Mixed-integer Bilevel Optimization for Capacity Planning with Rational Markets

Pablo Garcia-Herreros[a], Lei Zhang[b], Pratik Misra[c], Sanjay Mehta[c], and Ignacio E. Grossmann[a]

[a]Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, USA

[b]Department of Chemical Engineering, Tsinghua University, Beijing, China

[c]Air Products and Chemicals, Inc. Allentown, USA

April 30, 2015

Abstract

We formulate the capacity expansion planning as a bilevel optimization to model the hierarchical decision structure involving industrial producers and consumers. The formulation is a mixed-integer bilevel linear program in which the upper level maximizes the profit of a producer and the lower level minimizes the cost paid by markets. The upper-level problem includes mixed-integer variables that establish the expansion plan; the lower level problem is an LP that decides demands assignments. We reformulate the bilevel optimization as a single-level problem using two different approaches: KKT reformulation and duality-based reformulation. We analyze the performance of these reformulations and compare their results with the expansion plans obtained from the traditional single-level formulation. For the solution of large-scale problems, we propose improvements on the duality-based reformulation that allows reducing the number of variables and constraints. The formulations and the solution methods are illustrated with examples from the air separation industry.

Keywords: Capacity planning; Bilevel optimization; Rational markets.

1 Introduction

Capacity expansion is one of the most important strategic decisions for industrial gas companies. In this industry, most of the markets are served by local producers because of the competitive advantage given by the location of the production facilities. The dynamics of the industrial gas markets imply that companies must anticipate demand increases in order to plan their capacity expansion, maintain supply availability, and avoid regional incursion of new producers. The selection of the right investment and distribution plan plays a critical role for companies in this environment. A rigorous approach based on mathematical modeling and optimization offers the possibility to find the investment and distribution plan that yields the greatest economic benefit.

A rather large body of literature has been published on capacity planning problems in several industries [20]. Since the late 1950s, capacity expansion planning has been studied to develop models and solution approaches for diverse applications in the process industries [30], communication networks [7], electric power services [24], and water resource systems [25]. Sahinidis et al. [31] proposed a comprehensive MILP model for long range planning of process networks. Van den Heever and Grossmann [35] used disjunctive programming to extend this methodology to multi-period design and planning of nonlinear chemical processes. An MILP formulation that integrates scheduling with capacity planning for product development was presented by Maravelias and Grossmann [21]. Sundaramoorthy et al. [34] proposed a two-stage stochastic programming formulation for the integration of capacity and operations planning. In summary, capacity planning is considered a central problem for enterprise-wide optimization, a topic for which comprehensive reviews are available [15, 16].

Despite the importance of capacity expansion in industry, the study of the problem in a competitive environment has not received much attention. Soyster and Murphy [33] formulated a capacity planning problem for a perfectly competitive market. However, perfect competition is a strong assumption. A more realistic hypothesis is to assume an oligopolistic market as presented by Murphy and Smeers [23]. Game theory models have also been used [37] for the supply chain planning in cooperative and competitive environments.

The competition between two players whose decisions are made sequentially can be modeled as a Stackelberg game [36]. A Stackelberg competition is an extensive game with perfect information in which the leader chooses his actions before the follower has the opportunity to play. It is known that the most interesting equilibria of such games correspond to the solution of a bilevel optimization problem [26].

Bilevel optimization problems are mathematical programs with optimization problems in the constraints [4]. They are suitable to model problems in which two independent decision makers try to optimize their own objective functions [6, 2]. We present a mixed-integer linear bilevel formulation for the capacity expansion planning of an industrial gas company operating in a competitive environment. The purpose of the upper-level problem is to determine the investment and distribution plan that maximizes the Net Present Value (NPV). The response of markets that can choose among different producers is modeled in the lower-level as a linear programming (LP) problem. The lower-level objective function is selected to represent the rational behavior of the markets.

Solution approaches for bilevel optimization problems with lower-level LPs leverage the fact that optimal solutions occur at vertexes of the region described by upper and lower level constraints. They rely on vertex enumeration, directional derivatives, penalty terms, or optimality conditions [29]. The most direct approach is to reformulate the bilevel optimization as a single-level problem using the optimality conditions of the lower-level LP. The classic reformulation using Karush-Kuhn-Tucker (KKT) conditions maintains linearity of the problem except for the introduction of complementarity constraints [12, 1, 3]. An equivalent reformulation replaces the lower level problem by its primal and dual constraints, and guarantees optimality by enforcing strong duality [22, 14].

Strategic investment planning for electric power networks has been the most prolific application of bilevel optimization models. Motto et al. [22] implemented the duality-based reformulation for the analysis of electric grid security under disruptive threat. This bilevel problem was originally formulated by Salmeron et al. [32] with the purpose of identifying the interdictions that maximize network disruptions. A bilevel formulation for the expansion of transmission networks was developed by Garces et al. [14] to maximize the average social welfare over a set of lower-level problems representing different market clearing scenarios; they also implemented the duality-based reformulation. Ruiz et al. [27] modeled electricity markets as an Equilibrium Problem with Equilibrium Constraints (EPEC) in which competing producers maximize their profit in the upper level and a market operator maximizes social welfare in the lower level; they use the duality-based reformulation to guarantee optimality of the lower level problem and obtain an equilibrium solution by jointly formulating the KKT conditions of all producers. A similar strategy that includes the combination of duality-based and KKT reformulations was used by Huppmann and Egerer [18] to solve a three-level optimization problem that models the roles of independent system operators, regional planners, and supra-national coordination in the European energy system.

Another interesting application of bilevel optimization is the facility location problem in a duopolistic environment. The model presented by Fischer [11] selects facilities among a set of candidate locations and considers selling prices as optimization variables, which leads to a nonlinear bilevel formulation. The problem is simplified to a linear discrete bilevel formulation under the assumption that Nash equilibrium is reached for the prices. The solution to the discrete bilevel optimization problem is obtained using a heuristic algorithm.

Bilevel optimization models have also found application in chemical engineering. Clark and Westerberg [9] presented a nonlinear bilevel programming approach for the design of chemical processes and proposed algorithms to solve them. In their formulation, the upper level optimizes the process design and the lower level models thermodynamic equilibrium by minimizing Gibbs free energy. Burgard and Maranas [5] used bilevel optimization to test the consistency of experimental data obtained from metabolic networks with hypothesized objective functions. In the upper level, the problem minimizes the square deviation of the fluxes predicted by the metabolic model with respect to experimental data, whereas the lower level quantifies the individual importance of the fluxes. A bilevel programming model for supply chain optimization under uncertainty was developed by Ryu et al. [28]; the conflicting interests of production and distribution operations in a supply chain are modeled using separate objective functions. They reformulate the bilevel problem in single-level after finding the solution of the lower-level problem as parametric functions of the upper-level variables and the uncertain parameters. Chu and You [8] presented an integrated scheduling and dynamic optimization problem for batch processes. The scheduling problem, formulated in the upper level, is subject to the processing times and costs determined by the nonlinear dynamic lower-level problem. The bilevel formulation is transformed to a single level problem by replacing the lower-level with piece-wise linear response functions. They assert that the bilevel formulation can be used as a distributed optimization approach whose solutions can easily adapt to variation in the problem's parameters.

The novelty of our research resides on the application of bilevel optimization for capacity expansion planning in a competitive environment. Bilevel programming for these kind of problems can be seen as a risk mitigation strategy given the significant influence of external decision makers in the economic success of investment plans. In particular, we propose a mathematical model that includes a rational market behavior beyond the classic game theoretical models. The investment plans obtained from this approach are found to be less sensitive to changes in the business environment in comparison to the single-level formulations.

In order to solve the challenging bilevel formulation, we test the KKT and the duality-based reformulations with an illustrative example, a middle-size example, and an industrial example. The results show the advantages of the duality-based reformulation in terms of computational effort. Despite the efficiency obtained with this reformulation, we found necessary to implement two additional improvement strategies to solve large-scale instances.

The remaining paper is organized as follows. In Section 2, we describe the problem. In Section 3, we present the single-level capacity planning formulation. Section 4 presents the bilevel capacity planning problem with rational markets. In Section 5, we develop two reformulations that allow solving the bilevel optimization problem. Section 6, presents a small example that illustrates the proposed formulations. Subsequently, in Section 7 we evaluate the performance of the proposed reformulations with a middle-size example. In section 8, we elaborate on solution approaches for large-scale bilevel capacity planning problems. Section 9 presents an industrial example. Finally, in Section 10 we present our analysis and conclusions.

2 Problem statement

A company that produces and commercializes industrial products in a given geographic region is interested in developing an investment plan to expand its capacity in anticipation of future demand increase. The company operates some facilities with limited production capacity. Existing facilities are eligible for capacity expansion and other locations are candidates to open new facilities. The construction and expansion of facilities requires the investment of capital to develop the project and install new production lines. The potential increases in production capacity are assumed to be discrete and the corresponding investments are given by fixed costs. Based on the available capacity in the facilities, the company allocates production to market demands. Figure 1 shows a schematic representation of a region with several industrial producers and gas markets.

The company obtains revenue from selling its products at fixed prices in each market. The goal of the company is to find the investment plan that maximizes the Net Present Value (NPV) of its profit during a finite time horizon. The NPV is calculated by applying the appropriate discount factor to the income received from sales and the expenses related to investment, production, maintenance, and transportation costs.

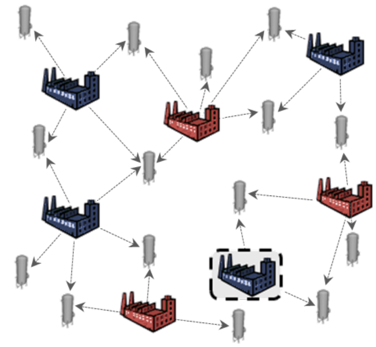

Figure 1: Superstructure of regional gas markets.

3 Single-level Capacity Planning with Captive Markets

The basic model to plan the capacity expansion of a company serving industrial markets assumes that all market demands are willing to buy the products at the price offered. In this context, markets are regarded as captive. The capacity expansion planning with captive markets can be formulated as the single-level Mixed-Integer Linear Program (MILP) presented in Eqns. (1) - (7).

$$\begin{aligned}
\max \quad & \sum_{t \in T} \sum_{i \in I^1} \sum_{j \in J} \sum_{k \in K} \frac{1}{(1+R)^t} P_{t,i,j,k} y_{t,i,j,k} \\
& - \sum_{t \in T} \sum_{i \in I^1} \frac{1}{(1+R)^t} \left(A_{t,i} v_{t,i} + B_{t,i} w_{t,i} \right) \\
& - \sum_{t \in T} \sum_{i \in I^1} \sum_{k \in K} \frac{1}{(1+R)^t} \left(E_{t,i,k} x_{t,i,k} + F_{t,i,k} \sum_{j \in J} y_{t,i,j,k} \right) \\
& - \sum_{t \in T} \sum_{i \in I^1} \sum_{j \in I} \sum_{k \in K} \frac{1}{(1+R)^t} G_{t,i,j,k} y_{t,i,j,k} && && (1) \\
\text{s.t.} \quad & w_{t,i} = V_{0,i} + \sum_{t' \in T'_t} v_{t',i} && \forall\ t \in T, i \in I^1 && (2) \\
& x_{t,i,k} \le w_{t,i} && \forall\ t \in T, i \in I^1, k \in K && (3) \\
& c_{t,i,k} = C_{0,i,k} + \sum_{t' \in T'_t} H_{i,k} x_{t',i,k} && \forall\ t \in T, i \in I^1, k \in K && (4) \\
& \sum_{j \in J} y_{t,i,j,k} \le c_{t,i,k} && \forall\ t \in T, i \in I^1, k \in K && (5) \\
& \sum_{i \in I^1} y_{t,i,j,k} \le D_{t,j,k} && \forall\ t \in T, j \in J, k \in K && (6) \\
& v_{t,i},\ w_{t,i},\ x_{t,i,k} \in \{0,1\}; \quad c_{t,i,k},\ y_{t,i,j,k} \in \mathbb{R}^+ && \forall\ t \in T, i \in I^1, j \in J, k \in K && (7)
\end{aligned}$$

where T, I^1, J, and K are respectively, the index sets for time (t), production facilities of the decision maker (i^1), markets (j), and products (k). We also define T' as the subset of time periods T in which expansions are allowed, and T'_t as the subset of time periods before t in which expansions are allowed. Formally, $T'_t = \{t' : t' \in T', t' \le t\}$.

