Mixed-Integer Bilevel Optimization with Nonconvex Quadratic Lower-Level Problems: Complexity and a Solution Method

Immanuel Bomze, Andreas Horländer, Martin Schmidt

Abstract. We study bilevel problems with a convex quadratic mixed-integer upper-level, integer linking variables, and a nonconvex quadratic, purely continuous lower-level problem. We prove Σ_2^p-hardness of this class of problems, derive an iterative lower- and upper-bounding scheme, and show its finiteness and correctness in the sense that it computes globally optimal points or proves infeasibility of the instance. To this end, we make use of the Karush–Kuhn–Tucker conditions of the lower-level problem for the lower-bounding step, since these conditions are only necessary but not sufficient in our setting. Moreover, integer no-good cuts as well as a simple optimality cut are used to obtain finiteness of the method. Finally, we illustrate the applicability of our approach by the first large-scale numerical experiment for this class of problems in the literature.

1. Introduction

Bilevel optimization is an important tool for modeling decision-making processes with hierarchical interactions between agents. While this is necessary in many applications such as in counter-terrorism (Wang et al. 2016), energy market design (Grimm et al. 2019), or revenue management (Labbé and Violin 2013), it also renders these types of problems very hard to solve—both in theory (Hansen et al. 1992; Jeroslow 1985) as well as in practice (Thürauf et al. 2024).

Fortunately, the area of bilevel optimization made some significant advances over the last years and decades so that we are able today to solve significantly harder and larger instances compared to what we were able to solve 10 or 15 years ago. One of the main ways on how to propel the field (besides simply trying to solve larger instances) has been to study bilevel optimization problems that have increasingly complicated lower-level problems. Examples are the inclusion of integer aspects in the lower-level problem (DeNegre 2011; DeNegre and Ralphs 2009; Fischetti et al. 2018; Kleinert et al. 2021b), the incorporation of stochastic or robust models for addressing uncertainty (Beck et al. 2023a), or the consideration of nonconvex nonlinearities (see below for a detailed discussion of the relevant literature).

In this paper, we focus on the latter aspect. We study the computational complexity of mixed-integer bilevel problems with integer linking variables and a purely continuous but nonconvex lower-level problem, and design an algorithm to solve these problems. In particular, we consider the case of a quadratic lower-level problem, where the lower level has a polyhedral feasible set but a nonconvex quadratic objective function. This nonconvexity is the main reason for the hardness of the problem, and we thus focus on this aspect both in terms of computational complexity as well as with the design of our solution method.

Date: June 19, 2025.

2020 *Mathematics Subject Classification.* 91A65, 90C11, 90C26, 68Q15.

Key words and phrases. Bilevel optimization, Nonconvex lower levels, Mixed-integer optimization, Complexity.

This paper offers three main contributions:

(i) We prove that the problems under consideration are Σ_2^p-hard in general, by a restriction to a special variant of the bilevel knapsack problem studied by Caprara et al. (2014); see Theorem 9.

(ii) We derive an iterative procedure that successively tightens lower and upper bounds. The lower-bounding step consists in a KKT-based single-level relaxation of the problem, which is only a relaxation and not a reformulation since the KKT conditions are only necessary and not sufficient in general. Finiteness and correctness of the method is further ensured by using integer no-good cuts and a simple optimality cut; see Theorem 17.

(iii) We present a numerical study showcasing the applicability of our approach; see Section 5. To the best of our knowledge, this is the first numerical study for this problem class using a large test set also including larger instances—compared to other instances used in the literature that often have not more than 5 to 10 variables.

Very closely related to our contribution is the method developed by Mitsos et al. (2008), where a purely continuous bilevel problem is considered which may contain nonconvex nonlinearities. The approach has been generalized later to the mixed-integer setting by Mitsos (2010). The key idea presented by Mitsos et al. (2008) is a lower-bounding procedure using an iterative tightening of a relaxation of an optimal-value function constraint, which leads to a finitely terminating algorithm when included in a Blankenship–Falk like algorithm (Blankenship and Falk 1976). This approach can be extended by the KKT conditions of the lower level (which are only necessary and not sufficient in the nonconvex setting), which is also what we do in this paper. However, their strategy of guaranteeing finite termination is completely different to ours. Moreover, our approach does not need comparably strong assumptions as Mitsos et al. (2008) do; see Assumption 3 in the cited paper, which excludes instances that have x-dependent equality constraints in the lower level; see Page 477 in Mitsos et al. (2008) for a brief discussion of their assumptions. However, there is a later follow-up paper by Mitsos and some other co-authors (Djelassi et al. 2019) in which they present an extension of the former methods with which it is then also possible to tackle x-dependent lower-level equality constraints. Still, the size of the instances considered in the numerical results is as small as reported above.

Let us mention that Mitsos and Barton (2006) showed that one cannot work with convex relaxations—such as the αBB underestimator as introduced in Androulakis et al. (1995)—of the lower-level problem. Other papers presenting methods for solving bilevel problems with nonconvexities caused by continuous nonlinear functions in the lower-level problem are the contributions about the branch-and-sandwich approach (Kleniati and Adjiman 2011; Kleniati and Adjiman 2014a; Kleniati and Adjiman 2015; Kleniati and Adjiman 2014b) or the value-function approach presented by Lozano and Smith (2017). Finally, let us also mention Gaar et al. (2022), Horländer et al. (2024), and Kleinert et al. (2021a), who all consider (M)IQP bilevel problems as well, but all with a convex lower-level problem. Hence, the derived solution methodologies are not comparable.

The remainder of this paper is structured as follows. In Section 2 we state the problem under consideration for which we prove its Σ_2^p-hardness in Section 3. Our solution approach is discussed in detail in Section 4, and we present the numerical results in Section 5. Finally, we conclude in Section 6, where we also sketch some open problems for future research.

2. Problem Statement

We study mixed-integer quadratic bilevel problems of the form

$$\begin{aligned} \min_{x,y} \quad & q_u(x,y) = \frac{1}{2}x^\top H_u x + c_u^\top x + \frac{1}{2}y^\top G_u y + d_u^\top y \\ \text{s.t.} \quad & Ax + By \geq a, \\ & x_i \in \mathbb{Z} \cap [x_i^-, x_i^+] \quad \text{for all } i \in I \subseteq [n_x] := \{1, \dots, n_x\}, \\ & x_i \in \mathbb{R} \quad \text{for all } i \in [n_x] \setminus I, \\ & y \in \arg\min_{\bar{y}} \left\{ q_\ell(\bar{y}) = \frac{1}{2}\bar{y}^\top G_\ell \bar{y} + d_\ell^\top \bar{y} \colon Cx_I + D\bar{y} \geq b, \ \bar{y} \in \mathbb{R}^{n_y} \right\}, \end{aligned} \tag{1}$$

where $H_u \in \mathbb{Q}^{n_x \times n_x}$ as well as $G_u \in \mathbb{Q}^{n_y \times n_y}$ are symmetric and positive semidefinite matrices and $G_\ell \in \mathbb{Q}^{n_y \times n_y}$ is symmetric but possibly indefinite. Furthermore, we have vectors $c_u \in \mathbb{Q}^{n_x}$, $d_u, d_\ell \in \mathbb{Q}^{n_y}$, matrices $A \in \mathbb{Q}^{m_u \times n_x}$, $B \in \mathbb{Q}^{m_u \times n_y}$, $C \in \mathbb{Q}^{m_\ell \times |I|}$, $D \in \mathbb{Q}^{m_\ell \times n_y}$, finite bounds $x_i^- \leq x_i^+$ for the integer variables x_i, $i \in I$, as well as right-hand side vectors $a \in \mathbb{Q}^{m_u}$ and $b \in \mathbb{Q}^{m_\ell}$. The variables x contain the integer (x_I) and continuous ($x_{[n_x] \setminus I}$) upper-level variables and y denotes the (continuous) lower-level variables. In this setup, the upper-level problem is a convex mixed-integer quadratic problem (MIQP) and for fixed integer linking variables x_I, the lower level is a possibly nonconvex quadratic problem (QP). Note that we presented both upper- and lower-level constraints by only using inequality constraints. This is just for the ease of better reading. All techniques presented in the following are also applicable to problems that have equality constraints or both types of constraints as well. This class of problems covers applications in which the lower-level problem is concave, e.g., due to economies of scale. Moreover, to give another example, it also allows to address clique-related problems in the lower-level problem via the Motzkin–Straus approach (Motzkin and Straus 1965) and its regularizations (Bomze 1997; Hungerford and Rinaldi 2019).

In what follows, we make the following assumptions.

Assumption 1. *All linking variables x_I are bounded integers.*

Assumption 2. *The bilevel constraint region*

$$P := \{(x,y) \colon Ax + By \geq a, \ Cx_I + Dy \geq b, \ x_i \in \mathbb{Z} \cap [x_i^-, x_i^+] \text{ for all } i \in I\}.$$

is non-empty and compact.

If some integer upper-level variables in Problem (1) do not appear in the lower-level problem, the corresponding columns in the matrix C would be set to zero. The following terminology is standard in bilevel optimization (Beck and Schmidt 2023; Dempe 2002; Dempe et al. 2015).

Definition 3 (High-point relaxation, rational reaction set, bilevel feasibility)**.** *The* high-point relaxation *of the bilevel problem* (1) *is given as*

$$\min\{q_u(x,y) \colon (x,y) \in P\}$$

Moreover, for any $\hat{x} \in \mathbb{R}^{n_x}$ the set

$$S(\hat{x}) := \arg\min_{\bar{y}} \{q_\ell(\bar{y}) \colon C\hat{x}_I + D\bar{y} \geq b, \ \bar{y} \in \mathbb{R}^{n_y}\}$$

is called rational reaction set *of the follower and every point $(x,y) \in P$ with $y \in S(x)$ is called* bilevel-feasible.

3. Σ_2^p-Hardness

In this section, we show that the bilevel problem (1) is Σ_2^p-hard.[1] The proofs of all results in this section can be found in Appendix A. In what follows, we consider the problem

$$\begin{aligned} \min_{x\in\mathbb{Z},y} \quad & -bx - c^\top y \\ \text{s.t.} \quad & x^- \le x \le x^+, \\ & y \in \arg\min_{\bar y} \left\{ -d^\top \bar y + M\left(\sum_{i=1}^{n_y} \bar y_i - \bar y_i^2\right) : \; -d^\top \bar y \ge -x, \; \bar y \in [0,1]^{n_y} \right\}, \end{aligned} \tag{2}$$

where $0 \neq d \in \mathbb{Z}_+^{n_y}$, $c \in \mathbb{Z}_+^{n_y}$, $b, x^-, x^+ \in \mathbb{Z}_+$, and $M > 0$. Problem (2) is a special case of Problem (1). The quadratic term $M(\sum_{i=1}^{n_y} \bar y_i - \bar y_i^2)$ in the lower-level objective function can be restated as $M(\bar y^\top Q \bar y + e^\top \bar y)$, where e is the vector of all ones and $Q = \operatorname{diag}(-e)$. It is strictly concave, hence nonconvex, and corresponds to a penalty term that is positive if $\bar y$ is fractional. Obviously, Problem (2) is equivalent to

$$\begin{aligned} \max_{x\in\mathbb{Z},y} \quad & bx + c^\top y \\ \text{s.t.} \quad & x^- \le x \le x^+, \\ & y \in \arg\max_{\bar y} \left\{ d^\top \bar y - M\left(\sum_{i=1}^{n_y} \bar y_i - \bar y_i^2\right) : \; d^\top \bar y \le x, \; \bar y \in [0,1]^{n_y} \right\}, \end{aligned} \tag{3}$$

with the same range specification for the parameters $d, c, b, x^\pm, M$. We will show that Problem (3) covers a special variant of a bilevel knapsack problem for which its Σ_2^p-hardness is shown in the literature. From this, Σ_2^p-hardness of the problem class considered in this paper follows immediately.

