, \\ & x \geq 0, \\ & y = 0, \\ & \max_{y} \quad \sum_{i=1}^{n_u} y_i \\ & \text{s.t.} \quad y \leq u, \\ & \qquad\; w \leq \mathbf{1} - u, \end{aligned}$$

where $y \in \mathbb{R}^{n_u}$. In this formulation, the integrality constraints are no more required, as they are enforced by the upper level constraints $y = 0$, together with the lower level optimality conditions.

Other equivalences between classical combinatorial problems (e.g. bilinear disjoint programming, generalized linear complementarity, traveling salesman, multicriteria optimization, nonconvex quadratic programming) and bilevel programs have been showed (see e.g. Marcotte and Savard 2005). The interest in these reformulations goes beyond the complexity issue and indeed has been instrumental in developing efficient algorithms based on

the concept of *embedded algorithms* (see e.g. Audet et al. 1997 and Marcotte et al. 2004). In all these classes of problems, the bilevel formulation involves a lower level that admits extremal solutions, a property that allows the development of methods that guarantee a global optimum (see Sects. 3.2, 3.3 and 3.4). In contrast, the *nonlinear* instances of bilevel programming have mostly stimulated research on local optimality results (see Sects. 3.5, 3.6 and 3.7). Indeed, a number of authors have derived *necessary optimality conditions* for bilevel programming problems. Among them are Dempe (1992a, 1992b), Yezza (1996), Outrata (1993), Ye and Zhu (1995) and Scheel and Scholtes (2000), using tools from nonsmooth analysis (see e.g. Clarke 1990). Savard and Gauvin (1994) and Vicente and Calamai (1995) developed optimality conditions based on the geometry of the induced region into account: the former authors adapted the notion of *steepest descent* to the case of bilevel programs, while the latter generalized first- and second-order optimality conditions to the case of bilevel programs involving quadratic strictly convex lower-level problems. For this class of problems, the set of feasible directions is the union of a finite number of convex polyhedral cones.

However, due to the inherent difficulty of manipulating the mathematical objects involved in all these optimality conditions, they have few practical use.

3.2 Extreme-point approaches for the linear case

An important property of *linear* bilevel programs, i.e., programs where all functions involved are linear and the set X is polyhedral, is that their solution set, whenever it is nonempty, contains at least one vertex of the constraint region defined by the polyhedron

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

Hence a wide class of methods for solving linear BLPPs is based on vertex enumeration.

The first method using such an approach was proposed by Candler and Townsley (1982) for solving BLPPs with no upper-level constraints and with unique lower-level solutions. Their algorithm explores a decreasing number of bases of the lower-level problem, but was shown to be relatively slow in subsequent numerical tests. Bialas and Karwan (1984) introduced the Kth best method, which considers bases of the relaxed problem (2.2) sorted in increasing order of upper-level objective function values. The method stops at the lowest index K corresponding to a *rational basis*. Such a basis is clearly globally optimal.

Similar vertex enumeration methods were introduced by Papavassilopoulos (1982), with the difference that all extreme points considered belong to the induced region IR, and that separation techniques are used to explore the adjacent vertices.

Related contributions are those by Chen and Florian (1992), Chen et al. (1992) and Tuy et al. (1993).

3.3 Branch-and-bound

When the lower-level problem is convex and regular, it can be replaced by its Karush-Kuhn-Tucker (KKT) conditions, yielding the single-level reformulation of problem (2.1):

$$\min_{x \in X, y, \lambda} \quad F(x, y) \tag{3.1a}$$

$$\text{s.t.} \quad G(x, y) \leq 0, \tag{3.1b}$$

$$g(x, y) \leq 0, \tag{3.1c}$$

$$\lambda_i \geq 0, \quad i = 1, \ldots, m_2, \tag{3.1d}$$

$$\lambda_i \; g_i(x, y) = 0, \quad i = 1, \ldots, m_2, \tag{3.1e}$$

$$\nabla_y \mathcal{L}(x, y, \lambda) = 0, \tag{3.1f}$$

where

$$\mathcal{L}(x, y, \lambda) = f(x, y) + \sum_{i=1}^{m_2} \lambda_i g_i(x, y)$$

is the Lagrangean function associated with the lower-level problem.

Even under suitable convexity assumptions on the functions F, G and the set X, the above mathematical program is not easy to solve, due mainly to the nonconvexities that occur in the complementarity and Lagrangean constraints. While the Lagrangean constraint is linear in certain important cases (linear or convex quadratic functions), the complementarity constraint is intrinsically combinatorial, and is best addressed by enumeration algorithms, such as branch-and-bound.

In the branch-and-bound scheme, the root node of the tree corresponds to problem (3.1) from which constraint (3.1e) is removed. At a generic node of the branch-and-bound tree that does not satisfy the complementarity constraints, separation is performed in the following manner: two children nodes are constructed, one with $\lambda_i = 0$ as an additional constraint, and the other with the constraint $g_i(x, y) = 0$. The optimal values of these problems yield lower bounds valid for the corresponding subtree.

In the absence of upper-level constraints, a rational solution can be computed by solving the lower-level problem resulting from setting x to the partial optimal solution of the relaxed problem. Note that, in contrast with standard branch-and-bound implementations, feasible (i.e., rational) solutions are then generated at *every* node of the implicit enumeration tree. The upper bound is updated accordingly.

