Computational Optimization and Applications (2024) 87:835–891
https://doi.org/10.1007/s10589-023-00543-7

Enhancements of discretization approaches for non-convex mixed-integer quadratically constrained quadratic programming: Part I

Benjamin Beach[1] · Robert Burlacu[2] · Andreas Bärmann[3] · Lukas Hager[3] · Robert Hildebrand[1]

Received: 3 November 2022 / Accepted: 25 November 2023 / Published online: 30 January 2024

Abstract

We study mixed-integer programming (MIP) relaxation techniques for the solution of non-convex mixed-integer quadratically constrained quadratic programs (MIQCQPs). We present MIP relaxation methods for non-convex continuous variable products. In this paper, we consider MIP relaxations based on separable reformulation. The main focus is the introduction of the enhanced separable MIP relaxation for non-convex quadratic products of the form $z = xy$, called *hybrid separable* (HybS). Additionally, we introduce a logarithmic MIP relaxation for univariate quadratic terms, called *sawtooth relaxation*, based on Beach (Beach in J Glob Optim 84:869–912, 2022). We combine the latter with HybS and existing separable reformulations to derive MIP relaxations of MIQCQPs. We provide a comprehensive theoretical analysis of these

B. Beach and R. Hildebrand are supported by AFOSR grant FA9550-21-0107. Furthermore, R. Hildebrand was also partially supported by ONR Grant N00014-20-1-2156, L. Hager acknowledges financial support by the Bavarian Ministry of Economic Affairs, Regional Development and Energy through the Center for Analytics—Data—Applications (ADA-Center) within the framework of "BAYERN DIGITAL II".

✉ Lukas Hager
lukas.hager@fau.de

Benjamin Beach
bben6@vt.edu

Robert Burlacu
robert.burlacu@iis.fraunhofer.de

Andreas Bärmann
andreas.baermann@math.uni-erlangen.de

Robert Hildebrand
rhil@vt.edu

[1] Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA, USA

[2] Fraunhofer Institute for Integrated Circuits IIS, 90411 Nuremberg, Germany

[3] Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany

techniques, underlining the theoretical advantages of HybS compared to its predecessors. We perform a broad computational study to demonstrate the effectiveness of the enhanced MIP relaxation in terms of producing tight dual bounds for MIQCQPs. In Part II, we study MIP relaxations that extend the MIP relaxation *normalized multi-parametric disaggregation technique* (NMDT) (Castro in J Glob Optim 64:765–784, 2015) and present a computational study which also includes the MIP relaxations from this work and compares them with a state-of-the-art of MIQCQP solvers.

Keywords Quadratic programming · MIP relaxations · Discretization · Binarization · Piecewise linear approximation

1 Introduction

In this work, we study relaxations of general mixed-integer quadratically constrained quadratic programs (MIQCQPs). More precisely, we consider discretization techniques for non-convex MIQCQPs that allow for relaxations of the set of feasible solutions based on mixed-integer programming (MIP) formulations. To this end, we study a number of MIP formulations that form relaxations of the quadratic equations $z = x^2$ and $z = xy$. These MIP relaxations can then be applied to MIQCQPs by introducing auxiliary variables and constraints for each quadratic term to form a relaxation of the overall problem. In particular, we consider the strength of various MIP relaxations applied directly to a given problem, which is the simplest approach to enable the solution of MIQCQPs via an MIP solver. Our focus here is to analyze these approaches both theoretically and computationally with respect to the quality of the dual bound they deliver for MIQCQPs. Dual bounds give a lower bound for the optimal value in a minimization problem. The term comes from the so-called dual program, which can also be used to determine such bounds.

Background MIQCQPs naturally arise in the solution of many real-world optimization problems, stemming e.g. from the contexts of power supply systems [2], gas networks [19, 27], water management [23] or pooling/mixing [6, 10, 15, 30, 31]. See [25, 37] and the references therein for more examples. For the solution of such problems, there are a number of different approaches, which differ in case the problems are convex or non-convex. Within this work, we focus on the most general case, i.e. non-convex MIQCQPs, and only require finite upper and lower bounds on the variables.

In the literature, a variety of solution techniques for non-convex MIQCQPs exists. The most prominent class among them are *McCormick*-based techniques, see e.g. [12–14, 16, 35, 36]. For quadratic programs, in particular, convexification can be applied to bivariate monomials xy by introducing a new variable $z = xy$ and constructing the convex hull over the bounds on x and y. This yields the so-called *McCormick relaxation*, which is the smallest convex set containing the feasible set of the equation $z = xy$ for given finite bounds on x and y. This relaxation is known to be a polytope described by four linear inequalities (see [34]), and it is tighter the smaller the a priori known bounds on x and y are. Hence, one standard solution approach is *spatial branch-and-bound*, where the key idea is to split the domain recursively into two subregions.

For instance, one can choose the two subregions where $x \leqslant \bar{x}$ and $x \geqslant \bar{x}$, respectively, for some value $\bar{x}$. By branching on subregions, we can improve the convexification of the feasible region by adding valid inequalities to the subproblems. Thus, applying spatial branch-and-bound in conjunction with convexification (such as McCormick Relaxations) sequentially tightens the relaxation of the problem.

Alternatively, similar effects can be achieved through some kind of *binarization*. This is a general term that describes the conversion of continuous or integer variables into binary variables. By branching on these new binary variables, we also partition the space into subproblems in a way that simulates spatial branch-and-bound. The binarization of the partition makes the resulting problem a piecewise linear (p.w.l.) relaxation of the original problem with binary auxiliary variables. McCormick-based methods can differ in the way the partition and the binarization are performed. The partition can be performed purely on one variable or on both variables, equidistantly or non-equidistantly. The binarization can be done linearly or logarithmically in the number of partition elements, see [32, 40]. In a broader sense, (axial-)spatial branching for bilinear terms can also be seen as a piecewise McCormick linearization approach. Here, the partition is not performed a priori, but rather an initial partition is refined via branching on continuous variables. An overview of spatial-branching techniques can be found in [8].

Another common idea for linearizing variable products is to use *quadratic convex reformulations* as in [7, 9, 21, 22, 26]. This technique transforms the non-convex parts of the problem into univariate terms via reformulations. In [7], the authors apply *diagonal perturbation* to convexify the quadratic matrices. The resulting univariate quadratic correction terms are then linearized by introducing new variables and constraints of the form $z_i = x_i^2$, which are then approximated by p.w.l. functions. The binarization of the univariate p.w.l. functions is done logarithmically by using the so-called *sawtooth* function, introduced in [42]. An advantage of this approach is that only linearly many expressions of the form $z_i = x_i^2$ have to be linearized instead of quadratically many equations of the form $z_{ij} = x_i x_j$, with respect to the dimension of the original quadratic matrix. This approach yields a convex MIQCQP relaxation instead of the MIP relaxation obtained via direct modeling using bilinear terms. See also [1] that adapts the branch and bound approach αBB [3] to general twice differentiable objectives by providing convex reformulations via perturbations.

A further set of approaches relies on *separable reformulations* of the non-convex variable products, as done e.g. in [5]. Here, each term of the form xy is reformulated as a sum of separable univariate terms, for example using the equivalent reformulation $xy = 1/2(x^2 + y^2 - (x-y)^2) = 1/2(r + s - t)$ with $r = x^2$, $s = y^2$, and $t = (x-y)^2$ as described by [4]. The univariate constraints, here equations of the form $r = x^2$, $s = y^2$, and $t = (x-y)^2$, are then relaxed. Again, this approach can be combined with a logarithmic encoding of the univariate linear segments, as in [7, 22]. In [5], the authors analyze the following possible reformulations:

$$\begin{aligned}
&\text{Bin1}: xy = (1/2(x+y))^2 - (1/2(x-y))^2\,,\\
&\text{Bin2}: xy = 1/2\left((x+y)^2 - x^2 - y^2\right),\\
&\text{Bin3}: xy = 1/2\left(x^2 + y^2 - (x-y)^2\right).
\end{aligned}$$

They prove that MIP-based approximations of each of these univariate reformulations require fewer binary variables than a bivariate MIP-based approximation that guarantees the same maximal approximation error, if this prescribed error is small enough. However, this comes at the cost of weaker linear programming (LP) relaxations.

Alternatively, one can also obtain an MIP relaxation of xy directly via a bivariate p.w.l. relaxation, see e.g. [5, 11, 27, 40]. One way to do this is to perform a triangulation of the domain, which defines a p.w.l. approximation of the variable product. This p.w.l. approximation can then easily be converted into a relaxation of the feasible set by axis-parallel shifting, which yields a p.w.l. underestimator and overestimator. Bivariate p.w.l. approximations can also be binarized using (logarithmically-many) binary variables, see e.g. [27, 32, 40].

Contribution We compare different MIP relaxation approaches, both known ones, and a new one, in terms of the dual bound, they impose for non-convex MIQCQPs. We extend the separable approximation approaches Bin2 and Bin3 from [5] to MIP relaxations for $z = xy$. Additionally, we introduce a novel MIP relaxation for $z = xy$ called *hybrid separable* (HybS) that is based on a sophisticated combination of Bin2 and Bin3 that allows us to relax only linearly-many univariate quadratic terms (in the dimension of the quadratic matrix). In a theoretical analysis, we show that HybS has theoretical advantages, such as fewer binary variables and better LP relaxations compared to Bin2 and Bin3. We combine HybS, Bin2, and Bin3 with an MIP relaxation, called *sawtooth relaxation*, for $z = x^2$ that requires only logarithmically-many binary variables with respect to the relaxation error. Thus, we can obtain MIP relaxations for MIQCQPs. The sawtooth relaxation is an extension of the sawtooth approximation from [7], which has the strong property of *hereditary sharpness*. The hereditary sharpness of an MIP formulation means that the formulation is tight in the space of the original variables, even after branching on integer variables. We can show that the sawtooth relaxation is also hereditary sharp.

Finally, we perform an extensive numerical study where we generate MIP relaxations of non-convex MIQCQPs. Foremost, we test the different relaxation techniques in their ability to generate tight dual bounds for the original quadratic problems. We will see that HybS has a clear advantage over its predecessors Bin2 and Bin3. This effect becomes even more apparent on dense instances.

We present Part II of this work in a separate paper, where we study MIP relaxations that are distinctly different and are extensions of the *normalized multiparametric disaggregation technique* (NMDT) [13]. We provide further theoretical and computational analyses. The NMDT uses a combination of McCormick envelopes and selective discretization of variables; it was useful in some applications to chemical engineering. In addition, we perform a comparison of HybS with NMDT-based methods and Gurobi as an MIQCQP solver.

Outline We proceed as follows. In Sect. 2, we introduce several useful concepts and notations used throughout the work. In Sect. 3, we present core formulations used repeatedly in our linear relaxations of quadratic terms. In Sect. 4, we introduce the new MIP relaxation HybS for equations of the form $z = xy$. In Sect. 5, we prove various properties about the strengths of this MIP relaxation focusing on volume, sharpness, and optimal choice of breakpoints. In Appendix B we prove that the sawtooth relaxation is hereditarily sharp. In Sect. 6, we present our computational study.

2 MIP formulations

In this work, we study relaxations of general mixed-integer quadratically constrained quadratic programs (MIQCQPs), which are defined as

$$\begin{aligned} \min\ & \boldsymbol{x}^\top Q_0 \boldsymbol{x} + \boldsymbol{c}^0 \cdot \boldsymbol{x}, \\ \text{s.t. } & \boldsymbol{x}^\top Q_j \boldsymbol{x} + \boldsymbol{c}^j \cdot \boldsymbol{x} + b_j \leqslant 0 && j = 1, \dots, m, \\ & x_i \in [\underline{x}_i, \bar{x}_i] && i = 1, \dots, k, \\ & x_l \in \{0, 1\} && l = k+1, \dots, n, \end{aligned} \tag{1}$$

for $Q_0, Q_j \in \mathbb{R}^{n \times n}$, $\boldsymbol{c}^0, \boldsymbol{c}^j \in \mathbb{R}^n$ and $b_j \in \mathbb{R}$, $j = 1, \dots m$.

Throughout this article, we use the following convenient notation: for any two integers $i \leqslant j$, we define $[\![i, j]\!] := \{i, i+1, \dots, j\}$, and for an integer $i \geqslant 1$ we define $[\![i]\!] := [\![1, i]\!]$. We will denote sets using capital letters but also use capital letters for matrices, some functions, and the number of layers L. We typically denote variables using lowercase letters and vectors of variables using boldface. For a vector $\boldsymbol{u} = (u_1, \dots, u_n)$ and some index set $I \subseteq [\![n]\!]$, we write $\boldsymbol{u}_I := (u_i)_{i \in I}$. Thus, e.g. $\boldsymbol{u}_{[\![i]\!]} = (u_1, \dots, u_i)$. Furthermore, we introduce the following notation: for a function $F \colon X \to \mathbb{R}$ and a subset $B \subseteq X$, let $\mathrm{gra}_B(F)$, $\mathrm{epi}_B(F)$ and $\mathrm{hyp}_B(F)$ denote the *graph*, *epigraph* and *hypograph* of the function F over the set B, respectively. That is,

$$\begin{aligned} \mathrm{gra}_B(F) &:= \{(\boldsymbol{u}, z) \in B \times \mathbb{R} : z = F(\boldsymbol{u})\}, \\ \mathrm{epi}_B(F) &:= \{(\boldsymbol{u}, z) \in B \times \mathbb{R} : z \geqslant F(\boldsymbol{u})\}, \\ \mathrm{hyp}_B(F) &:= \{(\boldsymbol{u}, z) \in B \times \mathbb{R} : z \leqslant F(\boldsymbol{u})\}. \end{aligned}$$

In the following, we introduce the concept of MIP formulations as well as properties regarding MIP formulations which will be used later on.

We will study mixed-integer linear sets, so-called *mixed-integer programming (MIP) formulations*, of the form

$$P^{\mathrm{IP}} := \{(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{z}) \in \mathbb{R}^{d+1} \times [0,1]^p \times \{0,1\}^q : A(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{z}) \leqslant b\}$$

for some matrix A and vector b of suitable dimensions. The *linear programming (LP) relaxation* or *continuous relaxation* P^{LP} of P^{IP} is given by

$$P^{\mathrm{LP}} := \{(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{z}) \in \mathbb{R}^{d+1} \times [0,1]^p \times [0,1]^q : A(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{z}) \leqslant b\}.$$

We will often focus on the projections of these sets onto the variables $\boldsymbol{u}$, i.e.

$$\mathrm{proj}_{\boldsymbol{u}}(P^{\mathrm{IP}}) := \{\boldsymbol{u} \in \mathbb{R}^{d+1} : \exists (\boldsymbol{v}, \boldsymbol{z}) \in [0,1]^p \times \{0,1\}^q \quad \text{s.t.} \quad (\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{z}) \in P^{\mathrm{IP}}\}. \tag{2}$$

The corresponding *projected linear relaxation* $\mathrm{proj}_{\boldsymbol{u}}(P^{\mathrm{LP}})$ onto the $\boldsymbol{u}$-space is defined accordingly.

In order to assess the quality of an MIP formulation, we will work with several possible measures of formulation strength. First, we define notions of sharpness, as in [7, 29]. These relate to the tightness of the LP relaxation of an MIP formulation. Whereas properties such as total unimodularity guarantee an LP relaxation to be a complete description for the mixed-integer points in the full space, we are interested here in LP relaxations that are tight descriptions of the mixed-integer points in the projected space.

Definition 1 (Sharpness) We say that the MIP formulation P^{IP} is *sharp* if

$$\text{proj}_{\boldsymbol{u}}(P^{\text{LP}}) = \text{conv}(\text{proj}_{\boldsymbol{u}}(P^{\text{IP}}))$$

holds. Further, we call it *hereditarily sharp* if, for all $I \subseteq [\![q]\!]$ and $\hat{\boldsymbol{z}} \in \{0,1\}^{|I|}$, we have

$$\text{proj}_{\boldsymbol{u}}(P^{\text{LP}}|_{z_I=\hat{z}}) = \text{conv}\left(\text{proj}_{\boldsymbol{u}}(P^{\text{IP}}|_{z_I=\hat{z}})\right).$$

Sharpness expresses a tightness at the root node of a branch-and-bound tree. Hereditarily sharp means that fixing any subset of binary variables to 0 or 1 preserves sharpness, and therefore this means sharpness is preserved throughout a branch-and-bound tree.

In this article, we study certain non-polyhedral sets $U \subseteq \mathbb{R}^{d+1}$ and will develop MIP formulations P^{IP} to form relaxations of U in the projected space, as defined in the following.

Definition 2 (MIP relaxation) For a set $U \subseteq \mathbb{R}^{d+1}$ we say that an MIP formulation P^{IP} is an *MIP relaxation* of U if

$$U \subseteq \text{proj}_{\boldsymbol{u}}(P^{\text{IP}}).$$

Given a function $F\colon [0,1]^d \to \mathbb{R}$, we will mostly consider

$$U = \text{gra}_{[0,1]^d}(F) \subseteq \mathbb{R}^{d+1}.$$

In particular, we will focus on either

$$U = \{(x,z) \in [0,1]^2 : z = x^2\} \quad \text{or} \quad U = \{(x,y,z) \in [0,1]^3 : z = xy\}.$$

We now define several quantities to measure the error of an MIP relaxation.

Definition 3 (Error) For an MIP relaxation P^{IP} of a set $U \subseteq \mathbb{R}^{d+1}$, let $\bar{\boldsymbol{u}} \in \text{proj}_{\boldsymbol{u}}(P^{\text{IP}})$. We then define the *pointwise error* of $\bar{\boldsymbol{u}}$ as

$$\mathcal{E}(\bar{\boldsymbol{u}}, U) := \min\{|\boldsymbol{u}_{d+1} - \bar{\boldsymbol{u}}_{d+1}| : \boldsymbol{u} \in U, \boldsymbol{u}_{[\![d]\!]} = \bar{\boldsymbol{u}}_{[\![d]\!]}\}.$$

We next define the following two error measures for P^{IP} w.r.t. U:

1. The *maximum error* of P^{IP} w.r.t. U is defined as

$$\mathcal{E}^{\max}(P^{\mathrm{IP}}, U) := \max_{\bar{\boldsymbol{u}} \in \mathrm{proj}_{\boldsymbol{u}}(P^{\mathrm{IP}})} \mathcal{E}(\bar{\boldsymbol{u}}, U).$$

2. The *average error* of P^{IP} w.r.t. U is defined as

$$\mathcal{E}^{\mathrm{avg}}(P^{\mathrm{IP}}, U) := \mathrm{vol}(\mathrm{proj}_{\boldsymbol{u}}(P^{\mathrm{IP}}) \setminus U).$$

Via integral calculus, the second, volume-based error measure can be interpreted as the average pointwise error width of all points $\boldsymbol{u} \in \mathrm{proj}_{\boldsymbol{u}}(P^{\mathrm{IP}})$. Note that whenever the volume of U is zero (i.e. it is a lower-dimensional set), the average error just reduces to the volume of $\mathrm{proj}_{\boldsymbol{u}}(P^{\mathrm{IP}})$.

Both of the defined error quantities for an MIP relaxation P^{IP} can also be used to measure the tightness of the corresponding LP relaxation P^{LP}. In Sect. 5.3.2, we use these to compare formulations when P^{LP} is not sharp.

3 Core relaxations

In the definition of the MIP relaxations studied in this work, we repeatedly make use of several "core" formulations for specific sets of feasible points. They are introduced in the following.

For our relaxations of MIQCQPs, we will frequently need to consider terms of the form $z = xy$ for continuous or integer variables x and y within certain bounds $D_x := [\underline{x}, \bar{x}]$ and $D_y := [\underline{y}, \bar{y}]$, respectively. To this end, we introduce the function $F\colon D \to \mathbb{R},\ F(x, y) = xy$, $D := D_x \times D_y$, and refer to the set of feasible solutions to the equation $z = xy$ via the graph of F, i.e. $\mathrm{gra}_D(F) = \{(x, y, z) \in D \times \mathbb{R} : z = xy\}$. In order to simplify the exposition, we will, for example, often write $\mathrm{gra}_D(xy)$ or refer to a relaxation of the equation $z = xy$ instead of $\mathrm{gra}_D(F)$. We will do this similarly for the epigraph and hypograph of F as well as for the univariate function $f\colon D_x \to \mathbb{R},\ f(x) = x^2$ and equations of the form $z = x^2$, for example.

3.1 McCormick envelopes

The convex hull of the equation $z = xy$ for $(x, y) \in D$ is given by a set of linear equations known as the McCormick envelope. See [34].

$$\mathcal{M}(x, y) := \left\{(x, y, z) \in [\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}] \times \mathbb{R} : (4)\right\}. \tag{3}$$

$$\begin{aligned} \underline{x} \cdot y + x \cdot \underline{y} - \underline{x} \cdot \underline{y} \leqslant \quad & z \quad \leqslant \bar{x} \cdot y + x \cdot \underline{y} - \bar{x} \cdot \underline{y}, \\ \bar{x} \cdot y + x \cdot \bar{y} - \bar{x} \cdot \bar{y} \leqslant \quad & z \quad \leqslant \underline{x} \cdot y + x \cdot \bar{y} - \underline{x} \cdot \bar{y}. \end{aligned} \tag{4}$$

3.2 Sawtooth-based MIP formulations

We next recall an MIP formulation for approximating equations of the form $z = x^2$ that requires only logarithmically-many binary variables in the number of linear segments. It makes use of an elegant p.w.l. formulation for $\text{gra}_{[0,1]}(x^2)$ from [42] using the recursively defined *sawtooth* function presented in [39] to formulate the approximation of $\text{gra}_{[0,1]}(x^2)$, as described in [7].