First we show that for every $M > 0$, a globally optimal solution $\hat y$ to the x-parameterized lower level of Problem (3) has at most one fractional component.

Lemma 4. *Let x be feasible for the upper level of Problem (3) and let $\hat y$ be a globally optimal solution to the x-parameterized lower level. Assume that there exists an index j such that $\hat y_j \in (0,1)$. Then, it holds $\hat y_i \in \{0,1\}$ for all $i \neq j$.*

In the special case of $\|d\|_\infty = 1$, we show that if $\hat y$ is a globally optimal solution to the x-parameterized lower level, then there exists no j with $\hat y_j \in (0,1)$.

Lemma 5. *Assume that $\|d\|_\infty = 1$. Let x be feasible for the upper level of (3). Then, every globally optimal solution to the x-parameterized lower level is binary.*

For $\|d\|_\infty \ge 2$ we use Lemma 4 to show that a fractional component of the globally optimal solution to the x-parameterized lower level of Problem (3) cannot be arbitrarily close to zero or one.

Lemma 6. *Assume that $\|d\|_\infty \ge 2$. Let x be feasible for the upper level of (3) and let $\hat y$ be a globally optimal solution to the x-parameterized lower level. Let j be an index such that $\hat y_j \in (0,1)$. Then, it holds*

$$d^\top \hat y = x$$

and

$$\frac{1}{\|d\|_\infty} \le \hat y_j \le 1 - \frac{1}{\|d\|_\infty}.$$

With Lemma 6 we immediately get the following result.

[1] For the ease of better readability, we avoid to state "the decision variant of the bilevel problem" in what follows.

Lemma 7. *Assume that* $\|d\|_\infty \geq 2$. *Let* $(\hat{x}, \hat{y})$ *be bilevel-feasible for Problem* (3) *with* $\hat{y}_j \in (0,1)$ *for one index* j. *Then, it holds*

$$\sum_{i=1}^{n_y} \hat{y}_i - \hat{y}_i^2 \geq \frac{1}{\|d\|_\infty} - \frac{1}{\|d\|_\infty^2}.$$

Lemma 7 states a lower bound for the penalization in the lower-level objective function of a bilevel-feasible point that is not integer. Now, we consider the bilevel problem

$$\begin{aligned}
\max_{x \in \mathbb{Z}, y} \quad & bx + c^\top y \\
\text{s.t.} \quad & x^- \leq x \leq x^+, \\
& y \in \arg\max_{\bar{y}} \left\{ d^\top \bar{y} : d^\top \bar{y} \leq x, \bar{y} \in \{0,1\}^{n_y} \right\},
\end{aligned} \tag{4}$$

where $0 \neq d \in \mathbb{Z}_+^{n_y}$, $c \in \mathbb{Z}_+^{n_y}$, and $b, x^-, x^+ \in \mathbb{Z}_+$. The problem has been introduced in Dempe and Richter (2000) and its Σ_2^p-hardness has been shown in Section 3.1 in Caprara et al. (2014), where it is called “DeRi”.

We show the equivalence of Problem (3) and Problem (4) for M being sufficiently large but still of polynomial size (in the size of the input data of the problem).

Theorem 8. *Let* $\mathcal{B}_1$ *be the set of bilevel-feasible points of Problem* (3) *and let* $\mathcal{B}_2$ *be the set of bilevel-feasible points of Problem* (4). *If* $\|d\|_\infty = 1$, *then for any* $M > 0$ *it holds* $\mathcal{B}_1 = \mathcal{B}_2$. *Furthermore, if* $\|d\|_\infty \geq 2$, *then, for*

$$M > \frac{n_y \|d\|_\infty^3}{\|d\|_\infty - 1},$$

it holds $\mathcal{B}_1 = \mathcal{B}_2$.

To sum up, we have shown that Problem (3) is equivalent to Problem (2), which is a special case of Problem (1). Furthermore, there exists a constant $M > 0$ of polynomial size (in the input data of the problem) so that Problem (3) is equivalent to Problem (4); see Theorem 8. Hence, we have the following result.

Theorem 9. *Problem* (1) *is* Σ_2^p*-hard.*

4. Solution Approach

We tackle Problem (1) by using an iterative method, which improves the lower and upper bound for the optimal objective function value until global optimality is certified. In Section 4.1 and 4.2 we describe how we get lower and upper bounds before we then state the algorithm in Section 4.3 and prove its correctness.

4.1. Computing Lower Bounds.

To compute lower bounds for Problem (1), we replace the lower level with its KKT conditions. This yields the optimization problem

$$\begin{aligned}
\min_{x,y,\lambda} \quad & q_u(x,y) = \frac{1}{2} x^\top H_u x + c_u^\top x + \frac{1}{2} y^\top G_u y + d_u^\top y \\
\text{s.t.} \quad & Ax + By \geq a, \quad Cx_I + Dy \geq b, \\
& x_i \in \mathbb{Z} \cap [x_i^-, x_i^+] \quad \text{for all } i \in I, \\
& G_\ell y + d - D^\top \lambda = 0, \quad \lambda \geq 0, \\
& \lambda^\top (Cx_I + Dy - b) = 0.
\end{aligned} \tag{5}$$

For Problem (5), the following property holds.

Proposition 10. *Problem (5) is a relaxation of problem (1) in the following sense: Let (x^*, y^*) be a globally optimal solution to the bilevel problem (1) and let $(\hat{x}, \hat{y}, \hat{\lambda})$ be a globally optimal solution to the KKT relaxation (5). Then, it holds $q_u(\hat{x}, \hat{y}) \leq q_u(x^*, y^*)$. Furthermore, if $(\hat{x}, \hat{y})$ is bilevel-feasible, it holds $q_u(\hat{x}, \hat{y}) = q_u(x^*, y^*)$, i.e., $(\hat{x}, \hat{y})$ is bilevel-optimal.*

Proof. We define the set of KKT points of the x-parameterized lower level as follows:

$$\begin{aligned}\mathrm{KKT}(x) := \Big\{ & y \in \mathbb{R}^{n_y} : Dy \geq b - Cx_I, \\ & \exists \lambda \in \mathbb{R}^{m_u}_+ : G_\ell y + d - D^\top \lambda = 0,\ \lambda^\top (Cx_I + Dy - b) = 0 \Big\},\end{aligned}$$

Since all constraints of the lower-level problem are linear, the KKT theorem leads to $S(x) \subseteq \mathrm{KKT}(x)$ for all x. The remainder is evident. □

Proposition 10 states that we obtain a lower bound for the optimal objective value of the bilevel problem (1) by solving the KKT relaxation (5), which is an MIQP with complementarity constraints. We handle the latter KKT complementarity conditions via SOS1-type constraints; see, e.g., Kleinert and Schmidt (2023).

In our iterative method, we improve this lower bound by successively reducing the feasible region of Problem (5). We do this by adding cutting planes of the form $h(x, y) \geq 0$. Here, h is a function (in our application, linear) that satisfies $h(x, y) < 0$ for all $(x, y) \in V'$ and $h(x, y) \geq 0$ for all $(x, y) \notin V'$, where V' is a certain subset of P. In practice, V' will, among others, contain points that we already visited in previous iterations. Note that a solution to Problem (5) that additionally contains cutting planes[2] still provides a lower bound as long as the cuts do not cut off all bilevel-optimal points. This is formalized in the following result.

Proposition 11. *Let k be a positive integer. For $i = 1, \dots, k$, let $V_i \subseteq P$ and let h_i be a function that separates V_i and $P \setminus V_i$, i.e.,*

$$h_i(x, y) < 0 \text{ for all } (x, y) \in V_i \quad \text{and} \quad h_i(x, y) \geq 0 \text{ for all } (x, y) \in P \setminus V_i\,.$$

Consider Problem (5) with the additional restrictions $h_i(x, y) \geq 0$ for $i = 1, \dots, k$, and let $(\hat{x}, \hat{y}, \hat{\lambda})$ be a globally optimal solution to that problem. If there exists a point $(x^*, y^*) \notin V := \bigcup_{i=1}^k V_i$ that is globally optimal for the bilevel problem (1), then it holds

$$q_u(\hat{x}, \hat{y}) \leq q_u(x^*, y^*)\,.$$

From Corollary 10 we know that if a globally optimal solution to the KKT relaxation (5) is bilevel-feasible, it is also globally optimal for the bilevel problem (1). On the other hand, if it is not bilevel-feasible, we remove it from the feasible region of Problem (5) by using a cutting plane. If we re-solve the problem, we get a lower bound that is not smaller than the lower bound obtained by the previous solution. Hence, in an iterative method that adds cutting planes to Problem (5) in each iteration, we successively improve the lower bound. The derivation of the cutting planes is discussed in Section 4.3.

4.2. **Computing Upper Bounds.** Every bilevel-feasible point provides an upper bound (and every optimal solution to Problem (5), which is bilevel-feasible, is also bilevel-optimal by Proposition 10). However, if a globally optimal solution $(\hat{x}, \hat{y}, \hat{\lambda})$ to the KKT relaxation (5) is bilevel-infeasible, we can still use it to obtain an upper

[2]For linear h, incorporation of these cuts can easily be embedded into our general framework by suitable extension of A and b, and no constraint qualifications are needed.

bound for the optimal objective function value of Problem (1). Therefore, we first solve the $\hat{x}$-parameterized lower level (a nonconvex QP) to obtain

$$\phi(\hat{x}) := \min_y \left\{ \frac{1}{2} y^\top G_\ell y + d_\ell^\top y \colon C\hat{x}_I + Dy \geq b,\ y \in \mathbb{R}^{n_y} \right\}.$$

Afterward, we apply a refinement procedure as done in, e.g., Fischetti et al. (2018). Hence, we solve the restricted high-point relaxation, a QP with an additional nonconvex quadratic constraint (QCQP):

$$\begin{aligned} \min_{x,y} \quad & q_u(x, y) \\ \text{s.t.} \quad & (x, y) \in P, \\ & x_I = \hat{x}_I, \\ & q_\ell(y) \leq \phi(\hat{x}). \end{aligned} \tag{6}$$

If the refinement problem (6) is feasible, we obtain a bilevel-feasible point, and, hence, an upper bound. Otherwise, every point (x, y) with $x_I = \hat{x}_I$ is bilevel-infeasible.

4.3. **The Algorithm.** The proposed method successively generates upper bounds for the optimal objective function value of Problem (1) by computing bilevel-feasible points. Furthermore, it iteratively improves the lower bound. For the former, in each iteration, we compute a bilevel-feasible point as explained in Section 4.2 if possible and for the latter, we shrink the feasible region of Problem (5) by adding cutting planes in form of integer no-good cuts (INGCs) as explained in Section 4.1. We use the relative optimality gap[3]

$$\gamma := \frac{|U - L|}{10^{-10} + |U|} \tag{7}$$

as a stopping criterion, where L is the best known lower bound and U is the incumbent, i.e., the best known upper bound.