The first term in expression (1) represents the income obtained from sales. Income is proportional to demand assignments ($y_{t,i,j,k}$) according to the price paid by the markets ($P_{t,i,j,k}$). The second term includes the cost of opening new facilities and the maintenance cost of open facilities. The binary variable deciding if a new facility is open at location i at time period t is $v_{t,i}$; parameter $A_{t,i}$ determines the fixed cost to build a new facility. The binary variable $w_{t,i}$ indicates if facility i is open at time period t; if the facility is open at period t, a fixed cost $B_{t,i}$ must be paid for maintenance. The third term includes expansion and production costs. The expansion of production capacity for product k in facility j at period t is decided with binary variable $x_{t,i,k}$; the cost of expansions is given by parameter $E_{t,i,k}$. Production costs are proportional to demand assignments ($y_{t,i,j,k}$) according to their unit production cost ($F_{t,i,k}$). Finally, the last term represents the transportation cost from production facilities to markets. Transportation is proportional to demand assignments ($y_{t,i,j,k}$) according to the unit transportation cost ($G_{t,i,j,k}$). All terms are discounted in every time period with an interest rate (R).

Constraint set (2) is used to model the maintenance cost of facilities during the time periods when they are open; the binary parameter V_i^0 indicates the facilities that are initially open. Constraint set (3) requires capacity expansions to take place only at open facilities. Constraint set (4) determines the production capacity of facilities according to the expansion decisions; parameters $C_{0,i,k}$ indicate the initial capacities and H_k is the magnitude of the potential capacity expansion. Constraints (5) bound the demand assignments according to the production capacities. Finally, constraints set (6) bounds demand assignments according to market demands. The domains of the variables are given by expressions (7).

4 Bilevel Capacity Planning with Rational Markets

The most intuitive way to model a competitive environment is to assume that the markets have the possibility to select their providers according to their own interest. The rational behavior of the markets can be modeled with a mathematical program that optimizes their objective function. The behavior of the markets is included in the constraints of the capacity planning problem, yielding a bilevel optimization formulation. In this formulation, the upper-level problem is intended to find the optimal capacity expansion plan by selecting the investments that maximize the NPV of the leader. The lower-level represents the response of markets that select production facilities as providers with the unique interest of satisfying their demands at lowest cost.

The formulation presented in Section 3 is modified to ensure that market demands are completely satisfied. This is achieved by transforming constraint set (6) into equality constraints. This change is necessary to avoid the market cost from dropping to zero by leaving all demands unsatisfied. Additionally, the set of potential providers is expanded to include facilities from independent producers. We assume that the initial capacity of all production facilities is large enough to satisfy all market demands regardless of the expansion plan of the leader. This assumption is also useful to avoid unprofitable investments in capacity expansions driven by the need to maintain feasibility of the problem.

The products offered by the competing producers are considered homogeneous and the markets have no other preference for producers than price. Cases in which the markets have no preference between two or more facilities are resolved by the upper level according to the interest of the leader; this modeling framework is known as the *optimistic approach* [19]. In our model, the *optimistic approach* is a key assumption because all facilities controlled by the leader offer the same price to each market. Therefore, the optimization problem of the markets is degenerate. However, the markets are only concerned about selecting the producer that offers the lowest price and they are indifferent to the facility from which they are served; consequently, the leader is free to choose the facilities it uses to satisfy its demands.

The bilevel optimization problem for the capacity expansion planning in a competitive environment is presented in Eqn. (8) - (18).

$$\begin{aligned}
\max_{v,w,x} \quad & \sum_{t\in T}\sum_{i\in I^1}\sum_{j\in J}\sum_{k\in K}\frac{1}{(1+R)^t}P_{t,i,j,k}y_{t,i,j,k} \\
& -\sum_{t\in T}\sum_{i\in I^1}\frac{1}{(1+R)^t}\left(A_{t,i}v_{t,i}+B_{t,i}w_{t,i}\right) \\
& -\sum_{t\in T}\sum_{i\in I^1}\sum_{k\in K}\frac{1}{(1+R)^t}\left(E_{t,i,k}x_{t,i,k}+F_{t,i,k}\sum_{j\in J}y_{t,i,j,k}\right) \\
& -\sum_{t\in T}\sum_{i\in I^1}\sum_{j\in J}\sum_{k\in K}\frac{1}{(1+R)^t}G_{t,i,j,k}y_{t,i,j,k} && (8)\\
\text{s.t.} \quad & w_{t,i}=V_i^0+\sum_{t'\in T_t'}v_{t',i} & \forall\, t\in T, i\in I^1 \quad & (9)\\
& x_{t,i,k}\le w_{t,i} & \forall\, t\in T, i\in I^1, k\in K \quad & (10)\\
& c_{t,i,k}=C_{i,k}^0+\sum_{t'\in T_t'}H_{i,k}x_{t',i,k} & \forall\, t\in T, i\in I^1, k\in K \quad & (11)\\
& \min_{y} \quad \sum_{t\in T}\sum_{i\in I}\sum_{j\in J}\sum_{k\in K}\frac{1}{(1+R)^t}P_{t,i,j,k}y_{t,i,j,k} && (12)\\
& \text{s.t.} \quad \sum_{j\in J}y_{t,i,j,k}\le c_{t,i,k} & \forall\, t\in T, i\in I^1, k\in K \quad & (13)\\
& \qquad \sum_{j\in J}y_{t,i,j,k}\le C_{i,k}^0 & \forall\, t\in T, i\in I^2, k\in K \quad & (14)\\
& \qquad \sum_{i\in I}y_{t,i,j,k}=D_{t,j,k} & \forall\, t\in T, j\in J, k\in K \quad & (15)\\
& \qquad y_{t,i,j,k}\in\mathbb{R}^+ & \forall\, t\in T, i\in I, j\in J, k\in K \quad & (16)\\
& c_{t,i,k}\in\mathbb{R}^+ & \forall\, t\in T, i\in I^1, j\in J, k\in K \quad & (17)\\
& v_{t,i},\ w_{t,i},\ x_{t,i,k}\in\{0,1\} & \forall\, t\in T, i\in I^1, k\in K \quad & (18)
\end{aligned}$$

where I is the set of all production facilities, $I^1\subset I$ is the subset of facilities controlled by the leader, and $I^2\subset I$ is the subset of facilities controlled by the competitors. It should be noted that Eqns. (8) - (11) are identical to Eqns. (1) - (4) in the single-level formulation. However, in the bilevel formulation the upper-level decision maker only controls variables $v_{t,i}$, $w_{t,i}$, $x_{t,i,k}$, and

$c_{t,i,k}$. Demand assignment decisions ($y_{t,i,j,k}$) are controlled by the lower level with the objective of minimizing the cost paid by the markets according to Eqn. (12). Eqns. (13) and (14) constrain the production capacity of the facilities; Eqn. (15) enforces demand satisfaction in every time period. The domains of the variables are presented in Eqns. (16) - (18). It is important to note that upper-level variables only take discrete values and all lower-level variables are continuous. This attribute of the model is crucial for the reformulations that we propose.

5 Reformulation as a Single-level Optimization Problem

An optimistic bilevel program with a convex and regular lower-level can be transformed into a single-level optimization problem using its optimality conditions [10]. The key property of convex programs is that their KKT conditions are necessary and sufficient to characterize their corresponding global optimal solutions. In the case of linear programs, KKT optimality conditions are equivalent to the satisfaction of primal feasibility, dual feasibility, and strong duality [13]. Based on this equivalence, we derive two single-level reformulations for the capacity planning problem in a competitive environment.

5.1 KKT Reformulation

The classic reformulation for bilevel programs with a lower-level LP is to replace the lower lower-level problem by its KKT conditions. In the case of the capacity planning in a competitive environment, the KKT reformulation is obtained by introducing constraints that guarantee the stationarity conditions, primal feasibility, dual feasibility, and complementary slackness for the cost minimization problem modeling markets behavior. The resulting reformulation is presented in Eqns. (19) - (32).

$$
\begin{aligned}
\max \quad & \sum_{t\in T}\sum_{i\in I^1}\sum_{j\in J}\sum_{k\in K}\frac{1}{(1+R)^t}P_{t,i,j,k}y_{t,i,j,k} \\
& -\sum_{t\in T}\sum_{i\in I^1}\frac{1}{(1+R)^t}\left(A_{t,i}v_{t,i}+B_{t,i}w_{t,i}\right) \\
& -\sum_{t\in T}\sum_{i\in I^1}\sum_{k\in K}\frac{1}{(1+R)^t}\left(E_{t,i,k}x_{t,i,k}+F_{t,i,k}\sum_{j\in J}y_{t,i,j,k}\right) \\
& -\sum_{t\in T}\sum_{i\in I^1}\sum_{j\in I}\sum_{k\in K}\frac{1}{(1+R)^t}G_{t,i,j,k}y_{t,i,j,k} && (19)\\
\text{s.t.} \quad & w_{t,i}=V_i^0+\sum_{t'\in T'_t}v_{t',i} \qquad \forall\, t\in T, i\in I^1 && (20)\\
& x_{t,i,k}\le w_{t,i} \qquad \forall\, t\in T, i\in I^1, k\in K && (21)\\
& c_{t,i,k}=C_{i,k}^0+\sum_{t'\in T'_t}H_{i,k}x_{t',i,k} \qquad \forall\, t\in T, i\in I^1, k\in K && (22)\\
& \sum_{j\in J}y_{t,i,j,k}\le c_{t,i,k} \qquad \forall\, t\in T, i\in I^1, k\in K && (23)
\end{aligned}
$$

$$\sum_{j \in J} y_{t,i,j,k} \leq C^0_{i,k} \qquad \forall\ t \in T, i \in I^2, k \in K \tag{24}$$

$$\sum_{i \in I} y_{t,i,j,k} = D_{t,j,k} \qquad \forall\ t \in T, j \in J, k \in K \tag{25}$$

$$\frac{1}{(1+R)^t} P_{t,i,k} + \lambda_{t,j,k} + \mu_{t,i,k} - \gamma_{t,i,j,k} = 0 \qquad \forall\ t \in T, i \in I, j \in J, k \in K \tag{26}$$