Let L be an positive integer and let F^L be the piecewise linear interpolation of x^2 at uniformly spaced breakpoints $\frac{i}{2^L}$ for $i = 0, 1, \ldots, 2^L$; see Fig. 1. This function has a convenient recursive definition [39, 42]. To this end, define the "tooth" function $G\colon [0, 1] \to [0, 1]$, $G(x) = \min\{2x, 2(1 - x)\}$. Subsequently, we define compositions of the tooth function

$$G^j := \underbrace{G \circ G \circ \ldots \circ G}_{j}. \tag{5}$$

Under this notation, we can formally define the function $F^L\colon [0, 1] \to [0, 1]$,

$$F^L(x) := x - \sum_{j=1}^{L} 2^{-2j} G^j(x). \tag{6}$$

We summarize useful information from [7, 42] about the approximation F^L. These properties will be used in our analysis of the models that we propose.

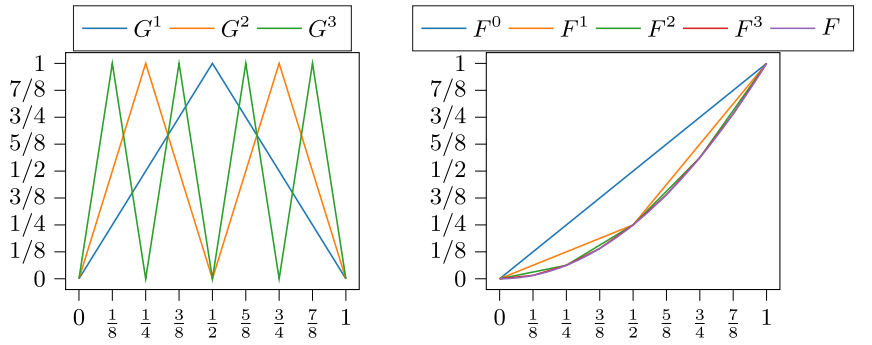
Proposition 1 ([7, 42]) *The function F^L satisfies the following properties:*

1. *The function F^L is the piecewise linear interpolation of x^2 at uniformly spaced breakpoints $\frac{i}{2^L}$ for $i = 0, 1, \ldots, 2^L$; see Fig. 1. The shifted piecewise linear function $F^L - 2^{-2L-2}$ has each affine part being the tangent to x^2 at the midpoint $\frac{i}{2^L} + \frac{1}{2^{L+1}}$; see Fig. 2.*
2. *It holds $0 \leqslant F^L(x) - x^2 \leqslant 2^{-2L-2}$ for all $x \in [0, 1]$.*
   Equivalently, $0 \leqslant x^2 - (F^L(x) - 2^{-2L-2}) \leqslant 2^{-2L-2}$ for all $x \in [0, 1]$.
3. *It holds $F^L(x) - 2^{-2L-2} = x^2$ if and only if $x = \frac{i}{2^L} + \frac{1}{2^{L+1}}$ with $i = 0, 1, \ldots, 2^L - 1$.*
4. *The function F^L is convex on the interval $[0, 1]$.*

Following [7], we create an MIP formulation to encode this piecewise linear function. We create variables g_j to represent the output of a "sawtooth" function of x and binary variables $\boldsymbol{\alpha} \in \{0, 1\}^L$ that represent decision in $G(x)$ that either $2x \leqslant 2(1-x)$ or $2(1-x) \leqslant 2x$. In particular, we design the formulation such when $\boldsymbol{\alpha} \in \{0, 1\}^L$, the relationship between g_j and g_{j-1} is $g_j = \min\{2g_{j-1}, 2(1 - g_{j-1})\}$ for $j = 1, \ldots, L$,

To this end, we define a formulation parameterized by the depth $L \in \mathbb{N}$:

$$S^L := \left\{ (x, \boldsymbol{g}, \boldsymbol{\alpha}) \in [0, 1] \times [0, 1]^{L+1} \times \{0, 1\}^L : (8) \right\}. \tag{7}$$



(a) The sawtooth functions G^j for $j = 1, 2, 3$.

(b) The successive piecewise linear approximations (interpolations) of $F(x) = x^2$.

Fig. 1 An illustration of the functions G^j and F^L that underlie the construction of our MIP formulations

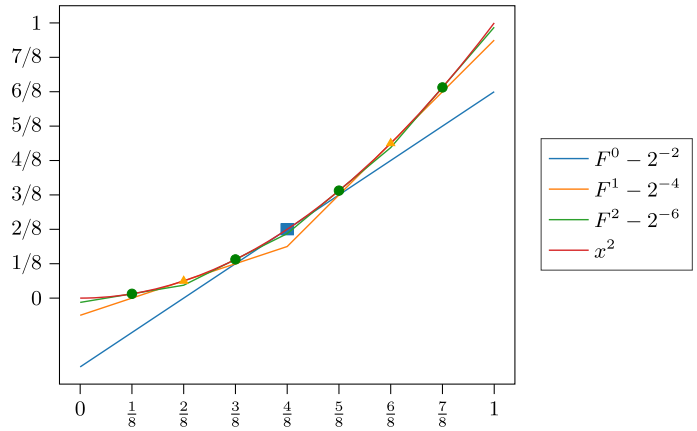


Fig. 2 The successive piecewise linear approximations of x^2 shifted down to be underestimators. The markers indicate the places where the underestimators coincide with x^2 and in fact, show that the affine segments are tangent lines to the function. The inequality $z \geqslant F^L(x) - 2^{-2L-2}$ in fact creates 2^L tangent lower bounds

$$\begin{aligned} g_0 &= x, \\ 2(g_{j-1} - \alpha_j) \leqslant g_j &\leqslant 2g_{j-1} \qquad j = 1, \ldots, L, \\ 2(\alpha_j - g_{j-1}) \leqslant g_j &\leqslant 2(1 - g_{j-1}) \qquad j = 1, \ldots, L. \end{aligned} \tag{8}$$

Using the relationships (5) and (6) between x and $\boldsymbol{g}$, any constraint of the form $z = x^2$ can be approximated via the function

$$\begin{aligned} f^L : [0, 1] \times [0, 1]^{L+1} &\to [0, 1], \\ f^L(x, \boldsymbol{g}) &= x - \sum_{j=1}^{L} 2^{-2j} g_j, \end{aligned} \tag{9}$$

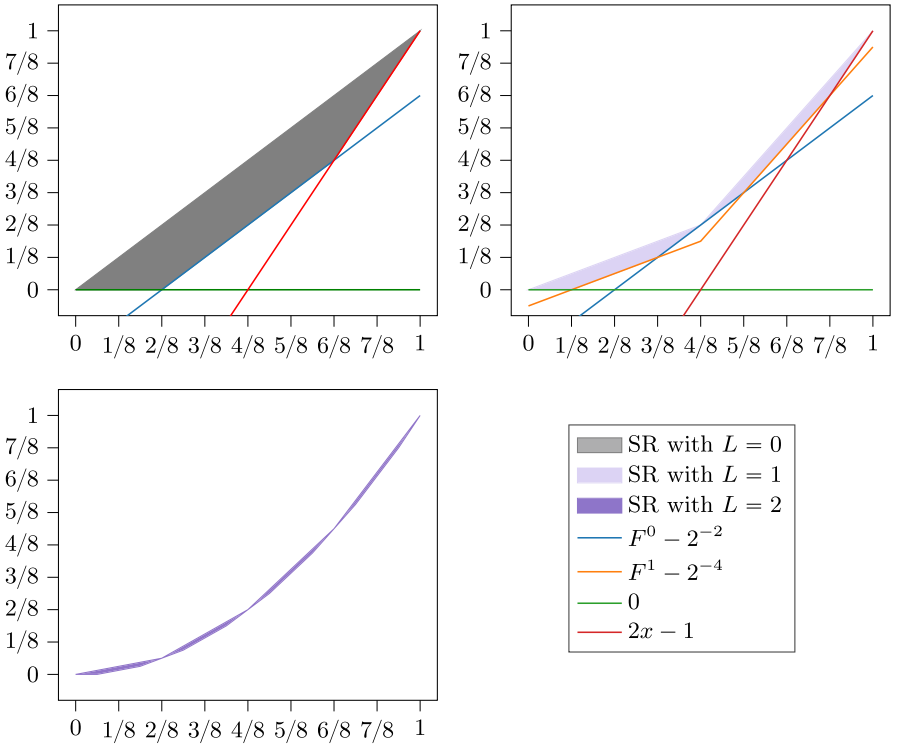


Fig. 3 The sawtooth relaxation from Definition 5 at depths $L = 0, 1, 2$. The shaded region is the relaxation. Some additional inequalities are plotted to help visualize the inequalities with respect to the functions F^j

for an integer $L \geqslant 0$. We use the above definitions to give an MIP formulation that approximates equations of the form $z = x^2$.

Definition 4 (Sawtooth Approximation, [7]) Given some $L \in \mathbb{N}$, the *depth-L sawtooth approximation* for $z = x^2$ on the interval $x \in [0, 1]$ is given by

$$\left\{(x, z) \in [0, 1]^2 : \exists (\boldsymbol{g}, \boldsymbol{\alpha}) \in [0, 1]^{L+1} \times \{0, 1\}^L : z = f^L(x, \boldsymbol{g}),\ (x, \boldsymbol{g}, \boldsymbol{\alpha}) \in S^L\right\}. \tag{10}$$

The set (10) is a compact approximation of $\text{gra}_{[0,1]}(x^2)$ in terms of the number of variables and constraints.

Based on the sawtooth approximation, we can now present the sawtooth relaxation for $z = x^2$ from [7], illustrated in Fig. 3, which arises by shifting each approximating function F^j, $j = 0, \ldots, L$, down by its maximum error 2^{-2j-2} (established in Proposition 1, Item 2) and then adding additional outer-approximation cuts to x^2 at $x = 0$ and $x = 1$ (Fig. 4).

Definition 5 (Sawtooth Relaxation, SR [7]) Given some $L \in \mathbb{N}$, the *depth-L sawtooth relaxation* for $z = x^2$ on the interval $x \in [0, 1]$ is given by

$$\left\{(x, z) \in [0, 1] \times \mathbb{R} : \exists (\boldsymbol{g}, \boldsymbol{\alpha}) \in [0, 1]^{L+1} \times \{0, 1\}^L : (12)\right\}. \tag{11}$$

$$\begin{aligned} z &\leqslant f^L(x, \boldsymbol{g}), \\ z &\geqslant f^j(x, \boldsymbol{g}) - 2^{-2j-2} \quad j = 0, \ldots, L \\ z &\geqslant 0, \quad z \geqslant 2x - 1, \\ (x, \boldsymbol{g}, \boldsymbol{\alpha}) &\in S^L. \end{aligned} \tag{12}$$

Remark 1 (Transformation to General Bounds) To this point, the sawtooth MIP formulations were presented for x^2 with $x \in [0, 1]$. However, all sawtooth-based MIP formulations can be extended to general intervals $x \in [\underline{x}, \bar{x}]$ by mapping $[\underline{x}, \bar{x}]$ to $[0, 1]$ via the substitution $\hat{x} = \frac{x - \underline{x}}{\bar{x} - \underline{x}} \in [0, 1]$ and applying the sawtooth formulation to model the equation

$$\hat{z} = \hat{x}^2 = \left(\frac{x - \underline{x}}{\bar{x} - \underline{x}}\right)^2 = \frac{x^2 - 2x\underline{x} + \underline{x}^2}{(\bar{x} - \underline{x})^2} = \frac{z - 2x\underline{x} + \underline{x}^2}{(\bar{x} - \underline{x})^2} = \frac{z - \underline{x}(2x - \underline{x})}{(\bar{x} - \underline{x})^2}.$$

Thus, for general intervals, we first apply the approximation to $\hat{z} = \hat{x}^2$, then add the equations

$$\hat{x} = \frac{x - \underline{x}}{\bar{x} - \underline{x}}, \quad \hat{z} = \frac{z - \underline{x}(2x - \underline{x})}{(\bar{x} - \underline{x})^2}.$$

In our computational study in Sect. 6, these constraints are implemented as defining expressions for $\hat{x}$ and $\hat{z}$, and the MIP formulations are constructed for $\hat{x}$ and $\hat{z}$ then. See Appendix A for the generalized MIP formulations under this transformation. ◇

Now, we consider the LP relaxation of S^L, where each variable α_j is relaxed to the interval $[0, 1]$. Then, via the constraints (8), we see that the weakest lower bounds on each g_j w.r.t. g_{j-1} can be attained via setting $\alpha_j = g_{j-1}$, yielding a lower bound of 0. Thus, after projecting out $\boldsymbol{\alpha}$, the LP relaxation of S^L in terms of just x and $\boldsymbol{g}$ can be stated as

$$T^L = \left\{(x, \boldsymbol{g}) \in [0, 1] \times [0, 1]^{L+1} : (13)\right\}.$$

$$\begin{aligned} g_0 &= x, \\ g_j &\leqslant 2(1 - g_{j-1}) \quad j = 1, \ldots, L, \\ g_j &\leqslant 2g_{j-1} \quad\quad\; j = 1, \ldots, L. \end{aligned} \tag{13}$$

The sawtooth relaxation (11) is sharp by Theorem 1 (proved later in this work), which follows in much the same way as the sharpness of the sawtooth approximation (10), as established in [7, Theorem 1]. Thus, the LP relaxation of the sawtooth relaxation (11) yields the same lower bound on z as the MIP version due to sharpness and the convexity of F^L. This allows us to define an LP outer approximation for inequalities of the form $z \geqslant x^2$:

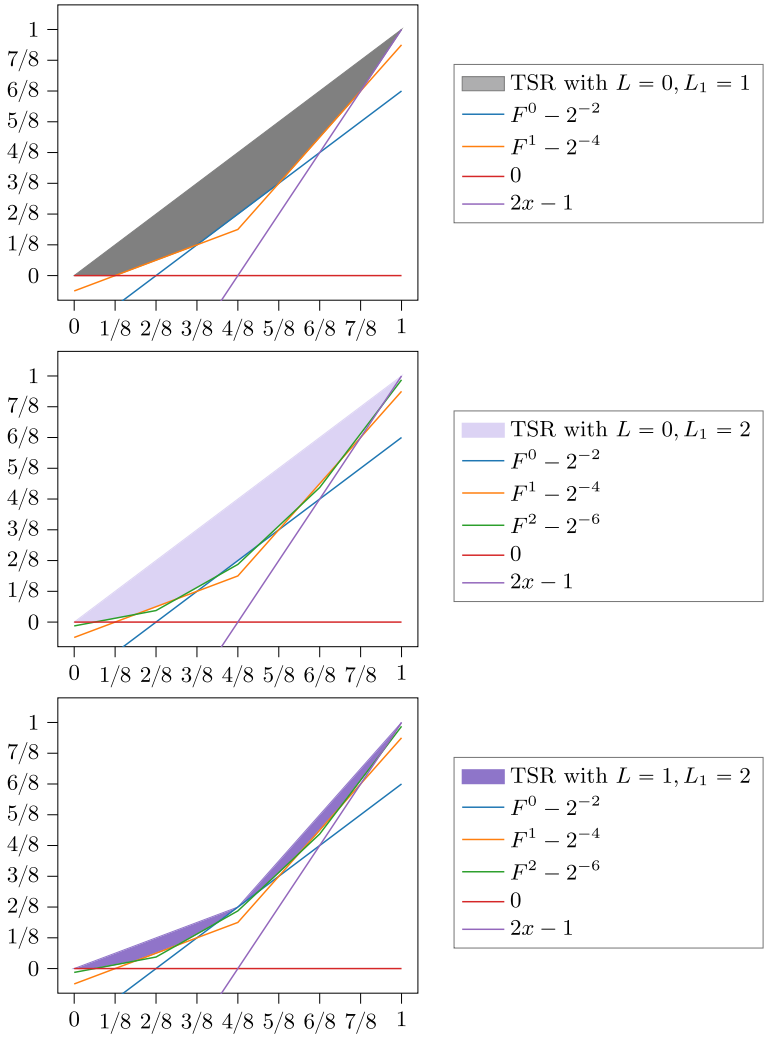


Fig. 4 The tightened sawtooth relaxations R^{L,L_1} from Definition 7 for the pairs $(L, L_1) = (0, 1), (0, 2), (1, 2)$. By increasing L_1 beyond L, we tighten the lower bound by creating more inequalities. This is done by only adding linearly-many variables and inequalities in the extended formulation to gain exponentially-many equally spaced cuts in the projection

Definition 6 (Sawtooth Epigraph Relaxation, SER) Given some $L \in \mathbb{N}$, the *depth-L sawtooth epigraph relaxation* for $z \geqslant x^2$ on the interval $x \in [0, 1]$ is given by

$$Q^L := \left\{ (x, z) \in [0, 1] \times \mathbb{R} : \exists \boldsymbol{g} \in [0, 1]^{L+1} : (15) \right\}. \tag{14}$$

$$\begin{aligned} z &\geqslant f^j(x, \boldsymbol{g}) - 2^{-2j-2} \quad j = 0, \ldots, L, \\ z &\geqslant 0, \quad z \geqslant 2x - 1, \\ (x, \boldsymbol{g}) &\in T^L. \end{aligned} \tag{15}$$

We will prove in Proposition 2 that the maximum error for the sawtooth epigraph relaxation is 2^{-2L-4}.

Finally, we combine the depth-L sawtooth relaxation (11) with the depth-L_1 sawtooth epigraph relaxation (14) for some $L_1 \geqslant L$ to obtain a sawtooth relaxation which is stronger in the lower bound, but uses the same number of binary variables.

Definition 7 (Tightened Sawtooth Relaxation, TSR) Given some $L, L_1 \in \mathbb{N}$ with $L_1 \geqslant L$, the *tightened sawtooth relaxation* for $z = x^2$ on the interval $x \in [0, 1]$ with upper-bounding depth L and lower-bounding depth L_1 is given by

$$R^{L,L_1} := \{(x, z) \in [0, 1] \times \mathbb{R} : \exists (\boldsymbol{g}, \boldsymbol{\alpha}) \in [0, 1]^{L_1+1} \times \{0, 1\}^L : (17)\}. \tag{16}$$

$$z \leqslant f^L(x, \boldsymbol{g}_{[\![0,L]\!]}), \tag{17a}$$

$$(x, \boldsymbol{g}_{[\![0,L]\!]}, \boldsymbol{\alpha}) \in S^L, \tag{17b}$$

$$\begin{cases} (x, \boldsymbol{g}) \in T^{L_1}, & \text{(17c)} \\ z \geqslant f^j(x, \boldsymbol{g}) - 2^{-2j-2} \quad j = 0, \ldots, L_1, & \text{(17d)} \\ z \geqslant 0, & \text{(17e)} \\ z \geqslant 2x - 1. & \text{(17f)} \end{cases}$$

We connect the last constraints with a brace since there are all defining constraints for Q^{L_1}. Since we define Q^{L_1} in the projection space (x, z), we cannot simply write $(x, \boldsymbol{g}) \in Q^{L_1}$ since we need the same $\boldsymbol{\alpha}$ and $\boldsymbol{g}$ to apply to the other constraints as well.

We will prove in Theorem 1 that the tightened sawtooth relaxation is also sharp, and in Theorem 2 that it is hereditarily sharp.

4 MIP relaxations for non-convex MIQCQPs

In this section, we focus on MIP relaxations for bilinear equations of the form $z = xy$. For convenience, we define a *completely dense* MIQCQP as an MIQCQP for which all terms of the form x_i^2 and $x_i x_j$ appear in either the objective or in some constraint. The novel formulation *HybS* presented herein is an extension of existing formulations Bin2 and Bin3, designed to significantly reduce the number of binary variables required to reach the same level of relaxation accuracy compared to its original predecessors *Bin2* and *Bin3* for completely dense MIQCQPs, which will also be introduced in the following.

4.1 Separable MIP relaxations

We present three MIP relaxations based on separable reformulations. A separable reformulation turns a multivariate expression into a sum of univariate functions. To

this end, we make use of the reformulation approaches Bin2 and Bin3, given via

$$\begin{aligned} \text{Bin2} : xy &= \tfrac{1}{2}((x+y)^2 - x^2 - y^2), \\ \text{Bin3} : xy &= \tfrac{1}{2}(x^2 + y^2 - (x-y)^2), \end{aligned}$$

see e.g. [5], and combine them with the sawtooth relaxation (16) to derive MIP relaxations for the occurring equations of the form $z = xy$. While the following MIP relaxations on Bin2 and Bin3 are natural extensions of the MIP approximations studied in [5] to MIP relaxations, we will also combine both reformulations to a new formulation in which the MIP relaxation requires significantly less binary variables if it is used to solve problems of the form (1) As a reminder, in the definitions below, the notation $\mathcal{M}$ is used to describe the McCormick envelope.

Remark 2 In [5], Bin1: $xy = (1/2(x+y))^2 - (1/2(x-y))^2$ is also discussed as a possible separable reformulation. However, for completely dense MIQCQPs, Bin1 requires a number of binary variables that is by a factor of roughly 2 greater than that required for Bin2 and Bin3. This is due to the fact that for each bivariate product $x_i x_j$, we need to discretize both $(1/2(x_i + x_j))^2$ and $(1/2(x_i - x_j))^2$ instead of only one of the two squares for Bin2 and Bin3. Therefore, we omit Bin1 in the following. ◇

Definition 8 (Bin2) The MIP relaxation Bin2 of $z = xy, x, y \in [0,1]^2$, with a lower-bounding depth $L_1 \in \mathbb{N}$ and an upper-bounding depth $L \in \mathbb{N}$, is defined as follows:

$$\begin{aligned} p &= x + y \\ z &= 1/2(z_p - z_x - z_y) \\ (x, y, z) &\in \mathcal{M}(x, y) \\ (x, z_x),\ (y, z_y),\ (p, z_p) &\in R^{L,L_1} \\ x, y \in [0,1], &\quad p \in [0,2]. \end{aligned} \tag{18}$$

Definition 9 (Bin3) The MIP relaxation Bin3 of $z = xy, x, y \in [0,1]^2$, with a lower-bounding depth $L_1 \in \mathbb{N}$ and an upper-bounding depth of $L \in \mathbb{N}$, is defined as follows:

$$\begin{aligned} p &= x - y \\ z &= 1/2(z_x + z_y - z_p) \\ (x, y, z) &\in \mathcal{M}(x, y) \\ (x, z_x),\ (y, z_y),\ (p, z_p) &\in R^{L,L_1} \\ x, y \in [0,1], &\quad p \in [-1,1]. \end{aligned} \tag{19}$$

Note that we apply the tightened sawtooth relaxation R^{L,L_1}, defined in (16), not only to $x, y \in [0,1]$, but also to the variable p, where the domain is either $[0,2]$ or $[-1,1]$. This is done by following the transformation in Remark 1 to map p and z_p to the interval $[0,1]$ and then applying (16) to the transformed variables.