Remark 12. *To apply integer no-good cuts (DeNegre and Ralphs 2009; Fischetti et al. 2018; Schmidt and Thürauf 2024; Wolsey 2020), we first represent the variables x_i for $i \in I$ with binary vectors v^i, i.e.,*

$$x_i = -2^{s_i} v^i_{s_i} + \sum_{k=0}^{s_i - 1} 2^k v^i_k,$$

where

$$s_i := \left\lceil \log_2 \left(\max \left\{ |x_i^-|, |x_i^+| \right\} \right) + 1 \right\rceil.$$

Now, let $(\hat{x}, \hat{y}, \hat{\lambda})$ be a point that is feasible for the KKT relaxation (5), and let $\{\hat{v}^i : i \in I\}$ be the binary vectors representing $\hat{x}_I$. After we perform a refinement step as explained in Section 4.2, we use the inequality

$$\sum_{i \in I} \left(\sum_{k \in S_i : \hat{v}^i_k = 0} v^i_k + \sum_{k \in S_i : \hat{v}^i_k = 1} (1 - v^i_k) \right) \geq 1 \tag{8}$$

with $S_i := \{0, \dots, s_i\}$ to cut off every point (x, y) with $x_I = \hat{x}_I$. Inequality (8) is then called an integer no-good cut for $\hat{x}_I$.

[3] This is motivated by `https://www.ibm.com/docs/en/icos/22.1.1?topic=parameters-relative-mip-gap-tolerance`.

Note that the integer no-good cut in (8) may cut off multiple points. This can be done because after performing the refinement step, none of the points (x, y) with $x_I = \hat{x}_I$ can improve the upper bound U anymore. In addition to these cuts, we add an optimality cut of the form $q_u(x, y) \leq U$ to Problem (5). Adding both types of inequalities to the KKT relaxation leads to the problem

$$
\begin{aligned}
\min_{x,y,\lambda} \quad & q_u(x,y) = \frac{1}{2} x^\top H_u x + c_u^\top x + \frac{1}{2} y^\top G_u y + d_u^\top y \\
\text{s.t.} \quad & Ax + By \geq a, \quad Cx_I + Dy \geq b, \\
& x_i \in \mathbb{Z} \cap [x_i^-, x_i^+] \text{ for all } i \in I, \\
& G_\ell y + d - D^\top \lambda = 0, \quad \lambda \geq 0, \\
& \lambda^\top (Cx_I + Dy - b) = 0, \\
& \text{INGC (8) on } \tilde{x}_I \quad \text{for all } \tilde{x}_I \in X, \\
& q_u(x,y) \leq U.
\end{aligned}
\tag{9}
$$

The set X contains the values of the linking variables of points that are feasible for Problem (5) and gets updated throughout the procedure. The optimality cut in the last constraint is optional; see Section 5.3.1 for a thorough discussion. Note that Problem (9) is, in general, not a relaxation of the bilevel problem (1) anymore, because the optimality- and integer no-good cuts may separate bilevel-feasible points. However, we never cut off points that are bilevel-feasible and that can improve the incumbent in the respective iteration. Hence, we get the following corollary.

Corollary 13. *Let (x^*, y^*) with $x_I^* \notin X$ be a globally optimal solution to the bilevel problem (1) and let $(\hat{x}, \hat{y}, \hat{\lambda})$ be a globally optimal solution to Problem (9). Then, it holds $q_u(\hat{x}, \hat{y}) \leq q_u(x^*, y^*)$.*

Corollary 13 is a special case of Proposition 11, where each function $h_i(x, y)$ represents one integer no-good cut and $V = \{(x, y) : x_I \in X\}$. Note that the optimality cut does not separate bilevel-optimal points. Hence, its presence in Problem (9) does not interfere with the statement in Corollary 13. Based on Problem (9) we propose the procedure in Algorithm 1 to solve the bilevel problem (1).

In Line 2 of Algorithm 1 we solve Problem (9). Note that this problem is the KKT relaxation (5) for $k = 0$. By Assumption 2, we know that if it is feasible, we get a solution (x^k, y^k, λ^k) in Line 4. However, for some iteration k, Problem (9) with $X := X_k$ might be infeasible. This can happen when X_k contains every bilevel-optimal point. In that case, U equals the optimal objective function value of Problem (1), see Line 13, and the optimality cut prevents $q_u(x^k, y^k)$ from exceeding U. Moreover, the last bilevel-feasible point that lead to an update of U in a previous iteration is bilevel-optimal. Hence, the algorithm proceeds in Line 17 and verifies that U is finite. It then terminates in Line 22 with a globally optimal solution to Problem (1).

Otherwise, i.e., if Problem (9) is feasible, the value $q_u(x^k, y^k)$ is always a lower bound on the optimal upper-level objective function value; see Corollary 13. Hence, we update the lower bound L in Line 4. Then, we solve the x^k-parameterized lower-level problem in Line 5 to obtain an optimal response $\bar{y}^k$ for the given x^k. It is possible that (x^k, y^k) is already bilevel-feasible, i.e., $q_\ell(y^k) = q_\ell(\bar{y}^k)$. Then, the point is also bilevel-optimal, see Proposition 14 below, and the algorithm terminates. The computed point $(x^k, \bar{y}^k)$ is bilevel-feasible if it satisfies the upper-level constraints. Moreover, the x^k-parameterized lower-level could have multiple globally optimal solutions that lead to different values in the upper-level objective. To address both subjects, we perform a refinement step in Line 9 of Algorithm 1; see Section 4.2

Algorithm 1: Iterative method to solve the bilevel problem (1)

Input: Problem (1), $L = -\infty$, $U = \infty$, $\gamma = \infty$, $k = 0$, and $X_0 = \emptyset$.
Output: A globally optimal solution to the bilevel problem (1) or an indication of infeasibility.

1 **if** $\gamma > 0$ **then**
2   Solve Problem (9) with $X = X_k$ to global optimality.
3   **if** *Problem* (9) *is feasible* **then**
4   Let (x^k, y^k, λ^k) be the solution to Problem (9). Set $L \leftarrow q_u(x^k, y^k)$.
5   Solve the x^k-parameterized lower-level problem to compute an optimal follower's response $\bar{y}^k$ as well as $\phi(x^k)$.
6   **if** (x^k, y^k) *is bilevel-feasible, i.e.,* $q_\ell(y^k) = q_\ell(\bar{y}^k)$ **then**
7   Set $(x^*, y^*) \leftarrow (x^k, y^k)$ and go to Line 22.
8   **else**
9   Solve the refinement problem (6).
10   **if** *the refinement problem* (6) *is feasible* **then**
11   denote the obtained bilevel-feasible point $(\bar{x}, \bar{y})$. It holds $\bar{x}_I = x_I^k$.
12   **if** $q_u(\bar{x}, \bar{y}) < U$ **then**
13   Set $U \leftarrow q_u(\bar{x}, \bar{y})$ and $(x^*, y^*) \leftarrow (\bar{x}, \bar{y})$.
14   **if** $U < \infty$ **then**
15   Set $\gamma \leftarrow |U - L|/(10^{-10} + |U|)$.
16   Set $X_{k+1} \leftarrow X_k \cup \{x_I^k\}$ and $k \leftarrow k + 1$. Go to Line 2.
17   **else if** $U < \infty$ **then**
18   Go to Line 22.
19   **else**
20   **return** *"The bilevel problem* (1) *is infeasible."*
21 **else**
22   **return** *the globally optimal solution* (x^*, y^*) *to the bilevel problem* (1).

and Fischetti et al. (2018) for further details. Therefore, we solve the restricted high-point relaxation (6). If it is feasible, then we obtain a bilevel-feasible point in Line 11 and we improve the incumbent U if possible; see Line 13. Note that this also updates the optimality cut given in Problem (9). Otherwise, every pair (x, y) with $x_I = x_I^k$ and y being an optimal solution to the x_I^k-parameterized lower level violates the upper-level constraints. Hence, there exists no bilevel-feasible point $(\bar{x}, \bar{y})$ with $\bar{x}_I = x_I^k$. In both cases, we checked every point (x, y) with $x_I = x_I^k$ for bilevel feasibility. Hence, we use an integer no-good cut as described in (8) to remove all points (x, y) with $x_I = x_I^k$ from our search space. We do this by setting X_{k+1} to $X_k \cup \{x_I^k\}$ in Line 16 of Algorithm 1.

The algorithm terminates if the optimality gap γ as defined in (7) is zero at the start of any iteration; see Line 1. In that case, the last point (x^*, y^*) that improved our incumbent U is bilevel-optimal; see Line 22. Once the method updates the upper bound in Line 13, γ gets updated in every iteration; see Line 15. The algorithm also terminates if for a solution (x^k, y^k, λ^k) obtained in Line 4, the point (x^k, y^k) is bilevel-feasible; see Lines 6–7 and Proposition 14 below. In both cases, the algorithm terminates with a globally optimal solution to the bilevel problem (1). However, the algorithm also detects infeasibility of the bilevel problem. If Problem (9) with $X := X_k$ is infeasible for some iteration k, then the algorithm checks in Line 17 if $U < \infty$, i.e., if a bilevel-feasible point exists. If this is the case, then (x^*, y^*), i.e., the last point that improved the incumbent, is a globally optimal solution to the bilevel problem (1). Otherwise, the bilevel problem (1) is infeasible; see Line 20. Note that the check $\gamma > 0$ in the very first line can also be replaced by $\gamma \geq \varepsilon$ in order to be able to terminate earlier with ε-optimal points.

First, we prove that the termination criterion in Lines 6–7 is correct.

Proposition 14. *If for any iteration $k \geq 0$ the solution (x^k, y^k) to Problem (9) with $X := X_k$ is bilevel-feasible, then it is bilevel-optimal.*

Proof. Let (x^*, y^*) be a bilevel-optimal solution. If $x_I^* \in X_k$ for iteration k, i.e., we already identified the point (x^*, y^*) as bilevel-feasible in a previous iteration $k' < k$, then we have $U = q_u(x^*, y^*)$. This is because $q_u(x^*, y^*)$ yields the best possible upper bound. Due to the optimality cut $q_u(x, y) \leq U$, in iteration k of Algorithm 1, we exclude every non-optimal bilevel-feasible point in Problem (9) and, hence, (x^k, y^k) has to be bilevel-optimal as well. On the other hand, if $x_I^* \notin X_k$, then it holds $q_u(x^k, y^k) \leq q_u(x^*, y^*)$ due to Corollary 13 and $q_u(x^*, y^*) \leq q_u(x^k, y^k)$ because (x^k, y^k) is bilevel-feasible. Hence, it is bilevel-optimal. □

Note, that Proposition 14 does not hold true in general if we omit the optimality cut in Problem (9), which we illustrate using the following example.