$$\mu_{t,i,k} \left(\sum_j y_{t,i,j,k} - c_{t,i,k} \right) = 0 \qquad \forall\ t \in T, i \in I^1, k \in K \tag{27}$$

$$\mu_{t,i,k} \left(\sum_j y_{t,i,j,k} - C^0_{i,k} \right) = 0 \qquad \forall\ t \in T, i \in I^2, k \in K \tag{28}$$

$$\gamma_{t,i,j,k} y_{t,i,j,k} = 0 \qquad \forall\ t \in T, i \in I, j \in J, k \in K \tag{29}$$

$$y_{t,i,j,k},\ \mu_{t,i,k},\ \gamma_{t,i,j,k} \in \mathbb{R}^+ \qquad \forall\ t \in T, i \in I, j \in J, k \in K \tag{30}$$

$$\lambda_{t,j,k} \in \mathbb{R}, \qquad \forall\ t \in T, j \in J, k \in K \tag{31}$$

$$c_{t,i,k} \in \mathbb{R}^+ \qquad \forall\ t \in T, i \in I^1, k \in K \tag{32}$$

$$v_{t,i},\ w_{t,i},\ x_{t,i,k} \in \{0,1\} \qquad \forall\ t \in T, i \in I^1, k \in K \tag{33}$$

where $\mu_{t,i,k}$, $\lambda_{t,j,k}$, and $\gamma_{t,i,j,k}$ are the Lagrange multipliers of the lower-level constraints presented in Eqns. (13) - (14), (15), and (16), respectively. The upper-level problem is kept unchanged as shown in Eqns. (19) - (22). Constraints (23) - (25) ensure primal feasibility of the lower level; the constraints presented in (26) are the stationary conditions for the lower level; Eqns. (27) and (28) represent the complementary conditions corresponding to inequalities (13) and (14); the constraints (29) are the complementary conditions corresponding to the domain of the lower-level variables presented in Eqn. (16). The domains are presented in Eqns. (30) - (33).

The main disadvantage associated to this reformulation is the introduction of non-linear complementary constraints. In order to avoid the solution of a nonconvex Mixed-Integer Non-Linear Program (MINLP), the complementary constraints can be formulated as disjunctions that are transformed into mixed-integer constraints [17]. In particular, we rewrite Eqns. (23) and (24) as equality constraints by introducing the slack variables $s_{t,i,k}$,

$$\sum_{j \in J} y_{t,i,j,k} + s_{t,i,k} = c_{t,i,k} \quad \forall t \in T, i \in I^1, k \in K \tag{34}$$

$$\sum_{j \in J} y_{t,i,j,k} + s_{t,i,k} = C^0_{i,k} \quad \forall t \in T, i \in I^2, k \in K \tag{35}$$

and use the Big-M reformulation to express that either constraints (34) and (35) are active or the corresponding multipliers ($\mu_{t,i,k}$) are zero. The Big-M constraints modeling this disjunction are presented in Eqn. (36) using binary variable $z^1_{t,i,k}$.

$$\begin{aligned} & s_{t,i,k} \leq M z^1_{t,i,k} \\ & \mu_{t,i,k} \leq M \left(1 - z^1_{t,i,k}\right) \qquad \forall t \in T, i \in I, k \in K \\ & z^1_{t,i,k} \in \{0,1\} \end{aligned} \tag{36}$$

Similarely, the Big-M reformulation of constraint set (29) is presented in Eqn. (37).

$$
\begin{aligned}
& y_{t,i,j,k} \leq M z^2_{t,i,k} \\
& \gamma_{t,i,j,k} \leq M \left(1 - z^2_{t,i,k}\right) && \forall t \in T, i \in I, j \in J, k \in K \qquad (37)\\
& z^2_{t,i,j,k} \in \{0,1\}
\end{aligned}
$$

The result of replacing constraints (23), (24), (27), (28) and (29) by (34), (35), (36), and (37) is a single-level MILP formulation that is equivalent to the bilevel formulation presented in Section 4.

5.2 Duality-based Reformulation

The alternative reformulation for the bilevel capacity planning problem is obtained by introducing constraints that guarantee the satisfaction of strong duality [22, 14]. This is achieved by replacing the lower-level problem described by Eqns. (12) - (16) with its primal and dual constraints, and equating their objective functions. The dual formulation corresponding to the lower-level LP is presented by Eqns. (38) - (41).

$$
\begin{aligned}
\max \quad & \sum_{t\in T}\sum_{k\in K}\left[\sum_{j\in J} D_{t,j,k}\lambda_{t,j,k} - \sum_{i\in I^1} c_{t,i,k}\mu_{t,i,k} - \sum_{i\in I^2} C^0_{i,k}\mu_{t,i,k}\right] && (38)\\
\text{s.t.} \quad & \lambda_{t,j,k} - \mu_{t,i,k} \leq \frac{1}{(1+R)^t} P_{t,i,j,k} && \forall\ t \in T, i \in I, j \in J, k \in K \quad (39)\\
& \mu_{t,i,k} \in \mathbb{R}^+ && \forall\ t \in T, i \in I, k \in K \quad (40)\\
& \lambda_{t,j,k} \in \mathbb{R}, && \forall\ t \in T, j \in J, k \in K \quad (41)
\end{aligned}
$$

The resulting duality-based reformulation is presented in Eqns. (42) - (54).

$$
\begin{aligned}
\max \quad & \sum_{t\in T}\sum_{i\in I^1}\sum_{j\in J}\sum_{k\in K} \frac{1}{(1+R)^t} P_{t,i,j,k} y_{t,i,j,k} \\
& - \sum_{t\in T}\sum_{i\in I^1} \frac{1}{(1+R)^t}\left(A_{t,i}v_{t,i} + B_{t,i}w_{t,i}\right) \\
& - \sum_{t\in T}\sum_{i\in I^1}\sum_{k\in K} \frac{1}{(1+R)^t}\left(E_{t,i,k}x_{t,i,k} + F_{t,i,k}\sum_{j\in J} y_{t,i,j,k}\right) \\
& - \sum_{t\in T}\sum_{i\in I^1}\sum_{j\in I}\sum_{k\in K} \frac{1}{(1+R)^t} G_{t,i,j,k} y_{t,i,j,k} && (42)\\
\text{s.t.} \quad & w_{t,i} = V^0_i + \sum_{t'\in T'_t} v_{t',i} && \forall\ t \in T, i \in I^1 \quad (43)\\
& x_{t,i,k} \leq w_{t,i} && \forall\ t \in T, i \in I^1, k \in K \quad (44)\\
& c_{t,i,k} = C^0_{i,k} + \sum_{t'\in T'_t} H_{i,k} x_{t',i,k} && \forall\ t \in T, i \in I^1, k \in K \quad (45)
\end{aligned}
$$

$$\sum_{t\in T}\sum_{i\in I}\sum_{j\in J}\sum_{k\in K}\frac{1}{(1+R)^t}P_{t,i,j,k}y_{t,i,j,k}$$
$$=\sum_{t\in T}\sum_{k\in K}\left[\sum_{j\in J}D_{t,j,k}\lambda_{t,j,k}-\sum_{i\in I^1}c_{t,i,k}\mu_{t,i,k}-\sum_{i\in I^2}C^0_{i,k}\mu_{t,i,k}\right] \tag{46}$$
$$\sum_{j\in J}y_{t,i,j,k}\le c_{t,i,k} \qquad \forall\ t\in T, i\in I^1, k\in K \tag{47}$$
$$\sum_{j\in J}y_{t,i,j,k}\le C^0_{i,k} \qquad \forall\ t\in T, i\in I^2, k\in K \tag{48}$$
$$\sum_{i\in I}y_{t,i,j,k}= D_{t,j,k} \qquad \forall\ t\in T, j\in J, k\in K \tag{49}$$
$$\lambda_{t,j,k}-\mu_{t,i,k}\le\frac{1}{(1+R)^t}P_{t,i,j,k} \qquad \forall\ t\in T, i\in I, j\in J, k\in K \tag{50}$$
$$y_{t,i,j,k},\ \mu_{t,i,k}\in\mathbb{R}^+ \qquad \forall\ t\in T, i\in I, j\in J, k\in K \tag{51}$$
$$\lambda_{t,j,k}\in\mathbb{R}, \qquad \forall\ t\in T, j\in J, k\in K \tag{52}$$
$$c_{t,i,k}\in\mathbb{R}^+ \qquad \forall\ t\in T, i\in I^1, k\in K \tag{53}$$
$$v_{t,i},\ w_{t,i},\ x_{t,i,k}\in\{0,1\} \qquad \forall\ t\in T, i\in I^1, k\in K \tag{54}$$

The upper-level problem represented by Eqns. (42) - (45) remains unchanged in the duality-based reformulation. Strong duality is enforced by equating the primal and dual objective functions as presented in Eqn. (46). Lower-level primal constraints (47) and (49) are kept in the formulation to guarantee primal feasibility. Dual feasibility of the lower level is ensured with constraints (50).

It must be noted that this reformulation yields a Mixed-Integer Nonlinear Program (MINLP). The nonlinearity arises from the dual objective function in the right hand side of Eqn. (46), because of the product of upper-level variable $c_{t,i,k}$ and lower-level dual variable $\mu_{t,i,k}$. Fortunately, the problem can be posed as a MILP because the variable $c_{t,i,k}$ only takes values in discrete increments as indicated by Eqn. (45). The linearization procedure is based on eliminating variable $c_{t,i,k}$ from the formulation by replacing it according to Eqn. (45). The resulting bilinear terms are products of continuous variables ($\mu_{t,i,k}$) and binary variables ($x_{t',i,k}$). Therefore, they can be modeled with a set of mixed-integer constraints by including a continuous variable ($u_{t',t,i,k}$) for each bilinear term.

The resulting linearized MILP formulation is obtained after replacing Eqn. (46) with Eqn. (55),

$$\sum_{t\in T}\sum_{i\in I}\sum_{j\in J}\sum_{k\in K}\frac{1}{(1+R)^t}P_{t,i,j,k}y_{t,i,j,k}$$
$$=\sum_{t\in T}\sum_{k\in K}\left[\sum_{j\in J}D_{t,j,k}\lambda_{t,j,k}-\sum_{i\in I}C^0_{i,k}\mu_{t,i,k}-\sum_{i\in I^1}\sum_{t'\in T'_t}H_{i,k}u_{t,t',i,k}\right] \tag{55}$$

and introducing the mixed-integer constraints in Eqns. (56) - (57).

$$u_{t,t',i,k}\ge\mu_{t,i,k}-M\left(1-x_{t',i,k}\right) \qquad t\in T, t'\in T'_t, i\in I^1, k\in K \tag{56}$$
$$u_{t,t',i,k}\in\mathbb{R}^+ \tag{57}$$

It is important to note that only the two terms presented in Eqns. (56) and (57) are necessary to linearize the bilinear terms because they are sufficient to bound the values of $u_{t,t',i,k}$ in the improving direction of the objective function.

6 Illustrative Example

Both MILP reformulations of the bilevel capacity planning problem are implemented to solve a small case study from the air separation industry. The illustrative example considers two existing facilities of the leader, one candidate location for a new facility of the leader, and a single facility of the competition. Facilities controlled by the leader and the competitor must satisfy the demand of 15 markets for a single commodity. The problem has a time horizon of 3 years divided in 12 time periods (quarters of year). In this time horizon, the leader is allowed to execute investment decisions in time periods 1, 5, and 12. Capacity expansion is achieved by installing additional production lines with capacity of 9,000 ton/period. The complete dataset for this illustrative example is presented in Appendix A. All examples use a discount rate (R) of 3% per time period.