We now combine Bin2 and Bin3 to derive an MIP relaxation for $z = xy$ based on bounding z in the following two ways:

$$\begin{aligned} z &\leqslant 1/2(x^2 + y^2 - (x-y)^2), \\ z &\geqslant 1/2((x+y)^2 - x^2 - y^2), \end{aligned}$$

and then replacing each right-hand side with proper upper and lower bounds. We choose this setting so that we only have to model lower bounds for the $(x-y)^2$- and $(x+y)^2$-terms and can thus apply the sawtooth epigraph relaxation (14) to circumvent the use of binary variables for these terms. To this end, we introduce the continuous auxiliary variables p_1, p_2, z_x, z_y, z_{p_1}, z_{p_2} and z to obtain an equivalent relaxation for $z = xy$:

$$p_1 = x + y,\ p_2 = x - y, \tag{20a}$$
$$z_x \leqslant x^2,\ z_y \leqslant y^2, \tag{20b}$$
$$z_{p_1} \geqslant p_1^2,\ z_{p_2} \geqslant p_2^2, \tag{20c}$$
$$z \leqslant z_x + z_y - z_{p_1},\ z \geqslant z_{p_2} - z_x - z_y. \tag{20d}$$

Finally, we replace x^2 and y^2 in the non-convex constraints (20b) with a sawtooth relaxation (17a) of depth L and p_1^2 and p_2^2 in the convex constraints (20c) by a sawtooth epigraph relaxation (17f) with depth L_1 to obtain a relaxation of $z = xy$ in (20d). The resulting model is especially interesting as, in contrast to Bin2 and Bin3, it does not require binary variables to model equations of the form $p_1^2 = (x+y)^2$ and $p_2^2 = (x-y)^2$, since we only need to incorporate lower bounds as used in Q^L.

Definition 10 (Hybrid Separable HybS) Let $x, y \in [0,1]$, and let $L, L_1 \in \mathbb{N}$. The following MIP relaxation for $z = xy$, which combines the relaxations Bin2 and Bin3, is called the *hybrid separable* MIP relaxation, in short HybS, with a lower-bounding depth of L_1 and an upper-bounding depth of L:

$$\begin{aligned} p_1 = x + y, \quad p_2 = x - y \\ (x, z_x),\ (y, z_y) \in R^{L,L_1} \\ (p_1, z_{p_1}),\ (p_2, z_{p_2}) \in Q^{L_1} \\ 1/2(z_{p_1} - z_x - z_y) \leqslant z \leqslant 1/2(z_x + z_y - z_{p_2}) \\ (x, y, z) \in \mathcal{M}(x, y) \\ x, y \in [0,1], \quad p_1 \in [0,2], \quad p_2 \in [-1,1]. \end{aligned} \tag{21}$$

As Q^{L_1} in (21) is originally defined for variables in $[0,1]$, we again use the transformation from Remark 1 to extend it to other domains.

Note that, when some constraint of an MIQCQP has a completely dense quadratic matrix, the number of (20c)-type constraints is quadratic in the dimension of x. Thus, the number of binary variables for Bin2 and Bin3 is in $O(n^2 L)$, while the formulation HybS requires only nL binary variables. As we will show in Sect. 5, the formulation HybS also has a strictly tighter LP relaxation than that of either formulation Bin2 or Bin3. This implies a smaller volume of the projected LP relaxation as well. We also note, however, that the MIP relaxation is not strictly tighter. For example, let $L = L_1 = 1$ and consider the point $(x, y) = (\frac{1}{4}, \frac{3}{4})$. The upper bound on $z = xy$ produced by the MIP relaxation Bin2 at this point is $z \leqslant \frac{3}{16}$, i.e. the exact value. The MIP relaxation HybS (as well as Bin3), however, has a weaker upper bound of $z \leqslant \frac{1}{4}$ at this point.

Table 1 A summary of characteristics of the different MIP relaxations. Binary variables and constraints are given in the worst-case, in which every possible quadratic term must be modeled, for example if some matrix Q_i is completely dense

MIP relax	# Bin. variables	# Constraints	Max. error	Avg. error
HybS	nL	$n(\frac{1}{2}(5n-3)+2n(L+L_1))$	2^{-2L-2}	$\frac{1}{3}2^{-2L}$
Bin2	$\frac{1}{2}(n^2+1)L$	$n(\frac{1}{2}(3n-1)+(n+1)(L+L_1))$	2^{-2L-1}	$\frac{1}{2}2^{-2L}$
Bin3	$\frac{1}{2}(n^2+1)L$	$n(\frac{1}{2}(3n-1)+(n+1)(L+L_1))$	2^{-2L-1}	$\frac{1}{2}2^{-2L}$

When we apply any of the separable formulations Bin2, Bin3 and HybS to compute dual bounds for MIQCQPs in Sect. 6, all original univariate quadratic terms of the form x_i^2 (i.e. those not resulting from any reformulations) are modeled via the tightened sawtooth relaxation (16).

Remark 3 We can alternatively obtain a convex mixed-integer quadratic relaxation of $z = xy$ by directly incorporating the convex quadratic constraints $z_x \leqslant x^2$, $z_y \leqslant y^2$, $z_{p_1} \geqslant p_1^2$ and $z_{p_2} \geqslant p_2^2$ in (20) exactly instead of using p.w.l. relaxations. This variation could be implemented using a convex solver instead of a linear solver. ◇

Remark 4 (Binary Variables and Dense MIQCQPs) When modeling Problem (1) using the MIP relaxations Bin2 and Bin3 at depth L, we have L binary variables created whenever the tightened sawtooth relaxation R^{L,L_1} is used. For Bin2, we need the relaxations $(x_i, z_{x_i}) \in R^{L,L_1}$ and $(p_{ij}, z_{p_{ij}}) \in R^{L,L_1}$ for all pairs $i \neq j$, where $p_{ij} = x_i + x_j$. Note that $p_{ij} = p_{ji}$. Thus, we need $(n + \frac{1}{2}(n-1)^2)L = \frac{1}{2}(n^2+1)L$ binary variables.

We have the same result for Bin3, where instead we have $p_{ij} = x_i - x_j$ for all pairs $i \neq j$. Although this means $p_{ij} \neq p_{ji}$, we still have $p_{ij}^2 = p_{ji}^2$. Thus, a careful implementation also has $\frac{1}{2}(n^2+1)L$ binary variables.

HybS uses significantly fewer binary variables as it only requires $(x_i, z^{x_i}) \in R^{L,L_1}$ for each i. Hence, there are only nL binary variables. Surprisingly, this relaxation halves the error bound from Bin2 and Bin3. The strength in this approach is gained without quadratically-many binary variables by using the tightening set Q^{L_1} with the p_1-and p_2-variables. ◇

5 Theoretical analysis

In this section, we give a theoretical analysis of the presented MIP relaxations for the equation $z = xy$ over $x, y \in [0, 1]$ as well as the equation $z = x^2$ over $x \in [0, 1]$, respectively, in order to allow for a comparison of structural properties between them (Fig. 5). In particular, we will analyze their maximum and average errors, formulation strengths, i.e. (hereditary) sharpness and LP relaxation volumes, as well as the optimal placement of breakpoints to minimize average errors. The results we will arrive at are summarized in Table 1.

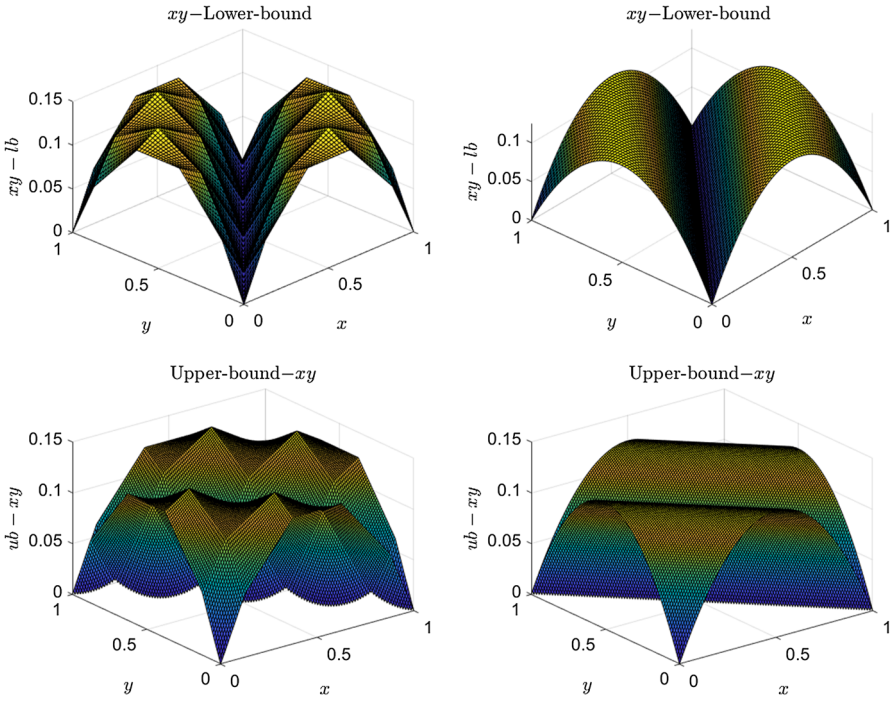


Fig. 5 Maximum overestimation and maximum underestimation of the MIP relaxation Bin2 defined in (18). In the left column, we show the case $L = L_1 = 1$. In the right column, we show $L = 1$ and $L_1 \to \infty$

5.1 Maximum error

We start the error analysis by discussing the maximum errors of the presented MIP relaxations.

5.1.1 Core formulations

First, we discuss the maximum errors of the core formulations from Sect. 3.1. For the sawtooth approximation (10), the maximum error is an overestimation by 2^{-2L-2}, see [7]. The maximum error of the sawtooth epigraph relaxation is 2^{-2L-4}, which we prove in the following. The tightened sawtooth relaxation stated in (16) uses the sawtooth approximation for overestimation while the lower bound, which is incident with the sawtooth epigraph relaxation (14), gains an extra layer of accuracy, with a maximum error of 2^{-2L-4}. Due to the overestimator, the (tightened) sawtooth relaxation has the same maximum error of 2^{-2L-2} as the sawtooth approximation.

Proposition 2 (Error of the sawtooth epigraph relaxation) *The maximum error of the sawtooth epigraph relaxation Q^L for $z \geqslant x^2$ with $x \in [0, 1]$ defined in (14) is 2^{-2L-4}.*

Proof The lower-bounding inequalities on z induced by the (x, z)-projection of the sawtooth epigraph relaxation, i.e. $\text{proj}_{x,z}(Q^L)$, are exactly the supporting valid lin-

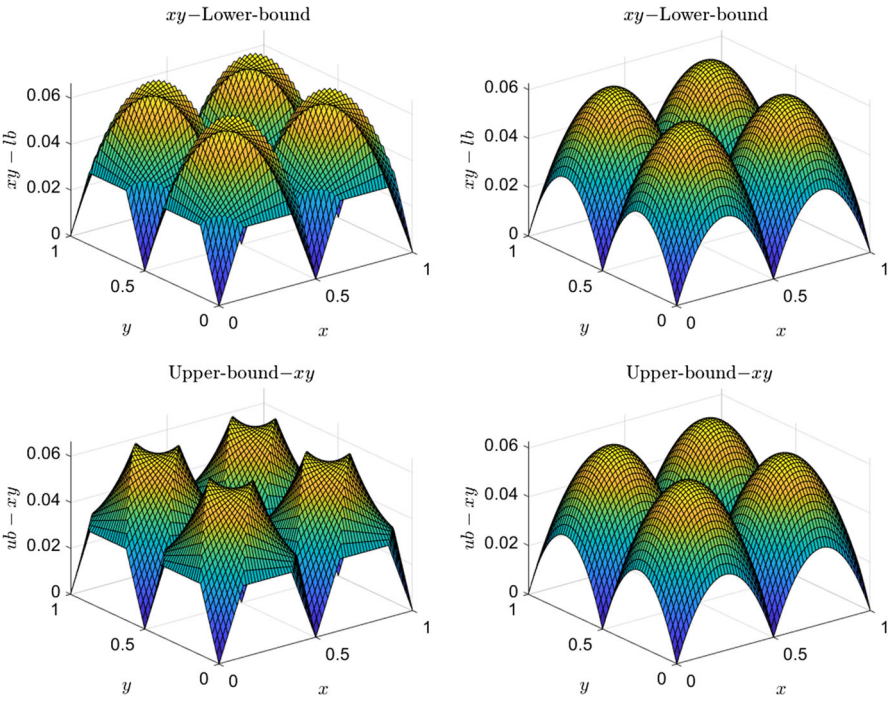


Fig. 6 Maximum overestimation and maximum underestimation of the MIP relaxation HybS defined in (21). In the left column, we show the case $L = L_1 = 1$. In the right column, we show $L = 1$ and $L_1 \to \infty$

ear inequalities to $z \geqslant x^2$ at the points $x_k := \frac{k}{2^{L+1}}$, $k = 0, \ldots, 2^L$; see Proposition 1. The maximum error is attained at the intersection of two consecutive linear segments on the boundary of the feasible region defined by these inequalities, i.e. at $(\bar{x}_k, z_k) := (\frac{x_k + x_{k+1}}{2}, x_k x_{k+1}) = ((k + \frac{1}{2})2^{-L-1}, k(k+1)2^{-2L-2})$. Thus, the maximum error is given by

$$\mathcal{E}^{\max}(Q^L, \mathrm{epi}_{[0,1]}(x^2)) = \left((k + \tfrac{1}{2})2^{-L-1}\right)^2 - k(k+1)2^{-2L-2} = 2^{-2L-4},$$

independent of the choice of k. □

In addition to the sawtooth-based formulations, we use McCormick relaxations as core formulations to form MIP relaxations of MIQCQPs. For the McCormick relaxation of the equation $z = xy$ over the box domain $[\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}]$, the maximum under- and overestimation is $\frac{1}{4}(\bar{x} - \underline{x})(\bar{y} - \underline{y})$, attained at $(x, y) = (\frac{1}{2}(\underline{x} + \bar{x}), \frac{1}{2}(\underline{y} + \bar{y}))$, see e.g. [33, page 23].

5.1.2 Separable MIP relaxations

In order to generate MIP relaxations of MIQCQPs with either the Bin2, Bin3, or the HybS approach, we need to discretize univariate quadratic terms and products of variables.

Univariate Quadratic Terms in MICQCP's. First, for univariate quadratic terms, i.e., $z = x^2$, in MIQCQPs, we use the tightened sawtooth relaxation to discretize in either approach. The tightened sawtooth relaxation has a maximum error of 2^{-2L-2}, as shown in Proposition 1.

Bivariate Products in MICQCP's. Second, for bivariate products, i.e., $z = xy$, in MIQCQPs, we use a different separable reformulation in each approach. In the following, we derive upper bounds, purely depending on L, and lower bounds, depending on L and L_1, on the maximum errors for variable products. Depending on the reformulation, we have to address two different maximum error scenarios in the bounds on z.

We start with the maximum error in the relaxations for z in which x^2 and y^2 are overestimated and p^2 is underestimated. This applies to the upper and lower bound on z in HybS, the lower bound on z in Bin2, and the upper bound on z in Bin3. In each of these cases, the maximum overestimation of both $z_x = x^2$ and $z_y = y^2$ with the sawtooth relaxation is 2^{-2L-2}, occurring at the grid centers $x_k = y_k = (k + \frac{1}{2})2^{-L}$, $k = 0, \ldots, 2^L - 1$. If we combine these points, x_k and y_k, with a point on the graph of p^2, i.e. $z_p = p^2$, this point has an approximation error 0 and we obtain a lower bound for the maximum error in the relaxation of $z = xy$. Namely, if P^{IP}_{L,L_1} denotes either of the MIP relaxations Bin2, Bin3 or HybS of $\text{gra}_{[0,1]^2}(xy)$ with depths L, L_1, we have

$$\begin{aligned}\mathcal{E}^{\max}(P^{\text{IP}}_{L,L_1}, \text{gra}_{[0,1]^2}(xy))) &\geqslant \tfrac{1}{2}(((x_k^2 + 2^{-2L-2}) - x_k^2) + ((y_k^2 + 2^{-2L-2}) - y_k^2) \\ &\quad + ((p^2 + 0) - p^2)) \\ &\geqslant \tfrac{1}{2}\left(2^{-2L-2} + 2^{-2L-2} + 0\right) \\ &= 2^{-2L-2},\end{aligned}$$

independent of the choice of k. This yields the following proposition.

Proposition 3 *The maximum error in the MIP relaxations* Bin2*,* Bin3 *and* HybS *for* $z = xy$ *with* $x, y \in [0, 1]$ *is at least* 2^{-2L-2}.

Furthermore, the maximum underestimation of p^2 is 2^{-2L_1-2} (twice the domain width, which means the error quadruples). This means we have an upper bound of

$$\frac{1}{2}(2^{-2L-2} + 2^{-2L-2} + 2^{-2L_1-2}) = 2^{-2L-2} + 2^{-2L_1-3}$$

on the maximum error in the lower bound on z in Bin2, the upper bound on z in Bin3 and both the upper and lower bound on z in HybS. We can use this observation to give an upper bound on the maximum error in the MIP relaxation HybS for $z = xy$. See Fig. 6 for the maximum over- and underestimation of the HybS MIP relaxation.

Proposition 4 *The maximum error in the MIP relaxation* HybS *for* $z = xy$ *with* $x, y \in [0, 1]$ *is at most* $2^{-2L-2} + 2^{-2L_1-3}$.

Next, we consider the upper bound on z in Bin2 and the lower bound on z in Bin3. Here, we are interested in the overestimation of p^2 and the underestimation of x^2 and y^2. The maximum overestimation of p^2 is 2^{-2L} (again, doubling the domain width quadruples the error). Combined with the maximum underestimation of the sawtooth relaxation for x^2 and y^2 of 2^{-2L_1-4}, this yields an upper bound on the maximum error on z of

$$\frac{1}{2}(2^{-2L} + 2^{-2L_1-4} + 2^{-2L_1-4}) = 2^{-2L-1} + 2^{-2L_1-4}$$

in terms of overestimation in Bin2 and underestimation in Bin3. Thus, we obtain the following upper bound for the maximum error in Bin2 and Bin3. See Fig. 5 for the maximum over- and underestimation of the Bin2 MIP relaxation.

Proposition 5 *The maximum error in the MIP relaxations* Bin2 *and* Bin3 *for* $z = xy$ *with* $x, y \in [0, 1]$ *is at most* $2^{-2L-1} + 2^{-2L_1-3}$.

In summary, we have the same lower bound for the maximum error of 2^{-2L-2} in Bin2, Bin3 and HybS. However, the known upper bound $2^{-2L-1} + 2^{-2L_1-4}$ in HybS is slightly better than that of Bin2 and Bin3 with $2^{-2L-1} + 2^{-2L_1-3}$.

Remark 5 In the MIP relaxations Bin2, Bin3, and HybS, increasing L_1 does not introduce any new binary variables. Therefore, we note that in our computations in Sect. 6 we choose L_1 to be significantly larger than L, such that the maximum error depends primarily on L. As L_1 increases to infinity, the maximum errors in all three MIP relaxations converge to 2^{-2L-2}. ◇

5.2 Average error and minimizing the average error

In this section, we will study the average error of an MIP relaxation by computing the volume enclosed by the projected MIP relaxation as an additional measure of its relaxation quality.

First, we compute the volumes of all presented MIP relaxations. Then we prove that the uniform discretizations, which are used by definition in each MIP formulation in this article, are indeed optimal in terms of minimizing the volume of the projected MIP relaxation if the number of discretization points is fixed (i.e. if L and L_1 are fixed).

In all separable formulations, we use the sawtooth relaxation (11) for equations of the form $z = x^2$. In [7, Propostion 6], the authors show that the volume of this relaxation $R^{L,L}$ is $3/16 \cdot 2^{-2L}$. Furthermore, from [7, Proposition 5] it follows that for any fixed number of breakpoints a uniform discretization minimizes the volume of the sawtooth epigraph relaxation.