Example 15. *Consider the problem*

$$\begin{aligned} \min_{x \in \{0,1\}, y} \quad & x + y^2 \\ \text{s.t.} \quad & y \in \arg\min \left\{ -\bar{y}^2 \colon y \geq -x, -2y \geq 3x - 2 \right\}; \end{aligned} \tag{10}$$

see Figure 1 for an illustration. For $x = 0$, the follower can choose $y \in [0, 1]$ and for $x = 1$ it is $y \in [-1, -0.5]$. The KKT conditions of the lower level are given by

$$\begin{aligned} -2y - \lambda_1 + 2\lambda_2 = 0, \\ \lambda_1, \lambda_2 \geq 0, \\ y + x \geq 0, \ -2y - 3x + 2 \geq 0, \\ \lambda_1(y + x) = 0, \ \lambda_2(-2y - 3x + 2) = 0, \end{aligned}$$

where λ_1 and λ_2 are the respective Lagrange multipliers. For $x = 0$, the KKT points are given by $(y, \lambda_1, \lambda_2) \in \{(0, 0, 0), (1, 0, 1)\}$. For $x = 1$, there is one KKT point given by $(y, \lambda_1, \lambda_2) = (-1, 2, 0)$. Therefore, the KKT relaxation (5) has three feasible points out of which on only $(1, 0, 1)$ and $(-1, 2, 0)$ correspond to bilevel-feasible points. The first one, $(1, 0, 1)$ for $x = 0$ is the uniquely determined bilevel-optimal point. When we solve the KKT relaxation (5) in Line 2 in iteration $k = 0$ of Algorithm 1, we get the solution $(x^0, y^0) = (0, 0)$, which is bilevel-infeasible. Then, we apply the refinement step in Line 9 and compute the bilevel-feasible point $(\bar{x}, \bar{y}) = (0, 1)$ that is also globally optimal for the bilevel-problem. We update the incumbent to $U = 1$ and set $(x^*, y^*) = (0, 1)$, see Line 13. Then, we add the no-good cut $x \geq 1$ (note, that x is already binary) in Line 16. Now, if we consider Problem (9) in iteration $k = 1$ with $X = \{1\}$ and with the optimality cut $q_u(x, y) \leq 1$, the problem is infeasible. Hence, we terminate with the bilevel-optimal solution $(x^*, y^*) = (0, 1)$; see Lines 17 and 22. However, if we omit the optimality cut in Problem (9), the point $(x, y) = (1, -1)$ remains feasible in iteration $k = 1$ and this will then be its solution. Since the point is also bilevel-feasible, the algorithm would wrongly terminate with the non-optimal solution $(x^*, y^*) = (1, -1)$; see Lines 6–7 and 22.

Since we use an integer no-good cut in every iteration of Algorithm 1, it is easy to see that the feasible region of problem (9) is reduced.

Corollary 16. *Let $\mathcal{F}^k$ denote the feasible region of problem (9) in iteration k. Then, it holds $\mathcal{F}^{k+1} \subsetneq \mathcal{F}^k$.*

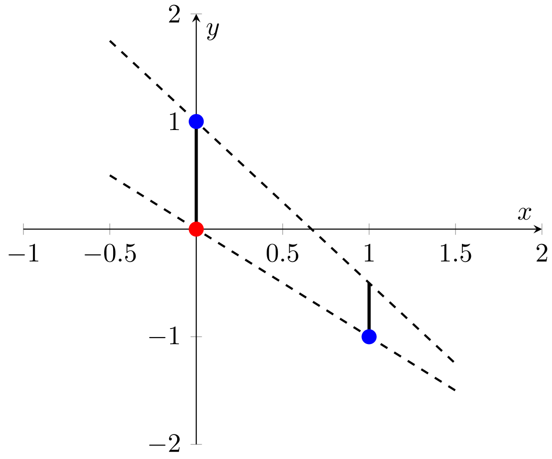

FIGURE 1. The feasible region of Problem (10). The point $(0, 0)$ is the bilevel-infeasible solution of the KKT relaxation. The points $(0, 1)$ and $(1, -1)$ are bilevel-feasible.

From this, it is obvious that the lower bound L is non-decreasing. With this at hand, we can prove the correctness of the method.

Theorem 17. *Algorithm 1 terminates after finitely many steps either with a globally optimal solution of the bilevel problem (1) or with a correct indication of infeasibility.*

Proof. All upper-level variables that appear in the lower level, i.e., x_i with $i \in I$ are bounded integers; see Assumption 1. Hence, the set

$$T := \left\{ x_I : x_i \in \mathbb{Z} \cap [x_i^-, x_i^+] \text{ for all } i \in I \right\}$$

is finite.

Let us first consider the case in which the bilevel problem (1) is infeasible. Then, the refinement problem that we solve in Line 9 is also infeasible. Hence, we never update the incumbent U since we never reach Line 13. However, we still apply an integer no-good cut that removes all points (x, y) with $x_I = x_I^k$ in every iteration. After at most $|T| < \infty$ many iterations, Problem (9) with $X := X_{|T|}$ becomes infeasible and the algorithm terminates in Line 20 with a correct indication of infeasibility.

Now assume that the bilevel problem is feasible. If a solution (x^k, y^k) to Problem (9) with $X := X_k$ is not bilevel-feasible, we apply the refinement procedure in Line 9. Among all points $(x, y) \in P$ with $x_I = x_I^k$ we compute a bilevel-feasible point that minimizes the upper-level objective function value, if any exists. If the computed point improves the incumbent, we store it and update U in Line 13. Otherwise, there exists no point $(x, y) \in P$ with $x_I = x_I^k$ that is bilevel-feasible and improves the incumbent. In any case, we remove the points $(x, y) \in P$ with $x_I = x_I^k$ by using an integer no-good cut on the linking variables of x^k in Line 16. Therefore, we cannot exclude bilevel-feasible points that would further improve our incumbent. After $k^* \leq |T| < \infty$ many iterations, Problem (9) with $X := X_{k^*}$ becomes either infeasible or has a solution that is bilevel-feasible. In the first case, we know that we observed every point $(x, y) \in P$ because infeasibility can only result from the integer no-good cuts. Since the bilevel problem is feasible, we already obtained a bilevel-optimal point in a previous iteration and it is given by (x^*, y^*). In the latter case, the algorithm terminates correctly with (x^{k^*}, y^{k^*}) as a bilevel-optimal solution as shown in Proposition 14. □

5. Numerical Results

In this section we present the numerical results obtained with our method, applied to bilevel test instances from the literature. To the best of our knowledge, there is no publicly available solver that is tailored for mixed-integer quadratic bilevel problems with a nonconvex QP in the lower level. Note that there are methods that can solve the more general case of mixed-integer nonconvex bilevel problems; see, e.g., Mitsos (2010). However, their computational study is limited to only a few and small instances, and we are not aware of any publicly available implementation that we could use for a comparison. Since the high-point relaxation, i.e., Problem (9) without the KKT conditions of the lower-level problem, is the most classic relaxation in bilevel optimization (Kleinert et al. 2021b), we take this alternative relaxation as the baseline for our numerical study. Therefore, we compare the following two approaches:

KKT: Algorithm 1 with Problem (9) including the KKT conditions of the lower-level problem;

HPR: Algorithm 1 with Problem (9) not including the KKT conditions of the lower-level problem.

The implementation of both methods is publicly available at `https://github.com/AndreasHorlaender/Solver-for-MIQP-QP-Bilevel-Problems`. In Section 5.1, we briefly describe our computational setup. Then, in Section 5.2, we describe the instance sets used for our experiments. Finally, Section 5.3 contains the discussion of the numerical results.

5.1. **Hardware and Software Setup.** We implemented Algorithm 1 in Python 3.9.7 and we use Gurobi 11.0.2 to solve all occurring single-level problems. In our computational study we used the default settings of Gurobi. All computations have been executed on the high performance cluster "Elwetritsch" at the TU Kaiserslautern, which is part of the "Alliance of High Performance Computing Rheinland-Pfalz" (AHRP). We use a single Intel XEON SP 6126 core with 2.6 GHz and a maximum of 30 GB RAM.

5.2. **Test Instances.** For our numerical tests we considered problems of the form

$$\min_{x\in\mathbb{Z}^{n_x},y} \quad c_x^\top x + c_y^\top y \quad \text{s.t.} \quad Ax + By \geq a,\ y \in S(x),$$

where $S(x)$ is the set of optimal solutions of the x-parameterized lower-level problem

$$\min_{y\in\mathbb{R}^{n_y}} \quad \frac{1}{2} y^\top Q y + d^\top y \quad \text{s.t.} \quad Cx + Dy \geq b.$$

Therefore, we used MILP-MILP problem instances from the BOBILib, which can be found under `https://bobilib.org`; see Thürauf et al. (2024) for further details. Other instance libraries such as BASBLib (Paulavicius and Adjiman 2017) or BOLIB (Zhou et al. 2018) are not suitable for our purposes since they only include purely continuous bilevel instances. An overview of the used instance sets used in our numerical study is given in Table 1. We modified the instances by adding a nonconvex quadratic term to the lower-level objective function. Therefore, the symmetric and indefinite matrix Q is generated using the MATLAB function sprandsym. It takes the input (n_y, density, eigenvalues), where the density is computed according to the MATLAB function[4] of Kleinert and Schmidt (2021), and the eigenvalues are randomly chosen integers in $[-1000, 1000]$. Since the entries of the computed matrix are fractional, we round them to obtain integer data.

[4]The MATLAB function can be found under `https://github.com/m-schmidt-math-opt/qp-bilevel-matrix-generator`.

For our numerical tests, we excluded the instance sets clique, inter-clique, generalized, kp, inter-fire, and or, since almost none of those instances (extended with indefinite lower-lever objective matrices) can be solved within a time limit of 2 hours. Hence, our test set consists of instances of the classes qbmkp, denegre, xuwang, xularge, inter-assig, and inter-kp; see Thürauf et al. (2024) for the details. Moreover, we excluded those instances which both methods can solve within 1 second. They are labeled "Trivial" in Table 1. Instances that neither KKT nor HPR can solve within the time limit of 2 hours are labeled with "Time limit". We also did this for instances, where the square matrix Q cannot be computed within the time limit. In the latter case, we removed the instances from our test set. This affects 30 out of 31 instances in xularge that are labeled with "Time limit"; see Table 1. All non-trivial instances for which either approach terminates within the time limit are labeled "Solvable". This includes the instances for which infeasibility is detected within the time limit.

Table 1. Overview of the test instances per instance class. The first group is the subset of instance test sets that we use in our numerical study.