The computational statistics of the single-level capacity planning with captive markets and the reformulations of the capacity planning in a competitive environment are presented in Table 1. All MILP problems were implemented in GAMS 24.4.1 and solved using GUROBI 6.0.0 on an Intel Core i7 CPU 2.93 Ghz processor with 4 GB of RAM.

Table 1: Model statistics for the illustrative example.

Model statistic	Single-level with captive markets	KKT reformulation	Dualilty-base reformulation
Number of constraints:	225	1,473	682
Number of continuous variables:	420	996	636
Number of binary variables:	48	480	48
LP relaxation at rootnode:	110	110	101
Final incumbent value:	110	97	97
Final optimality gap:	0.00%	0.00%	0.00%
Number of B&B nodes:	1	262	1
Solution time:	0.01 s	0.63 s	0.19 s

Table 1 shows the number of constraints and variables for the proposed formulations. It can be observed that the KKT reformulation is significantly larger than the duality-based reformulation; in particular, it requires 10 times more binary variables because of the complementarity constraints. The growth in the number of binary variables does not have much impact for the solution time of this small example, but it is likely to complicate the solution of larger instances.

The solutions obtained from the optimization problems establish the investment plan for the leader. The plan obtained from the formulation with captive markets does not expand any facilities in the time horizon. The optimal investment plan obtained from the bilevel formulation (both reformulations) expands facility 1 in the first time period. The bilevel optimal demand assignments in the first time period of this illustrative example are presented in Fig. 2; it can be observed that some markets have dual sourcing because of the capacity limitations of production facilities. Table 2 compares the income, investment costs, and operating costs for the single-level and bilevel expansion plans. In order to quantify the potential regret of implementing an expansion plan that ignores the decision criterion of markets, the expansion plan obtained from the single-level formulation is also evaluated in an environment of rational markets.

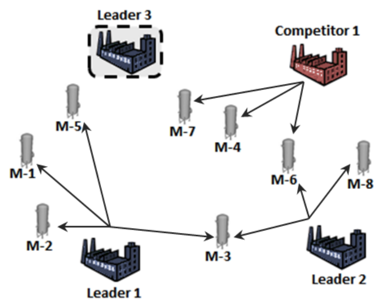

Figure 2: Optimal demand assignments obtained in the first time period of the illustrative example using the bilevel formulation.

Table 2: Results of the single-level and bilevel expansion plans for the illustrative example.

Term in objective function	Single-level with captive markets	Single-level with rational markets	Bilevel with rational markets
Income from sales (MM$):	354	345	398
Investment in new facilities (MM$):	0	0	0
Expansion cost (MM$):	0	0	29
Maintenance cost (MM$):	31	31	31
Production cost (MM$):	139	139	162
Transportation cost (MM$):	74	118	79
NPV (MM$):	110	57	97
Market cost (MM$):	523	510	508

Table 2 shows the benefits of the expansion plan obtained from the bilevel formulation when markets are considered rational. The single-level formulation with captive markets predicts a level of income that is not attainable with rational markets. The bilevel formulation offers the lowest cost for the markets with a small deterioration of the leader's NPV in comparison to what could be obtained with captive markets. When markets are considered rational, the NPV obtained with the bilevel expansion plan is MM$40 higher than the one obtained by the single-level expansion plan; this measure of regret accounts for 41% of the potential NPV.

7 Middle-size Instances

From the illustrative example presented in Section 6, we observe that the KKT and the duality-based reformulations yield exactly the same results. Despite the difference in formulation sizes shown in Table 1, both reformulation solve the illustrative example in approximately the same time. In order to predict the performance of the reformulations on large-scale instances, we use a middle-size example of the capacity planning problem.

The example comprises the production and distribution of one product to 15 markets. Initially, the leader has three production facilities with capacities equal to 27,000 ton/period, 13,500 ton/period, and 31,500 ton/period. The leader also considers the possibility of opening a new facility at a candidate location. We evaluate the investment decisions in a time horizon of 5 years divided in 20 time periods.

We analyze two instances that share the same data but allow different timing for the investment decisions. In the first instance (Middle-size 1), the leader is allowed to open the new facility and expand capacities in every fourth time period. In the second instance (Middle-size 2), the leader is allowed to execute the investments only every eight time periods. In both cases, capacity must be expanded in discrete increments of 9,000 ton/period. The investment costs associated with opening the new facility and expanding production capacity grow in time according to inflation; the maintenance cost of open facilities also increase with time.

Market demands in each time period vary during the time horizon. Fig. 3 shows the trajectory of the demands for the middle-size example. The selling prices offered by the leader to the markets are presented in Fig. 4; each market is offered a different price based on their proximity to the production facilities of the leader. Unit production costs at the facilities controlled by the leader are presented in Fig. 5; they show the characteristic seasonal variation caused by the electricity cost. Other cost coefficients of the example are not revealed by confidentiality reasons, but they maintain the same magnitudes presented in Appendix A.

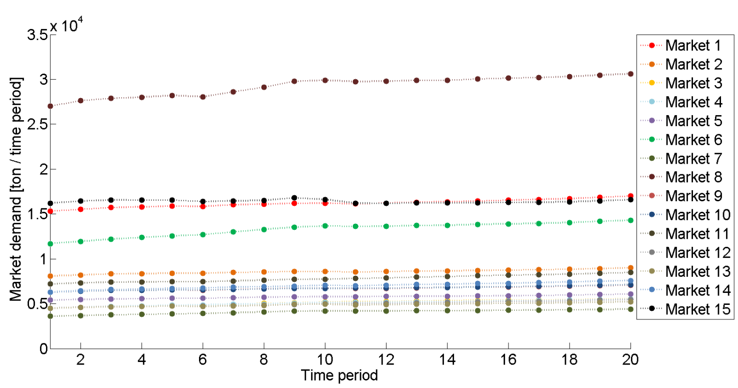

Figure 3: Evolution of market demands in the middle-size instances.

The computational statistics for the two middle-size instances of the capacity planning with rational markets are presented in Table 3. The KKT and the duality-based reformulations were

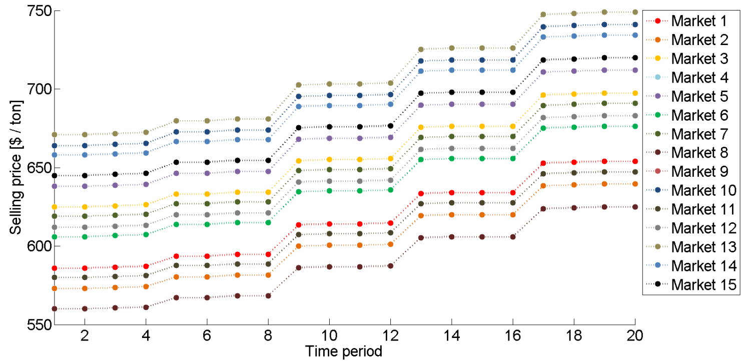

Figure 4: Evolution of selling prices in the middle-size instances.

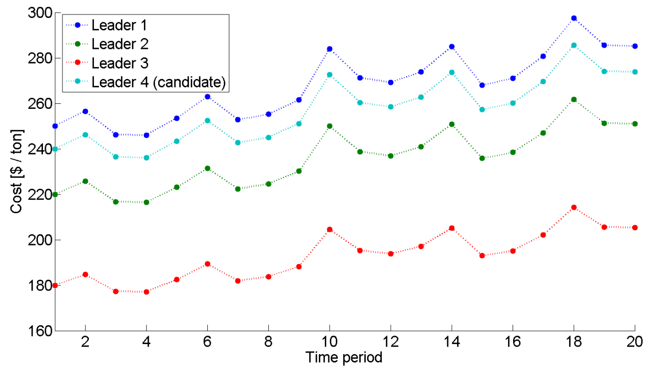

Figure 5: Evolution of production costs in the middle-size instances.

implemented in GAMS 24.4.1 and solved using GUROBI 6.0.0.

Table 3 demonstrates the benefits of the duality-based reformulation in comparison to the KKT reformulation. The time required to solve both instances using the duality-based reformulation is less than 1 second, whereas the KKT reformulation requires a few minutes for each instance. Interestingly, the KKT reformulation takes longer to solve the second middle-size instance that has fewer investment options. The reason behind this counter-intuitive behavior is that the solver takes longer to find a feasible solution to the problem.

The significant difference in solution time for both reformulations is explained by the number of constraints and variables in the problem. The KKT reformulation requires in both instances 2,240 additional binary variables to model complementarity conditions. The growth in the number of binary variables has a severe impact in the solution time of the problem.

Table 4 compares the income, investment costs, and operating costs of the proposed expansion plans. It shows that the expansion plan obtained for the first instance produces a slightly higher NPV when compared with the plan obtained for the second instance. This result can be anticipated because the feasible region of the first instance contains the feasible region of the second instance completely. However, the additional restrictions for the execution of investment decisions in the

Table 3: Model statistics for middle-size instances.

Model statistic	Middle-size 1 KKT reformulation	Middle-size 1 Dualilty-base reformulation	Middle-size 2 KKT reformulation	Middle-size 2 Dualilty-base reformulation
Number of constraints:	7,200	2,961	7,192	2,857
Number of continuous variables:	4,860	2,965	4,860	2,763
Number of binary variables:	2,345	105	2,335	95
LP relaxation at rootnode:	372	346	346	324
Final incumbent value:	316	316	308	308
Final optimality gap:	0.01%	0.00%	0.01%	0.00%
Number of B&B nodes:	1	11,367	16,786	1
Solution time:	157 s	0.83 s	282 s	0.73 s

Table 4: Results of the bilevel expansion plans for the middle-size instances.

Term in objective function	Middle-size 1	Middle-size 2
Income from sales (MM$):	895	885
Investment in new facilities (MM$):	0	0
Expansion cost (MM$):	85	82
Maintenance cost (MM$):	94	94
Production cost (MM$):	315	313
Transportation cost (MM$):	85	88
NPV (MM$):	316	308
Market cost (MM$):	1,319	1,319

second instance only produces a decrease of 1.1% in its NPV.

The investment plans obtained from the bilevel formulation do not invest to open the new facility in any of the instances. In the first instance, the plan expands facilities 2 and 3 in the first time period, and facility 3 in the fifth time period. The optimal capacities and production levels at the facilities controlled by the leader in first instance are presented in Fig. 6; arrows indicate the time periods in which capacity is expanded. We can observe in Fig. 6 that all production facilities have high utilization. The expanded capacities in facilities 2 and 3 are used as soon as they are available; facility 1 experiences a temporary decrease in its production because of the capacity increase at facility 3, but it returns to full utilization with demand growth. The bilevel expansion plan obtained for the second instance is very similar to the plan obtained for the first instance; it expands facilities 2 and 3 in the first time period, and delays the second expansion of facility 3 until the ninth time period. In both instances, investment and maintenance cost are equal for all facilities controlled by the leader; therefore, the expansion trends observed are good indicators of the competitiveness of facilities with respect production and transportation cost.