Algorithms based on this idea were proposed by Bard and Falk (1982) and Fortuny-Amat and McCarl (1981) for solving linear bilevel programming problems. The approach was adapted by Bard and Moore (1990) to linear-quadratic problems and by Al-Khayal et al. (1992), Bard (1988) and Edmunds and Bard (1991) to the quadratic case.

Combining branch-and-bound, monotonicity principles and penalties similar to those used in mixed-integer programming, Hansen et al. (1992) have developed a code capable of solving medium-sized linear bilevel programs.[4] Thoai et al. (2002) have developed a similar scheme for mathematical programs with linear complementarity constraints.

3.4 Complementary pivoting

The first approach using *complementary pivots* is that of Bialas et al. (1980) for solving linear BLPPs. Their algorithm—named *Parametric Complementary Pivot (PCP) Algorithm*—is based on the reformulation (3.1) of a linear bilevel program using the KKT optimality conditions for the lower-level problem. At each iteration, the algorithm computes a feasible point (x, y) for the original problem such that the upper-level objective $F(x, y)$ takes a value at most equal to α, and where constraint (3.1f) is perturbed by adding a term $\varepsilon H y$, where H is a negative definite matrix and ε is sufficiently small so that the solution to the original problem is not modified. The parameter α is updated after each iteration and the process is

[4] I.e., of the order of 200 variables and 200 constraints.

repeated until no feasible (x, y) can be found. However Ben-Ayed and Blair (1990) showed that this algorithm does not always converge to the optimal solution.

Let us also mention the contributions of Júdice and Faustino (1988, 1992, 1994), who introduced the so-called *sequential linear complementarity problem* (LCP) for solving linear and linear-quadratic bilevel programming problems. Note that their approach may actually be viewed as a combination of the techniques described in Sects. 3.2 and 3.3, namely vertex enumeration and branch-and-bound methods.

3.5 Descent methods

Assuming that, for any x, the optimal solution of the lower-level problem is unique and defines y as an implicit function $y(x)$ of x, problem (2.1) may be viewed solely in terms of the upper-level variables $x \in \mathbb{R}^{n_1}$. Given a feasible point x, an attempt is made to find a feasible (rational) direction $d \in \mathbb{R}^{n_1}$ along which the upper-level objective decreases. A new point $x + \alpha d$ $(\alpha > 0)$ is computed so as to ensure a reasonable decrease in F while maintaining feasibility for the bilevel problem. However, a major issue is the availability of the gradient (or a sub-gradient) of the upper-level objective, $\nabla_x F(x, y(x))$, at a feasible point. Applying the chain rule of differentiation, we have, whenever $\nabla_x y(x)$ is well defined:

$$\nabla_x F(x, y(x)) = \nabla_x F(x, y) + \nabla_y F(x, y)\ \nabla_x y(x),$$

where the functions are evaluated at the current iterate. Kolstad and Lasdon (1990) have proposed a method for approximating this gradient.

Another line of attack is that of Savard and Gauvin (1994), for problems where no upper-level constraints are present and where the lower-level constraints are rewritten as:

$$\begin{aligned} g_i(x, y) &\le 0, \quad i \in I, \\ g_j(x, y) &= 0, \quad j \in J. \end{aligned}$$

The authors first show that an upper-level descent direction at a given point x is a vector $d \in \mathbb{R}^{n_1}$ such that

$$\nabla_x F(x, y^*)d + \nabla_y F(x, y^*)\ w(x, d) < 0, \tag{3.2}$$

where $y^* = y(x)$ and $w \in \mathbb{R}^{n_2}$ is a solution of the program

$$\begin{aligned} \min_{w} \quad & (d^T, w^T)\nabla^2_{xy}\mathcal{L}(x, y^*, \lambda)\ (d, w) \\ \text{s.t.} \quad & \nabla_y g_i(x, y^*)w \le -\nabla_x g_i(x, y^*)d, \quad i \in I(x), \\ & \nabla_y g_j(x, y^*)w = -\nabla_x g_j(x, y^*)d, \quad j \in J, \\ & \nabla_y f(x, y^*)w = -\nabla_x f(x, y^*)d + \nabla_x \mathcal{L}(x, y^*, \lambda)d, \end{aligned} \tag{3.3}$$

with $I(x) = \{i \in I : g_i(x, y^*) = 0\}$ and

$$\mathcal{L}(x, y, \lambda) = f(x, y) + \sum_{i \in I(x) \cup J} \lambda_i g_i(x, y)$$

is the Lagrangean of the lower-level problem with respect to the active constraints. The steepest descent then coincides with the optimal solution of the *linear-quadratic bilevel*

program

$$\begin{aligned} \min_{d} \quad & \nabla_x F(x, y^*)d + \nabla_y F(x, y^*)w(x, d) \\ \text{s.t.} \quad & \|d\| \leq 1, \\ & w(x, d) \text{ solves the quadratic program (3.3)}, \end{aligned} \tag{3.4}$$

for which exact algorithms exist, such as those by Bard and Moore (1990) or Jaumard et al. (2000).

Alternatively, Vicente et al. (1994) proposed a descent method for convex quadratic bilevel programs, i.e., problems where both objectives are quadratic, and where constraints are linear. They extend the work of Savard and Gauvin (1994) by solving problem (3.4) using the sequential LCP method of Júdice and Faustino (1994), and propose a way to compute exact stepsizes. Motivated by the fact that checking local optimality in the sequential LCP approach is very difficult, Vicente et al. (1994) have designed a hybrid algorithm using both the abovementioned features and a pivot step strategy that enforces the complementarity constraints.