Next, we consider the volumes for the MIP relaxations of $z = xy$. We start by showing that Bin2, Bin3 and HybS induce a grid structure in terms of relaxation error and have constant volumes over the resulting grid pieces. While the grid structure for

HybS is obvious, we have yet to show it for Bin2 and Bin3. From [5, Table 4], we further know that for $L, L_1 \to \infty$ the z-values in the projected LP relaxation of Bin2 (18) are bounded from below by the convex function $C_2^L : [\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}] \to \mathbb{R}$ and from above by the concave function $C_2^U : [\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}] \to \mathbb{R}$,

$$C_2^L(x, y) = \frac{1}{2}((x + y)^2 - (\bar{x} + \underline{x})x + \bar{x}\underline{x} - (\bar{y} + \underline{y})y + \bar{y}\underline{y}), \tag{22}$$

$$C_2^U(x, y) = \frac{1}{2}((\underline{x} + \bar{x} + \underline{y} + \bar{y})(x + y) - (\underline{x} + \underline{y})(\bar{x} + \bar{y}) - x^2 - y^2). \tag{23}$$

The same holds for Bin3 (19) and the convex and concave functions $C_3^L : [\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}] \to \mathbb{R}$ and $C_3^U : [\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}] \to \mathbb{R}$,

$$C_3^L(x, y) = \frac{1}{2}(x^2 + y^2 - (\bar{x} + \underline{x} - \bar{y} - \underline{y})(x - y) + (\bar{x} - \bar{y})(\bar{x} - \underline{y})), \tag{24}$$

$$C_3^U(x, y) = \frac{1}{2}((\underline{x} + \bar{x})x - \underline{x}\bar{x} + (\underline{y} + \bar{y})y - \underline{y}\bar{y} - (x - y)^2). \tag{25}$$

As the upper bound on the z-value in HybS is the same as that for Bin2 and the lower bound is the same as that for Bin3, the respective projected LP relaxations P^{LP}_{L,L_1} in the limit for Bin2, Bin3 and HybS are

$$\text{[Bin2]:} \quad \lim_{L,L_1\to\infty}(\text{proj}_{x,y,z}(P^{\text{LP}}_{L,L_1})) = \{(x, y, z) \in [0, 1]^2 \times \mathbb{R} : C_2^L(x, y) \leqslant z \leqslant C_2^U(x, y)\}, \tag{26}$$

$$\text{[Bin3]:} \quad \lim_{L,L_1\to\infty}(\text{proj}_{x,y,z}(P^{\text{LP}}_{L,L_1})) = \{(x, y, z) \in [0, 1]^2 \times \mathbb{R} : C_3^L(x, y) \leqslant z \leqslant C_3^U(x, y)\}, \tag{27}$$

$$\text{[HybS]:} \quad \lim_{L,L_1\to\infty}(\text{proj}_{x,y,z}(P^{\text{LP}}_{L,L_1})) = \{(x, y, z) \in [0, 1]^2 \times \mathbb{R} : C_3^L(x, y) \leqslant z \leqslant C_2^U(x, y)\}. \tag{28}$$

In the following discussion, we will let $L_1 \to \infty$ in all three formulations. This simplifies the proofs considerably and is relevant in so far as in our computations we use a relatively high value of $L_1 = 10$, which has a resulting maximum error below the standard accuracy of state-of-the-art MIP solvers (10^{-6}) and yet has no influence on the number of binary variables and uses only $O(L_1)$ constraints. Although for different values of L_1 the volumes are different, the hierarchy of MIP relaxations that we establish is independent of this choice. We start with the volume of the MIP relaxation HybS.

Proposition 6 *Let* $P^{\text{IP}}_{(L_x,L_y),L_1}$ *be the MIP relaxation* HybS *from (21) without the McCormick inequalities, where we now allow for independent discretization depths* L_x *and* L_y *to overestimate* x^2 *and* y^2*, respectively (i.e. with* $(x, z_x) \in R^{L_x,L_1}$ *and*

$(y, z_y) \in R^{L_y, L_1}$*), i.e.*

$$\begin{gathered} p_1 = x + y, \quad p_2 = x - y \\ (x, z_x) \in R^{L_x, L_1} \ (y, z_y) \in R^{L_y, L_1} \\ (p_1, z_{p_1}), \ (p_2, z_{p_2}) \in Q^{L_1} \\ 1/2(z_{p_1} - z_x - z_y) \leqslant z \leqslant 1/2(z_x + z_y - z_{p_2}) \\ x, y \in [0, 1], \quad p_1 \in [0, 2], \quad p_2 \in [-1, 1]. \end{gathered}$$

Then the volume of $P^{\mathrm{IP}}_{(L_x, L_y), L_1}$ *converges to the same value over each grid piece of the form* $[k_x 2^{-L_x}, (k_x + 1)2^{-L_x}] \times [k_y 2^{-L_y}, (k_y + 1)2^{-L_y}]$*, where* $k_x \in [\![0, 2^{L_x}]\!]$ *and* $k_y \in [\![0, 2^{L_y}]\!]$ *for* $L_1 \to \infty$*. Furthermore, for the total volume of* $P^{\mathrm{IP}}_{(L_x, L_y), L_1}$*, we have*

$$\lim_{L_1 \to \infty} \mathrm{vol}\left(\mathrm{proj}_{x,y,z}(P^{\mathrm{IP}}_{(L_x, L_y), L_1})\right) = \tfrac{1}{6}(2^{-2L_x} + 2^{-2L_y}).$$

Proof Since $F^{L_1} \to x^2$ uniformly over $[0, 1]$ as $L_1 \to \infty$, we have

$$\begin{aligned} &\lim_{L_1 \to \infty} \{(p, z_p) \in [0, 1] \times \mathbb{R} : (p, z_p) \in Q^{L_1}\} \\ &\quad = \{(p, z_p) \in [0, 1] \times \mathbb{R} : (p, z_p) \in \mathrm{epi}_{[0,1]}(p^2)\} \end{aligned}$$

under Hausdorff distance. In HybS, we have $(p_1, z_{p_1}), (p_2, z_{p_2}) \in Q^{L_1}$ (via the transformation in Remark 1) as well as $p_1 = x + y$ and $p_2 = x - y$. Thus, we have in the limit, as $L_1 \to \infty$:

$$z_{p_1} \geqslant (x + y)^2 \text{ and } z_{p_2} \geqslant (x - y)^2.$$

Furthermore, since $F^L(x) \geqslant x^2$ for all $x \in [0, 1]$, $L \in \{L_x, L_y\}$, and $(x, z_x) \in R^{L_x, L_1}$, $(y, z_y) \in R^{L_y, L_1}$, we obtain

$$z_x \leqslant F^{L_x}(x) \text{ and } z_y \leqslant F^{L_y}(y).$$

Therefore, the inequality

$$1/2(z_{p_1} - z_x - z_y) \leqslant z \leqslant 1/2(z_x + z_y - z_{p_2})$$

from (21) implies the following in the limit:

$$1/2((x + y)^2 - F^{L_x}(x) - F^{L_y}(y)) \leqslant z \leqslant 1/2(F^{L_x}(x) + F^{L_y}(y) - (x - y)^2).$$

Now we apply these inequalities to grid pieces of the form $[\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}]$. Let $\underline{x} := k_x 2^{-L_x}$, $\bar{x} := (k_x + 1)2^{-L_x}$, $\underline{y} := k_y 2^{-L_y}$ and $\bar{y} := (k_y + 1)2^{-L_y}$, and define $w_x := \bar{x} - \underline{x} = 2^{-L_x}$ as well as $w_y := \bar{y} - \underline{y} = 2^{-L_y}$. Then, as $F^{L_x}(x) = -(\bar{x} + \underline{x})x + \bar{x}\underline{x}$ for

$x \in [\underline{x}, \bar{x}]$ and $F^{L_x}(y) = -(\bar{y} + \underline{y})y + \bar{y}\underline{y}$ for $y \in [\underline{y}, \bar{y}]$, the above bounds on z are exactly the envelopes $C_2^L(x, y)$ for the lower bound and $C_3^U(x, y)$ for the upper bound, respectively. Thus, by Proposition 11, which is proved later, the volume of $\text{proj}_{x,y,z}(P^{\text{IP}}_{(L_x,L_y),L_1})$ over the grid piece is

$$\tfrac{1}{6}(w_x w_y^3 + w_y w_x^3) = \tfrac{1}{6}2^{-(L_x+L_y)}(2^{-2L_x} + 2^{-2L_y})$$

in the limit. Note that this does not depend on the choice of k_x and k_y (and thus the choice of grid piece).

Since we have $2^{L_x L_y}$ grid pieces overall, the total volume in the limit is then given by

$$\begin{aligned}\lim_{L_1\to\infty} \text{vol}(\text{proj}_{x,y,z}(P^{\text{IP}}_{(L_x,L_y),L_1})) &= 2^{L_x L_y} 2^{-(L_x+L_y)}(2^{-2L_x} + 2^{-2L_y}) \\ &= \tfrac{1}{6}(2^{-2L_x} + 2^{-2L_y}).\end{aligned}$$

which finishes the proof. □

The following proposition establishes the volumes of the MIP relaxations and grid structure for the MIP relaxations Bin2 and Bin3. As this derivation is extensive, we prove it in Appendix C.

Proposition 7 *Let P^{IP}_{L,L_1} be either the MIP relaxation* Bin2 *from* (18) *or* Bin3 *from* (19). *Then the volume of P^{IP}_{L,L_1} converges to the same value over each grid piece of the form $[k2^{-(L-1)}, (k+1)2^{-(L-1)}] \times [k2^{-(L-1)}, (k+1)2^{-(L-1)}]$, where $k \in [\![0, 2^L]\!]$. Furthermore, for the total volume we have*

$$\lim_{L_1\to\infty} \text{vol}\left(\text{proj}_{x,y,z}(P^{\text{IP}}_{L,L_1})\right) = \tfrac{1}{2}2^{-2L}.$$

Now that we have calculated the average error, i.e. the volume of the MIP relaxations, for uniform breakpoints, we show that among all possible breakpoint choices, uniform placement of breakpoints minimizes the average error. For $z = x^2$ and the sawtooth functions, this has already been shown in [7]; for equations $z = xy$ it still has to be shown. We prove average error minimization for uniform breakpoint placement in HybS and do not consider the formulations Bin2 and Bin3 here, as they are hard to analyze in this respect, which is also mentioned in [5] for approximations. In Proposition 6, we have shown that HybS has a grid structure where on each grid piece, the average error is $\frac{1}{6}(w_x w_y^3 + w_y w_x^3)$, where w_x and w_y are the widths of the grid piece in x- and y-direction respectively. In the following, we consider a piecewise relaxation defined via these grid pieces and show that the total average error is minimized by a uniform breakpoint placement, as is the result of HybS.

Proposition 8 *Let $0 = x_0 < x_1 < \ldots < x_n = 1$ and $0 = y_0 < y_1 < \ldots < y_m = 1$ be sets of breakpoints. For each grid piece $[x_{i-1}, x_i] \times [y_{j-1}, y_j]$, consider a relaxation of $\text{gra}_{[0,1]^2}(xy)$ with average error $\frac{1}{6}(w_{x_i} w_{y_j}^3 + w_{y_j} w_{x_i}^3)$, where $w_{x_i} := x_i - x_{i-1}$ and $w_{y_j} := y_j - y_{j-1}$ are the widths of the grid piece with $i \in [\![n]\!]$ and $j \in [\![m]\!]$. Then a*

uniform spacing of these breakpoints minimizes the average error overall piecewise relaxations of this form.

Proof The problem of minimizing the average error of a piecewise relaxation of this form can be formulated as

$$\begin{aligned} \min\ & \tfrac{1}{6}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{m}(w_{x_i}w_{y_j}^3 + w_{y_j}w_{x_i}^3) \\ \text{s.t. } & \textstyle\sum_{i=1}^{n} w_{x_i} = 1 \\ & \textstyle\sum_{j=1}^{m} w_{y_j} = 1 \\ & w_{x_i} \geqslant 0 \qquad\qquad i = 1,\ldots,n \\ & w_{y_j} \geqslant 0 \qquad\qquad j = 1,\ldots,m. \end{aligned} \tag{29}$$

The objective function in (29) sums the average errors over the single grid pieces while the constraints ensure that all single grid lengths sum up to 1 and are greater than or equal to 0. The objective function can be rewritten to

$$\begin{aligned} & \tfrac{1}{6}\sum_{i=1}^{n}\sum_{j=1}^{m}(w_{x_i}w_{y_j}^3 + w_{y_j}w_{x_i}^3) \\ & = \tfrac{1}{6}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}(w_{x_i}w_{y_j}^3) + \sum_{i=1}^{n}\sum_{j=1}^{m}(w_{y_j}w_{x_i}^3)\right) \\ & = \tfrac{1}{6}\left(\sum_{i=1}^{n} w_{x_i}\sum_{j=1}^{m} w_{y_j}^3 + \sum_{j=1}^{m} w_{y_j}\sum_{i=1}^{n} w_{x_i}^3\right) \\ & = \tfrac{1}{6}\left(1\cdot\sum_{j=1}^{m} w_{y_j}^3 + 1\cdot\sum_{i=1}^{n} w_{x_i}^3\right) \\ & = \tfrac{1}{6}\sum_{j=1}^{m} w_{y_j}^3 + \tfrac{1}{6}\sum_{i=1}^{n} w_{x_i}^3. \end{aligned}$$

Thus, (29) decomposes into two independent problems where the respective optimal solutions $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$, can be composed to create $(\boldsymbol{x}^*, \boldsymbol{y}^*)$, which is optimal for the original problem (29). The subproblems are

$$\begin{aligned} \min\ & \textstyle\sum_{i=1}^{n} w_{x_i}^3 \\ \text{s.t. } & \textstyle\sum_{i=1}^{n} w_{x_i} = 1 \\ & w_{x_i} \geqslant 0 \qquad i = 1,\ldots,n \end{aligned} \tag{30}$$

and

$$\begin{aligned} \min\ & \textstyle\sum_{j=1}^{m} w_{y_j}^3 \\ \text{s.t. } & \textstyle\sum_{j=1}^{m} w_{y_j} = 1 \\ & w_{y_j} \geqslant 0 \qquad j = 1,\ldots,m. \end{aligned} \tag{31}$$

These are exactly the sawtooth-area optimization problems from [7, Proposition 5], such that a uniform placement of the breakpoints where each $w_{x_i} = \frac{1}{n}$ is optimal for (30), and $w_{y_j} = \frac{1}{m}$ is optimal for (31). Consequently, a uniform placement of grid points is optimal for (29) and the total volume is $\frac{1}{6}(\frac{1}{m^2} + \frac{1}{n^2})$. □

Remark 6 Let $P^{\text{IP}}_{L,L}$ be a depth-L HybS MIP relaxation of $\text{gra}_{[0,1]^2}(xy)$ from (21), with $L = L_1$. Since $P^{\text{IP}}_{L,L}$ satisfies the uniform spacing of breakpoints discussed in Proposition 8, we see that $P^{\text{IP}}_{L,L}$ is an optimal piecewise relaxation in the sense of minimizing the average error, attaining the average error of $\mathcal{E}^{\text{avg}}(P^{\text{IP}}_{L,L}, \text{gra}_{[0,1]^2}(xy)) = \frac{1}{3}2^{-2L}$. ◇

5.3 Formulation strength

In the previous section, we discussed the maximum and average errors incurred from using certain discretizations. We will now consider the strength of the resulting MIP relaxations by analyzing their LP relaxation. First, we will check for sharpness and later compare them via the volume of the projected LP relaxation. Sharpness means that the projected LP relaxation equals the convex hull of the set to be formulated. If we now consider the volume of a projected LP relaxation, it can minimally be the volume of the convex hull, which precisely holds if the formulation is sharp. If a formulation is not sharp, the volume of the projected LP relaxation measures how much a formulation deviates from sharpness. The volume of LP relaxation as a measure of formulation strength was previously used in [5].

5.3.1 Sharpness

We start with the core formulations from Sect. 3. It is well known that the McCormick relaxation yields the convex hull of the feasible set of $z = xy$ over box domains. Therefore, it is obviously sharp. In [7], it is shown that the sawtooth approximation for $z = x^2$ is sharp. We use this result to prove that sharpness also holds for the tightened sawtooth relaxation (16). See Fig. 4 for examples of this relaxation under different parameter choices.

Theorem 1 (Sharpness of the tightened sawtooth relaxation) *Consider the tightened sawtooth relaxation P^{IP}_{L,L_1} described in (16) in the space of $(x, z, \boldsymbol{g}, \boldsymbol{\alpha})$ for $L, L_1 \in \mathbb{N}$ with $L \leqslant L_1$. The MIP relaxation P^{IP}_{L,L_1} is sharp.*

Proof sketch Define

$$P^{\text{IP}+}_{L,L_1} := \{(x, z, \boldsymbol{g}, \boldsymbol{\alpha}) \in [0,1] \times \mathbb{R} \times [0,1]^{L_1+1} \times \{0,1\}^L : (17b, 17c, 17a),$$
$$P^{\text{IP}-}_{L,L_1} := \{(x, z, \boldsymbol{g}, \boldsymbol{\alpha}) \in [0,1] \times \mathbb{R} \times [0,1]^{L_1+1} \times \{0,1\}^L : (17b, 17c, 17d, 17f)\}.$$

Then P^{IP}_{L,L_1} is sharp if and only if both $P^{\text{IP}+}_{L,L_1}$ and $P^{\text{IP}-}_{L,L_1}$ are sharp. This simplification holds since $P^{\text{IP}}_{L,L_1} = P^{\text{IP}+}_{L,L_1} \cap P^{\text{IP}-}_{L,L_1}$ and since the upper bound $P^{\text{IP}+}_{L,L_1}$ strictly overestimates x^2, while the lower bound, $P^{\text{IP}-}_{L,L_1}$ strictly underestimates x^2, such that sharpness of the two can be considered separately.

Now, the sharpness of $P^{\text{IP}+}_{L,L_1}$ follows directly from the sharpness of the sawtooth approximation (10), which holds by [7, Theorem 1]. For the sharpness of $P^{\text{IP}-}_{L,L_1}$, the proof closely follows the proof of sharpness in [7, Theorem 1], except that, after choosing some fixed $x \in [0, 1]$, we frame the contradiction as follows:

1. Choose $\boldsymbol{g}^*$ as in [7, Theorem 1], and choose the minimum possible value of z^* given $\boldsymbol{g}^*$, such that z^* attains one of its lower bounds.
2. Observe that the chosen solution admits a feasible solution in P^{IP}_{L,L_1}, such that if it is minimal in the LP, then we are done.
3. Suppose for a contradiction that there exists a better z-minimal solution $(\hat{z}, \hat{\boldsymbol{g}})$ than the proposed solution $(z^*, \boldsymbol{g}^*)$, such that some incident lower bound must have been improved.
4. Observe that the improved incident lower bound must be of the form $z \geqslant f^j(x, \boldsymbol{g}^*) - 2^{-2L-2}$ for some $j \geqslant 0$, as the lower bounds 0 and $2x - 1$ do not change with the choice of g^*. Thus, $f^j(x, \boldsymbol{g}^*) - 2^{-2L-2} \geqslant f^j(x, \hat{\boldsymbol{g}}) - 2^{-2L-2}$
5. Show that $f^j(x, \hat{\boldsymbol{g}}) - f^j(x, \boldsymbol{g}^*) < 0$, a contradiction on the choice of $(\hat{y}, \hat{\boldsymbol{g}})$. Thus, the solution $(z^*, \boldsymbol{g}^*)$ was optimal to begin with, and therefore sharpness must hold.

The proof that $f^j(x, \boldsymbol{g}^*) - f^j(x, \boldsymbol{g}^*) < 0$ follows in exactly the same manner as [7, Theorem 1] and is thus omitted here. □

In [7], besides sharpness, it is further shown that the sawtooth approximation is also hereditarily sharp. The following theorem states that the same is true for the tightened sawtooth relaxation (16) and $z = x^2$.

Theorem 2 *The tightened sawtooth relaxation for $z = x^2$ is hereditarily sharp.*

As the proof of Theorem 2 takes up a significant amount of space, we moved it to Appendix B.

Next, we show that neither of the MIP relaxations Bin2, Bin3 nor HybS for $z = xy$ are sharp. That is, their projected LP relaxation does not equal $\mathcal{M}(x, y)$ for any $L, L_1 \in \mathbb{N}$. Note that we have included the McCormick inequalities in the definitions of Bin2, Bin3 and HybS to make the formulations stronger. The following proofs, however, refer to the fact that if one omits the McCormick inequalities in these formulations, then they are not sharp. Together with the McCormick inequalities, of course, they are sharp trivially.

Proposition 9 *Let P^{IP}_{L,L_1} be the MIP relaxation* HybS *for $z = xy$ stated in* (21)*. Then, without the inequalities from the McCormick envelope $\mathcal{M}(x, y)$, P^{IP}_{L,L_1} is not sharp for any $L, L_1 \in \mathbb{N}$.*

Proof Without the McCormick envelope, the HybS MIP relaxation P^{IP}_{L,L_1}, and its LP-relaxation P^{LP}_{L,L_1}, become strictly tighter as either L or L_1 increases. Thus, we have

$$\text{proj}_{x,y,z}(P^{\text{LP}}_{L,L_1}) \supseteq \lim_{L,L_1 \to \infty} \text{proj}_{x,y,z}(P^{\text{LP}}_{L,L_1})$$

and

$$\text{conv}(\text{proj}_{x,y,z}(P^{\text{IP}}_{1,1})) \supseteq \text{conv}(\text{proj}_{x,y,z}(P^{\text{IP}}_{L,L_1})) \quad \text{for any } L, L_1 \in \mathbb{N}.$$

We now show $\left(\lim_{L,L_1\to\infty} \text{proj}_{x,y,z}(P^{\text{LP}}_{L,L_1})\right) \setminus \text{conv}(\text{proj}_{x,y,z}(P^{\text{IP}}_{1,1})) \neq \emptyset$, which implies $\text{proj}_{x,y,z}(P^{\text{LP}}_{L,L_1}) \setminus \text{conv}(\text{proj}_{x,y,z}(P^{\text{IP}}_{L,L_1})) \neq \emptyset$, such that P^{IP}_{L,L_1} is not sharp for any $L, L_1 \in \mathbb{N}$. The argument works in the following manner:

$$\begin{aligned}
&\text{proj}_{x,y,z}(P^{\text{LP}}_{L,L_1}) \setminus \text{conv}(\text{proj}_{x,y,z}(P^{\text{IP}}_{L,L_1})) \\
&\supseteq \left(\lim_{L,L_1\to\infty} \text{proj}_{x,y,z}(P^{\text{LP}}_{L,L_1})\right) \setminus \text{conv}(\text{proj}_{x,y,z}(P^{\text{IP}}_{1,1})) \neq \emptyset \\
&\Rightarrow \text{proj}_{x,y,z}(P^{\text{LP}}_{L,L_1}) \setminus \text{conv}(\text{proj}_{x,y,z}(P^{\text{IP}}_{L,L_1})) \neq \emptyset \\
&\Rightarrow \text{proj}_{x,y,z}(P^{\text{LP}}_{L,L_1}) \neq \text{conv}(\text{proj}_{x,y,z}(P^{\text{IP}}_{L,L_1})).
\end{aligned}$$

To this end, we show that there exist points $(x, y, z) \in \lim_{L,L_1\to\infty} \text{proj}_{x,y,z}(P^{\text{LP}}_{L,L_1})$ with $(x, y, z) \notin \text{proj}_{x,y,z}(P^{\text{IP}}_{1,1})$. Observe that, for any L, the point (x, x) is feasible within the LP relaxation of the tightened sawtooth relaxation (16) for x^2, with $\alpha_i = g_{i-1}$, $g_i = 0$. Thus, for all $L, L_1 \geqslant 0$ and for all $\hat{x}, \hat{y} \in [0, 1]^2$, we have that P^{LP}_{L,L_1}, and thus also its limit $\lim_{L,L_1\to\infty} \text{proj}_{x,y,z}(P^{\text{LP}}_{L,L_1})$, admits the values $z_x = \hat{x}$, $z_y = \hat{y}$ and $z_{p_1} = (\hat{x} + \hat{y})^2$. Therefore, for $(x, y) = (0, \frac{1}{4})$, we obtain

$$z = \tfrac{1}{2}((x + y)^2 - x - y) = -\tfrac{3}{16},$$

such that $(0, \frac{1}{4}, -\frac{3}{16}) \in P^{\text{LP}}_{\infty,\infty}$.