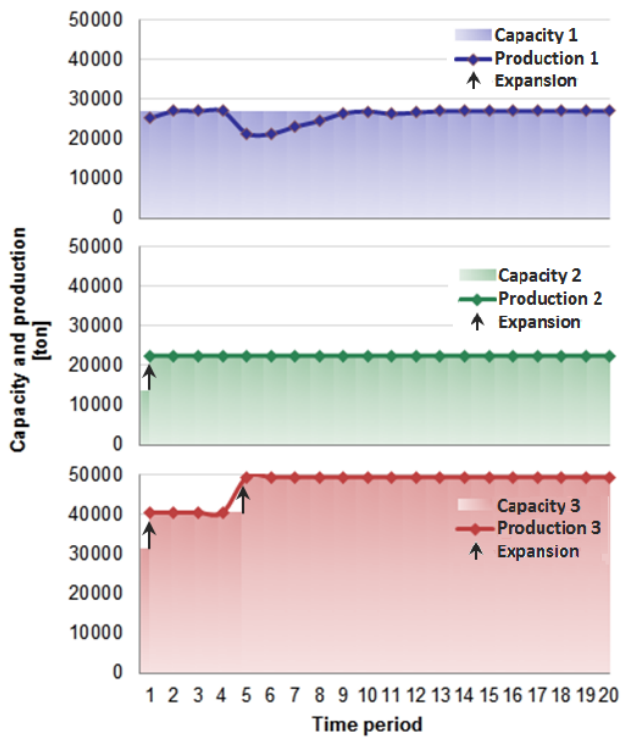

Figure 6: Capacity and production of the leader in the first instance of the middle-size example.

8 Solution Strategies for Large-scale problems

The implementation of the bilevel formulation for capacity planning problems in industrial instances requires developing a solution strategy for large-scale problems. The results obtained from the middle-size instance suggest that the KKT reformulation is not appropriate to solve large instances. Additionally, we can expect the duality-based reformulation to struggle solving large-scale instances given the relative weakness of its LP relaxation. Therefore, we propose an improved duality-based reformulation and a domain reduction scheme; these solution strategies are evaluated in Section 9 with an industrial example.

8.1 Strengthened Duality-based Reformulation

The LP relaxation of the duality-based reformulation can be strengthened by enforcing strong duality independently for each commodity in every time period. The justification for this modification comes from the observation that once the leader has fixed its capacity, the optimization problem of the follower can be decomposed by time period and commodity. Consequently, we can replace Eqn. (55) by its disaggregated version presented in Eqn. (58).

$$\sum_{i \in I} \sum_{j \in J} \frac{1}{(1+R)^t} P_{t,i,j,k} y_{t,i,j,k} = \sum_{j \in J} D_{t,j,k} \lambda_{t,j,k} - \sum_{i \in I} C^0_{i,k} \mu_{t,i,k} - \sum_{i \in I^1} \sum_{t'=1}^{t} H_{i,k} u_{t,t',i,k} \quad \forall t \in T, k \in K \tag{58}$$

Replacing Eqn. (55) by Eqn. (58) yields a modest improvement in the LP relaxation of the duality-based reformulation. In the first instance of the middle-size example presented in Section 7, the value of the LP relaxation is reduced from MM$346 to MM$343 (9.49% to 8.54% initial gap).

8.2 Domain Reduction for the Duality-based Reformulation

A clever strategy to reduce the size of the capacity planning problem with rational markets derives from the analysis of the feasible region of the bilevel optimization problem. In the bilevel optimization literature, the bilevel feasible region is called the inducible region [2]. In essence, the inducible region is the set of upper-level feasible solutions and their corresponding rational reactions in the lower-level problem. In order to describe the inducible region mathematically, we define the set of upper-level feasible solutions as the capacity expansion plans that satisfy upper-level constraints. This set of upper-level feasible solutions is represented in Eqn. (59),

$$(v, w, x, c) \in X \tag{59}$$

where X denotes the polyhedron described by upper-level constraints (9)-(11) and upper-level domains (17)-(18).

The rational reaction set for the follower is defined by expression (60) as a function of the upper-level variables,

$$\Psi(v, w, x, c) = \left\{ y : \arg\min_{y \in Y} \left[\sum_{t \in T} \sum_{i \in I} \sum_{j \in J} \sum_{k \in K} \frac{1}{(1+R)^t} P_{t,i,j,k} y_{t,i,j,k} \right] \right\} \tag{60}$$

where Y denotes the polyhedron described by lower-level constraints (13)-(16).

According to expressions (59) and (60), the inducible region of the bilevel capacity expansion problem is defined by expression (61).

$$IR = \{(v, w, x, c, y) : \ (v, w, x, c) \in X, \ y \in \Psi(v, w, x, c)\} \tag{61}$$

We know from our original assumptions that any expansion plan satisfying Eqn. (59) has a nonempty rational reaction set ($\Psi(v, w, x, c)$). However, not all demand assignments satisfying the lower-level constraints ($y \in Y$) are bilevel feasible because some of them might be suboptimal for all expansion plans of the leader. Hence, it is possible to reduce the dimension of the bilevel formulation by excluding from its domain those demand assignments ($y \in Y$) that are never optimal in the lower level.

The first step to identify demand assignments that are bilevel infeasible is to solve the lower-level problem with the production capacities of the leader fixed to their upper bounds. Once we know the optimal demand assignments in the lower-level problem with maximum capacity, we can infer which demands are never assigned to the leader. The intuition for this inference is that only the demands ($D_{t,j,k}$) that are assigned to the leader when the capacity is at its upper bound, can be assigned to the leader when its capacity is more constrained.

The idea behind the domain reduction is that demand assignments that are nonbasic in the optimal solution of the LP with maximum capacity, must remain nonbasic when capacity is reduced. Proposition 1 formalizes this idea. Its proof can be found in Appendix B.

Proposition 1. *A demand assignment ($y_{t,i,j,k}$) with positive reduced cost in the optimal solution of the lower-level problem with maximum capacity also has a positive reduced cost when capacities are reduced.*

For the implementation of the domain reduction strategy, it is important to remember that nonbasic variables are associated with positive reduced costs in the minimization problem. In order to identify nonbasic variables, we denote by $\mu^U_{t,i,k}$ and $\lambda^U_{t,j,k}$ the optimal dual solution of the lower-level problem with capacities of the leader are at their upper bound ($C^U_{t,i,k} \ \ \forall i \in I^1$). Then, according to Proposition 1, Eqn. (62) establishes valid upper bounds for the demand assignments in the bilevel capacity expansion problem.

$$y_{t,i,j,k} \leq \begin{cases} 0 & \text{if } \frac{1}{(1+R)^t} P_{t,i,j,k} + \mu^U_{t,i,k} - \lambda^U_{t,j,k} > 0 \\ D_{t,j,k} & \text{otherwise} \end{cases} \qquad \forall t \in T, i \in I^1, j \in J, k \in K \tag{62}$$

The range reduction proposed in expression (62) can have a significant impact in the size of the bilevel formulation because many assignment variables can be fixed if we determine that zero is their only bilevel feasible value. However, it is not the only advantage of the domain reduction strategy when we use the duality-based reformulation. If we analyze the lower-level LP in light of complementary slackness, we can conclude that expression (62) also implies that some dual constraints (50) are never active. In particular, those dual constraints (50) corresponding to the variables $y_{t,i,j,k}$ that can be fixed to zero are irrelevant in the duality-based formulation. Therefore, the domain reduction strategy proposed for the bilevel capacity expansion planning offers the double benefit of reducing the number of continuous variables and the number of constraints in the duality-based reformulation.

9 Industrial Example

The solution strategies proposed for large-scale instances are tested with a capacity planning problem for an air separation company. This large-scale example includes 3 existing facilities of the leader, 2 candidate facilities of the leader, and 5 facilities of competitors. Demands of 20 markets for 2 different commodities are considered in a time horizon of 20 years divided in 80 time periods. Two instances allowing different timing for the investment decisions are analyzed: the first instance allows investments every fourth time period and the second instance allows investments every eighth time period.

According to the formulation, the leader maximizes the NPV obtained during the 20-year time horizon. Markets select their providers by controlling the demand assignments with the objective of minimizing the discounted cost they pay. A discount rate (R) of 3% per time period is used in both objective functions. Cost coefficients and all other parameters are omitted because of confidentiality reasons.

The computational statistics for the original duality-based reformulation and the large-scale duality-based reformulation are presented in Table 5; the large-scale reformulation enforces strong duality for each commodity in every time period and implements the domain reduction strategy to fix variables and eliminate constraints. Table 5 shows that both instances of the industrial example have a significant number of constraints, continuous and discrete variables. However, if we compare the original and the large-scale duality-based reformulations, we observe a reduction between 13% and 17% in the number of continuous variables and constraints.

The performance of both reformulations is also presented in Table 5; we observe a significant difference in the performance of the original and the large-scale duality-based reformulations. A major advantage of the large-scale reformulation is related to its LP relaxation at the rootnode. This improvement derives partially from disaggregating strong duality, and more importantly from excluding demand assignments that are bilevel infeasible. In the first industrial instance the LP relaxation gap is reduced from 34.9% to 3.9%, whereas in the second industrial instance the reduction is from 34.4% to 3.7%.

Even after implementing the proposed strategies for large-scale problems, the industrial instances are still difficult to solve using GUROBI 6.0.0. For our industrial example, only the second instance was solved to the desired optimality gap of 1% with the large-scale duality-based reformulation. However, if we compare the best solutions obtained for both industrial instances, we observe that allowing more frequent expansions in the first instance produces a NPV that is MM$45 higher,

Table 5: Model statistics for industrial instances.

Model statistic	Industrial 1		Industrial 2	
	Original duality-based	large-scale duality-based	Original duality-based	large-scale duality-based
Number of constraints:	46,601	40,025	42,501	35,925
Number of continuous variables:	46,000	39,905	42,520	35,265
Number of binary variables:	640	640	520	520
LP relaxation at rootnode:	4,289	2,906	4,002	2,851
Final incumbent value:	2,662	2,791	2,689	2,746
Final optimality gap:	33.2%	1.27%	26.4%	0.98%
Solution time:	60 min*	60 min*	60 min*	5 min

* Time limit reached

Table 6: Results of the bilevel expansion plans for the industrial instances.

Term in objective function	Industrial 1	Industrial 2
Income from sales (MM$):	5,984	5,888
Investment in new facilities (MM$):	0	0
Expansion cost (MM$):	439	411
Maintenance cost (MM$):	215	215
Production cost (MM$):	2,100	2,072
Transportation cost (MM$):	439	444
NPV (MM$):	2,791	2,746
Market cost (MM$):	10,545	10,546

which accounts for 1.6% of the potential profit. Table 6 presents in detail the terms in the objective function for the best solutions obtained; the table shows that allowing more frequent expansions in the first instance generates a more dynamic expansion plan that can capture a higher market share. However, the optimal number of expansions is the same for both instances and none of them includes investments in new facilities.

The optimal capacity and production levels at facilities controlled by the leader in the first industrial instance are presented in Figs. 7 and 8 for commodity 1 and 2, respectively; the figures show that utilization of the production capacities is high for all the facilities being expanded. The only capacity that is not expanded in the entire time horizon is the production capacity of commodity 1 at facility 1; the utilization of this production capacity fluctuates according to the available capacity at the facilities 2 and 3. The expansion trends observed preserve a close relation with the competitiveness of facilities that is mainly determined by their production and distribution costs.