Finally, let us mention the work of Falk and Liu (1995), who present a bundle method where the decrease of the upper-level objective is monitored according to subdifferential information obtained from the lower-level problem. They call the resulting setup a *leader predominate algorithm*, according to the role played by the leader in the sequential decision making process.

3.6 Penalty function methods

Penalty methods constitute another important class of algorithms for solving *nonlinear* BLPPs, although they are generally limited to computing stationary points and local minima.

An initial step in this direction was achieved by Aiyoshi and Shimizu (1981, 1984) and Shimizu and Aiyoshi (1981). Their approach consists in replacing the lower-level problem (2.1c–2.1d) by the penalized problem

$$\min_{y} \quad p(x, y, r) = f(x, y) + r\phi(g(x, y)), \tag{3.5}$$

where r is a positive scalar, ϕ is a continuous penalty function that satisfies

$$\begin{aligned} &\phi(g(x, y)) > 0 \quad \text{if } y \in \operatorname{int} S(x), \\ &\phi(g(x, y)) \to +\infty \quad \text{if } y \to \operatorname{bd} S(x), \end{aligned} \tag{3.6}$$

and int $S(x)$ and bd $S(x)$ denote the relative interior and the relative boundary of $S(x) = \{y : g(x, y) \leq 0\}$, respectively. Problem (2.1) is then transformed into:

$$\begin{aligned} \min_{x \in X, y} \quad & F(x, y^*(x, r)) \\ \text{s.t.} \quad & G(x, y^*(x, r)) \leq 0, \\ & p(x, y^*(x, r), r) = \min_{y} p(x, y, r). \end{aligned} \tag{3.7}$$

Shimizu and Aiyoshi (1981) proved that the sequence $\{(x^k, y^*(x^k, r^k))\}$ of optimal solutions to (3.7) converges to the solution of (2.1). The main drawback of this method is that

solving (3.7) for a fixed value of r requires the global solution at every update of the upper-level variables. Each subproblem is not significantly easier to solve than the original bilevel program.

Ishizuka and Aiyoshi (1992) proposed a double penalty method in which both objective functions (2.1a) and (2.1c) are penalized. They still use the augmented lower-level objective (3.5) and the penalty function ϕ characterized by (3.6) but replace the lower-level problem by its stationarity condition $\nabla_y p(x, y, r) = 0$, thus transforming (2.1) into the single-level program

$$\begin{aligned} \min_{x \in X, y} \quad & F(x, y) \\ \text{s.t.} \quad & G(x, y) \leq 0, \\ & \nabla_y p(x, y, r) = 0, \\ & g(x, y) \leq 0. \end{aligned} \tag{3.8}$$

Note that the last constraint restricts the domain of the function p. For a given r, problem (3.8) is solved using a second penalty function applied to the constraints.

A more recent contribution, by Case (1999), follows up on ideas of Bi et al. (1991), who themselves extend a technique proposed in Bi et al. (1989) for linear bilevel programs. Their approach is based on (3.1), that is, a bilevel program for which the lower-level problem has been replaced by its Karush-Kuhn-Tucker conditions. Their method involves a penalty function of the form

$$p(x, y, \lambda, \mu) = F(x, y) + \mu v(x, y, \lambda),$$

where μ is a positive *penalty parameter* and the upper-level objective $F(x, y)$ is augmented by a weighted, nonnegative penalty function associated with the current iterate. More precisely, Case (1999) builds a penalty function $v(x, y, \lambda)$ with respect to the ℓ_1 norm, defined as the sum of the terms associated with each constraint of the single-level problem (3.1). The resulting algorithm involves the minimization of the penalty function $p(x, y, \lambda, \mu)$ for a fixed value of μ. In view of the complex structure of the latter function, the authors develop a *trust-region method*, where the *model* for p (see Sect. 3.7) is obtained by replacing each component function of $p(x, y, \lambda, \mu)$ by its second-order Taylor expansion around the current iterate.

3.7 Trust-region methods

Trust-region algorithms are iterative methods based on the approximation of the original problem by a *model* around the current iterate. More specifically, let us consider the unconstrained problem

$$\min_x f(x).$$

Given the iterate x_k obtained at iteration k, one constructs a model m_k that approximates the objective function within a *trust region* usually defined as a ball (according to some norm) of radius Δ_k centered at x_k. The solution s_k to the *trust-region subproblem*

$$\begin{aligned} \min_s \quad & m_k(x_k + s) \\ \text{s.t.} \quad & \|s\| \leq \Delta_k \end{aligned}$$

is then computed. One then evaluates the quality of the model through the ratio of the *actual reduction* over the *predicted reduction*[5]

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}. \tag{3.9}$$

If ρ_k is large enough ($\rho_k > \eta_2$ for some $0 < \eta_2 < 1$), the trial point is accepted as the next iterate ($x_{k+1} = x_k + s_k$), and the trust region radius may be increased. The trial point is also accepted for smaller values of ρ_k that satisfy the condition $\rho_k \in [\eta_1, \eta_2)$, with $0 < \eta_1 < \eta_2$, but in this case the trust-region radius is *not* increased (it might be decreased). Finally, if ρ_k is too small, the trial point is dismissed ($x_{k+1} = x_k$) and the trust-region radius is decreased. The rules for modifying Δ_k are as follows:

$$\Delta_{k+1} \in \begin{cases} [\Delta_k, \infty) & \text{if } \rho_k \geq \eta_2, \\ [\gamma_2 \Delta_k, \Delta_k] & \text{if } \rho_k \in [\eta_1, \eta_2), \\ [\gamma_1 \Delta_k, \gamma_2 \Delta_k] & \text{if } \rho_k < \eta_1, \end{cases}$$

where $0 < \gamma_1 \leq \gamma_2 < 1$ are predefined parameters. For an in-depth study and a comprehensive reference on trust-region methods we refer the reader to the monograph of Conn et al. (2000).