Next, in order to prove $(0, \frac{1}{4}, -\frac{3}{16}) \notin \text{conv}(\text{proj}_{x,y,z}(P^{\text{IP}}_{1,1}))$, we show $\min\{z : (y, z) \in \text{proj}_{y,z}(P^{\text{IP}}_{1,1}|_{x=0})\} = -\frac{1}{8}$. If this holds, then we have $\min\{z : (y, z) \in \text{conv}(\text{proj}_{y,z}(P^{\text{IP}}_{1,1}|_{x=0}))\} = -\frac{1}{8}$, such that $(0, \frac{1}{4}, -\frac{3}{16}) \notin \text{conv}(\text{proj}_{x,y,z}(P^{\text{IP}}_{1,1}))$. We derive a representation of $\text{proj}_{y,z}(P^{\text{IP}}_{1,1}|_{x=0})$ that becomes an LP after branching spatially at $y = \frac{1}{2}$ to resolve the upper bound on z_y. We then minimize z over both branches via solving an MIP.

Let $x = 0$. Then the bounds on z, z_x, z_y, z_{p_1} within $\text{proj}_{x,y,z}(P^{\text{IP}}_{1,1})$ are

$$\begin{aligned}
z_x &\leqslant 0, \quad z_y \leqslant y - \tfrac{1}{4}\min\{2y, 2(1 - y)\} = \max\{\tfrac{y}{2}, \tfrac{3y-1}{2}\} \\
z_{p_1} &\geqslant 4\left(\tfrac{y}{2} - \tfrac{1}{4}\min\{2\tfrac{y}{2}, 2(1 - \tfrac{y}{2}) - \tfrac{1}{16}\}\right) = \max\{y - \tfrac{1}{4}, 3y - \tfrac{9}{4}\} \\
z_{p_1} &\geqslant 4(\tfrac{y}{2} - \tfrac{1}{4}) = 2y - 1 \\
z_{p_1} &\geqslant 0 \\
z_{p_1} &\geqslant 4(2\tfrac{y}{2} - 1) = 4(y - 1) \\
z &\geqslant z_{p_1} - z_x - z_y \\
y &\in [0, 1].
\end{aligned}$$

Note that the two pieces of the upper bound on z_y meet at $y = \frac{1}{2}$. Using this to separately minimize z over the above set, once over $y \in [0, \frac{1}{2}]$ and once over $y \in [\frac{1}{2}, 1]$, e.g. using an MIQCQP solver, we obtain two globally minimizing solutions with $z = -\frac{1}{8}$, namely at $y = \frac{1}{4}$ and at $y = \frac{3}{4}$. Thus, we conclude that $(0, \frac{1}{4}, -\frac{3}{16}) \notin \text{conv}(\text{proj}_{x,y,z}(P^{\text{IP}}_{1,1}))$, such that P^{IP}_{L,L_1} is not sharp for any $1 \leqslant L \leqslant L_1$. □

Proposition 10 *Let P^{IP}_{L,L_1} be either of the two MIP relaxations* Bin2(18) *or* Bin3(19). *Then, without the inequalities from the McCormick envelope $\mathcal{M}(x, y)$, P^{IP}_{L,L_1} is not sharp for any $L, L_1 \in \mathbb{N}$.*

Proof Since Bin2 (18) has the same lower-bounding constraints as HybS, the proof follows directly from Proposition 9. Moreover, for Bin3 (19), the proof follows in exactly the same way as the proof of Proposition 9, except for the upper-bounding version of the same point, $(x, y, z) = (0, \frac{1}{4}, \frac{3}{8})$, and acting on the upper-bounding constraints from (21) and maximizing z instead. As the proof is very similar, with the corresponding upper bound $z = \frac{1}{8}$ on $\text{proj}_{y,z}(P^{\text{IP}}_{1,1}|_{x=0})$, we omit it here. □

5.3.2 LP relaxation volume

Having proved that none of the separable MIP relaxations is sharp, which implies that they are also not hereditarily sharp, we now turn to consider the volume of projected LP relaxations.

For $L = L_1$, the volume for the tightened sawtooth formulation (7) is $\frac{3}{16}2^{-2L}$, which has been shown in [7]. For general L_1, by integrating over the overapproximation and underapproximation errors separately with the same analysis as in [7], we can derive a general volume of $\frac{1}{6}2^{-2L} + \frac{1}{48}2^{-2L_1}$. We omit the precise calculation here.

In our analysis of the separable MIP relaxations, we only consider the limits for $L, L_1 \to \infty$. This allows us to evaluate the volumes independently of the underlying discretizations. For the additional volumes resulting from discretization errors, we refer to [7, Appendix], where the volume over the error function of the sawtooth approximation is given. We start with HybS.

Proposition 11 *Let P^{LP}_{L,L_1} be the LP relaxation of the MIP relaxation* HybS *stated in* (21) *over the general domain $[\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}]$. Without the McCormick envelope constraints, the volume of the limit of the projected LP relaxation $\lim_{L,L_1\to\infty} \text{proj}_{x,y,z}(P^{\text{LP}}_{L,L_1})$ is $\frac{1}{6}(w_x w_y^3 + w_y w_x^3)$, where $w_x = \bar{x} - \underline{x}$ and $w_y = \bar{y} - \underline{y}$.*

Proof The z-values in the projected LP relaxation of (21) are bounded by the convex function C_2^L and the concave function C_3^U, which are stated above in (22) and (25), respectively. The volume of the projected LP relaxation (21) is then calculated via integration:

$$\int_{\underline{x}}^{\bar{x}} \int_{\underline{y}}^{\bar{y}} (C_3^U(x, y) - C_2^L(x, y))dydx = \tfrac{1}{6}(w_x w_y^3 + w_y w_x^3).$$

□

Proposition 12 *Let P^{LP}_{L,L_1} be the LP relaxation of either the MIP relaxation* Bin2 *or* Bin3 *stated in* (18) *and* (19) *over the domain $[\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}]$. Without the McCormick envelope constraints, the volume of the limit of the projected LP relaxation is*

$$\lim_{L,L_1\to\infty} \text{vol}(\text{proj}_{x,y,z}(P^{\text{LP}}_{L,L_1})) = \frac{1}{12} w_x w_y (2w_x^2 + 3w_x w_y + 2w_y^2),$$

where $w_x = \bar{x} - \underline{x}$ and $w_y = \bar{y} - \underline{y}$.

Proof The z-values in the projected LP relaxation of (18) and (19) are bounded by the convex function C_2^L and the concave function C_3^U, which are stated above in (22) and (25), respectively. The volume calculation is then done via integration:

$$\int_{\underline{x}}^{\bar{x}} \int_{\underline{y}}^{\bar{y}} (C_3^U(x,y) - C_3^L(x,y))dydx = \int_{\underline{x}}^{\bar{x}} \int_{\underline{y}}^{\bar{y}} (C_2^U(x,y) - C_2^L(x,y))dydx$$
$$= \frac{1}{12} w_x w_y (2w_x^2 + 3w_x w_y + 2w_y^2).$$

□

We use Proposition 11 and Proposition 12 to prove that HybS yields strictly tighter LP relaxations than Bin2 and Bin3.

Proposition 13 *Without the McCormick envelope constraints, the LP relaxation of the MIP relaxation* HybS *in the limit as* $L, L_1 \to \infty$ *is strictly tighter than that of* Bin2 *or* Bin3. *Moreover, the volume of the projected LP relaxation of formulation* HybS *in the limit as* $L, L_1 \to \infty$ *is smaller by* $\frac{1}{4} w_x^2 w_y^2$.

Proof In [5, Appendix, Proposition 2] it has been shown that C_2^L is a tighter convex underestimator than C_3^L and that C_3^U is a tighter concave overestimator than C_2^U for $z = xy$. Thus, since the HybS approach converges to C_2^L as an underestimator and C_3^L as an overestimator, it is strictly tighter than either of Bin2 or Bin3. The volume calculation can again be done via integration:

$$\int_{\underline{x}}^{\bar{x}} \int_{\underline{y}}^{\bar{y}} (C_2^U(x,y) - C_2^L(x,y))dydx - \int_{\underline{x}}^{\bar{x}} \int_{\underline{y}}^{\bar{y}} (C_3^U(x,y) - C_2^L(x,y))dydx$$
$$= \int_{\underline{x}}^{\bar{x}} \int_{\underline{y}}^{\bar{y}} (C_3^U(x,y) - C_3^L(x,y))dydx - \int_{\underline{x}}^{\bar{x}} \int_{\underline{y}}^{\bar{y}} (C_3^U(x,y) - C_2^L(x,y))dydx$$
$$= \frac{1}{4} w_x^2 w_y^2 > 0.$$

□

6 Computational results

In the previous sections, we have shown the theoretical advantages of HybS compared to Bin2 and Bin3, most importantly that it requires fewer binary variables to model MIP relaxations of variable products with the same accuracy. As the density of quadratic matrices in MIQCQPs increases, this advantage becomes larger, leading to a maximum of $\mathcal{O}(n)$ binary variables for HybS and $\mathcal{O}(n^2)$ binary variables for Bin2 and Bin3; see Table 1. In general, the number of binary variables of an MIP relaxation is crucial for its solution time. Hence, the theoretical results suggest that the HybS formulation yields MIP relaxations that are faster to solve than the Bin2 and Bin3 relaxations. Consequently, shorter run times or better primal and dual bounds after certain run time

limits can be expected. To analyze these MIP relaxations for $z = xy$, it is preferable to use a model for the x^2 terms that requires as few binaries as possible. Otherwise, the impact of fewer binaries for HybS might not be that noticeable, since the difficulty of the various MIP models might then be more determined by the MIP formulations of the x^2 terms. The sawtooth relaxation does exactly that with its logarithmic number of binary variables. Furthermore, we proved that it is also a hereditary sharp formulation. In the computational study, we first compare both run times and dual bounds of the MIP relaxations. MIP relaxations are primarily used to deliver dual bounds for the MIQCQPs. The best dual bound of an MIP relaxation is then a valid dual bound for the MIQCQP. However, with increasing accuracy of the relaxations, the solution times also increase. Therefore, both the run time (for coarser relaxations) and the best dual bounds (for finer relaxations) are important measures if we want to compare different MIP relaxations with the same accuracy.

Complementary to this, in a second part of the study we investigate to what extent the MIP solutions can serve as a starting point to find feasible solutions to the MIQCQP. A common heuristic approach is to fix any integer variables from the original problem according to the MIP solution and solve the resulting QCQP to local optimality. The starting points of the continuous variables of the original problem again correspond to the values of the MIP solution. As before, our theoretical results imply that the HybS relaxations are generally more likely to find MIP solutions after certain run time limits due to the smaller number of binary variables. Presumably, this translates to a higher probability of finding feasible solutions to the MIQCQP using the heuristic approach. In detail, we solve MIP relaxations using either HybS, Bin2, or Bin3 in combination with the sawtooth relaxation using Gurobi [28] and a callback function that uses the non-linear programming (NLP) solver IPOPT [41] to find local optimal solutions for the QCQP.

All instances were solved in Python 3.8.3, via Gurobi 9.5.1 and IPOPT 3.12.13 on the 'Woody' cluster, using the "Kaby Lake" nodes with two Xeon E3-1240 v6 chips (4 cores, HT disabled), running at 3.7 GHZ with 32 GB of RAM. For more information, see the Woody Cluster Website of Friedrich-Alexander-Universität Erlangen-Nürnberg. The global relative optimality tolerance in Gurobi was set to the default value of 0.01%, for all MIPs and MIQCQPs.

6.1 Study design

In the following, we explain the design of our study and go into detail regarding the instance set as well as the various parameter configurations.

Instances. We consider a three-part benchmark set of 60 instances: 20 non-convex boxQP instances from [7, 17, 22] and earlier works, 20 AC optimal power flow (ACOPF) instances from the NESTA benchmark set (v0.7.0) (see [18]), previously used in [2], and 20 MIQCQP instances from the QPLIB [24]. In Appendix D links that contain download options and detailed descriptions of the instances can be found. For an overview of the IDs of all instances, see Table 8. The benchmark set is equally divided into 30 sparse and 30 dense instances. We call an instance dense if either the objective function and/or at least one quadratic function in the constraint set is of the

Table 2 In the study, we consider the parameters cuts, depth, and formulation on 60 MIQCQP instances and thus solve $(2 \cdot 4) \cdot 3 \cdot 60 = 1440$ MIP relaxations

Depth	Formulation	Instances
$L = 1, 2, 4, 6$	Bin2	boxQP (20 instances)
$L_1 = L$	Bin3	ACOPF (20 instances)
Tightened	HybS	QPLIB (20 instances)
$L = 1, 2, 4, 6$		
$L_1 = \max\{2, 1.5L\}$		

Table 3 Maximum error for different values of L

	HybS	Bin2/Bin3
L = 1	2e−02	3e−02
L = 2	5e−03	8e−03
L = 4	3e−04	5e−04
L = 6	2e−05	3e−05

form $\boldsymbol{x}^\top Q\boldsymbol{x}$, where $x \in \mathbb{R}^n$ are all variables of the problem and $Q \in \mathbb{R}^{n,n}$ is a matrix with at least 25% of its entries being nonzero.

Parameters. For each instance, we solve the resulting MIP relaxation of each method from Sect. 4 using various approximation depths of $L \in \{1, 2, 4, 6\}$ and a time limit of 8 h. In Sect. 6.1, we have listed the maximum errors associated with each L, which are derived from the values in Table 1. All sawtooth and separable MIP relaxations are solved once with $L_1 = L$ and once with a tightened underestimator version for univariate quadratic terms where $L_1 = \max\{2, 1.5L\}$. This tightening is done as described in Definition 7 by adding linear cuts and without introducing further binary variables. In the separable methods HybS, Bin2, and Bin3 this leads to a tightening of the relaxation of $z = xy$ terms as well as of $z = x^2$ terms in the original MIQCQP. We refer to the tightened MIP relaxations as T-HybS, T-Bin2, and T-Bin3. Table 2 gives an overview of the different parameters in our study. In total, we have 24 parameter configurations for 60 original problems, which means that we solve 1440 MIP instances. In Table 3, we list the maximum error in our different models under changing L.

Callback function. Solving all MIP relaxations, we use a callback function with the local NLP solver IPOPT that works as follows: given any MIP-feasible solution, the callback function fixes any integer variables from the original problem (before applying any of the discretization techniques from this work) according to this solution and then solves the resulting QCCP, the original MIQCQP with fixed binaries, locally via IPOPT in an attempt to find a feasible solution for the original MIQCQP problem.

6.2 Number of binaries

In advance of the results of the study, we provide another table that shows, how many binary variables can be saved relatively with HybS compared to Bin2 and Bin3. In

Table 4 Average number of binary variables per instance and the relative percentage of binary variables in HybS models compared to those of Bin2 and Bin3

	Sparse			Dense		
	Bin2/Bin3	HybS	rel (%)	Bin2/Bin3	HybS	rel (%)
L=1	318	231	72.8	987	61	6.2
L=2	579	406	70.2	1972	119	6.1
L=4	1102	756	68.6	3942	236	6.0
L=6	1625	1106	68.0	5912	352	6.0

Table 4 we specify how many variables occur on average with each method in the MIP relaxation models. Apart from a few original variables of the MIQCQPs, the main part of the binary variables comes from the MIP relaxations of quadratic terms. Since Bin2 and Bin3 require exactly the same number of binary variables for each univariate or bivariate MIP relaxation, only Bin2 is listed in Table 4. The table shows that HybS requires close to two-thirds of the binary variables on the sparse instances. The difference is much greater on the dense instances, where HybS requires only nearly 6% of the binary variables of Bin2 and Bin3. Both numbers are in line with our theoretical findings. Assuming, we had an MIQCQP instance with only one variable product $x_i x_j$ and we would set $L = 1$, then there would be three binary variables each for Bin2 and Bin3, while we would need only two for HybS. The fact that this effect is significantly stronger for dense instances stems from the quadratic increase of binary variables in dense matrices for Bin2 and Bin3 compared to the linear increase for HybS.

6.3 Results

In the following, we present the results of our study at a detailed level. In particular, we aim to answer the following questions regarding run times, dual bounds, and the ability to find feasible solutions for the MIQCQPs:

- Is our enhanced method HybS computationally superior to its predecessors Bin2 or Bin3?
- Is it beneficial to use tightened versions of the MIP relaxations HybS, Bin2, and Bin3, i.e., to choose $L_1 > L$?

We point out that in Part II of this work, we also present a more detailed comparison with different MIP relaxation methods and the state-of-art MIQCQP solver Gurobi.

6.3.1 Run times

We start with a discussion on the run times for the different methods. Here, we use the shifted geometric mean, which is a common measure for comparing two different MIP-based solution approaches. The shifted geometric mean of n numbers $t_1, \ldots, t_n$ with shift s is defined as $\left(\prod_{i=1}^{n}(t_i + s)\right)^{1/n} - s$. It has the advantage that it is neither affected by very large outliers (in contrast to the arithmetic mean) nor by very small

Table 5 Shifted geometric mean for run times on all instances

	Bin2	T-Bin2	Bin3	T-Bin3	HybS	T-HybS
All						
$L = 1$	74.62	95.53	74.67	96.69	**31.00**	44.55
$L = 2$	174.87	265.15	271.16	265.70	**67.62**	77.07
$L = 4$	940.70	895.52	754.62	895.13	**172.59**	395.29
$L = 6$	1301.88	1485.40	1104.60	1484.55	**455.38**	859.92
Sparse						
$L = 1$	40.47	42.10	39.59	42.91	**33.66**	48.78
$L = 2$	63.64	81.66	93.12	81.88	**62.65**	66.49
$L = 4$	362.13	367.90	297.24	367.98	**154.53**	253.81
$L = 6$	499.46	602.40	487.41	601.63	**380.29**	441.66
Dense						
$L = 1$	236.27	443.88	245.83	444.68	**26.01**	36.77
$L = 2$	1020.66	2131.53	1818.35	2134.26	**77.82**	100.90
$L = 4$	3872.15	3348.79	2991.87	3344.09	**203.47**	761.74
$L = 6$	4850.41	5137.58	3396.35	5139.58	**583.77**	2145.94

Bold values indicate the best run time in each row

outliers (in contrast to the geometric mean). We use a typical shift $s = 10$. Moreover, we only include those instances in the computation of the shifted geometric mean, where at least one solution method delivered an optimal solution within the run time limit of 8 hours.

In Table 5, the shifted geometric mean values of the run times for solving the separable MIP relaxations on all instances are given. Here, HybS clearly outperforms all other methods, including its tightened variant T-HybS. HybS is at least a factor of two faster than (T-)Bin2 and (T-)Bin3. Tightening HybS, Bin2, and Bin3 results in comparable but slightly higher run times for Bin2 and Bin3 and partially in notably higher run times for HybS, e.g. by a factor of more than two in case of $L = 4$.

For sparse instances, the same picture emerges, although the benefit of HybS is not as great as before, see the second block in Table 5. Conversely, the advantage of HybS increases dramatically for dense instances. Here, HybS is at least a factor of five faster than (T-)Bin2 and (T-)Bin3, see the third block in Table 5. Tightening the three methods again leads to mostly slightly higher run times for Bin2 and Bin3 and to considerably higher run times for HybS.

6.3.2 Dual bounds

As mentioned before, MIP relaxations are primarily used to deliver (tight) dual bounds for MIQCQPs. Thus, we now compare the tightness of the dual bounds provided by the various methods. To this end, we compute relative optimality gaps $g_{p,s} := |d_{p,s} - b_p|/|b_p|$ for all methods s (with a certain L value) and instances p of the benchmark

Table 6 Arithmetic (left) and geometric (right) mean of relative optimality gaps (in %) on all instances for separable MIP relaxations

	BIN2	T-BIN2	BIN3	T-BIN3	HybS	T-HybS
All						
$L = 1$	65.04/8.39	47.32/8.84	46.35/8.35	47.33/8.84	46.13/7.94	**46.04/7.57**
$L = 2$	45.99/7.92	37.35/7.32	36.65/6.67	37.36/7.32	33.07/4.96	**32.33/4.50**
$L = 4$	45.07/4.36	40.86/4.04	35.53/4.24	51.89/4.08	**24.84/1.81**	31.42/1.90
$L = 6$	48.42/2.53	45.53/2.80	41.84/2.75	57.68/2.81	**32.97/1.05**	53.75/1.83
Sparse						
$L = 1$	24.30/14.34	**23.30/13.50**	23.73/13.88	23.30/13.50	23.85/14.01	23.53/13.70
$L = 2$	21.11/11.39	**20.33**/10.44	20.78/10.87	**20.33/10.43**	21.21/11.52	20.39/10.36
$L = 4$	15.18/3.06	**14.90/2.08**	14.92/2.45	14.87/**2.08**	14.93/2.19	15.04/2.13
$L = 6$	11.23/0.93	12.09/0.84	12.41/0.89	12.07/0.83	**10.91/0.72**	11.65/0.74
Dense						
$L = 1$	105.77/4.90	71.34/5.78	68.98/5.03	71.37/5.79	**68.40**/4.50	68.56/**4.19**
$L = 2$	70.88/5.50	54.36/5.13	52.52/4.09	54.40/5.13	44.94/2.14	**44.28/1.96**
$L = 4$	74.97/6.22	66.82/7.84	56.14/7.36	88.92/8.02	**34.76/1.49**	47.80/1.69
$L = 6$	85.61/6.89	78.97/9.34	71.27/8.54	103.28/9.51	**55.04/1.52**	95.86/4.56

Bold values indicate the best run time in each row

set, where $d_{p,s}$ is the corresponding best dual bound found by method s and b_p is the best-known primal bound for instance p.