Instance class	Total	Solvable (in %)	Time limit (in %)	Trivial (in %)
qbmkp	200	112 (56.00%)	78 (39.00%)	10 (5.00%)
denegre	50	34 (68.00%)	2 (4.00%)	14 (28.00%)
xuwang	100	93 (93.00%)	0 (0.00%)	7 (7.00%)
xularge	60	29 (48.33%)	31 (51.67%)	0 (0.00%)
inter-assig	24	24 (100.00%)	0 (0.00%)	0 (0.00%)
inter-kp	99	54 (54.55%)	45 (45.45%)	0 (0.00%)
clique	60	6 (10.00%)	54 (90.00%)	0 (0.00%)
inter-clique	80	1 (1.25%)	79 (98.75%)	0 (0.00%)
generalized	90	1 (1.11%)	89 (98.89%)	0 (0.00%)
kp	450	3 (0.67%)	447 (99.33%)	0 (0.00%)
inter-fire	72	0 (0.00%)	72 (100.00%)	0 (0.00%)
or	810	6 (0.74%)	801 (98.89%)	3 (0.37%)
Σ	2095	363 (17.33%)	1698 (81.05%)	34 (1.62%)

5.3. **Discussion of the Results.** In Algorithm 1, we solve Problem (9) in every iteration. We compare this approach to the case in which we omit the KKT conditions of the lower level in Problem (9) and use the high-point relaxation as a basis instead, i.e., we solve the problem

$$\begin{aligned} \min_{x,y} \quad & q_u(x,y) = \frac{1}{2}x^\top H_u x + c_u^\top x + \frac{1}{2}y^\top G_u y + d_u^\top y \\ \text{s.t.} \quad & Ax + By \geq a, \quad Cx_I + Dy \geq b, \\ & x_i \in \mathbb{Z} \cap [x_i^-, x_i^+] \quad \text{for all } i \in I, \\ & \text{INGC (8) on } \tilde{x}_I \quad \text{for all } \tilde{x}_I \in X, \\ & q_u(x,y) \leq U. \end{aligned} \tag{11}$$

In the former approach (called KKT in what follows), Problem (9) is a nonconvex mixed-integer program and its feasible region only consists of points (x, y, λ) such that (y, λ) is a KKT point of the x-parameterized lower level. In the latter approach (denoted by HPR in the following), Problem (11) is a mixed-integer convex-quadratic program with a feasible region that is a subset of the bilevel constraint region P (see

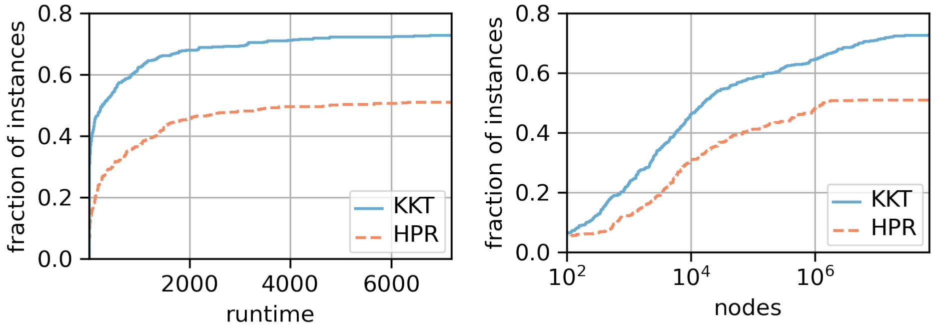

FIGURE 2. ECDFs of runtimes (left) and node counts (right).

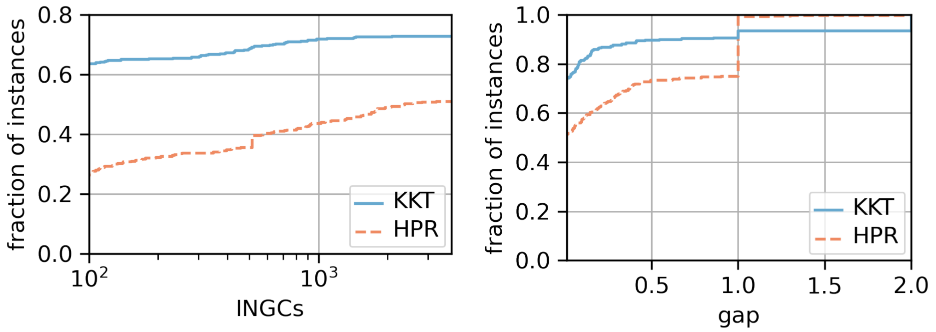

FIGURE 3. ECDFs of the number of INGCs (left) and gaps (right).

Assumption 2), i.e., it may contain points (x, y, λ) such that (y, λ) is not a KKT point of the x-parameterized lower level.

The empirical cumulative distribution functions (ECDFs) w.r.t. runtimes, node counts, number of INGCs, and (relative) gaps at termination of KKT and HPR are given in Figure 2 and 3. As illustrated in Figure 2, KKT clearly outperforms HPR w.r.t. node counts (see Figure 2) and the number of integer no-good cuts (see Figure 3). This can be expected since the feasible region of Problem (9) is, in general, much smaller than the one of Problem (11) used in HPR. However, the method KKT also outperforms method HPR w.r.t. the runtimes. Therefore, one can conclude that the benefit of having less iterations overall compensates for the issue that Problem (9) is harder to solve than Problem (11) due to the present nonconvexities.

In most of the instances that both methods cannot solve, the gaps are similarly small; see Figure 3 (right). In around 10 % of the instances, KKT does neither find an upper bound nor a lower bound, hence, no gap could be computed. However, for the same instances HPR finds both an upper and a lower bound. This is because in those instances, Problem (5) is already too hard to be solved in the time limit of 2 hours. In around 7 % of the instances, the method HPR finds a lower but no upper bound. In additional 17 % of instances, the method finds a lower bound of zero and a finite upper bound. In both cases, i.e., in approximately 24 % of the instances, we get a gap of 1. This is also reflected by the "jump" in Figure 3 (right).

We showcase the efficiency of the methods KKT and HPR compared to a full enumeration method in Figure 4. On the x-axis, we have $(\#\text{INGC})/2^{n_b}$, where #INGC is the number of integer no-good cuts used by either KKT or HPR and n_b is the number of binary variables needed to represent all integer linking variables of the instance. Hence, a full enumeration requires 2^{n_b} many iterations. The plot

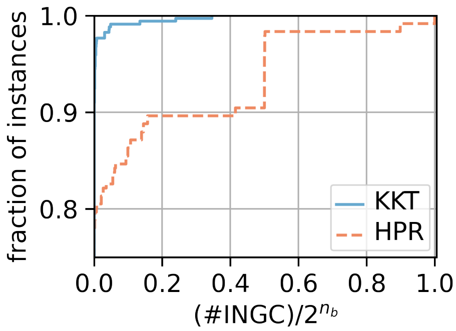

FIGURE 4. ECDF of $(\#\text{INGC})/2^{n_b}$ for those instances solved by the respective method KKT and HPR

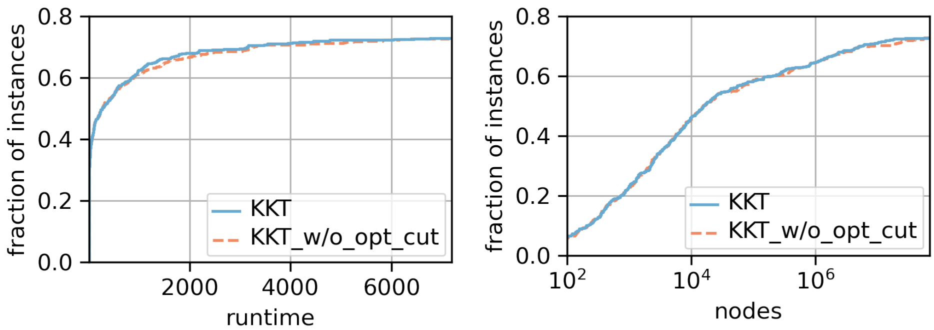

FIGURE 5. ECDFs of runtimes (left) and node counts (right).

shows that for roughly 90 % of all instances solved by the HPR approach, it requires less than 20 % of the iterations that are required by a full enumeration. Finally, for the KKT approach, the effect is even more pronounced. For roughly 99 % of all instances solved by the KKT approach, it requires less then 5 % of the iterations that are required by a full enumeration. This shows that the novel approach presented in this paper clearly outperforms a full enumeration.

5.3.1. *Optimality Cuts.* The optimality cut prevents a solution (x, y, λ) to Problem (9) to have a worse upper-level objective function value than the current incumbent U. However, the bilevel-feasible point computed in Line 11 of Algorithm 1 may violate this constraint. Hence, to test the efficiency of the optimality cut, we tested the KKT approach without using it in Problem (9). As mentioned in Example 15, one then has to be careful w.r.t. Lines 6 and 7 of Algorithm 1. In Line 6 we have to add the condition that $L \leq U$ has to hold, to avoid incorrect termination of the algorithm in some special cases; see Appendix B.

As shown in Figure 5, the optimality cut does not lead to a significant improvement w.r.t. runtimes and node counts. This is because the optimality cut only separates points that lead to a large upper-level objective function value, i.e., its absence does not change the overall solution to Problem (9), if any exists. However, we decided to keep it, since it also has no negative effect on the overall performance of the algorithm and, for some instances, it may lead to small improvements w.r.t. the runtimes.

6. Conclusion

We presented an iterative method to solve MIQP-QP bilevel problems with integer linking variables and a nonconvex and continuous QP in the lower level. The method is based on a single-level relaxation of the bilevel problem using the lower level's KKT conditions. Note that we discussed problems of Type (1), but our approach can also handle problems with more general upper levels, as long as a suitable solver for tackling the resulting single-level relaxation is at hand. Even nonlinear lower-level constraints can be considered if a reasonable constraint qualification can be guaranteed. The proposed method uses integer no-good cuts to separate points which do not improve an upper bound for the optimal upper-level objective function value. As shown in Section 5.2, the considered class of problems is highly challenging. Therefore, it would be nice to have different types of cutting planes (other than integer no-good cuts), tailored to separate bilevel-infeasible points, since this could significantly improve the performance of Algorithm 1.

Note that for other but related settings, there already exist approaches that iteratively use cutting planes to separate bilevel-infeasible points; see, e.g., Fischetti et al. (2018), Gaar et al. (2022), or Horländer et al. (2024). However, these methods all separate bilevel-infeasible extreme points of convex subproblems which are based on the high-point relaxation. For these, convexity of the subproblems is crucial and cannot be achieved when using Problem (5) as a basis, as done in Algorithm 1. Moreover, the mentioned methods are tailored for bilevel problems with integer variables in the lower level, and do not work in general for continuous lower-level variables, because they heavily rely on the usage of bilevel-free sets (see, e.g., Theorem 3 in Fischetti et al. (2018)) for their cut generation. This approach requires the existence of a suitable integer point $\hat{y}$ which may not be given in our setting. To this end, we would need to solve the following separation problem:

Input: A bilevel problem of the form (1) and a point $(\tilde{x}, \tilde{y}, \tilde{\lambda}) \in P \cap (\mathrm{KKT}(\tilde{x}) \setminus S(\tilde{x}))$.

Output: A function $h(x, y)$ such that $h(\tilde{x}, y) \geq 0$ for all $y \in S(\tilde{x})$ and $h(\tilde{x}, y) < 0$ for all $y \in \mathrm{KKT}(\tilde{x}) \setminus S(\tilde{x})$.

Here, $S(\tilde{x})$ is the rational reaction set of the follower given $\tilde{x}$ (see Definition 3) and $\mathrm{KKT}(\tilde{x})$ is the set of KKT points of the $\tilde{x}$-parameterized lower level (see the proof of Proposition 10). To the best of our knowledge, there are no approaches in the literature to solve this separation problem. Hence, any advancement in this respect would lead both to a better understanding on how to improve existing methods, and to new approaches tailored for this class of bilevel problems.

Finally, let us also mention that another important question for future research is to study the convergence of methods such as the one presented in this paper if the lower-level problem is only solved approximately. However, this might be rather tricky, see, e.g., Beck et al. (2023b), and is thus beyond the scope of this paper.