A trust-region algorithm was recently developed by Colson et al. (2005a) for solving nonlinear bilevel programs where the function G depends solely on the upper-level vector x. This is not the first attempt to solve bilevel programs by means of a trust-region methods. Indeed, a related approach has been proposed by Liu et al. (1998) for problems that do not involve upper-level constraints, and where the lower-level program is strongly convex and linearly constrained. Under suitable assumptions, convergence to a Clarke stationary point may be proved. No computational experience has been reported.

The algorithm in Colson et al. (2005a) is an iterative method which, given the current iterate or incumbent solution $(\bar{x}, \bar{y})$, is based on the *linear-quadratic bilevel* model

$$\begin{array}{lll} \min\limits_{x \in X, y} & F_m(x, y) & \\ \text{s.t.} & G_m(x) \leq 0, & \\ & \min\limits_{y} & f_m(x, y) \\ & \text{s.t.} & g_m(x, y) \leq 0, \end{array} \tag{3.10}$$

of problem (2.1), where F_m, G_m and g_m are linear models of F, G, g at $(\bar{x}, \bar{y})$ respectively, while f_m is a quadratic model of f at $(\bar{x}, \bar{y})$. The bilevel problem (3.10) thus defines the trust-region subproblem. This subproblem can be solved for its global solution either by using a specialized algorithm—e.g., Jaumard et al. (2000)—, either by reformulating it as a mixed integer program (see Marcotte and Savard 2005) and resorting to an off-the-shelf software.

Let (x^m, y^m) denote the solution of the subproblem, that may fail to be rational. In order to evaluate the true value of this solution, one must compute the lower-level reaction to x^m,

[5]Note that, if the model is not accurate, there could be a deterioration of the objective, i.e., the ratio could be negative.

i.e., the optimal solution of

$$\begin{aligned} \min_{y} \quad & f(x^m, y) \\ \text{s.t.} \quad & g(x^m, y) \leq 0, \end{aligned} \tag{3.11}$$

which is denoted by y^*. After computation of the ratio (3.9) of achieved versus predicted reduction

$$\rho_k = \frac{F(x_k, y_k) - F(x^m, y^*)}{F_m(x_k, y_k) - F_m(x^m, y^m)},$$

the algorithm updates both the current iterate and the trust-region radius, and the process is repeated until convergence occurs.

This algorithm has been tested on a set of test problems, including toll-setting problems described in Sect. 1. The good performance of the algorithm in terms of the quality of the solution (a global solution is frequently reached) is due to the accuracy of the model approximation, itself a bilevel program that can be solved for its global solution.

4 Mathematical programs with equilibrium constraints

Having reviewed the major developments in the field of bilevel programming, we would like to complete our survey by considering another important class of related problems, namely *Mathematical Programs with Equilibrium Constraints*, or MPECs. Actually, relationships between BLPPs and MPECs are so strong that some authors use the same terminology for both classes of problems, which may sometimes lead to confusion.

MPECs may be viewed as bilevel programs where the lower-level problem consists in a *variational inequality*. For a given function $\psi : \mathbb{R}^n \to \mathbb{R}^n$ and convex set $C \subseteq \mathbb{R}^n$, the vector $x^* \in C$ is said to be a solution of the variational inequality VI (ψ, C) if it satisfies

$$(x - x^*)^T \psi(x^*) \geq 0 \quad \text{for all} \quad x \in C. \tag{4.1}$$

Note that variational inequalities are mathematical programs that allow the modelling of many *equilibrium* phenomena encountered in engineering, physics, chemistry or economics, hence the origin of the name of MPEC. The reader is referred to the monograph of Luo et al. (1996) for a description of fields of application.

The general formulation of an MPEC is as follows:

$$\begin{aligned} \min_{x,y} \quad & F(x, y) \\ \text{s.t.} \quad & (x, y) \in Z \quad \text{and} \quad y \in \mathcal{S}(x), \end{aligned} \tag{4.2}$$

where $Z \subseteq \mathbb{R}^{n_1+n_2}$ is a nonempty closed set and $\mathcal{S}(x)$ is the solution set of the parameterized variational inequality

$$y \in \mathcal{S}(x) \quad \Leftrightarrow \quad y \in C(x) \quad \text{and} \quad (v - y)^T \psi(x, y) \geq 0 \quad \text{for all } v \in C(x) \tag{4.3}$$

defined over the closed convex set $C(x) \subset \mathbb{R}^{n_2}$. As for bilevel problems, the terms *upper-level* and *lower-level* variables are used to designate x and y respectively.