10 Conclusions

We have developed a novel formulation for capacity planning problems that considers markets as rational decision makers. The formulation is based on bilevel optimization, a framework that

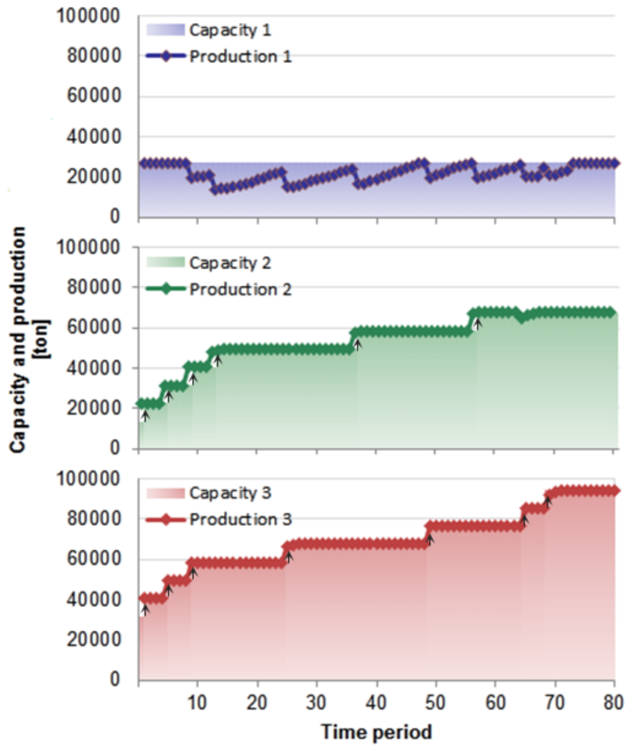

Figure 7: Capacity and production of commodity 1 at the facilities controlled by the leader in the first instance of the industrial example.

allows modeling the conflicting interests of producers and consumers. The expansion plans obtained from the bilevel formulation produce greater economic benefits when the producers operate in a competitive environment. In particular, the single-level formulation tends to overestimate the market share that can be obtained and might generate expensive investment plans that are less profitable.

The bilevel formulation for capacity planning is a challenging optimization problem. We have proposed two different approaches to reformulate it as a single-level MILP. The first approach ensures optimality of the lower-level problem through its KKT conditions. The second approach uses strong duality of LPs for the reformulation. In the middle-size instances, we have shown that the duality-based reformulation offers superior performance compared to the KKT reformulation; this result is explained by the large number of binary variables required in the KKT approach to linearize the complementarity constraints. The duality-based reformulation does not require the addition of binary variables but the strong duality condition gives rise to nonlinearities; for the case in which all upper-level variables are discrete, the nonlinearities can be avoided with the introduction of continuous variables and linear constraints.

Despite the relative advantage of the duality-based reformulation, the solution of large-scale instances of the bilevel capacity planning problem is still computationally demanding. We proposed two strategies to improve the duality-based reformulation. The first strategy leverages separability of the lower-level problem by disaggregating the strong duality constraint. The second strategy uses

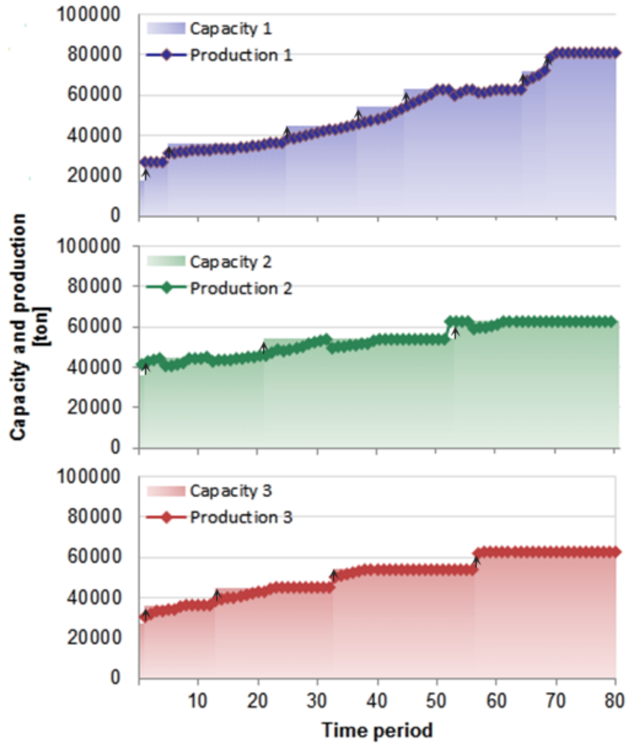

Figure 8: Capacity and production of commodity 2 at the facilities controlled by the leader in the first instance of the industrial example.

the topology of the bilevel feasible region to reduce the number of variables and constraints in the duality-based reformulation. The implementation of these strategies yields a significant reduction in the solution time of large-scale problems.

The bilevel formulation for capacity planning has shown to be useful for developing capacity expansion plans that considers markets as rational decision makers. This novel approach is more realistic than the traditional formulation because it models the dynamic nature of industrial markets. Furthermore, we have proposed an effective strategy to solve large-scale instances that allows using the bilevel capacity planning formulation in industrial applications.

References

[1] J. F. Bard and J. E. Falk. An explicit solution to the multi-level programming problem. *Computers & Operations Research*, 9:77–100, 1982.

[2] J. F. Bard and J. T. Moore. An algorithm for the discrete bilevel programming problem. *Naval Research Logistics (NRL)*, 39(3):419–435, 1992.

[3] W. F. Bialas and M. H. Karwan. On two-level optimization. *Automatic Control, IEEE Transactions on*, 27:211–214, 1982.

[4] J. Bracken and J. T. McGill. Mathematical programs with optimization problems in the constraints. *Operations Research*, 21:p37–44, 1973.

[5] Anthony P. Burgard and Costas D. Maranas. Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnology and Bioengineering*, 82:670–677, 2003.

[6] W. Candler and R. Townsley. A linear two-level programming problem. *Computers & Operations Research*, 9(1):59–76, 1982.

[7] S-G. Chang and B. Gavish. Telecommunications network topological design and capacity expansion: Formulations and algorithms. *Telecommunication Systems*, 1(1):99–131, 1993.

[8] Y. Chu and F. You. Integrated Scheduling and Dynamic Optimization by Stackelberg Game: Bilevel Model Formulation and Efficient Solution Algorithm. *Industrial & Engineering Chemistry Research*, 53 (13):5564–5581, 2014.

[9] P. A. Clark and A. W. Westerberg. Bilevel programming for steady-state chemical process design: I. fundamentals and algorithms. *Computers & Chemical Engineering*, 14:87–97, 1990.

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

[11] K. Fischer. Sequential discrete p-facility models for competitive location planning. *Annals of Operations Research*, 111:253–270, 2002.

[12] J. Fortuny-Amat and B. McCarl. A representation and economic interpretation of a two-level programming problem. *The Journal of the Operational Research Society*, 32:783–792, 1981.

[13] D. Gale, H. W. Kuhn, and A. W. Tucker. *Activity Analysis of Production and Allocation*, chapter Linear programming and the theory of games, page 317329. New York: Wiley, 1951.

[14] L. P. Garces, A. J. Conejo, R. Garcia-Bertrand, and R. Romero. A bilevel approach to transmission expansion planning within a market environment. *Power Systems, IEEE Transactions on*, 24:1513–1522, 2009.

[15] I. E. Grossmann. Enterprise-wide optimization: A new frontier in process systems engineering. *Aiche Journal*, 51:1846–1857, 2005.

[16] I. E. Grossmann. Advances in mathematical programming models for enterprise-wide optimization. *Computers & Chemical Engineering*, 47:2–18, 2012.

[17] I. E. Grossmann and C. A. Floudas. Active constraint strategy for flexibility analysis in chemical processes. *Computers & Chemical Engineering*, 11:675–693, 1987.

[18] D. Huppmann and J. Egerer. National-Strategic Investment in European Power Transmission Capacity. *DIW Berlin Discussion Paper No. 1379*, pages 1–23, 2014.

[19] P. Loridan and J. Morgan. Weak via strong Stackelberg problem: New results. *Journal of Global Optimization*, 8:263–287, 1996. ISSN 0925-5001.

[20] H. Luss. Operations research and capacity expansion problems: A survey. *Operations Research*, 30(5): 907–947, 1982.

[21] C. T. Maravelias and I. E. Grossmann. Simultaneous planning for new product development and batch manufacturing facilities. *Industrial & Engineering Chemistry Research*, 40(26):6147–6164, 2001.

[22] A. L. Motto, J. M. Arroyo, and F. D. Galiana. A Mixed-Integer LP Procedure for the Analysis of Electric Grid Security Under Disruptive Threat. *Power Systems, IEEE Transactions on*, 20:1357–1365, 2005.

[23] F. H. Murphy and Y. Smeers. Generation capacity expansion in imperfectly competitive restructured electricity markets. *Operations Research*, 53(4):646–661, 2005.

[24] F. H. Murphy, S. Sen, and A. L. Soyster. Electric utility capacity expansion planning with uncertain load forecasts. *A I I E Transactions*, 14(1):52–59, 1982.

[25] W. S. Nainis and Y. Y. Haimes. A multilevel approach to planning for capacity expansion in water resource systems. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-5(1):53–63, 1975.

[26] M. J. Osborne and A. Rubenstein. *A course in game theory*, chapter 6. Extensive games with perfect information, pages 89–115. Cambridge (MA): The MIT Press., 1994.

[27] C. Ruiz, AJ. Conejo, and Y. Smeers. Equilibria in an oligopolistic electricity pool with stepwise offer curves. *IEEE Transactions on Power Systems*, 27:752–761, 2012.

[28] J-H. Ryu, V. Dua, and E. N. Pistikopoulos. A bilevel programming framework for enterprise-wide process networks under uncertainty. *Computers & Chemical Engineering*, 28:1121–1129, 2004.

[29] G. K. D. Saharidis, A. J. Conejo, and G. Kozanidis. Exact solution methodologies for linear and (mixed) integer bilevel programming. In El-Ghazali Talbi, editor, *Metaheuristics for Bi-level Optimization*, volume 482 of *Studies in Computational Intelligence*, pages 221–245. Springer Berlin Heidelberg, 2013.

[30] N. V. Sahinidis and I. E. Grossmann. Reformulation of the Multiperiod MILP Model for Capacity Expansion of Chemical Processes. *Operations Research*, 40:S127–S144, 1992.

[31] N. V. Sahinidis, I. E. Grossmann, R. E. Fornari, and M. Chathrathi. Optimization model for long range planning in the chemical industry. *Computers & Chemical Engineering*, 13(9):1049–1063, 1989.

[32] J. Salmeron, K. Wood, and R. Baldick. Analysis of electric grid security under terrorist threat. *Power Systems, IEEE Transactions on*, 19:905–912, 2004.

[33] A. L. Soyster and F. H. Murphy. *Economic Behaviour of Electric Utilities*. Prentice-Hall, 1989.

[34] A. Sundaramoorthy, J. M. B. Evans, and P. I. Barton. Capacity planning under clinical trials uncertainty in continuous pharmaceutical manufacturing, 1: Mathematical framework. *Industrial & Engineering Chemistry Research*, 51(42):13692–13702, 2012.