Table 6 shows the arithmetic and geometric means of the relative optimality gaps for all 60 instances. Please note that we rounded each gap below 0.0001 to avoid multiplications by 0 for the geometric mean. First, the arithmetic mean decreases with higher L values but then starts to increase again. This pattern indicates the presence of more outliers with higher L values, leading to inconsistencies in the arithmetic mean. On the other hand, the geometric mean shows a tendency that with higher L values, we can expect tighter dual bounds for the considered instances. This trend is more consistent and reflects a more balanced view of overall performance. HybS often achieves the lowest geometric mean values, which indicates its superior performance. In summary, the geometric means in Table 6 emphasize the effectiveness of higher L values for tighter dual bounds, with HybS standing out as a particularly strong method based on the considered data. Comparing the tightened versions (T-Bin2, T-Bin3, and T-HybS) with their non-tightened counterparts, the results are mixed. The tightened versions yield similar optimality gaps, with some showing slightly better and others slightly worse performance depending on different L values. However, there is no clear trend, suggesting that there is generally no advantage to tightening the methods.

Dividing the benchmark set into sparse and dense instances, gives a similar picture for dense instances as on the full benchmark set, see the third block in Table 6. However, a different trend can be seen for sparse instances in Table 6. Here, for higher L values, both the arithmetic and geometric means consistently decrease, while HybS again

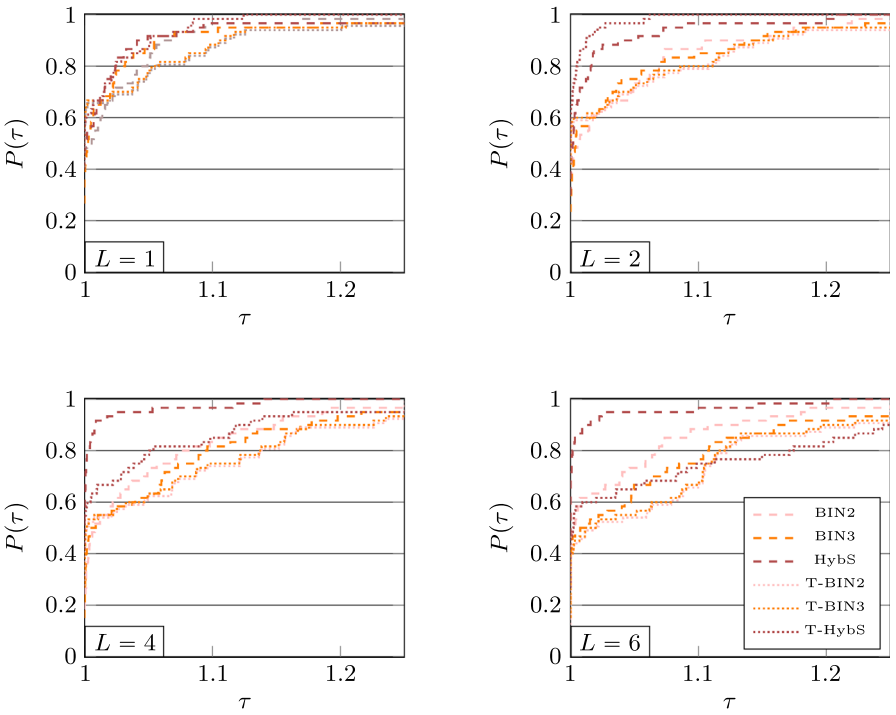


Fig. 7 Performance profiles to dual bounds of separable MIP relaxations on all instances

outperforms Bin2 and Bin3. In contrast to the full benchmark set, the tightening is now slightly beneficial for all three methods.

Additionally, we provide performance profile plots as proposed by Dolan and More [20] to illustrate the scaling of the dual bounds, see Figs. 7, 8 and 9. The intention here is to obtain a more sophisticated picture of how the various methods perform if we allow the dual bounds to lie within a given factor of the best overall dual bound. The performance profiles work as follows: Let $d_{p,s}$ again be the best dual bound obtained by MIP relaxation s for instance p after a certain time limit. With the performance ratio $r_{p,s} := d_{p,s} / \min_s d_{p,s}$, the performance profile function value $P(\tau)$ is the percentage of problems solved by approach s such that the ratios $r_{p,s}$ are within a factor $\tau \in \mathbb{R}$ of the best possible ratios. All performance profiles are generated with the help of *Perprof-py* by Siqueira et al. [38]. In addition to the performance profiles across all instances, we also show performance profiles for the dense and sparse subsets of the instance set. Please note that in minimization problems, the higher the value of a dual bound, the better it is. Since lower values are considered better in performance profiles, we simply take the inverse of the dual bound as the value to be compared.

In Fig. 7 the performance profiles of the separable MIP relaxations with regard to dual bounds using all instances can be seen. Starting with $L = 2$, the newly introduced methods HybS and T-HybS deliver significantly better dual bounds. Except for $L = 2$, where T-HybS dominates HybS, we do not obtain better dual bounds by tightening the separable MIP relaxations. With $L = 4$ and $L = 6$, HybS yields dual bounds that are

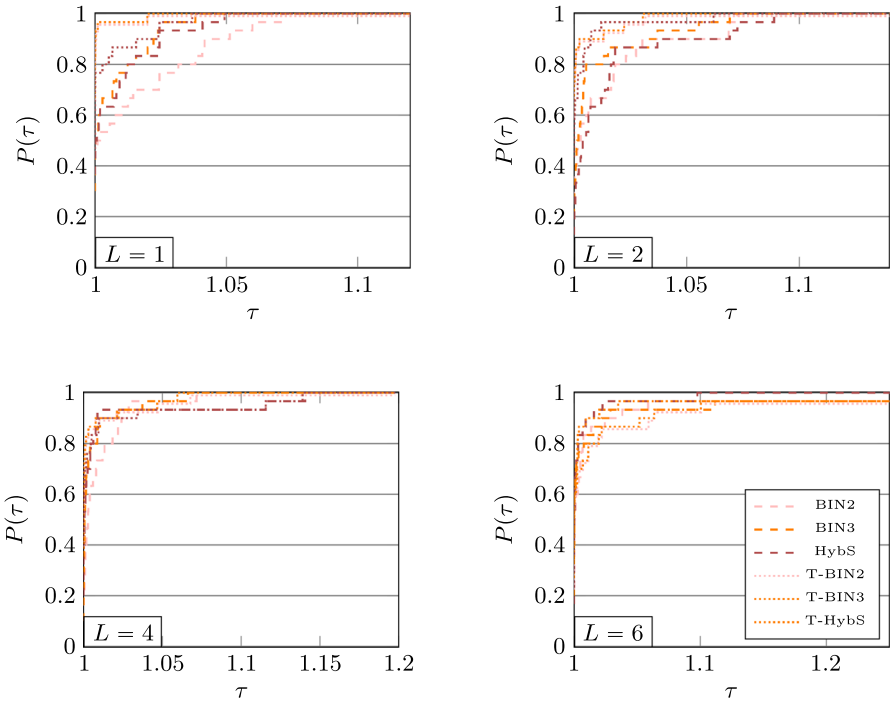


Fig. 8 Performance profiles to dual bounds of separable MIP relaxations on sparse instances

within a factor 1.05 of the overall best bounds among separable MIP relaxations for nearly all instances. The other methods require a corresponding factor of at least 1.2. In Figs. 8 and 9, we divide the benchmark set into sparse and dense instances again to obtain a more in-depth look at the benefits of HybS. For sparse instances, using HybS and T-HybS has no clear advantage, as Fig. 8 shows. However, with $L = 1$ and $L = 2$, the tightened variants deliver notably better dual bounds. For $L = 1$, the dual bounds computed with T-Bin2 and T-Bin3 are in almost all cases the overall best-found bounds. Their counterparts Bin2 and Bin3 are only able to provide the overall best bounds for about 50% of the instances. For $L = 2$, we see a similar picture. T-Bin2 and T-Bin3 deliver the best bounds for roughly 80% of the instances, while Bin2 and Bin3 achieve this only in 40% of the cases.

For dense instances, the picture is much clearer. Here, HybS and T-HybS are considerably better than Bin2, Bin3, and their tightened variants, particularly from $L = 2$ to $L = 6$; see Fig. 9. With $L = 2$, HybS and T-HybS are able to compute dual bounds that are within a factor 1.05 of the overall best bounds for nearly all instances. All other methods require a corresponding factor of more than 1.2. For $L = 4$ and $L = 6$, we obtain by HybS the best overall bounds for roughly 90% of all instances, while all other approaches provide the best bounds for less than 50% of the instances. With the exception of $L = 2$, where tightening HybS results in slightly better dual bounds, the tightened versions of the separable MIP relaxations attain significantly weaker dual bounds than their corresponding counterparts.

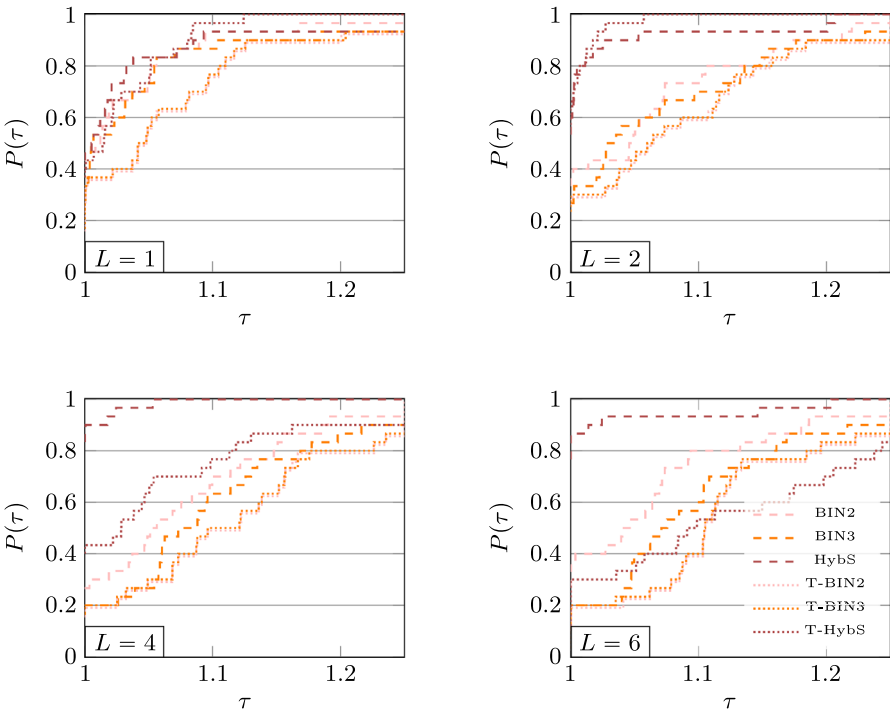


Fig. 9 Performance profiles to dual bounds of separable MIP relaxations on dense instances

6.3.3 Feasible solutions

Finally, we highlight some important results on primal bounds. Table 7 gives the number of feasible solutions that the separable MIP methods were able to find in combination with IPOPT as the local QCQP solver. The quality of the corresponding solutions is computed in terms of relative optimality gaps, where we used the best-known dual bounds from the literature or computed them using Gurobi and our methods. Regarding the ability to find feasible solutions, all separable methods perform quite similarly and find more feasible solutions with higher L values. With $L = 6$, HybS in combination with IPOPT is able to compute feasible solutions to the original MIQCQP for 51 out of 60 benchmark instances, 43 of which have a relative optimality gap below 1% and 40 of which are even globally optimal, i.e., which have a gap below 0.01%. All in all, HybS offers a slight advantage in terms of finding feasible solutions when coupled with IPOPT.

6.4 Discussion

All in all, the clear winner among the separable methods is HybS. For large L values, HybS provides the best bounds, the shortest run times, and finds in combination with IPOPT the most and best feasible solutions for the original MIQCQP instances. This

Table 7 Number of feasible solutions found with different relative optimality gaps

	Bin2	T-Bin2	Bin3	T-Bin3	HybS	T-HybS
$L = 1$	23/29/39	24/31/38	29/**33**/40	24/31/38	**31**/**33**/40	30/**33**/**43**
$L = 2$	28/32/39	**33**/33/38	32/35/43	**33**/33/38	32/**37**/**44**	32/36/42
$L = 4$	39/42/**51**	35/40/48	38/41/49	35/40/48	**41**/**44**/50	38/**44**/49
$L = 6$	**40**/**43**/46	37/42/45	39/42/47	37/42/46	**40**/**43**/**51**	38/**43**/50

Bold values indicate the best run time in each row
The first number corresponds to a gap of less than 0.01%, the second to a gap of less than 1% and the third number indicates the number of feasible solutions

advantage is especially noticeable on dense instances and consistent with the theoretical findings from Sect. 5. While in HybS the number of binary variables increases linearly in the number of variable products, it increases quadratically in Bin2 and Bin3. On the one hand, this results in short run times for the HybS models or better bounds after certain run time limits. On the other hand, with significantly fewer binaries we are more likely to find feasible solutions for the MIP relaxations. As the accuracy increases, the MIP relaxations lead to solutions with smaller and smaller MIQCQP feasibility violations. Therefore, at higher L values, we are more likely to find an MIQCQP feasible solution using the heuristic IPOPT approach, which coincides with Table 7.

Furthermore, based on the computational results, a tightening of the separable methods is not advisable, except for sparse instances with small L values. This is most likely due to the large number of additional constraints that are needed to underestimate p_1^2 and p_2^2; see Table 1.

In Part II of this work, we revisit the idea of tightening MIP relaxations for the *normalized multiparametric disaggregation technique* (NMDT) introduced in [13]. In addition, we perform a comparison of HybS with NMDT-based methods and Gurobi as an MIQCQP solver. To this end, we reuse the results of HybS from Part I.

7 Conclusion

We introduced an enhanced MIP relaxation for non-convex quadratic products of the form $z = xy$, called *hybrid separable* (HybS). We showed that HybS has clear theoretical advantages over its predecessors Bin2 and Bin3, all based on separable reformulation of xy to univariate quadratic terms. Most importantly, HybS requires a significantly lower number of binary variables and has a tighter linear programming relaxation. In addition to this enhanced MIP relaxation for $z = xy$, we introduced a hereditary sharp MIP relaxation called *sawtooth relaxation* for $z = x^2$ terms, which requires only a logarithmic number of binary variables with respect to the relaxation error. We combined the sawtooth relaxation and HybS to obtain MIP relaxations for MIQCQPs.

In a broad computational study, we compared HybS against its predecessors from the literature, which we again combined with the sawtooth relaxation for univariate

quadratic terms. We showed that HybS determines far better dual bounds, while also exhibiting shorter run times. Finally, HybS is also able to find high-quality solutions to the original quadratic problems when used in conjunction with a primal solution callback function and a local non-linear programming solver.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability All instances are publicly available, see https://github.com/joehuchette/quadratic-relaxation-experiments for boxQP instances, https://github.com/robburlacu/acopflib for ACOPF instances, and https://qplib.zib.de/ for QPLIB instances.

Declarations

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

A MIP relaxations on general intervals

In this section, we generalize the MIP relaxations for $\mathrm{gra}_{[0,1]^2}(xy)$ and $\mathrm{gra}^2_{[0,1]}(x^2)$ discussed in this article to general box domains $(x, y) \in [\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}]$ and $x \in [\underline{x}, \bar{x}]$, where $\underline{x} < \bar{x}$, $\underline{y} < \bar{y}$ and $\underline{x}, \bar{x}, \underline{y}, \bar{y} \in \mathbb{R}$. by giving explicit formulations for general bounds on x and y.

A.1 MIP relaxations for bivariate quadratic equations

First, we consider MIP relaxations for $z = xy$ and give an explicit model of HybS for general box domains. We omit the formulation of Bin2 and Bin3 here, as these work analogously to HybS.

In the HybS MIP relaxation, in addition to the variables x and y, we must also transform the variables $p_1 = x + y$ and $p_2 = x - y$ and their respective bounds. In the following, the sawtooth modeling $(x, z_x) \in R^{L,L_1}$, $(y, z_y) \in R^{L,L_1}$, $(p_1, z_{p_1}) \in Q^{L_1}$, $(p_2, z_{p_2}) \in Q^{L_1}$ is performed according to Remark 1. HybS (21) for general box domains then reads as follows:

$$
\begin{aligned}
p_1 &= x + y \\
p_2 &= x - y \\
(x, z_x) &\in R^{L,L_1}
\end{aligned}
$$

$$
\begin{aligned}
(y, z_y) &\in R^{L,L_1} \\
(p_1, z_{p_1}) &\in Q^{L_1} \\
(p_2, z_{p_2}) &\in Q^{L_1} \\
z_{p_1} &\geqslant (w_x + w_y)^2 f^j\left(\tfrac{p_1 - \underline{x} - \underline{y}}{w_x + w_y}, \boldsymbol{g}^{p_1}\right) + (\underline{x} + \underline{y})\,(2p_2 - \underline{x} - \underline{y}) \qquad j \in 0, \ldots, L_1 \\
z_{p_2} &\geqslant (w_x + w_y)^2 f^j\left(\tfrac{p_2 - \underline{x} + \bar{y}}{w_x + w_y}, \boldsymbol{g}^{p_2}\right) + (\underline{x} - \bar{y})\,(2p_2 - \underline{x} + \bar{y}) \qquad j \in 0, \ldots, L_1 \\
z_x &\leqslant w_x^2 f^L(\tfrac{x - \underline{x}}{w_x}, \boldsymbol{g}^x) + \underline{x}(2x - \underline{x}) \\
z_y &\leqslant w_y^2 f^L(\tfrac{y - \underline{y}}{w_y}, \boldsymbol{g}^y) + \underline{y}(2y - \underline{y}) \\
z &\geqslant \tfrac{1}{2}(z_{p_1} - z_x - z_y) \\
z &\leqslant \tfrac{1}{2}(z_x + z_y - z_{p_2}) \\
(x, y, z) &\in \mathcal{M}(x, y) \\
x &\in [\underline{x}, \bar{x}] \\
y &\in [\underline{y}, \bar{y}] \\
p_1 &\in [\underline{x} + \underline{y}, \bar{x} + \bar{y}] \\
p_2 &\in [\underline{x} - \bar{y}, \bar{x} - \underline{y}].
\end{aligned} \tag{32}
$$

A.2 MIP relaxations for univariate quadratic equations

In order to MIP relaxations for $z = x^2$ where $x \in [\underline{x}, \bar{x}]$ with $\underline{x} < \bar{x}$ and $\underline{x}, \bar{x} \in \mathbb{R}$, we introduce the auxiliary variable $\hat{x} \in [0, 1]$ and apply each original MIP relaxation to model $\hat{z} = \hat{x}^2$. In addition, we map $\hat{x}$ and $\hat{z}$ back to $[0, 1]$, yielding

$$
\hat{x} = \tfrac{x - \underline{x}}{w_x}, \quad \hat{z} = \tfrac{y - \underline{x}(2x - \underline{x})}{w_x^2}, \text{ with } x \in [\underline{x}, \bar{x}],
$$

cf. Remark 1. With this transformation, we are able to formulate the tightened sawtooth relaxation for $x \in [\underline{x}, \bar{x}]$. The tightened sawtooth relaxation (16) for general box domains then reads

$$
\{(x, z) \in [\underline{x}, \bar{x}] \times \mathbb{R} : \exists (\hat{x}, \hat{z}, \boldsymbol{g}, \boldsymbol{\alpha}) \in [0, 1] \times \mathbb{R} \times [0, 1]^{L_1 + 1} \times \{0, 1\}^L : (34)\} \tag{33}
$$

where the constraints are

$$
\begin{aligned}
\hat{x} &= \tfrac{x - \underline{x}}{w_x} \\
\hat{z} &= \tfrac{y - \underline{x}(2x - \underline{x})}{w_x^2} \\
(\hat{x}, \boldsymbol{g}_{[\![0,L]\!]}, \boldsymbol{\alpha}) &\in S^L(\hat{x}) \\
(\hat{x}, \boldsymbol{g}) &\in T^{L_1}(\hat{x}) \\
\hat{z} &\leqslant f^L(\hat{x}, \boldsymbol{g}_{[\![0,L]\!]}) \\
\hat{z} &\geqslant f^j(\hat{x}, \boldsymbol{g}) - 2^{-2j-2} \qquad j \in 0, \ldots, L_1 \\
\hat{z} &\geqslant 0 \\
\hat{z} &\geqslant 2\hat{x} - 1.
\end{aligned} \tag{34}
$$

We note that generalizing the sawtooth epigraph relaxation (14) works analogously.