The relationship between bilevel programming problems and MPECs may be illustrated by considering two particular cases. To this end, let us first assume that the mapping $\psi(x,\cdot)$ is the partial gradient map of a real-valued continuously differentiable function $f:\mathbb{R}^{n_1+n_2}\to\mathbb{R}$, that is,

$$\psi(x,y)=\nabla_y f(x,y).$$

Then, for any fixed x, the VI (4.3) characterizes the set of stationarity conditions of the optimization problem

$$\begin{aligned}\min_y \quad & f(x,y)\\ \text{s.t.} \quad & y\in C(x).\end{aligned}\tag{4.4}$$

Moreover, if the parameterized set $C(x)$ takes the form

$$C(x)=\{y\ :\ g(x,y)\le 0\},$$

then problem (4.4) is nothing but the lower-level problem (2.1c–2.1d). This shows that MPECs subsume bilevel programs provided the latter involves a *convex* and differentiable lower-level problem. Conversely, an MPEC can be formulated as a bilevel program by replacing the lower-level VI by an optimization problem. The latter can, for instance, be constructed around the *gap function* as defined by Auslender (1976) ($\alpha=0$) or Fukushima (1992) ($\alpha>0$), that is

$$H_\alpha(x,y)=\max_{z\in C(x)}\phi(x,y,z)\tag{4.5}$$

where

$$\phi(x,y,z)=\langle\psi(x,y),y-z\rangle-\frac{1}{2}\alpha\|y-z\|^2.$$

Alternatively, one may minimize the sum of the complementarity slackness terms associated with the KKT formulation of the variational inequality, under suitable regularity conditions on the lower-level constraints.

Many *active set* approaches have been proposed for solving MPECs, see Bard (1998). More recent methods deal with *constraint regularization* (e.g. Facchinei et al. 1996, Fukushima and Pang 1999, or Scholtes 2001), *implicit programming techniques* (Outrata 1994, Outrata et al. 1998), or techniques borrowed from constrained nonlinear programming. For instance, *filter methods* have been used by Fletcher and Leyffer (2002b), while *sequential quadratic programming* approaches have been proposed by Jian and Ralph (1997), Ralph (1998) and Fletcher et al. (2002). Marcotte and Zhu (1996) discuss algorithms based on *penalty functions*, exact or inexact, constructed around gap functions (4.5). A *trust-region* scheme has been developed by Scholtes and Stöhr (1999), which is closely related to the algorithm of Colson et al. (2005a) for nonlinear bilevel programming as presented in the previous section.

The reader interested in a comprehensive analysis of MPECs is referred to the monographs of Luo et al. (1996) and Outrata et al. (1998). Finally, we mention that an environment for modelling and testing MPEC problems based on the algebraic system GAMS has been implemented by Dirkse and Ferris (1999).

5 Perspectives and challenges

As evidenced in this survey, bilevel programming is the subject of important research efforts from the mathematical programming and operations research communities. Many classes of bilevel programs now have dedicated solution algorithms, and researchers have started to study more complicated instances—like bilevel programs with integer variables or without derivatives—which to our view is an indication that some maturity has been reached in the field.

It is nevertheless the case that challenges remain to be tackled, in particular concerning nonlinear bilevel problems. Besides the improvement of existing methods and derivation of proper convergence results, our feeling is that a promising approach would be to develop tools similar to those by Scholtes (2002) allowing to take advantage of the inherent combinatorial structure of bilevel problems. These ideas, combined with well-tried tools from nonlinear programming like sequential quadratic programming, should allow the development of a new generation of solution methods.

From a more practical point of view, we feel that the set of available bilevel programming test problems is relatively poor compared to those existing for other classes of mathematical programs. There exist some collections, like the MacMPEC collection maintained by Leyffer (2000) or the problems presented in Colson (2002), but no modelling language currently allows a user-friendly embedding of the two-level structure (in MacMPEC, for instance, problems are reformulated as single-level programs using optimality conditions). This issue, together with the development of a suitable modelling language, might trigger advances in the numerical solution of BLPPs.

References

Aiyoshi, E., & Shimizu, K. (1981). Hierarchical decentralized systems and its new solution by a barrier method. *IEEE Transactions on Systems, Man, and Cybernetics*, *11*, 444–449.

Aiyoshi, E., & Shimizu, K. (1984). A solution method for the static constrained Stackelberg problem via penalty method. *IEEE Transactions on Automatic Control*, *29*, 1111–1114.

Al-Khayal, F. A., Horst, R., & Pardalos, P. M. (1992). Global optimization of concave functions subject to quadratic constraints: an application in nonlinear bilevel programming. *Annals of Operations Research*, *34*, 125–147.

Anandalingam, G., & Friesz, T. (1992). Hierarchical optimization: an introduction. *Annals of Operations Research*, *34*, 1–11.

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

Auslender, A. (1976). *Optimisation: méthodes numériques*. Paris: Masson.

Bard, J. F. (1988). Convex two-level optimization. *Mathematical Programming*, *40*, 15–27.

Bard, J. F. (1998). *Practical bilevel optimization*. *Nonconvex optimization and its applications* (Vol. *30*) Dordrecht: Kluwer Academic.

Bard, J. F., & Falk, J. (1982). An explicit solution to the multi-level programming problem. *Computers and Operations Research*, *9*, 77–100.

Bard, J. F., & Moore, J. T. (1990). A branch and bound algorithm for the bilevel programming problem. *SIAM Journal of Scientific and Statistical Computing*, *11*, 281–292.

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

Ben-Ayed, O. (1993). Bilevel linear programming. *Computers and Operations Research*, *20*, 485–501.

Ben-Ayed, O., & Blair, C. (1990). Computational difficulties of bilevel linear programming. *Operations Research*, *38*, 556–560.