[35] S. A. Van den Heever and I. E. Grossmann. Disjunctive multiperiod optimization methods for design and planning of chemical process systems. *Computers & Chemical Engineering*, 23(8):1075–1095, 1999.

[36] H. von Stackelberg. *Market Structure and Equilibrium*. Springer, 2011.

[37] M. A. Zamarripa, A. M. Aguirre, C. A. Mendez, and A. Espuna. Improving supply chain planning in a competitive environment. *Computers & Chemical Engineering*, 42:178–188, 2012.

Appendix A: data for illustrative example

Table A.1 shows the cardinality of the datasets used for the three examples presented in the paper.

Table A.1: Summary of the datasets used in examples.

	Illustrative example	middle-size instance	Industrial instance
Existing facilities of the leader:	2	3	3
Candidate facilities of the leader:	1	1	2
Facilities of the competitors:	1	3	5
Markets:	8	15	20
Commodities:	1	1	2
Time periods:	12	20	80

The complete dataset for the illustrative example is presented in Tables A.2 - A.10. The initial production capacity of the facilities is presented in Table A.2.

Table A.2: Initial capacities for facilities in the illustrative example.

Facility	Commodity 1 [ton/period]
Leader 1	22,500
Leader 2	36,000
Leader 3	0
Competitor 1	36,000

Market demands for all time periods are presented in Table A.3.

Table A.3: Market demands ($D_{t,j,k}$) in the illustrative example.

Time	Market demand [ton/period]							
period	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8
1	15,300	8,100	4,500	4,500	5,400	11,700	3,600	27,000
2	15,500	8,200	4,600	4,600	5,500	11,900	3,700	27,600
3	15,700	8,300	4,600	4,700	5,500	12,200	3,800	27,900
4	15,800	8,400	4,700	4,700	5,600	12,400	3,800	28,000
5	15,900	8,400	4,800	4,800	5,600	12,600	3,900	28,200
6	15,900	8,400	4,800	4,900	5,600	12,700	3,900	28,100
7	16,000	8,500	4,900	5,000	5,700	13,000	4,000	28,600
8	16,100	8,500	5,000	5,000	5,700	13,300	4,100	29,100
9	16,200	8,600	5,100	5,100	5,800	13,500	4,200	29,800
10	16,200	8,600	5,200	5,100	5,800	13,700	4,200	29,900
11	16,100	8,500	5,300	5,200	5,800	13,600	4,200	29,700
12	16,200	8,600	5,300	5,200	5,800	13,600	4,200	29,800

Tables A.4 - A.10 present the cost coefficients for the objective function of the illustrative example. Table A.4 shows the cost ($A_{t,3}$) of opening the candidate production facility in different time periods. In the illustrative example, it is allowed to open the new facility only in time periods 1, 5, and 9.

Table A.4: Investment cost ($A_{t,3}$) for the leader to open facility 3 in the illustrative example.

Time period	Investment cost [MM$]
1	20.00
5	20.40
9	20.86

Table A.5 presents the maintenance cost per time period ($B_{t,i}$) incurred by open facilities.

Table A.5: Maintenance cost ($B_{t,i}$) in the illustrative example.

Time	Maintenance cost [MM$/period]		
period	Leader 1	Leader 2	Leader 3
1	1.000	2.000	3.000
2	1.005	2.010	3.015
3	1.010	2.020	3.030
4	1.013	2.026	3.039
5	1.020	2.040	3.060
6	1.029	2.058	3.087
7	1.032	2.064	3.096
8	1.035	2.070	3.105
9	1.043	2.086	3.129
10	1.049	2.098	3.147
11	1.054	2.108	3.162
12	1.058	2.116	3.174

Table A.6 presents the investment cost ($E_{t,i,1}$) associated to the expansion of production capacity by 9,000 ton/period. In the illustrative example, all facilities are assumed to have the same expansion cost and expansions are allowed only in time periods 1, 5, and 9.

Table A.6: Expansion cost ($E_{t,i,1}$) in the illustrative example.

Time	Expansion cost [MM$/9,000 ton]		
period	Leader 1	Leader 2	Leader 3
1	30.00	30.00	30.00
5	30.60	30.60	30.60
9	31.29	31.29	31.29

The production cost of facilities ($F_{t,i,1}$) in the illustrative example are presented in Table A.7.

Table A.7: Production cost ($F_{t,i,1}$) in the illustrative example.

Time period	Production cost [$/ton] Leader 1	Leader 2	Leader 3
1	250	220	180
2	257	226	185
3	246	217	177
4	246	216	177
5	254	223	183
6	263	231	189
7	253	222	182
8	255	225	184
9	262	230	188
10	284	250	204
11	271	239	195
12	269	237	194

The transportation cost from facilities to markets in each time period are calculated from the transportation costs at the initial time period and their growth rate, according to Eqn. A.1. Initial transportation costs ($G^0_{i,j,k}$) are presented in Table A.8; their growth rate (G^{Rt}_t) are presented in Table A.10.

$$G_{t,i,j,k} = G^0_{i,j,k} G^{Rt}_t \tag{A.1}$$

Table A.8: Transportation cost ($G_{t,i,j,1}$) in the illustrative example.

Market	Transportation cost [$/ton] Leader 1	Leader 2	Leader 3
1	26	325	234
2	13	299	260
3	65	195	325
4	104	130	156
5	78	260	221
6	208	195	46
7	195	169	59
8	234	169	0.4

Selling prices offered by facilities to markets are calculated from the selling prices at the initial time period and their growth rate according to Eqn. A.2. Initial selling prices ($P^0_{i,j,k}$) are presented in Table A.9; their growth rates (P^{Rt}_t) are presented in Table A.10.

$$P_{t,i,j,k} = P^0_{i,j,k} P^{Rt}_t \tag{A.2}$$

Table A.9: Initial selling prices ($P^0_{i,j,k}$) from facilities to markets in the illustrative example.

Market	Leader 1, 2 & 3 [$\$/ton$]	Competitor 1 [$\$/ton$]
1	586	615
2	573	726
3	625	785
4	664	633
5	638	794
6	606	619
7	619	606
8	560	580

Table A.10: Growth rates for transportation costs (G^{Rt}_t) and selling prices (P^{Rt}_t) in the illustrative example.

Time period	Growth rate for transportation	Grow rate [MM$] for selling prices
1	1.00	1
2	1.00	1
3	1.03	1.001
4	1.05	1.002
5	1.09	1.013
6	1.09	1.013
7	1.12	1.015
8	1.12	1.015
9	1.12	1.047
10	1.14	1.048
11	1.14	1.048
12	1.16	1.049

Appendix B: proof of Proposition 1

Proposition 1. A demand assignment ($y_{t,i,j,k}$) with positive reduced cost in the optimal solution of the lower-level problem with maximum capacity also has a positive reduced cost when capacities are reduced.

Proof. We want to prove that the optimal reduced cost of the leader's assignment variables cannot decrease when capacities are reduced from their maximum feasible value ($C^U_{t,i,k}$). For this analysis, we decompose the lower-level problems by time periods ($t \in T$) and by commodities ($k \in K$); the problem minimizing the cost paid by markets is decomposable since all terms in the objective function and constraints are indexed by (t, k). Intuitively, this means that we can solve independent problems to minimize the cost paid at time period t for commodity k. The lower-level problem resulting from this decomposition is presented in Eqns. (B.1)-(B.5).

$$\min \quad \frac{1}{(1+R)^t} \sum_{i \in I} \sum_{j \in J} P_{i,j} y_{i,j} \tag{B.1}$$

$$\text{s.t.} \quad \sum_{j \in J} y_{i,j} \le C_i \qquad \forall\, i \in I^1 \tag{B.2}$$

$$\sum_{j \in J} y_{i,j} \le C_i^0 \qquad \forall\, i \in I^2 \tag{B.3}$$

$$\sum_{i \in I} y_{i,j} = D_j \qquad \forall\, j \in J \tag{B.4}$$

$$y_{i,j} \in \mathbb{R}^+ \qquad \forall\, i \in I, j \in J \tag{B.5}$$

Similarly, the dual lower-level problem disaggregated by time periods and commodities is presented in Eqns. (B.6) - (B.10).

$$\max \quad \sum_{j \in J} D_j \lambda_j - \sum_{i \in I^1} C_i \mu_i - \sum_{i \in I^2} C_i^0 \mu_i \tag{B.6}$$

$$\text{s.t.} \quad \lambda_j - \mu_i \le \frac{1}{(1+R)^t} P_{i,j} \qquad \forall\, i \in I^1, j \in J \tag{B.7}$$

$$\lambda_j - \mu_i \le \frac{1}{(1+R)^t} P_{i,j} \qquad \forall\, i \in I^2, j \in J \tag{B.8}$$

$$\mu_i \in \mathbb{R}^+ \qquad \forall\, i \in I \tag{B.9}$$

$$\lambda_j \in \mathbb{R} \qquad \forall\, j \in J \tag{B.10}$$

We assume that the dual lower-level problem is bounded (and the primal lower-level problem is feasible). The condition that guarantees a finite solution for the dual of the lower-level problem is presented in Eqn. (B.11).

$$\sum_{j \in J} D_j \le \sum_{i \in I^1} C_i + \sum_{i \in I^2} C_i^0 \tag{B.11}$$

An important observation regarding dual variables μ_i ($i \in I^1$) is that they all have the same optimal value. It is the case because constraints (B.7) are identical for all facilities of the leader (facilities of the leader offer the same price to each market) and the coefficients of all μ_i have the same sign in the objective function.

We also note that the condition presented in Eqn. (B.12) must be satisfied by the optimal solution of the dual lower-level problem in order to get the largest values of λ_j allowed by dual constraints (B.7) - (B.8).

$$\lambda_j = \min_{i \in I} \left(\frac{1}{(1+R)^t} P_{i,j} + \mu_i \right) \qquad \forall j \in J \tag{B.12}$$

Using Eqn. (B.12), we can rewrite the dual lower-level problem as presented by Eqn. (B.13).

$$\max_{\mu_i \geq 0} \left\{ \sum_{j \in J} D_j \left[\min_{i \in I} \left(\frac{1}{(1+R)^t} P_{i,j} + \mu_i \right) \right] - \sum_{i \in I^1} C_i \mu_i - \sum_{i \in I^2} C_i^0 \mu_i \right\} \tag{B.13}$$

In order to prove that the optimal reduced costs of the leader's assignment variables cannot decrease when capacities are reduced, we divide the proof in four steps.

Step 1: optimal values of μ_i ($i \in I^1$) cannot be less than their optimal values obtained with maximum capacity.