B Proof of Theorem 2: hereditary sharpness of the tightened sawtooth relaxation

This section is devoted to proving Theorem 2 which states that the tightened sawtooth relaxation (16) for $z = x^2$ is hereditarily sharp. This is a similar, albeit, more difficult result than the related one in [7] regarding the original sawtooth approximation. It is not clear how to obtain the former as a corollary of the latter. Furthermore, we use the result of [7] to shorten the work needed here. Before we begin the proof, we first introduce some required notation and restate several helpful results from [7]. For integers $L_1 \geqslant L \geqslant 0$, let P^{IP}_{L,L_1} be the tightened sawtooth relaxation from (16) in the space of $(x, z, \boldsymbol{g}, \boldsymbol{\alpha})$ and let P^{LP}_{L,L_1} be its LP relaxation, where in the latter all α-variables are relaxed to the interval $[0, 1]$. For convenience, and to avoid the variable redundancy $g_0 = x$ throughout this section, we will omit the use of g_0 and use the abbreviated notation $\boldsymbol{g} = \boldsymbol{g}_{[\![1,L_1]\!]}$.

To further simplify the notation, we omit the subscript L, L_1 when the context is clear and simply write P^{IP} and P^{LP} instead of P^{IP}_{L,L_1} and P^{LP}_{L,L_1}.

Now let $I \subseteq [\![L]\!]$ be the index set of the binary variables $\boldsymbol{\alpha}$ which are fixed to given values $\underline{\boldsymbol{\alpha}} \in \{0, 1\}^I$. This can be thought of as considering the branch in a branch-and-bound tree where $\boldsymbol{\alpha} = \underline{\boldsymbol{\alpha}}$ holds. Then we wish to show that at this node in the tree, sharpness also holds. More precisely, the goal is to show that P^{IP} is sharp under the restriction $\boldsymbol{\alpha}_I = \underline{\boldsymbol{\alpha}}$, where $\boldsymbol{\alpha}_I = [\alpha_{i_1}, \ldots, \alpha_{i_{|I|}}]^\top$ and $I = \{i_1, \ldots, i_{|I|}\}$. Hereditary sharpness of P^{IP} then means

$$\mathrm{conv}(\mathrm{proj}_{x,z}(P^{\mathrm{IP}}|_{\boldsymbol{\alpha}_I = \underline{\boldsymbol{\alpha}}})) = \mathrm{proj}_{x,z}(P^{\mathrm{LP}}|_{\boldsymbol{\alpha}_I = \underline{\boldsymbol{\alpha}}}).$$

In order to show this result, we cover $P^{\mathrm{IP}}|_{\boldsymbol{\alpha}_I = \underline{\boldsymbol{\alpha}}}$ using the following two sets, which encapsulate the upper and lower bounds w.r.t. z, respectively:

$$\begin{aligned}
\hat{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}} &:= \{(x, z, \boldsymbol{g}, \boldsymbol{\alpha}) \in [0, 1]^2 \times [0, 1]^{L_1} \times \{0, 1\}^L : \boldsymbol{\alpha}_I = \underline{\boldsymbol{\alpha}},\ (17b, 17c, 17a)\}, \\
\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}} &:= \{(x, z, \boldsymbol{g}, \boldsymbol{\alpha}) \in [0, 1]^2 \times [0, 1]^{L_1} \times \{0, 1\}^L : \boldsymbol{\alpha}_I = \underline{\boldsymbol{\alpha}},\ (17b, 17c, 17d, 17f)\}.
\end{aligned} \tag{35}$$

Observation 1 *It holds* $P^{\mathrm{IP}}|_{\boldsymbol{\alpha}_I = \underline{\boldsymbol{\alpha}}} = \hat{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}} \cap \check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}$, *and the formulation* P^{IP} *is hereditarily sharp if and only if both* $\hat{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}$ *and* $\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}$ *are sharp.*

Sharpness of $\hat{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}$. This follows directly from [7, Theorem 3]: the theorem establishes hereditary sharpness of the sawtooth approximation (10), which has the same upper-bounding constraints on z as (16). Thus, it remains for us to show that $\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}$ is sharp.

Sharpness of $\check{P}^{\text{IP},\underline{\boldsymbol{\alpha}}}$. Before beginning the proof, we set up some helpful notation. First, we define the projections onto $(x, \boldsymbol{g}, \boldsymbol{\alpha})$:

$$\begin{aligned}\check{P}^{\text{IP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})} &:= \text{proj}_{x,\boldsymbol{g},\boldsymbol{\alpha}}(\check{P}^{\text{IP},\underline{\boldsymbol{\alpha}}}),\\ \check{P}^{\text{LP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})} &:= \text{proj}_{x,\boldsymbol{g},\boldsymbol{\alpha}}(\check{P}^{\text{LP},\underline{\boldsymbol{\alpha}}}).\end{aligned} \tag{36}$$

In particular, these variables must satisfy (17b) and (17c). We also define the corresponding projections onto x, namely

$$\check{X}^{\text{IP}} := \text{proj}_x(\check{P}^{\text{IP},\underline{\boldsymbol{\alpha}}}) \quad \text{and} \quad \check{X}^{\text{LP}} := \text{proj}_x(\check{P}^{\text{LP},\underline{\boldsymbol{\alpha}}}).$$

Next, we define the lower-bounding functions $\check{f}^j : [0,1] \times [0,1]^{L_1+1} \to [0,1]$,

$$\begin{aligned}\check{f}^j(x,\boldsymbol{g}) &= f^j(x,\boldsymbol{g}) - 2^{-2j-2} \quad j = 0,\dots,L_1,\\ \check{f}^{-1}(x,\boldsymbol{g}) &= 2x-1,\\ \check{f}^{-2}(x,\boldsymbol{g}) &= 0.\end{aligned} \tag{37}$$

Note that $\check{f}^{-1}$ and $\check{f}^{-2}$ do not actually depend on $\boldsymbol{g}$. Further, note that there is a slight abuse of the notation above, since technically f^j has the domain $[0,1] \times [0,1]^{j+1}$; however, we assume the reader will interpret the functional expressions as $f^j(x, \boldsymbol{g}_{[\![j]\!]})$ instead. We also define the lower-bounding functions $\check{F}^j : [0,1] \to [0,1]$,

$$\begin{aligned}\check{F}^j(x) &= F^j(x) - 2^{-2j-2} \quad j = 0,\dots,L_1,\\ \check{F}^{-1}(x) &= 2x-1,\\ \check{F}^{-2}(x) &= 0\end{aligned} \tag{38}$$

in terms of only x, based on the functions F^L from (6), as the j-th p.w.l. underestimator to $z = x^2$ in the construction of the sawtooth relaxation, as defined in Sect. 3.2. Further, define $\check{f} : [0,1] \times [0,1]^L \to [0,1]$ and $\check{F} : [0,1] \to [0,1]$ with

$$\check{f}(x,\boldsymbol{g}) = \max_{j \in [\![-2,L]\!]} \check{f}^j(x,\boldsymbol{g}) \quad \text{and} \quad \check{F}(x) = \max_{j \in [\![-2,L]\!]} \check{F}^j(x).$$

Observation 2 *The function $\check{F}$ is convex as it is the maximum of a finite set of convex functions.*

Finally, we define the following sets with respect to j:

$$\begin{aligned}\check{P}^{\text{IP},\underline{\boldsymbol{\alpha}}}_j &:= \{(x,z,\boldsymbol{g},\boldsymbol{\alpha}) : (x,\boldsymbol{g},\boldsymbol{\alpha}) \in \check{P}^{\text{IP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})},\ z \geqslant \check{f}^j(x,\boldsymbol{g})\}, \quad j = -2,\dots,L_1,\\ \check{P}^{\text{LP},\underline{\boldsymbol{\alpha}}}_j &:= \{(x,z,\boldsymbol{g},\boldsymbol{\alpha}) : (x,\boldsymbol{g},\boldsymbol{\alpha}) \in \check{P}^{\text{LP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})},\ z \geqslant \check{f}^j(x,\boldsymbol{g})\}, \quad j = -2,\dots,L_1,\end{aligned} \tag{39}$$

and have $\check{P}^{\text{IP},\underline{\boldsymbol{\alpha}}} = \bigcap_{j=-2}^{L_1} \check{P}^{\text{IP},\underline{\boldsymbol{\alpha}}}_j$ or, equivalently,

$$\check{P}^{\text{IP},\underline{\boldsymbol{\alpha}}} = \{(x,z,\boldsymbol{g},\boldsymbol{\alpha}) : (x,\boldsymbol{g},\boldsymbol{\alpha}) \in \check{P}^{\text{IP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})},\ z \geqslant \max_{j \in -2,\dots,L_1} \check{f}^j(x,\boldsymbol{g})\}.$$

This applies analogously to $\check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}$.

We now state some important results from [7] that establish bounds on each variable g_i within $\check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}$ and a closed-form optimal solution for $\boldsymbol{g}$ when minimizing z within $\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}$ or any $\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_j$.

Lemma 1 (Bounds in Projection, Lemma 3 from [7]) *For all $i \in [\![0, L]\!]$, we have $\mathrm{proj}_{g_i}(\check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}) = \mathrm{conv}(\mathrm{proj}_{g_i}(\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})})) =: [a_i, b_i] \neq \emptyset$. Furthermore, it holds that $[a_L, b_L] = [0, 1]$, and $[a_{i-1}, b_{i-1}]$ can be computed from $[a_i, b_i]$ as*

$$[a_{i-1}, b_{i-1}] = \begin{cases} \left[\frac{1}{2}a_i, \frac{1}{2}b_i\right], & \text{if } i \in I \text{ and } \bar{\alpha}_i = 0, \\ \left[1 - \frac{1}{2}b_i, 1 - \frac{1}{2}a_i\right], & \text{if } i \in I \text{ and } \bar{\alpha}_i = 1, \\ \left[\frac{1}{2}a_i, 1 - \frac{1}{2}a_i\right], & \text{if } i \notin I. \end{cases} \tag{40}$$

Note that in the last case, $a_{i-1} \leqslant \frac{1}{2}$ and $b_{i-1} \geqslant \frac{1}{2}$ hold.

Note that Lemma 1 with $i = 0$ and $g_0 = x$ yields $\check{X}^{\mathrm{LP}} = \mathrm{conv}(\check{X}^{\mathrm{IP}})$, via

$$\check{X}^{\mathrm{LP}} = \mathrm{proj}_x(\check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}) = \mathrm{conv}(\mathrm{proj}_x(\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}})) = \mathrm{conv}(\check{X}^{\mathrm{IP}}), \tag{41}$$

which has also been used in [7].

Next, we adapt Lemma 5 from [7], which establishes that, when minimizing or maximizing z within $P^{\mathrm{LP}}_{L,L}|_{\boldsymbol{\alpha}_I = \underline{\boldsymbol{\alpha}}}$ given a fixed value for $\mathring{x}$, each g_i can directly be computed from g_{i-1} and the bounds established in Lemma 1. In particular, for the sawtooth relaxation (i.e. $I = \emptyset$), when minimizing z over the MIP-feasible points with a fixed x, we find that $g_i = \min\{2g_{i-1}, 1 - 2g_{i-1}\}$. That is, the $\boldsymbol{g}$-variables take one of the two upper bounds that restrict them. However, in this section, we have fixed several of the $\boldsymbol{\alpha}$-variables and have thus changed the feasible domain for each $\boldsymbol{g}$-variable. Now, it could be that b_i becomes an additional upper bound.

Lemma 2 (Adapted from Lemma 5 from [7]) *Let a_i and b_i be defined as in Lemma 1 for all $i \in [\![L_1]\!]$ and let $\mathring{x} \in [a_0, b_0]$. Further, define $\boldsymbol{g}^*$ as*

$$\begin{aligned} g_0^* &:= \mathring{x} \\ g_i^* &:= \min\{b_i, 2g_{i-1}, 1 - 2g_{i-1}\} && i \in [\![L_1]\!] \setminus I \\ g_i^* &:= G^i(g_{i-1}) && i \in I, \end{aligned}$$

where, for $i \in I$, it holds $G^i(g_{i-1}) = 2g_{i-1}$ if $\alpha_i = 0$, and $G^i(g_{i=1}) = 2(1 - g_{i-1})$ otherwise. Then we have

$$\boldsymbol{g}^* \in \mathrm{argmin}\{z \,:\, (z, \boldsymbol{g}) \in \mathrm{proj}_{z,\boldsymbol{g}}(\check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}|_{x=\mathring{x}})\}, \tag{42a}$$

$$\boldsymbol{g}^* \in \mathrm{argmin}\{z \,:\, (z, \boldsymbol{g}) \in \mathrm{proj}_{z,\boldsymbol{g}}(\check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}_j|_{x=\mathring{x}})\} \quad \forall j \in [\![-2, L_1]\!]. \tag{42b}$$

That is, each g_i with unfixed α_i can take on one of its upper bounds w.r.t. g_{i-1} when minimizing z within $\check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}|_{x=\mathring{x}}$ and $\check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}_j|_{x=\mathring{x}}$. Furthermore, this choice is unique for

all $i \leqslant j$, i.e.

$$|\text{argmin}\{z \,:\, (z, \boldsymbol{g}_{[\![j]\!]}) \in \text{proj}_{z,\boldsymbol{g}_{[\![j]\!]}}(\check{P}_j^{\text{LP},\underline{\boldsymbol{\alpha}}})|_{x=\mathring{x}})\}| = 1.$$

Finally, there exists some $j \in [\![-2, L_1]\!]$ for which

$$\check{f}^j(\mathring{x}, \boldsymbol{g}^*) = \min\{z \,:\, (z, \boldsymbol{g}) \in \text{proj}_{z,\boldsymbol{g}}(\check{P}^{\text{LP},\underline{\boldsymbol{\alpha}}})|_{x=\mathring{x}})\}. \tag{43}$$

Proof The proofs of the optimality results (42a) and (42b) on $\boldsymbol{g}^*$ for $j \geqslant 1$ closely follow the structure of the proof of Theorem 1, with the same underlying reasoning as in the proof of [7, Lemma 5]. In fact, the uniqueness of the optimizer also follows from the proof. Thus, the details are omitted here. To establish the optimality results for $j \leqslant 0$, we observe that in this case $\check{f}^j$ is purely a function of x, such that the choice of $\boldsymbol{g}$ has no effect on $\check{f}^j$, and $\boldsymbol{g}^*$ is thus still optimal.

Finally, to fulfil (43), let $j_{\max} \in [\![-2, L_1]\!]$ be chosen such that

$$\max_{j\in[\![-2,L_1]\!]} \check{f}^j(\mathring{x}, \boldsymbol{g}^*) = \check{f}^{j_{\max}}(\mathring{x}, \boldsymbol{g}^*).$$

Then we have

$$\begin{aligned}\min\{z \,:\, (z, \boldsymbol{g}) \in \text{proj}_{z,\boldsymbol{g}}(\check{P}^{\text{LP},\underline{\boldsymbol{\alpha}}})|_{x=\mathring{x}})\} &= \max_{j\in[\![-2,L_1]\!]} \check{f}^j(\mathring{x}, \boldsymbol{g}^*) = \check{f}^{j_{\max}}(\mathring{x}, \boldsymbol{g}^*) \\ = \min\{z \,:\, (z, \boldsymbol{g}) \in \text{proj}_{z,\boldsymbol{g}}(\check{P}^{\text{LP},\underline{\boldsymbol{\alpha}}}_{j_{\max}})|_{x=\mathring{x}})\} &\leqslant \min\{z \,:\, (z, \boldsymbol{g}) \in \text{proj}_{z,\boldsymbol{g}}(\check{P}^{\text{LP},\underline{\boldsymbol{\alpha}}})|_{x=\mathring{x}})\},\end{aligned}$$

as required. □

The next auxiliary result we need is a lemma concerning reflections over $x = \frac{1}{2}$ in $\check{P}^{\text{IP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}$ and $\check{P}^{\text{LP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}$ for the case where α_1 is not fixed.

Lemma 3 *Let $L \geqslant 0$, let $\mathring{x} \in \check{X}^{\text{IP}}$ and assume $1 \notin I$, so that α_1 is not fixed. Then*

$$\text{proj}_{\boldsymbol{g},\boldsymbol{\alpha}_{[\![2,L]\!]}}(\check{P}^{\text{IP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}|_{x=\mathring{x}}) = \text{proj}_{\boldsymbol{g},\boldsymbol{\alpha}_{[\![2,L]\!]}}(\check{P}^{\text{IP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}|_{x=1-\mathring{x}}). \tag{44}$$

Furthermore,

$$\mathring{x}^2 - \check{f}^j(\mathring{x}, \boldsymbol{g}^*) = (1-\mathring{x})^2 - \check{f}^j(1-\mathring{x}, \boldsymbol{g}^*) \qquad \text{for all } j \in [\![0, L_1]\!]. \tag{45}$$

That is, the maximum errors from the lower bounds coincide. Similarly,

$$\mathring{x}^2 - \check{f}^{-2}(\mathring{x}, \boldsymbol{g}^*) = (1-\mathring{x})^2 - \check{f}^{-1}(1-\mathring{x}, \boldsymbol{g}^*), \tag{46}$$

$$\mathring{x}^2 - \check{f}^{-1}(\mathring{x}, \boldsymbol{g}^*) = (1-\mathring{x})^2 - \check{f}^{-2}(1-\mathring{x}, \boldsymbol{g}^*), \tag{47}$$

where $\boldsymbol{g}^*$ is defined on Lemma 2. Lastly,

$$\mathring{x}^2 - \check{f}(\mathring{x}, \boldsymbol{g}^*) = (1-\mathring{x})^2 - \check{f}(1-\mathring{x}, \boldsymbol{g}^*). \tag{48}$$

Proof Recall that $\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}$ is formed from the constraints in S^L and T^{L_1}, along with fixing binary variables $\boldsymbol{\alpha}_I = \underline{\boldsymbol{\alpha}}$. It is easy to check that $(\mathring{x}, \boldsymbol{g}, \boldsymbol{\alpha}) \in \check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}$ if an only if $(1 - \mathring{x}, \boldsymbol{g}, \bar{\boldsymbol{\alpha}}) \in \check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}$, where $\bar{\alpha}_1 := 1 - \alpha_1$ and $\bar{\alpha}_i := \alpha_i$ for $i \in I \backslash \{1\}$. Thus, (44) holds due to this correspondence.

For $j \in [\![0, L_1]\!]$, we have

$$\begin{aligned}
\mathring{x}^2 - \check{f}^j(\mathring{x}, \boldsymbol{g}^*) &= \mathring{x}^2 - \left(\mathring{x} - \sum_{i=1}^{j} 2^{-2i} \boldsymbol{g}_i^* - 2^{-2j-2} \right) \\
&= \left(1 - 2\mathring{x}\right) + \mathring{x}^2 - \left(\left(1 - 2\mathring{x}\right) + \mathring{x} - \sum_{i=1}^{j} 2^{-2i} \boldsymbol{g}_i^* - 2^{-2j-2} \right) \\
&= \left(1 - \mathring{x}\right)^2 - \left(1 - \mathring{x} - \sum_{i=1}^{j} 2^{-2i} \boldsymbol{g}_i^* - 2^{-2j-2} \right) \\
&= (1 - \mathring{x})^2 - \check{f}^j(1 - \mathring{x}, \boldsymbol{g}^*).
\end{aligned}$$

Thus (48) holds. Similarly, (47) holds as

$$\begin{aligned}
\mathring{x}^2 - \check{f}^{-1}(\mathring{x}, \boldsymbol{g}^*) &= \mathring{x}^2 - (2\mathring{x} - 1) \\
&= \left(1 - \mathring{x}\right)^2 \\
&= (1 - \mathring{x})^2 - \check{f}^{-2}(1 - \mathring{x}, \boldsymbol{g}^*).
\end{aligned}$$

Lastly, (46) holds by considering the substitution $\mathring{x} \leftarrow 1 - \mathring{x}$ from (47).

The same secondary result holds if $\check{f}^j(x, \boldsymbol{g})$ is replaced with $\check{f}(x, \boldsymbol{g})$. This follows since each constituting function (for the pair $j = -1$, $j = -2$) is symmetric about $x = \frac{1}{2}$ w.r.t. the maximum error; the pointwise maximum over the functions retains the same symmetry. Similarly, the same result holds if $I = \emptyset$, such that $\check{X} = [0, 1]$. □

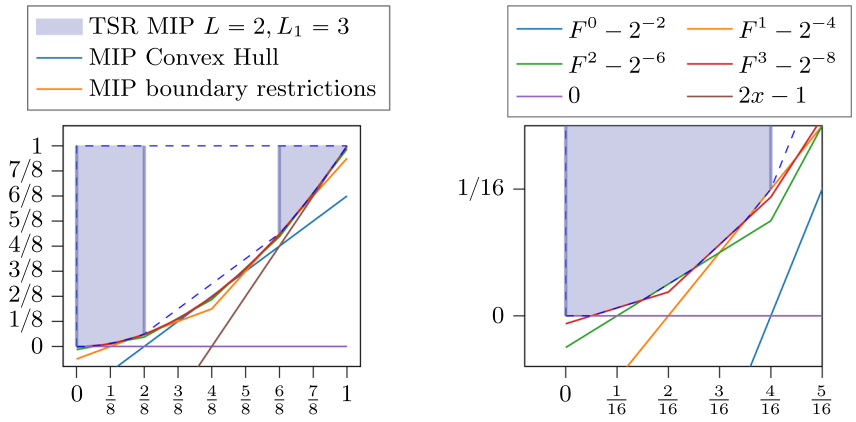
The following lemma formalizes the convex hull of convex functions whose domain is a finite union of closed and bounded intervals. By *gaps*, we refer to the open intervals in the convex hull of the domain but do not intersect the domain.