Bi, Z., Calamai, P. H., & Conn, A. R. (1989). An exact penalty function approach for the linear bilevel programming problem. Technical Report #167-O-310789, Department of Systems Design Engineering, University of Waterloo.

Bi, Z., Calamai, P. H., & Conn, A. R. (1991). An exact penalty function approach for the nonlinear bilevel programming problem. Technical Report #180-O-170591, Department of Systems Design Engineering, University of Waterloo.

Bialas, W., & Karwan, M. (1984). Two-level linear programming. *Management Science*, *30*, 1004–1020.

Bialas, W., Karwan, M., & Shaw, J. (1980). A parametric complementarity pivot approach for two-level linear programming. Technical Report 80-2, State University of New York at Buffalo, Operations Research Program.

Bracken, J., & McGill, J. (1973). Mathematical programs with optimization problems in the constraints. *Operations Research*, *21*, 37–44.

Bracken, J., & McGill, J. (1974). Defense applications of mathematical programs with optimization problems in the constraints. *Operations Research*, *22*, 1086–1096.

Bracken, J., & McGill, J. (1978). Production and marketing decisions with multiple objectives in a competitive environment. *Journal of Optimization Theory and Applications*, *24*, 449–458.

Brotcorne, L., Labbé, M., Marcotte, P., & Savard, G. (2001). A bilevel model for toll optimization on a multicommodity transportation network. *Transportation Science*, *35*, 1–14.

Candler, W., & Norton, R. (1977). Multilevel programming. Technical Report 20, World Bank Development Research Center, Washington D.C., USA.

Candler, W., & Townsley, R. (1982). A linear two-level programming problem. *Computers and Operations Research*, *9*, 59–76.

Case, L. M. (1999). An ℓ_1 penalty function approach to the nonlinear bilevel programming problem. PhD thesis, University of Waterloo, Ontario, Canada.

Chen, Y., & Florian, M. (1992). On the geometric structure of linear bilevel programs: a dual approach. Technical Report CRT-867, Centre de Recherche sur les Transports, Université de Montréal, Montréal, QC, Canada.

Chen, Y., Florian, M., & Wu, S. (1992). A descent dual approach for linear bilevel programs. Technical Report CRT-866, Centre de Recherche sur les Transports, Université de Montréal, Montréal, QC, Canada.

Clarke, F. H. (1990). *Optimization and nonsmooth analysis*. Philadelphia: SIAM.

Colson, B. (September 1999). Mathematical programs with equilibrium constraints and nonlinear bilevel programming problems. Master's thesis, Department of Mathematics, FUNDP, Namur, Belgium.

Colson, B. (October 2002). BIPA (BIlevel programming with approximation methods): software guide and test problems. Technical Report CRT-2002-38, Centre de Recherche sur les Transports, Université de Montréal, Montréal, QC, Canada.

Colson, B., Marcotte, P., & Savard, G. (March 2005a). A trust-region method for nonlinear programming: algorithm and computational experience. *Computational Optimization and Applications*, *30*.

Colson, B., Marcotte, P., & Savard, G. (2005b). Bilevel programming: a survey. *4OR*, *3*, 87–107.

Conn, A. R., Gould, N. I. M., & Toint, Ph. L. (2000). *Trust-region methods*. Philadelphia: SIAM.

Constantin, I., & Florian, M. (1995). Optimizing frequencies in a transit network: a nonlinear bi-level programming approach. *International Transactions in Operational Research*, *2*, 149–164.

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

Dempe, S. (1992a). A necessary and a sufficient optimality condition for bilevel programming problems. *Optimization*, *25*, 341–354.

Dempe, S. (1992b). Optimality conditions for bilevel programming problems. In P. Kall (Ed.), *System modelling and optimization* (pp. 17–24). Heidelberg: Springer.

Dempe, S. (2002). *Foundations of bilevel programming*. *Nonconvex optimization and its applications*, (Vol. *61*). Dordrecht: Kluwer Academic.

Dempe, S. (2003). Annotated bibliography on bilevel programming and mathematical programs with equilibrium constraints. *Optimization*, *52*, 333–359.

Dirkse, S. P., & Ferris, M. C. (1999). Modeling and solution environments for MPEC: gams & matlab. In M. Fukushima, & L. Qi (Eds.), *Reformulation: nonsmooth, piecewise smooth, semismooth and smoothing methods* (pp. 127–148). Dordrecht: Kluwer Academic.

Edmunds, T., & Bard, J. F. (1991). Algorithms for nonlinear bilevel mathematical programs. *IEEE Transactions on Systems, Man, and Cybernetics*, *21*, 83–89.

Facchinei, F., Jiang, H., & Qi, L. (1996). A smoothing method for mathematical programs with equilibrium constraints. Technical Report R03-96, Università di Roma "La Sapienza", Dipartimento di Informatica e Sistemistica.

Falk, J. E., & Liu, J. (1995). On bilevel programming, Part I : general nonlinear cases. *Mathematical Programming*, *70*, 47–72.

Fletcher, R., & Leyffer, S. (2002). Numerical experience with solving MPECs by nonlinear programming methods. Numerical Analysis Report NA/210, Department of Mathematics, University of Dundee, Dundee, Scotland.

Fletcher, R., Leyffer, S., Ralph, D., & Scholtes, S. (2002). Local convergence of SQP methods for Mathematical Programs with Equilibrium Constraints. Numerical Analysis Report NA/209, Department of Mathematics, University of Dundee, Dundee, Scotland.