Acknowledgements

All authors are indebted to the evaluation team (reviewer and editor) for their constructive comments, which helped to improve the paper. The second author thanks the DFG for their support within RTG 2126 "Algorithmic Optimization". The computations were executed on the high performance cluster "Elwetritsch" at the TU Kaiserslautern, which is part of the "Alliance of High Performance Computing Rheinland-Pfalz" (AHRP). We kindly acknowledge the support of RHRK. Finally, we thank Henri Lefebvre for carefully proof-reading some parts of the paper.

References

Androulakis, I. P., C. D. Maranas, and C. A. Floudas (1995). "αBB: A global optimization method for general constrained nonconvex problems." In: *Journal of Global Optimization* 7.4, pp. 337–363. DOI: 10.1007/BF01099647.

Beck, Y., I. Ljubić, and M. Schmidt (2023a). "A survey on bilevel optimization under uncertainty." In: *European Journal of Operational Research* 311.2, pp. 401–426. DOI: 10.1016/j.ejor.2023.01.008.

Beck, Y. and M. Schmidt (2023). *A Gentle and Incomplete Introduction to Bilevel Optimization*. Lecture notes. Trier University. URL: https://optimization-online.org/2021/06/8450/.

Beck, Y., M. Schmidt, J. Thürauf, and D. Bienstock (2023b). "On a Computationally Ill-Behaved Bilevel Problem with a Continuous and Nonconvex Lower Level." In: *Journal of Optimization Theory and Applications* 198, 428—447. DOI: 10.1007/s10957-023-02238-9.

Blankenship, J. W. and J. E. Falk (1976). "Infinitely constrained optimization problems." In: *Journal of Optimization Theory and Applications* 19.2, pp. 261–281. DOI: 10.1007/BF00934096.

Bomze, I. M. (1997). "Evolution towards the Maximum Clique." In: *Journal of Global Optimization* 10.2, pp. 143–164. DOI: 10.1023/A:1008230200610.

Caprara, A., M. Carvalho, A. Lodi, and G. J. Woeginger (2014). "A study on the computational complexity of the bilevel knapsack problem." In: *SIAM Journal on Optimization* 24.2, pp. 823–838. DOI: 10.1137/130906593.

Dempe, S. (2002). *Foundations of Bilevel Programming*. Springer. DOI: 10.1007/b101970.

Dempe, S., V. Kalashnikov, G. A. Pérez-Valdés, and N. Kalashnykova (2015). *Bilevel Programming Problems*. Springer. DOI: 10.1007/978-3-662-45827-3.

Dempe, S. and K. Richter (2000). *Bilevel programming with knapsack constraints*. Citeseer. URL: https://www.researchgate.net/publication/2376759_Bilevel_Programming_With_Knapsack_Constraints.

DeNegre, S. T. (2011). *Interdiction and discrete bilevel linear programming*. PhD thesis. Lehigh University. URL: http://coral.ie.lehigh.edu/~ted/files/papers/ScottDeNegreDissertation11.pdf.

DeNegre, S. T. and T. K. Ralphs (2009). "A branch-and-cut algorithm for integer bilevel linear programs." In: *Operations research and cyber-infrastructure*. Springer, pp. 65–78. DOI: 10.1007/978-0-387-88843-9_4.

Djelassi, H., M. Glass, and A. Mitsos (2019). "Discretization-based algorithms for generalized semi-infinite and bilevel programs with coupling equality constraints." In: *Journal of Global Optimization* 75.2, pp. 341–392. DOI: 10.1007/s10898-019-00764-3.

Fischetti, M., I. Ljubić, M. Monaci, and M. Sinnl (2018). "On the use of intersection cuts for bilevel optimization." In: *Mathematical Programming* 172.1, pp. 77–103. DOI: 10.1007/s10107-017-1189-5.

Gaar, E., J. Lee, I. Ljubić, M. Sinnl, and K. Tanınmış (2022). "SOCP-based disjunctive cuts for a class of integer nonlinear bilevel programs." In: *International Conference on Integer Programming and Combinatorial Optimization*. Springer, pp. 262–276. DOI: 10.1007/978-3-031-06901-7_20.

Grimm, V., T. Kleinert, F. Liers, M. Schmidt, and G. Zöttl (2019). "Optimal price zones of electricity markets: a mixed-integer multilevel model and global solution approaches." In: *Optimization Methods and Software* 34.2, pp. 406–436. DOI: 10.1080/10556788.2017.1401069.

Hansen, P., B. Jaumard, and G. Savard (1992). “New branch-and-bound rules for linear bilevel programming.” In: *SIAM Journal on Scientific and Statistical Computing* 13.5, pp. 1194–1217. DOI: 10.1137/0913069.

Horländer, A., I. Ljubić, and M. Schmidt (2024). *Using Disjunctive Cuts in a Branch-and-Cut Method to Solve Convex Integer Nonlinear Bilevel Problems.* URL: https://optimization-online.org/?p=25955.

Hungerford, J. T. and F. Rinaldi (2019). “A general regularized continuous formulation for the maximum clique problem.” In: *Mathematics of Operations Research* 44.4, pp. 1161–1173.

Jeroslow, R. G. (1985). “The polynomial hierarchy and a simple model for competitive analysis.” In: *Mathematical Programming* 32.2, pp. 146–164. DOI: 10.1007/BF01586088.

Kleinert, T., V. Grimm, and M. Schmidt (2021a). “Outer Approximation for Global Optimization of Mixed-Integer Quadratic Bilevel Problems.” In: *Mathematical Programming (Series B)*. DOI: 10.1007/s10107-020-01601-2.

Kleinert, T., M. Labbé, I. Ljubić, and M. Schmidt (2021b). “A Survey on Mixed-Integer Programming Techniques in Bilevel Optimization.” In: *EURO Journal on Computational Optimization* 9, p. 100007. DOI: 10.1016/j.ejco.2021.100007.

Kleinert, T. and M. Schmidt (2021). “Computing Feasible Points of Bilevel Problems with a Penalty Alternating Direction Method.” In: *INFORMS Journal on Computing* 33.1, pp. 198–215. DOI: 10.1287/ijoc.2019.0945.

– (2023). “Why there is no need to use a big-M in linear bilevel optimization: A computational study of two ready-to-use approaches.” In: *Computational Management Science*. DOI: 10.1007/s10287-023-00435-5.

Kleniati, P. M. and C. S. Adjiman (2011). “Branch-and-Sandwich: An Algorithm for Optimistic Bi-Level Programming Problems.” In: *21st European Symposium on Computer Aided Process Engineering*. Ed. by E. Pistikopoulos, M. Georgiadis, and A. Kokossis. Vol. 29. Computer Aided Chemical Engineering. Elsevier, pp. 602–606. DOI: 10.1016/B978-0-444-53711-9.50121-8.

Kleniati, P.-M. and C. S. Adjiman (2014a). “Branch-and-Sandwich: a deterministic global optimization algorithm for optimistic bilevel programming problems. Part II: Convergence analysis and numerical results.” In: *Journal of Global Optimization* 60.3, pp. 459–481. DOI: 10.1007/s10898-013-0120-8.

Kleniati, P.-M. and C. S. Adjiman (2015). “A generalization of the Branch-and-Sandwich algorithm: From continuous to mixed-integer nonlinear bilevel problems.” In: *Computers & Chemical Engineering* 72, pp. 373–386. DOI: 10.1016/j.compchemeng.2014.06.004.

Kleniati, P.-M. and C. S. Adjiman (2014b). “Branch-and-Sandwich: a deterministic global optimization algorithm for optimistic bilevel programming problems. Part I: Theoretical development.” In: *Journal of Global Optimization* 60.3, pp. 425–458. DOI: 10.1007/s10898-013-0121-7.

Labbé, M. and A. Violin (2013). “Bilevel programming and price setting problems.” In: *4OR* 11.1, pp. 1–30. DOI: 10.1007/s10288-012-0213-0.

Lozano, L. and J. C. Smith (2017). “A value-function-based exact approach for the bilevel mixed-integer programming problem.” In: *Operations Research* 65.3, pp. 768–786. DOI: 10.1287/opre.2017.1589.

Mitsos, A. (2010). “Global solution of nonlinear mixed-integer bilevel programs.” In: *Journal of Global Optimization* 47.4, pp. 557–582. DOI: 10.1007/s10898-009-9479-y.

Mitsos, A. and P. I. Barton (2006). *Issues in the development of global optimization algorithms for bilevel programs with a nonconvex inner program.* Tech. rep. URL:

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.703.4195&rep=rep1&type=pdf.

Mitsos, A., P. Lemonidis, and P. I. Barton (2008). "Global solution of bilevel programs with a nonconvex inner program." In: *Journal of Global Optimization* 42.4, pp. 475–513. DOI: 10.1007/s10898-007-9260-z.

Motzkin, T. S. and E. G. Straus (1965). "Maxima for Graphs and a New Proof of a Theorem of Turán." In: *Canadian Journal of Mathematics* 17, 533—540. DOI: 10.4153/CJM-1965-053-6.

Paulavicius, R. and C. S. Adjiman (2017). *BASBLib - a library of bilevel test problems, v2.2 [Data set].* DOI: 10.5281/zenodo.897966.

Schmidt, M. and J. Thürauf (2024). "An Exact Method for Nonlinear Network Flow Interdiction Problems." In: *SIAM Journal on Optimization* 34.4, pp. 3623–3652. DOI: 10.1137/22M152983X.

Thürauf, J., T. Kleinert, I. Ljubić, T. Ralphs, and M. Schmidt (2024). *BOBILib: Bilevel Optimization (Benchmark) Instance Library.* Tech. rep. URL: https://optimization-online.org/?p=27063.

Wang, Z., Y. Yin, and B. An (2016). "Computing Optimal Monitoring Strategy for Detecting Terrorist Plots." In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*. Ed. by D. Schuurmans and M. P. Wellman. AAAI Press, pp. 637–643. DOI: 10.1609/aaai.v30i1.10028.

Wolsey, L. A. (2020). *Integer Programming.* John Wiley & Sons, Ltd. DOI: 10.1002/9781119606475.

Zhou, S., A. B. Zemkoho, and A. Tin (2018). *BOLIB: Bilevel Optimization LIBrary of test problems.* URL: https://arxiv.org/abs/1812.00230.

Appendix A. Proofs

For some of the following proofs, we will make use of the following lemma.

Lemma 18. *Let $\alpha \in [0,1]$ and $y \in [0, \frac{1}{2}]$. Then, it holds*

$$\alpha y - \alpha^2 y^2 \le y - y^2.$$

Proof. It holds

$$\begin{aligned} y \le \frac{1}{2} \le \frac{1}{1+\alpha} &\implies y(1-\alpha^2) \le 1-\alpha \\ \implies \alpha - \alpha^2 y \le 1 - y &\implies \alpha y - \alpha^2 y^2 \le y - y^2 . \end{aligned}$$ ☐

We now provide the proofs of the theoretical results of Section 3.