We assume that C_i^U is the upper bound of the coefficient of dual variable μ_i in Eqn. (B.6), and we denote by (μ_i^U, λ_j^U) the corresponding optimal solution of the dual lower-level problem. Now, let us assume that the coefficients of μ_i are reduced by ΔC_i, and let us denote by (μ_i^*, λ_j^*) the optimal dual solution corresponding to capacities $C_i^* = C_i^U - \Delta C_i$. If we consider that Eqn. (B.13) is a maximization problem, we can establish the sequence of inequalities (B.14)-(B.17).

$$\sum_{j \in J} D_j \left[\min_{i \in I} \left(\frac{1}{(1+R)^t} P_{i,j} + \mu_i^* \right) \right] - \sum_{i \in I^1} C_i^U \mu_i^* - \sum_{i \in I^2} C_i^0 \mu_i^* \tag{B.14}$$

$$\leq \sum_{j \in J} D_j \left[\min_{i \in I} \left(\frac{1}{(1+R)^t} P_{i,j} + \mu_i^U \right) \right] - \sum_{i \in I^1} C_i^U \mu_i^U - \sum_{i \in I^2} C_i^0 \mu_i^U \tag{B.15}$$

$$\leq \sum_{j \in J} D_j \left[\min_{i \in I} \left(\frac{1}{(1+R)^t} P_{i,j} + \mu_i^U \right) \right] - \sum_{i \in I^1} \left(C_i^U - \Delta C_i \right) \mu_i^U - \sum_{i \in I^2} C_i^0 \mu_i^U \tag{B.16}$$

$$\leq \sum_{j \in J} D_j \left[\min_{i \in I} \left(\frac{1}{(1+R)^t} P_{i,j} + \mu_i^* \right) \right] - \sum_{i \in I^1} \left(C_i^U - \Delta C_i \right) \mu_i^* - \sum_{i \in I^2} C_i^0 \mu_i^* \tag{B.17}$$

We note that $\sum_{i \in I^1} \Delta C_i \mu_i^*$ is the difference between expressions (B.17) and (B.14). Similarly, the difference between expressions (B.16) and (B.15) is $\sum_{i \in I^1} \Delta C_i \mu_i^U$. Hence, we can infer that $\sum_{i \in I^1} \Delta C_i \mu_i^* \geq \sum_{i \in I^1} \Delta C_i \mu_i^U$. Since dual variables μ_i must have the same optimal value for all $i \in I^1$, then $\mu_i^* \geq \mu_i^U$ for all $i \in I^1$.

Step 2: optimal values of μ_i ($i \in I^2$) cannot be less than their optimal values obtained with maximum capacity.

In order to continue with the argument, let us define ϵ_i according to Eqn. (B.18).

$$\epsilon_i = \mu_i^* - \mu_i^U \tag{B.18}$$

By optimality of Eqn. (B.15), we know that any deviation of μ_i^U from their optimal values yields a lower bound as presented in Eqns. (B.19) - (B.20).

$$\sum_{j\in J} D_j \left[\min_{i\in I}\left(\frac{1}{(1+R)^t}P_{i,j} + \mu_i^U + \min_{i'}[\epsilon_{i'}]\right)\right] - \sum_{i\in I^1} C_i^U(\mu_i^U + \min_{i'}[\epsilon_{i'}]) - \sum_{i\in I^2} C_i^0(\mu_i^U + \min_{i'}[\epsilon_{i'}]) \tag{B.19}$$

$$\leq \sum_{j\in J} D_j \left[\min_{i\in I}\left(\frac{1}{(1+R)^t}P_{i,j} + \mu_i^U\right)\right] - \sum_{i\in I^1} C_i^U \mu_i^U - \sum_{i\in I^2} C_i^0 \mu_i^U \tag{B.20}$$

subtracting (B.20) from (B.19), we obtain inequality (B.21),

$$-\sum_{j\in J} D_j \min_{i'}[\epsilon_{i'}] + \sum_{i\in I^1} C_i^U \min_{i'}[\epsilon_{i'}] + \sum_{i\in I^2} C_i^0 \min_{i'}[\epsilon_{i'}] \geq 0 \tag{B.21}$$

which implies $\min_{i\in I}[\epsilon_i] \geq 0$ according to inequality (B.11).

Step 3: if capacities of the leader are reduced, optimal values of μ_i ($i \in I^2$) cannot increase faster than the values of μ_i ($i \in I^1$).

We want to show that $\max_{i\in I}[\epsilon_i] = \max_{i\in I^1}[\epsilon_i]$. Since all dual variables μ_i have the same optimal value for all $i \in I^1$, we denote by μ_1^U their optimal value in the problem with maximum capacity and by ϵ_1 their optimal deviation when capacities of the leader are reduced by ΔC_i.

By optimality of Eqn. (B.15), we can deduce inequality (B.22).

$$\begin{aligned}&\sum_{j\in J} D_j \left\{\min_{i\in I}\left(\frac{1}{(1+R)^t}P_{i,j} + \mu_i^U + \epsilon_i - \epsilon_1\right)\right\} - \sum_{i\in I^1} C_i^U\left(\mu_i^U + \epsilon_i - \epsilon_1\right) - \sum_{i\in I^2} C_i^0\left(\mu_i^U + \epsilon_i - \epsilon_1\right)\\ &\leq \sum_{j\in J} D_j \left\{\min_{i\in I}\left(\frac{1}{(1+R)^t}P_{i,j} + \mu_i^U\right)\right\} - \sum_{i\in I^1} C_i^U \mu_i^U - \sum_{i\in I^2} C_i^0 \mu_i^U\end{aligned} \tag{B.22}$$

which implies inequality (B.23),

$$\begin{aligned}&\sum_{j\in J} D_j \left\{\min_{i\in I}\left(\frac{1}{(1+R)^t}P_{i,j} + \mu_i^U + \epsilon_i\right)\right\} - \sum_{i\in I^1} C_i^U\left(\mu_i^U + \epsilon_i\right) - \sum_{i\in I^2} C_i^0\left(\mu_i^U + \epsilon_i\right)\\ &\leq \sum_{j\in J} D_j \left\{\min_{i\in I}\left(\frac{1}{(1+R)^t}P_{i,j} + \mu_i^U + \epsilon_1\right)\right\} - \sum_{i\in I^1} C_i^U\left(\mu_i^U + \epsilon_1\right) - \sum_{i\in I^2} C_i^0\left(\mu_i^U + \epsilon_1\right)\end{aligned} \tag{B.23}$$

By optimality, we also know that inequality (B.24) must be satisfied.

$$\begin{aligned}&\sum_{j\in J} D_j \left\{\min_{i\in I}\left(\frac{1}{(1+R)^t}P_{i,j} + \mu_i^U + \epsilon_1\right)\right\} - \sum_{i\in I^1} \left(C_i^U - \Delta C_i\right)\left(\mu_i^U + \epsilon_1\right) - \sum_{i\in I^2} C_i^0\left(\mu_i^U + \epsilon_1\right)\\ &\leq \sum_{j\in J} D_j \left\{\min_{i\in I}\left(\frac{1}{(1+R)^t}P_{i,j} + \mu_i^U + \epsilon_i\right)\right\} - \sum_{i\in I^1} \left(C_i^U - \Delta C_i\right)\left(\mu_i^U + \epsilon_i\right) - \sum_{i\in I^2} C_i^0\left(\mu_i^U + \epsilon_i\right)\end{aligned} \tag{B.24}$$

Furthermore, an upper bound on the right-hand side of inequality (B.24) is presented in Eqn. (B.25).

$$\sum_{j\in J} D_j \left\{ \min_{i\in I}\left(\frac{1}{(1+R)^t}P_{i,j}+\mu_i^U+\epsilon_i\right)\right\} - \sum_{i\in I^1}\left(C_i^U-\Delta C_i\right)\left(\mu_i^U+\epsilon_i\right) - \sum_{i\in I^2} C_i^0\left(\mu_i^U+\epsilon_i\right)$$
$$\leq \sum_{j\in J} D_j \left\{ \min_{i\in I}\left(\frac{1}{(1+R)^t}P_{i,j}+\mu_i^U\right)+\max_i[\epsilon_i]\right\} - \sum_{i\in I^1}\left(C_i^U-\Delta C_i\right)\left(\mu_i^U+\epsilon_i\right) - \sum_{i\in I^2} C_i^0\left(\mu_i^U+\epsilon_i\right) \tag{B.25}$$

If we subtract the left-hand side of (B.24) from the right-hand side of (B.25), we can infer inequality (B.26),

$$\sum_{j\in J} D_j\left\{\max_i[\epsilon_i]-\epsilon_1\right\} - \sum_{i\in I^2} C_i^0(\epsilon_i-\epsilon_1)\geq 0 \tag{B.26}$$

Now, let us assume that $\max_{i\in I}[\epsilon_i]>\epsilon_1$. Then, for $i'=\text{argmax}[\epsilon_i]$, inequality (B.27) must be satisfied.

$$C_{i'}^0 \leq \frac{\sum_{j\in J}\left\{\max_i[\epsilon_i]-\epsilon_1\right\}-\sum_{i\in I^2\setminus\{i'\}} C_i^0\left(\epsilon_i-\epsilon_1\right)}{(\epsilon_{i'}-\epsilon_1)} \tag{B.27}$$

But we have not imposed any restrictions on the capacity of the competitors. Therefore, $\epsilon_1=\max_i[\epsilon_i]$.

Step 4: reduced costs of assignment variables for the leader cannot decrease when its capacities are reduced.

A necessary condition for optimality of a minimization linear program is that the reduced cost of the nonbasic variables must be nonnegative. Therefore, optimal demand assignments to the leader that are nonbasic ($y_{i,j}^U=0 \quad i\in I^1$) in the problem with maximum capacity must have nonnegative reduced costs as presented by inequality (B.28).

$$r_{i,j}^U=\frac{1}{(1+R)^t}P_{i,j}+\mu_i^U-\lambda_j^U\geq 0 \qquad\qquad \forall(i,j)\in\left\{(i,j):\ i\in I^1,\ j\in J,\ y_{i,j}^U=0\right\} \tag{B.28}$$

Using Eqn. (B.12), we can rewrite the reduced cost ($r_{i,j}$) for nonbasic variables $y_{i,j}$ only in terms of dual variables μ_i,

$$r_{i,j}^U=\frac{1}{(1+R)^t}P_{i,j}+\mu_i^U-\min_{i'\in I}\left(\frac{1}{(1+R)^t}P_{i,j}+\mu_i^U\right)\geq 0 \quad \forall(i,j)\in\left\{(i,j):\ i\in I^1,\ j\in J,\ y_{i,j}^U=0\right\} \tag{B.29}$$

Recall that the lower-level problem is degenerate because the leader offers a single price to each market from all facilities. This degeneracy implies that some assignment variables are nonbasic but their reduced costs are strictly equal to zero. In order to keep in the bilevel problem the degenerate assignments, we restrict the domain reduction to variables with strictly positive reduced costs in the lower-level problem with maximum capacity.

In Step 3, we established that dual variables μ_i ($i\in I^2$) cannot increase more than dual variables μ_i ($i\in I^1$) when production capacities of the leader are reduced from C_i^U to $C_i^U-\Delta C_i$. Then, according to inequality (B.30), the reduced cost of the variables of the leader cannot decrease when capacities are reduced.

$$\frac{1}{(1+R)^t}P_{i,j}+\mu_i^U-\min_{i'\in I}\left(\frac{1}{(1+R)^t}P_{i,j}+\mu_i^U\right)$$
$$\leq\frac{1}{(1+R)^t}P_{i,j}+\mu_i^U+\epsilon_i-\min_{i'\in I}\left(\frac{1}{(1+R)^t}P_{i,j}+\mu_i^U+\epsilon_i\right)\qquad\forall(i,j)\in\big\{(i,j):\ i\in I^1,\ j\in J\big\}\tag{B.30}$$

Therefore, variables $y_{i,j}$ ($i\in I^1$) with positive reduced cost in the lower-level problem with maximum capacity have positive reduced cost regardless of the leader's expansion strategy.

□