Florian, M., & Chen, Y. (1995). A coordinate descent method for the bi-level o-d matrix adjustment problem. *International Transactions in Operational Research*, *2*, 165–179.

Fortuny-Amat, J., & McCarl, B. (1981). A representation and economic interpretation of a two-level programming problem. *Journal of the Operational Research Society*, *32*, 783–792.

Fukushima, M. (1992). Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems. *Mathematical Programming*, *53*, 99–110.

Fukushima, M., & Pang, J.-S. (1999). Complementarity constraint qualifications and simplified B-stationarity conditions for mathematical programs with equilibrium constraints. *Computational Optimization and Applications*, *13*, 111–136.

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: a guide to the theory of NP-completeness*. New York: Freeman.

Hansen, P., Jaumard, B., & Savard, G. (September 1992). New branch-and-bound rules for linear bilevel programming. *SIAM Journal on Scientific and Statistical Computing*, *13*, 1194–1217.

Haurie, A., Loulou, R., & Savard, G. (1992). A two player game model of power cogeneration in new england. *IEEE Transactions on Automatic Control*, *37*, 1451–1456.

Hearn, D. W., & Ramana, M. V. (1998). Solving congestion toll pricing models. In P. Marcotte (Ed.), *Equilibrium and advanced transportation modelling* (pp. 109–124). Dordrecht: Kluwer Academic.

Hobbs, B. F., & Nelson, S. K. (1992). A nonlinear bilevel model for analysis of electric utility demand-side planning issues. *Annals of Operations Research*, *34*, 255–274.

Hsu, S., & Wen, U. (1989). A review of linear bilevel programming problems. In *Proceedings of the National Science Council, Republic of China, Part A: Physical Science and Engineering* (Vol. 13, pp. 53–61).

Ishizuka, Y., & Aiyoshi, E. (1992). Double penalty method for bilevel optimization problems. *Annals of Operations Research*, *34*, 73–88.

Jaumard, B., Savard, G., & Xiong, J. (January 2000). A new algorithm for the convex bilevel programming problem. Technical Report, Ecole Polytechnique de Montréal.

Jeroslow, R. G. (1985). The polynomial hierarchy and a simple model for competitive analysis. *Mathematical Programming*, *32*, 146–164.

Jian, H., & Ralph, D. (December 1997). Smooth SQP methods for mathematical programs with nonlinear complementarity constraints. Manuscript, Department of Mathematics and Statistics, University of Melbourne.

Júdice, J. J., & Faustino, A. (1988). The solution of the linear bilevel programming problem by using the linear complementarity problem. *Investigação Operacional*, *8*, 77–95.

Júdice, J. J., & Faustino, A. (1992). A sequential LCP method for bilevel linear programming. *Annals of Operations Research*, *34*, 89–106.

Júdice, J. J., & Faustino, A. (1994). The linear-quadratic bilevel programming problem. *Information Systems and Operational Research*, *32*, 87–98.

Kara, B. Y., & Verter, V. (2004). Designing a road network for hazardous materials transportation. *Transportation Science*, *38*, 188–196.

Kolstad, C. D. (1985). A review of the literature on bi-level mathematical programming. Technical Report LA-10284-MS, Los Alamos National Laboratory, Los Alamos, New Mexico, USA.

Kolstad, C. D., & Lasdon, L. S. (1990). Derivative estimation and computational experience with large bilevel mathematical programs. *Journal of Optimization Theory and Applications*, *65*, 485–499.

Labbé, M., Marcotte, P., & Savard, G. (1998). A bilevel model of taxation and its applications to optimal highway pricing. *Management Science*, *44*, 1595–1607.

Larsson, T., & Patriksson, M. (1998). Side constrained traffic equilibrium models–traffic management through link tolls. In P. Marcotte, S. Nguyen (Eds.), *Equilibrium and advanced transportation modelling* (pp. 125–151). Dordrecht: Kluwer Academic.

LeBlanc, L. J. (1973). Meathematical programming algorithms for large scale network equilibrium and network design problems. PhD thesis, Northwestern University, Evanston, Illinois.

Leyffer, S. MacMPEC–AMPL collection of Mathematical Programs with Equilibrium Constraints. Available at http://www-unix.mcs.anl.gov/~leyffer/MacMPEC/.

Liu, G., Han, J., & Wang, S. (May 1998). A trust region algorithm for bilevel programming problems. *Chinese Science Bulletin*, *43*, 820–824.

Loridan, P., & Morgan, J. (1996). Weak via strong Stackelberg problem: new results. *Journal of Global Optimization*, *8*, 263–287.

Luo, Z.-Q., Pang, J.-S., & Ralph, D. (1996). *Mathematical programs with equilibrium constraints*. Cambridge: Cambridge University Press.