Proof of Lemma 4. Assume that there exists an index $k \neq j$ such that $\hat{y}_k \in (0,1)$. It holds

$$\hat{y}_j - \hat{y}_j^2 + \hat{y}_k - \hat{y}_k^2 > 0$$

and, since $\hat{y}$ is globally optimal for the lower level, $d_j > 0$ and $d_k > 0$ holds. We define

$$\begin{aligned} \zeta &:= \min\left\{\hat{y}_j + \frac{d_k}{d_j}\hat{y}_k, 1\right\}, \ \xi := \max\left\{0, \hat{y}_k + \frac{d_j}{d_k}\hat{y}_j - \frac{d_j}{d_k}\right\}, \\ \rho &:= \min\left\{\hat{y}_k + \frac{d_j}{d_k}\hat{y}_j, 1\right\}, \ \theta := \max\left\{0, \hat{y}_j + \frac{d_k}{d_j}\hat{y}_k - \frac{d_k}{d_j}\right\}. \end{aligned}$$

Note that ζ and ξ cannot both be fractional simultaneously. Also the simultaneous case

$$\hat{y}_j + \frac{d_k}{d_j}\hat{y}_k > 1 \ \text{ and } \ \hat{y}_k + \frac{d_j}{d_k}\hat{y}_j < \frac{d_j}{d_k},$$

which would imply $(\zeta,\xi) = (1,0)$, is impossible. The same holds for θ and ρ. Therefore, it holds

$$d_j\zeta + d_k\xi = d_j\theta + d_k\rho = d_j\hat{y}_j + d_k\hat{y}_k,$$

i.e., we can replace $\hat{y}_j$ and $\hat{y}_k$ either with ζ and ξ or with θ and ρ and obtain the same values in the linear part of the lower-level objective function value. We now show that either

$$\zeta - \zeta^2 + \xi - \xi^2 < \hat{y}_j - \hat{y}_j^2 + \hat{y}_k - \hat{y}_k^2$$

or

$$\theta - \theta^2 + \rho - \rho^2 < \hat{y}_j - \hat{y}_j^2 + \hat{y}_k - \hat{y}_k^2$$

holds. This means that we can always choose either ζ and ξ or θ and ρ so that the penalization of the fractionality in the lower-level objective function gets smaller than the one for $\hat{y}_j$ and $\hat{y}_k$. To this end, we distinguish the four following cases.

Case 1: $\hat{y}_j \geq \frac{1}{2}$ and $\hat{y}_k \leq \frac{1}{2}$: It holds $\frac{1}{2} \leq \hat{y}_j < \zeta$ and $\xi < \hat{y}_k \leq \frac{1}{2}$ and, hence,

$$\zeta - \zeta^2 + \xi - \xi^2 < \hat{y}_j - \hat{y}_j^2 + \hat{y}_k - \hat{y}_k^2.$$

Case 2: $\hat{y}_j \leq \frac{1}{2}$ and $\hat{y}_k \geq \frac{1}{2}$: We have $\theta < \hat{y}_j \leq \frac{1}{2}$ and $\frac{1}{2} \leq \hat{y}_k < \rho$ and, hence,

$$\theta - \theta^2 + \rho - \rho^2 < \hat{y}_j - \hat{y}_j^2 + \hat{y}_k - \hat{y}_k^2.$$

Case 3: $\hat{y}_j \leq \frac{1}{2}$ and $\hat{y}_k \leq \frac{1}{2}$: In case $d_k \leq d_j$, it holds $\zeta = \hat{y}_j + \frac{d_k}{d_j}\hat{y}_k$, $\xi = 0$, and, hence

$$\begin{aligned}
\zeta - \zeta^2 + \xi - \xi^2 &= \zeta - \zeta^2 \\
&= \hat{y}_j + \frac{d_k}{d_j}\hat{y}_k - \left(\hat{y}_j + \frac{d_k}{d_j}\hat{y}_k\right)^2 \\
&= \hat{y}_j - \hat{y}_j^2 + \frac{d_k}{d_j}\hat{y}_k - \frac{d_k^2}{d_j^2}\hat{y}_k^2 - 2\hat{y}_j\frac{d_k}{d_j}\hat{y}_k \\
&= \hat{y}_j - \hat{y}_j^2 + (1 - 2\hat{y}_j)\frac{d_k}{d_j}\hat{y}_k - \frac{d_k^2}{d_j^2}\hat{y}_k^2 \\
&< \hat{y}_j - \hat{y}_j^2 + \frac{d_k}{d_j}\hat{y}_k - \frac{d_k^2}{d_j^2}\hat{y}_k^2 \\
&\leq \hat{y}_j - \hat{y}_j^2 + \hat{y}_k - \hat{y}_k^2
\end{aligned}$$

due to Lemma 18 with $\alpha = d_k/d_j$ and $y = \hat{y}_k$. In the opposite case $d_k > d_j$, we have $\rho = \hat{y}_k + \frac{d_j}{d_k}\hat{y}_j$, $\theta = 0$, and, hence

$$\begin{aligned}
\rho - \rho^2 + \theta - \theta^2 &= \rho - \rho^2 \\
&= \hat{y}_k + \frac{d_j}{d_k}\hat{y}_j - \left(\hat{y}_k + \frac{d_j}{d_k}\hat{y}_j\right)^2 \\
&= \hat{y}_k - \hat{y}_k^2 + \frac{d_j}{d_k}\hat{y}_j - \frac{d_j^2}{d_k^2}\hat{y}_j^2 - 2\hat{y}_k\frac{d_j}{d_k}\hat{y}_j \\
&= \hat{y}_k - \hat{y}_k^2 + (1 - 2\hat{y}_k)\frac{d_j}{d_k}\hat{y}_j - \frac{d_j^2}{d_k^2}\hat{y}_j^2 \\
&< \hat{y}_k - \hat{y}_k^2 + \frac{d_j}{d_k}\hat{y}_j - \frac{d_j^2}{d_k^2}\hat{y}_j^2 \\
&\leq \hat{y}_k - \hat{y}_k^2 + \hat{y}_j - \hat{y}_j^2,
\end{aligned}$$

also due to Lemma 18 with $\alpha = d_j/d_k$ and $y = \hat{y}_j$.

Case 4: $\hat{y}_j \geq \frac{1}{2}$ and $\hat{y}_k \geq \frac{1}{2}$: In case $d_k \leq d_j$, it holds $\rho = 1$, $\theta = \hat{y}_j + \frac{d_k}{d_j}\hat{y}_k - \frac{d_k}{d_j}$, leading to

$$\begin{aligned}
\rho - \rho^2 + \theta - \theta^2 &= \theta - \theta^2 \\
&= \hat{y}_j + \frac{d_k}{d_j}\hat{y}_k - \frac{d_k}{d_j} - \left(\hat{y}_j + \frac{d_k}{d_j}\hat{y}_k - \frac{d_k}{d_j}\right)^2 \\
&= \hat{y}_j - \hat{y}_j^2 + \frac{d_k}{d_j}\hat{y}_k - \frac{d_k}{d_j} - \frac{d_k^2}{d_j^2}\hat{y}_k^2 - \frac{d_k^2}{d_j^2} - 2\hat{y}_j\frac{d_k}{d_j}\hat{y}_k + 2\frac{d_k}{d_j}\hat{y}_j + 2\frac{d_k^2}{d_j^2}\hat{y}_k \\
&= \hat{y}_j - \hat{y}_j^2 + (1 - 2\hat{y}_j)\frac{d_k}{d_j}\hat{y}_k - \left(\frac{d_k}{d_j} - 2\frac{d_k}{d_j}\hat{y}_j\right) + \frac{d_k^2}{d_j^2}\left(2\hat{y}_k - 1 - \hat{y}_k^2\right) \\
&= \hat{y}_j - \hat{y}_j^2 + (2\hat{y}_j - 1)\frac{d_k}{d_j}(1 - \hat{y}_k) - \frac{d_k^2}{d_j^2}(1 - \hat{y}_k)^2 \\
&< \hat{y}_j - \hat{y}_j^2 + \frac{d_k}{d_j}(1 - \hat{y}_k) - \frac{d_k^2}{d_j^2}(1 - \hat{y}_k)^2 \\
&\leq \hat{y}_j - \hat{y}_j^2 + (1 - \hat{y}_k) - (1 - \hat{y}_k)^2 \\
&= \hat{y}_j - \hat{y}_j^2 + \hat{y}_k - \hat{y}_k^2,
\end{aligned}$$

because of $1 - \hat{y}_j \leq 1/2$ and Lemma 18 with $\alpha = d_k/d_j$ and $y = 1 - \hat{y}_k$. In the opposite case $d_k > d_j$, we have $\zeta = 1$, $\xi = \hat{y}_k + \frac{d_j}{d_k}\hat{y}_j - \frac{d_j}{d_k}$, which implies

$$\begin{aligned}
\zeta - \zeta^2 + \xi - \xi^2 &= \xi - \xi^2 \\
&= \hat{y}_k + \frac{d_j}{d_k}\hat{y}_j - \frac{d_j}{d_k} - \left(\hat{y}_k + \frac{d_j}{d_k}\hat{y}_j - \frac{d_j}{d_k}\right)^2 \\
&= \hat{y}_k - \hat{y}_k^2 + \frac{d_j}{d_k}\hat{y}_j - \frac{d_j}{d_k} - \frac{d_j^2}{d_k^2}\hat{y}_j^2 - \frac{d_j^2}{d_k^2} - 2\hat{y}_k\frac{d_j}{d_k}\hat{y}_j + 2\frac{d_j}{d_k}\hat{y}_k + 2\frac{d_j^2}{d_k^2}\hat{y}_j \\
&= \hat{y}_k - \hat{y}_k^2 + (1 - 2\hat{y}_k)\frac{d_j}{d_k}\hat{y}_j - \left(\frac{d_j}{d_k} - 2\frac{d_j}{d_k}\hat{y}_k\right) + \frac{d_j^2}{d_k^2}\left(2\hat{y}_j - 1 - \hat{y}_j^2\right) \\
&= \hat{y}_k - \hat{y}_k^2 + (2\hat{y}_k - 1)\frac{d_j}{d_k}(1 - \hat{y}_j) - \frac{d_j^2}{d_k^2}(1 - \hat{y}_j)^2 \\
&< \hat{y}_k - \hat{y}_k^2 + \frac{d_j}{d_k}(1 - \hat{y}_j) - \frac{d_j^2}{d_k^2}(1 - \hat{y}_j)^2 \\
&\leq \hat{y}_k - \hat{y}_k^2 + (1 - \hat{y}_j) - (1 - \hat{y}_j)^2 \\
&= \hat{y}_k - \hat{y}_k^2 + \hat{y}_j - \hat{y}_j^2,
\end{aligned}$$

because of $1 - \hat{y}_j \leq 1/2$ and Lemma 18 with $\alpha = d_j/d_k$ and $y = 1 - \hat{y}_j$.