Marcotte, P. (1986). Network design problem with congestion effects: a case of bilevel programming. *Mathematical Programming*, *34*, 23–36.
Marcotte, P., & Savard, G. (2005). Bilevel programming: a combinatorial perspective. In: D. Avis, A. Hertz, & O. Marcotte (Eds.), *Graph theory and combinatorial optimization*. Boston: Kluwer Academic.
Marcotte, P., & Zhu, D. L. (1996). Exact and inexact penalty methods for the generalized bilevel programming problem. *Mathematical Programming*, *74*, 141–157.
Marcotte, P., Savard, G., & Semet, F. (2004). A bilevel programming approach to the travelling salesman problem. *Operations Research Letters*, *32*, 240–248.
Migdalas, A. (1995). Bilevel programming in traffic planning: models, methods and challenge. *Journal of Global Optimization*, *7*, 381–405.
Migdalas, A., Pardalos, P. M., & Värbrand, P. (Eds.) (1997). *Multilevel optimization: algorithms and applications, nonconvex optimization and its applications* (Vol. *20*). Dordrecht: Kluwer Academic.
Nash, J. (1951). Non-cooperative games. *Annals of Mathematics*, *54*, 286–295.
Outrata, J. (1993). Necessary optimality conditions for Stackelberg problems. *Journal of Optimization Theory and Applications*, *76*, 305–320.
Outrata, J. (1994). On optimization problems with variational inequality constraints. *SIAM Journal on Optimization*, *4*, 340–357.
Outrata, J., & Kocvara, M. (2006). Effective reformulations of the truss topology design problem. *Optimization and Engineering*, *7*, 201–219.
Outrata, J., Kočvara, M., & Zowe, J. (1998). *Nonsmooth approach to optimization problems with equilibrium constraints: theory, applications and numerical results, Nonconvex optimization and its applications* (Vol. 28). Dordrecht: Kluwer Academic.
Pang, J. S., & Trinkle, J. C. (1996). Complementarity formulations and existence of solutions of dynamic multi-rigid-body contact problems with coulomb friction. *Mathematical Programming*, *73*, 199–226.
Papavassilopoulos, G. (1982). Algorithms for static Stackelberg games with linear costs and polyhedral constraints. In *Proceedings of the 21st IEEE Conference on Decisions and Control* (pp. 647–652).
Patriksson, M. (1994). *The traffic assignment problem—models and methods*. Utrecht: VSP BV.
Ralph, D. (November 1998). Optimization with equilibrium constraints: a piecewise SQP approach. Manuscript, Department of Mathematics and Statistics, University of Melbourne.
Savard, G. (April 1989). Contribution à la programmation mathématique à deux niveaux. PhD thesis, Ecole Polytechnique de Montréal, Université de Montréal, Montréal, QC, Canada.
Savard, G., & Gauvin, J. (1994). The steepest descent direction for the nonlinear bilevel programming problem. *Operations Research Letters*, *15*, 265–272.
Scheel, H., & Scholtes, S. (2000). Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity. *Mathematics of Operations Research*, *25*, 1–22.
Scholtes, S. (2001). Convergence properties of a regularisation scheme for mathematical programs with complementarity constraints. *SIAM Journal on Optimization*, *11*, 918–936.
Scholtes, S. (May 2002). Combinatorial structures in nonlinear programming. Optimization Online. Available on the Internet at the address http://www.optimization-online.org/DB_HTML/2002/05/477.html.
Scholtes, S., & Stöhr, M. (1999). Exact penalization of mathematical programs with equilibrium constraints. *SIAM Journal on Control and Optimization*, *37*, 617–652.
Sherali, H. D., Soyster, A. L., & Murphy, F. H. (1983). Stackelberg-Nash-Cournot equilibria: characterizations and computations. *Operations Research*, *31*, 253–276.
Shimizu, K., & Aiyoshi, E. (1981). A new computational method for Stackelberg and min-max problems by use of a penalty method. *IEEE Transactions on Automatic Control*, *26*, 460–466.
Shimizu, K., Ishizuka, Y., & Bard, J. F. (1997). *Nondifferentiable and two-level mathematical programming*. Dordrecht: Kluwer Academic.
Stackelberg, H. (1952). *The theory of market economy*. Oxford: Oxford University Press.
Thoai, N. V., Yamamoto, Y., & Yoshise, A. (May 2002). Global optimization method for solving mathematical programs with linear complementarity constraints. Discussion Paper No. 987, Institute of Policy and Planning Sciences, University of Tsukuba, Japan.
Tuy, H., Migdalas, A., & Värbrand, P. (1993). A global optimization approach for the linear two-level program. *Journal of Global Optimization*, *3*, 1–23.
Van Ackere, A. (1993). The principal/agent paradigm: characterizations and computations. *European Journal of Operations Research*, *70*, 83–103.
Vicente, L. N., & Calamai, P. H. (1994). Bilevel and multilevel programming: a bibliography review. *Journal of Global Optimization*, *5*, 291–306.
Vicente, L. N., & Calamai, P. H. (1995). Geometry and local optimality conditions for bilevel programs with quadratic strictly convex lower levels. In: D. Du, & M. Pardalos (Eds.), *Minimax and applications. Nonconvex optimization and its applications* (Vol. *4*, pp. 141–151). Dordrecht: Kluwer Academic.

Vicente, L. N., Savard, G., & Júdice, J. J. (1994). Descent approaches for quadratic bilevel programming. *Journal of Optimization Theory and Applications*, *81*, 379–399.

Vicente, L. N., Savard, G., & Júdice, J. J. (1996). The discrete linear bilevel programming problem. *Journal of Optimization Theory and Applications*, *89*, 597–614.

Wen, U., & Hsu, S. (1991). Linear bi-level programming problems—a review. *Journal of the Operational Research Society*, *42*, 125–133.

Ye, J. J., & Zhu, D. L. (1995). Optimality conditions for bilevel programming problems. *Optimization*, *33*, 9–27.

Yezza, A. (1996). First-order necessary optimality conditions for general bilevel programming problems. *JOTA*, *89*, 189–219.