Thus, we can replace $\hat{y}_j$ and $\hat{y}_k$ either with ζ and ξ or with θ and ρ to obtain a better lower-level objective function value. This holds for all $M > 0$. Therefore, $\hat{y}$ is not optimal for the lower level, which is a contradiction to our assumption. This proves that $\hat{y}_j$ and $\hat{y}_k$ cannot be both fractional at the same time. □

Proof of Lemma 5. If $\|d\|_\infty = 1$, then $d_i \in \{0, 1\}$ for $i \in [n_y]$. Let $J := \{i \in [n_y] : d_i = 1\}$. If $|J| \leq x$, then $d^\top y \leq x$ for all $y \in [0,1]^{n_y}$. Consequently, every point $\hat{y}$ with $\hat{y}_i = 1$ for $i \in J_x$ and $\hat{y}_i \in \{0, 1\}$ for $i \notin J_x$ is lower-level optimal. Now assume that at least x-many entries of d are one, i.e., $|J| \geq x$. Let J_x be any subset of J with $|J_x| = x$. Set $\hat{y}_i = 1$ for $i \in J_x$ and $\hat{y}_i = 0$ for $i \notin J_x$. Then, the lower-level objective function value is given by

$$d^\top \hat{y} - M\left(\sum_{i=1}^{n_y} \hat{y}_i - \hat{y}_i^2\right) = x.$$

Hence, $\hat{y}$ is globally optimal for the lower level. Furthermore, every point $\bar{y}$ with $d^\top \bar{y} = x$ and $\bar{y}_j \in (0,1)$ for one j yields a lower-level objective function value of

$$d^\top \bar{y} - M\left(\sum_{i=1}^{n_y} \bar{y}_i - \bar{y}_i^2\right) < x,$$

i.e., it is not globally optimal. □

Proof of Lemma 6. For the first part assume that $d^\top \hat{y} < x$. If $\hat{y}$ is globally optimal for the lower level, we know that $d_j > 0$ because if $d_j = 0$, then $\hat{y}_j \in \{0,1\}$ would lead to a better lower-level objective function value. Furthermore, we have

$$\sum_{i=1, i\neq j}^{n_y} d_i \hat{y}_i \leq d^\top \hat{y} - M(\hat{y}_j - \hat{y}_j^2),$$

i.e., the lower-level objective function value is at least as good as in the case in which we set $\hat{y}_j$ to zero. It follows that

$$d_j \hat{y}_j \geq M(\hat{y}_j - \hat{y}_j^2) \quad \text{and} \quad d_j \geq M(1 - \hat{y}_j). \tag{12}$$

Since $d^\top \hat{y} < x$ and $\hat{y}_j \in (0,1)$, we can find an $\varepsilon > 0$ such that $\hat{y}_j + \varepsilon < 1$ and

$$d^\top \hat{y} < d^\top \hat{y} + d_j \varepsilon \leq x.$$

With (12) we get that

$$\begin{aligned}
& M(\hat{y}_j + \varepsilon) > 0 \\
\iff\; & M(1 - \hat{y}_j) - M(1 - 2\hat{y}_j - \varepsilon) > 0 \\
\implies\; & d_j - M(1 - 2\hat{y}_j - \varepsilon) > 0 \\
\iff\; & d_j \varepsilon - M(\varepsilon - 2\varepsilon \hat{y}_j - \varepsilon^2) > 0 \\
\iff\; & d_j(\hat{y}_j + \varepsilon) - M(\hat{y}_j + \varepsilon - (\hat{y}_j + \varepsilon)^2) > d_j \hat{y}_j - M(\hat{y}_j - \hat{y}_j^2).
\end{aligned}$$

From the last inequality it follows that, if we replace $\hat{y}_j$ with $\hat{y}_j + \varepsilon$, we get a better lower-level objective function value, i.e., $\hat{y}$ is not globally optimal. This proves $d^\top \hat{y} = x$.

For the second part assume that there exists an index j such that $\hat{y}_j \in (0,1)$ and $\hat{y}_j < 1/\|d\|_\infty$. It holds $d_j > 0$ because if $d_j = 0$, then $\hat{y}_j \in \{0,1\}$ would lead to a better lower-level objective function value. With that we have

$$0 < d_j \hat{y}_j < \frac{d_j}{\|d\|_\infty} \leq 1.$$

As shown before, it holds $d^\top \hat{y} = x$, leading to

$$x - 1 < \sum_{i=1, i\neq j}^{n_y} d_i \hat{y}_i < x.$$

From $x \in \mathbb{Z}$ it follows that $\sum_{i=1,i\neq j}^{n_y} d_i \hat{y}_i \notin \mathbb{Z}$ and, hence, there exists at least one index $k \neq j$ with $\hat{y}_k \in (0,1)$ and $d_k > 0$. With Lemma 4 we get that $\hat{y}$ cannot be optimal which is a contradiction to our assumption. Thus, it has to hold

$$\frac{1}{\|d\|_\infty} \leq \hat{y}_j.$$

Assume now that there exists an index j such that $\hat{y}_j \in (0,1)$ and $\hat{y}_j > 1 - 1/\|d\|_\infty$. Again, we have $d_j > 0$ because if $d_j = 0$, then $\hat{y}_j \in \{0,1\}$ would lead to a better lower-level objective function value. Furthermore, it holds

$$d_j - 1 \leq d_j \left(1 - \frac{1}{\|d\|_\infty}\right) < d_j \hat{y}_j < d_j.$$

We showed that $d^\top \hat{y} = x$. Consequently, we have

$$x - d_j < \sum_{i=1, i\neq j}^{n_y} d_i \hat{y}_i < x - d_j + 1,$$

i.e., $\sum_{i=1,i\neq j}^{n_y} d_i \hat{y}_i \notin \mathbb{Z}$. Again, there exists at least one index $k \neq j$ with $\hat{y}_k \in (0,1)$ and $d_k > 0$. Using Lemma 4 we see that $\hat{y}$ cannot be optimal and, thus,

$$\hat{y}_j \leq 1 - \frac{1}{\|d\|_\infty}$$

holds. □

Proof of Lemma 7. From Lemma 4 we know that $\hat{y}_j$ is the only fractional component in $\hat{y}$. Assume that $\hat{y}_j \leq 1/2$. Then, with Lemma 6 we have

$$\frac{1}{\|d\|_\infty} - \frac{1}{\|d\|_\infty^2} \leq \hat{y}_j - \hat{y}_j^2 = \sum_{i=1}^{n_y} \hat{y}_i - \hat{y}_i^2.$$

Now assume that $\hat{y}_j \geq 1/2$. Then, it holds

$$\frac{1}{\|d\|_\infty} - \frac{1}{\|d\|_\infty^2} = 1 - \frac{1}{\|d\|_\infty} - \left(1 - \frac{1}{\|d\|_\infty}\right)^2 \leq \hat{y}_j - \hat{y}_j^2 = \sum_{i=1}^{n_y} \hat{y}_i - \hat{y}_i^2. \qquad \square$$

Proof of Theorem 8. For $\|d\|_\infty = 1$, Lemma 5 ensures $\mathcal{B}_1 = \mathcal{B}_2$.

Now assume that $\|d\|_\infty \geq 2$. First, we show that $\mathcal{B}_1 \subseteq \mathcal{B}_2$. Let $(\hat{x}, \hat{y})$ be a bilevel-feasible point for Problem (3). Assume that there exists an index j such that $\hat{y}_j \in (0,1)$. From Lemma 4 we know that $\hat{y}_i \in \{0,1\}$ for all $i \neq j$ and with Lemma 7, we get

$$\sum_{i=1}^{n_y} \hat{y}_i - \hat{y}_i^2 \geq \frac{1}{\|d\|_\infty} - \frac{1}{\|d\|_\infty^2}.$$

For $M > n_y \|d\|_\infty^3 / (\|d\|_\infty - 1)$, we have

$$\begin{aligned} d^\top \hat{y} - M \left(\sum_{i=1}^{n_y} \hat{y}_i - \hat{y}_i^2\right) &\leq d^\top \hat{y} - M \left(\frac{1}{\|d\|_\infty} - \frac{1}{\|d\|_\infty^2}\right) \\ &< d^\top \hat{y} - n_y \|d\|_\infty \\ &\leq 0 = \min\left\{d^\top y : y \in [0,1]^{n_y}\right\}. \end{aligned} \tag{13}$$

Any $\tilde{y} \in \{0,1\}^{n_y}$ satisfies $M\left(\sum_{i=1}^{n_y} \bar{y}_i - \bar{y}_i^2\right) = 0$ and, hence, leads to a better lower-level objective function value than $\hat{y}$. Consequently, $\hat{y}$ is not bilevel-feasible. With that, every point being bilevel-feasible for Problem (3) maximizes $d^\top y$ subject to $d^\top y \leq x$ and $y \in \{0,1\}^{n_y}$, i.e., it is bilevel-feasible for Problem (4).

Now we show that $\mathcal{B}_2 \subseteq \mathcal{B}_1$. Let $(\hat{x}, \hat{y})$ be bilevel-feasible for Problem (4). With (13) we get that $\hat{y}$ has to be optimal for the $\hat{x}$-parameterized lower level of Problem (3). Thus, $(\hat{x}, \hat{y})$ is bilevel-feasible for Problem (3). □

Appendix B. Adapted Method without Optimality Cut

Algorithm 2: Iterative method to solve the bilevel problem (1)

Input: Problem (1), $L = -\infty$, $U = \infty$, $\gamma = \infty$, $k = 0$, and $X_0 = \emptyset$.
Output: A globally optimal solution to the bilevel problem (1) or an indication of infeasibility.

1 **if** $\gamma > 0$ **then**
2   Solve Problem (9) with $X = X_k$ to global optimality.
3   **if** *Problem (9) is feasible* **then**
4   Let (x^k, y^k, λ^k) be the solution to Problem (9). Set $L \leftarrow q_u(x^k, y^k)$.
5   Solve the x^k-parameterized lower-level problem to compute an optimal follower‘s response $\bar{y}^k$ as well as $\phi(x^k)$.
6   **if** (x^k, y^k) *is bilevel-feasible, i.e.,* $q_\ell(y^k) = q_\ell(\bar{y}^k)$ **then**
7   **if** $L \leq U$ **then**
8   Set $(x^*, y^*) \leftarrow (x^k, y^k)$.
9   Go to Line 24.
10   **else**
11   Solve the refinement problem (6).
12   **if** *the refinement problem (6) is feasible* **then**
13   denote the obtained bilevel-feasible point $(\bar{x}, \bar{y})$. It holds $\bar{x}_I = x_I^k$.
14   **if** $q_u(\bar{x}, \bar{y}) < U$ **then**
15   Set $U \leftarrow q_u(\bar{x}, \bar{y})$ and $(x^*, y^*) \leftarrow (\bar{x}, \bar{y})$.
16   **if** $U < \infty$ **then**
17   Set $\gamma \leftarrow |U - L|/(10^{-10} + |U|)$.
18   Set $X_{k+1} \leftarrow X_k \cup \{x_I^k\}$ and $k \leftarrow k + 1$. Go to Line 2.
19   **else if** $U < \infty$ **then**
20   Go to Line 24.
21   **else**
22   **return** *“The bilevel problem (1) is infeasible.”*
23 **else**
24   **return** *the globally optimal solution* (x^*, y^*) *to the bilevel problem (1).*

(I. Bomze) University of Vienna, Faculty of Mathematics & Research Network Data Science, Oskar-Morgenstern-Platz 1, 1090 Wien, Austria
Email address: immanuel.bomze@univie.ac.at

(A. Horländer, M. Schmidt) Trier University, Department of Mathematics, Universitätsring 15, 54296 Trier, Germany
Email address: horlaender@uni-trier.de
Email address: martin.schmidt@uni-trier.de