Lemma 4 *Let $X \subseteq \mathbb{R}$ be a finite union of compact intervals, and let $F\colon \operatorname{conv}(X) \to \mathbb{R}$ be a convex function. For any $\bar{x} \in \operatorname{conv}(X) \setminus X$, define*

$$\bar{x}_- := \max\{x \in X : x < \bar{x}\} \quad \textit{and} \quad \bar{x}_+ := \min\{x \in X : x > \bar{x}\}.$$

Now define $F_X\colon \operatorname{conv}(X) \to \mathbb{R}$,

$$F_X(x) = \begin{cases} F(x), & \text{if } x \in X, \\ \lambda F(x_-) + (1 - \lambda) F(x_+), & \text{if } x \notin X, \text{for } x = \lambda x_- + (1 - \lambda) x_+, \\ & \text{with } \lambda \in (0, 1). \end{cases} \tag{49}$$



(a) MIP convex hull with lower bounds plotted.

(b) Zoomed in to show the interaction of bounds at $x = 4/16$.

Fig. 10 The projected MIP convex hull for $L = 2, L_1 = 3$ where we fix $\alpha_2 = 0$. In particular, note that at the boundary points $\partial \check{X}^{\text{IP}} = \{0, \frac{2}{8}, \frac{6}{8}, 1\}$, the tight lower-bounding inequalities are $z \geqslant 0$, $z \geqslant 2x - 1$ and $z \geqslant F^1 - 2^{-4}$. Thus, on the gap $(\frac{2}{8}, \frac{6}{8})$ the functions $\check{F}^2$, $\check{F}^3$ are not needed to describe the convex hull of the MIP

Then we have

$$\text{conv}(\text{epi}_X(F)) = \text{epi}_{\text{conv}(X)}(F_X).$$

This lemma is proved in Appendix C. We are now ready to prove Theorem 2.

We denote the boundary of the set X by ∂X.

Proof of Theorem 2 As discussed before, we only need to show that $\check{P}^{\text{IP},\underline{\boldsymbol{\alpha}}}_{L,L_1}$ is sharp to conclude that P^{IP}_{L,L_1} is hereditarily sharp. In particular, we need to show that

$$\text{conv}(\text{proj}_{x,z}(\check{P}^{\text{IP},\underline{\boldsymbol{\alpha}}}_{L,L_1})) = \text{proj}_{x,z}(\check{P}^{\text{LP},\underline{\boldsymbol{\alpha}}}_{L,L_1}).$$

Reduction to $L_1 = L$: Recall that $L_1 \geqslant L$ holds by definition.

Claim We claim that it suffices to reduce L_1 to L to conclude hereditary sharpness of P^{IP}_{L,L_1}.

Claim proof Assume that $L_1 > L$ holds. To construct $\check{P}^{\text{IP},\underline{\boldsymbol{\alpha}}}_{L,L_1}$ from $\check{P}^{\text{IP},\underline{\boldsymbol{\alpha}}}_{L,L_1-1}$, we simply maintain the same fixing $\boldsymbol{\alpha}_I = \underline{\boldsymbol{\alpha}}$, then add a new variable $g_{L_1} \geqslant 0$, together with the new constraints

$$g_{L_1} \leqslant 2g_{L_1-1}, \quad g_{L_1} \leqslant 2(1 - g_{L_1-1}), \qquad \text{(from (17}c\text{) via (13))}$$

$$z \geqslant x - \sum_{i=1}^{L_1} 2^{-2i} g_i - 2^{-2L_1-2}. \qquad \text{(from (17}d\text{))}$$

We then note the following:

1. It holds $\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{L,L_1} \subseteq \check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{L,L_1-1}$, since $L_1 > L_1 - 1$, and thus there are more inequalities used to define $\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{L,L_1}$.
2. We have $\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{L,L_1}|_{x\in\partial\check{X}^{\mathrm{IP}}} = \check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{L,L_1-1}|_{x\in\partial\check{X}^{\mathrm{IP}}}$. To see this, first notice that $\partial\check{X}^{\mathrm{IP}} \subseteq \{\frac{i}{2^L} : i \in [\![2^L]\!]\}$, since $I \subseteq [\![L]\!]$. Thus, for $L_1 > L$, the inequality $z \geqslant x - \sum_{i=1}^{L_1} 2^{-2i} g_i - 2^{-2L_1-2}$ is not tight at any of these points in $\partial\check{X}^{\mathrm{IP}}$; see Proposition 1, Item 3.
3. It follows from the previous equation that for any $\bar{x} \in \partial\check{X}^{\mathrm{IP}}$, we have

$$\mathrm{proj}_{x,z}(\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{L,L_1-1}|_{x=\bar{x}}) = \mathrm{proj}_{x,z}(\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{L,L_1}|_{x=\bar{x}}) = \{(x,z) : z \geqslant \check{F}(x), x = \bar{x}\}.$$

4. When we restrict to the domain $\mathrm{conv}(\check{X}^{\mathrm{IP}}) \setminus \check{X}^{\mathrm{IP}}$ and consider the convex hulls, we have equality as we reduce L_1, i.e.

$$\mathrm{conv}(\mathrm{proj}_{x,z}(\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{L,L_1-1})|_{x\in\mathrm{conv}(\check{X}^{\mathrm{IP}})\setminus\check{X}^{\mathrm{IP}}}) = \mathrm{conv}(\mathrm{proj}_{x,z}(\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{L,L_1})|_{x\in\mathrm{conv}(\check{X}^{\mathrm{IP}})\setminus\check{X}^{\mathrm{IP}}}).$$

   This is due to Item 2, the convexity of $\check{F}$ and Lemma 4.

Thus, the convex hull remains unchanged across the gaps in $\check{X}^{\mathrm{IP}}$, and since the LP relaxation does not weaken, sharpness in lower bound is maintained; see Fig. 10. This implies that $\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{L,L_1}$ is sharp if $\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{L,L_1-1}$ is sharp. The claim then holds by induction. ⋄

We now proceed to prove sharpness of $\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{L,L}$ by induction on L.

Base case: If $L = 0$, then there are no binary variables and, hence, nothing to branch on; therefore, the result holds trivially.

Induction on L: For the inductive step, we assume that $\check{P}^{\mathrm{IP},\tilde{\underline{\boldsymbol{\alpha}}}}_{L-1,L-1}$ is hereditarily sharp for all possible fixings of $\boldsymbol{\alpha}$-variables, and show that $\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{L,L}$ is hereditarily sharp.

We begin by observing that

$$\mathrm{proj}_{x,z}\left(\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{L,L}\right) = \mathrm{epi}_{\check{X}^{\mathrm{IP}}}(\check{F}).$$

By Lemma 4, it follows that

$$\mathrm{conv}(\mathrm{epi}_{\check{X}^{\mathrm{IP}}}(\check{F})) = \mathrm{epi}_{\mathrm{conv}(\check{X}^{\mathrm{IP}})}\left(\check{F}_{\check{X}^{\mathrm{IP}}}\right),$$

where $\check{F}_{\check{X}^{\mathrm{IP}}}$ is defined as in Lemma 4. Thus, proving Theorem 2 is equivalent to proving that

$$\mathrm{proj}_{x,z}(\check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}_{L,L}) = \mathrm{epi}_{\mathrm{conv}(\check{X}^{\mathrm{IP}})}(\check{F}_{\check{X}^{\mathrm{IP}}}).$$

In particular, it suffices to show that for any $\mathring{x} \in \mathrm{conv}(\check{X}^{\mathrm{IP}})$, we have

$$\check{F}_{\check{X}^{\mathrm{IP}}}(\mathring{x}) = \min_{\boldsymbol{g}\in\check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}_{L,L}|_{x=\mathring{x}}} \check{f}(\mathring{x},\boldsymbol{g}) \tag{51}$$

which we do in the following.

Case I : $\mathring{x} \in \check{X}^{\text{IP}}$. By Theorem 1, $P^{\text{IP}}_{L,L}$ is sharp (i.e. when $I = \emptyset$). Thus, the LP lower bounds on z coincide with the MIP lower bounds for MIP-feasible points $x \in \check{X}^{\text{IP}}$, such that we have $\text{proj}_{x,z}(\check{P}^{\text{LP},\boldsymbol{\alpha}}_{L,L})|_{x\in\check{X}^{\text{IP}}} = \text{epi}_{\check{X}^{\text{IP}}}(\check{F})|_{x\in\check{X}^{\text{IP}}}$. This implies (51).

Case II : $\mathring{x} \in \text{conv}(\check{X}^{\text{IP}})\backslash\check{X}^{\text{IP}}$. Let $\mathring{x}_-, \mathring{x}_+ \in \check{X}^{\text{IP}}$ as defined in Lemma 4. Since $\mathring{x} \notin \check{X}^{\text{IP}}$, it follows that $\mathring{x}_-, \mathring{x}_+ \in \partial\check{X}^{\text{IP}}$.

Case II.A : $1 \notin I$. Assume $1 \notin I$.

Case II.A.1 : $[\mathring{x}_-, \mathring{x}_+] \subseteq \partial\check{X}^{\text{IP}} \cap [0, 1/2]$. We make use of the induction hypothesis here. To this end, we will work with $L - 1$ layers. We will decorate variables and parameters from the smaller set using "˜".

Define $\tilde{\boldsymbol{\alpha}}{:=}\boldsymbol{\alpha}$ and $\tilde{I}{:=}\{i - 1 : i \in I\}$, i.e. the same variables α_i are fixed but with indices decremented by 1. Now, define the linear map

$$\Phi : [0, 1] \times [0, 1] \times [0, 1]^{L-1} \times [0, 1]^{L-1} \to [0, 1] \times [0, 1] \times [0, 1]^{L} \times [0, 1]^{L}$$

such that $(\tilde{x}, \tilde{z}, \tilde{\boldsymbol{g}}, \tilde{\boldsymbol{\alpha}}) \mapsto (x, z, \boldsymbol{g}, \boldsymbol{\alpha})$ is defined via

$$\begin{aligned} x &= \tfrac{\tilde{x}}{2}, \quad z = \tfrac{\tilde{z}}{4}, \\ g_1 &= \tilde{x}, \quad \boldsymbol{g}_{[\![2,L]\!]} = \tilde{\boldsymbol{g}}, \\ \alpha_1 &= \tilde{x}, \quad \boldsymbol{\alpha}_{[\![2,L]\!]} = \tilde{\boldsymbol{\alpha}}. \end{aligned} \tag{52}$$

For convenience, under the definitions above, we write $x = \Phi_x(\tilde{x})$, $z = \Phi_z(\tilde{z})$, $\boldsymbol{g} = \Phi_{\boldsymbol{g}}(\tilde{\boldsymbol{g}})$, and $\boldsymbol{\alpha} = \Phi_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{\alpha}})$, and note that $g_0 = x$ and $\tilde{g}_0 = \tilde{x}$.

Claim $\Phi\left(\check{P}^{\text{IP},\tilde{\boldsymbol{\alpha}}}_{L-1,L-1}\right) = \check{P}^{\text{IP},\boldsymbol{\alpha}}_{L,L}\Big|_{x\in\text{conv}(\check{X}^{\text{IP}}\cap[0,1/2])}$.

Claim proof Let $(\tilde{x}, \tilde{z}, \tilde{\boldsymbol{g}}, \tilde{\boldsymbol{\alpha}}) \in \check{P}^{\text{LP},\tilde{\boldsymbol{\alpha}}}_{L-1,L-1}$ such that $\tilde{z}$ is minimal, and let $(x, z, \boldsymbol{g}, \boldsymbol{\alpha}) = \Phi(\tilde{x}, \tilde{z}, \tilde{\boldsymbol{g}}, \tilde{\boldsymbol{\alpha}})$. We will show that $(x, z, \boldsymbol{g}, \boldsymbol{\alpha}) \in \check{P}^{\text{IP},\boldsymbol{\alpha}}_{L,L}\Big|_{x\in\text{conv}(\check{X}^{\text{IP}}\cap[0,1/2])}$. To do so, we reference the formula (35), and show that Constraints (17b), (17c), (17d) and (17f) hold for $(x, z, \boldsymbol{g}, \boldsymbol{\alpha})$.

Since $\tilde{z}$ is minimal, we have $\tilde{z} = \check{\tilde{f}}^{j}(\tilde{x}, \tilde{\boldsymbol{g}})$ for some j. We claim that $z = \check{f}^{j'}(x, \boldsymbol{g})$ for some j'.

If $j \geqslant 0$, then, noting that $\frac{1}{4}\tilde{x} = \frac{1}{2}\tilde{x} - \frac{1}{4}\tilde{x} = x - \frac{1}{4}g_1$, we have

$$\begin{aligned} z &= \Phi_z(\tilde{z}) \\ &= \tfrac{1}{4}(\check{f}^{j}(\tilde{x}, \tilde{\boldsymbol{g}})) \\ &= \tfrac{1}{4}\left(\tilde{x} - \textstyle\sum_{i=1}^{j} 2^{-2i}\tilde{g}_i - 2^{-2j-2}\right) \\ &= x - \tfrac{1}{4}g_1 - \tfrac{1}{4}\left(\textstyle\sum_{i=1}^{j} 2^{-2i}\tilde{g}_i - 2^{-2j-2}\right) \\ &= x - \textstyle\sum_{i=1}^{j+1} 2^{-2i}g_i - 2^{-2(j+1)-2} = \check{f}^{j+1}(x, \boldsymbol{g}). \end{aligned}$$

If $j = -1$, we have

$$z = \Phi_z(\tilde{z}) = \tfrac{1}{4}(\check{f}^{-1}(\tilde{x}, \tilde{\boldsymbol{g}})) = \tfrac{1}{4}(2\tilde{x} - 1) = x - \tfrac{1}{4} = \check{f}^{0}(x, \boldsymbol{g}).$$

Finally, if $j = -2$, then

$$z = \Phi_z(\tilde{z}) = \tfrac{1}{4}(\check{f}^{-1}(\tilde{x}, \tilde{\boldsymbol{g}})) = 0 = \check{f}^{-2}(x, \boldsymbol{g}).$$

Thus, we have that $\Phi_z(\tilde{z}) \geqslant \check{f}^j(\Phi_x(\tilde{x}), \Phi_{\boldsymbol{g}}(\tilde{\boldsymbol{g}}))$ for all $j \neq 1$, where the absence of $\check{f}^{-1}(x, \boldsymbol{g})$ is due to the fact that $\check{f}^{-1}(x, \boldsymbol{g}) \leqslant 0$ for $x \in [0, \frac{1}{2}]$, such that that the corresponding bound is inactive on $\Phi_x(\overset{\tilde{\check{}}}{X}^{\mathrm{LP}})$.

Note that the above calculations also imply that, for all $\tilde{j} \in [\![-2, L-1]\!]$ and for all $(\tilde{x}, \tilde{\boldsymbol{g}}) \in \mathrm{proj}_{x,\boldsymbol{g}}(\tilde{P}^{\mathrm{LP}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})})$, we have for some $j \in [\![-2, L]\!]$ that $\Phi_z(\overset{\tilde{\tilde{j}}}{\check{f}}(\tilde{x}, \tilde{\boldsymbol{g}})) = \check{f}^j(\Phi_x(\tilde{x}), \Phi_{\boldsymbol{g}}(\boldsymbol{g}))$. Further, since each $\tilde{j}$ maps to a unique j (with only the inactive $j = -1$ skipped), this implies that $\Phi_z(\overset{\tilde{\tilde{j}}}{\check{f}}(\tilde{x}, \tilde{\boldsymbol{g}})) = \check{f}(\Phi_x(\tilde{x}), \Phi_{\boldsymbol{g}}(\boldsymbol{g}))$. Thus, we can conclude that (17d) and (17f) hold.

Next, we argue that $(x, \boldsymbol{g}, \boldsymbol{\alpha}) \in \mathrm{proj}_{x,\boldsymbol{g},\boldsymbol{\alpha}_{[\![2,L]\!]}}(\check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})})$. This implies in particular that (17b) as well as (17c) hold and that we have $\boldsymbol{\alpha}_I = \underline{\boldsymbol{\alpha}}_I$.

Since $g_1 = \tilde{x} = 2x$, we observe that $\check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}$ can be written as the set of points $(x, \boldsymbol{g}, \boldsymbol{\alpha}) \in [0,1] \times [0,1]^L \times [0,1]^L$ such that

$$\begin{aligned}
g_0 &= x && \\
g_i &= 2g_{i-1} && i = 1 \text{ or } i \in I, \underline{\alpha}_i = 0 \\
g_i &= 2(1 - g_{i-1}) && i \in I, \underline{\alpha}_i = 1 \\
|g_{i-1} - \alpha_i| \leqslant g_i &\leqslant \min(2g_{i-1}, 2(1 - g_{i-1})) && i \in [\![L]\!] \setminus I,\ i \geqslant 2 \\
\boldsymbol{\alpha}_I &= \underline{\boldsymbol{\alpha}}_I && \\
x, g_i, \alpha_i &\in [0,1] && i \in [\![L]\!].
\end{aligned}$$

In this form, it is straightforward to confirm $(x, \boldsymbol{g}, \boldsymbol{\alpha}) \in \check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}$ from the corresponding form for $\tilde{P}^{\mathrm{LP}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}$: since the indices for both the map on $\boldsymbol{g}$ and on the shift from $\tilde{I}$ to I are shifted by 1 in the same direction, with the same choice of $\underline{\boldsymbol{\alpha}}$, all equality constraints on g_i, $i \in \tilde{I}$, are preserved through the mapping. Further, the relationship between each g_i and g_{i-1} is likewise preserved, as the corresponding α_i is the same, and finally the choice of g_1 is feasible given x. Thus, all constraints are satisfied, such that $(x, \boldsymbol{g}, \boldsymbol{\alpha}) \in \check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}$, yielding for the choice of z above that $(x, z, \boldsymbol{g}, \boldsymbol{\alpha}) \in \check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}_{L,L}|_{x \in \mathrm{conv}(\check{X}^{\mathrm{IP}} \cap [0,1/2])}$.

Further, from the form for $\check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}$ above, we observe that $\Phi_x(\overset{\tilde{\check{}}}{X}^{\mathrm{IP}}) = \check{X}^{\mathrm{IP}} \cap [0, \frac{1}{2}]$ and $\Phi_x(\overset{\tilde{\check{}}}{X}^{\mathrm{LP}}) = \mathrm{conv}(\check{X}^{\mathrm{IP}} \cap [0, \frac{1}{2}])$. To show the first part, we have already shown that $\Phi_x(\overset{\tilde{\check{}}}{X}^{\mathrm{IP}}) \subseteq \check{X}^{\mathrm{IP}} \cap [0, \frac{1}{2}]$. To prove the other direction, we simply reverse the map for any $(x, \boldsymbol{g}, \boldsymbol{\alpha}) \in \check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}|_{x \in [0,1/2]}$, ignoring α_1: letting $\tilde{x} = g_1 = \frac{x}{2}$, $\tilde{\boldsymbol{g}} = \boldsymbol{g}_{[\![2,L]\!]}$ and $\tilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_{[\![2,L]\!]}$, it is easy to confirm $(\tilde{x}, \tilde{\boldsymbol{g}}, \tilde{\boldsymbol{\alpha}}) \in \tilde{P}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}$.

To show that $\mathrm{proj}_x\left(\Phi(\check{P}^{\mathrm{LP},\tilde{\underline{\boldsymbol{\alpha}}}}_{L-1,L-1})\right) = \mathrm{conv}(\check{X}^{\mathrm{IP}} \cap [0, \frac{1}{2}])$, we observe that $\mathrm{conv}(\check{X}^{\mathrm{IP}})|_{x \in [0,1/2]}$ is a closed interval with boundary points in $\check{X}^{\mathrm{IP}} \cap [0, \frac{1}{2}] = \Phi_x(\overset{\tilde{\check{}}}{X}^{\mathrm{IP}})$,

such that $\text{conv}(\check{X}^{\text{IP}} \cap [0, \frac{1}{2}]) = \text{conv}(\Phi_x(\tilde{\check{X}}^{\text{IP}})) = \Phi_x(\text{conv}(\tilde{\check{X}}^{\text{IP}})) = \Phi_x(\tilde{\check{X}}^{\text{LP}})$, since Φ is linear in x. ◇

We now show two facts:

Claim 1 Let $\mathring{x} \in \tilde{\check{X}}^{\text{LP}}$ and $\tilde{z}^* \in \text{argmin}\{\tilde{z} : (\tilde{z}, \tilde{\boldsymbol{g}}) \in \text{proj}_{\tilde{z},\tilde{\boldsymbol{g}}}(\check{P}^{\text{LP},\tilde{\boldsymbol{\alpha}}}_{L-1,L-1}|_{\tilde{x}=\mathring{x}})\}$ with the corresponding solution $\tilde{\boldsymbol{g}}^*$ defined in Lemma 2. Then

$$\left(\tfrac{1}{4}\tilde{z}^*, \Phi_{\boldsymbol{g}}(\tilde{\boldsymbol{g}}^*)\right) \in \text{argmin}\left\{z : (z, \boldsymbol{g}) \in \text{proj}_{z,\boldsymbol{g}}(\check{P}^{\text{LP},\boldsymbol{\alpha}}_{L,L}|_{x=\Phi_x(\tilde{x})})\right\}.$$

Claim 2 We have $\tilde{z} = \tilde{\check{F}}_{\check{X}^{\text{IP}}}(\tilde{x})$ if and only if $\Phi_z(\tilde{z}) = \check{F}_{\check{X}^{\text{IP}}}(\Phi_x(\tilde{x}))$, such that $\check{F}_{\check{X}^{\text{IP}}}(\Phi_x(\tilde{x})) = 4\tilde{\check{F}}_{\check{X}^{\text{IP}}}(\tilde{x})$.

By the sharpness of $\check{P}^{\text{IP},\tilde{\boldsymbol{\alpha}}}_{L-1,L-1}$, these facts then imply that

$$\check{F}_{\check{X}^{\text{IP}}}(\Phi_x(\tilde{x})) = 4\tilde{\check{F}}_{\check{X}^{\text{IP}}}(\tilde{x}) = 4 \min_{\boldsymbol{g} \in \check{P}^{\text{LP},\tilde{\boldsymbol{\alpha}}}_{L-1,L-1}|_{x=\tilde{x}}} (\check{f}(\mathring{x}, \boldsymbol{g})) = \min_{\boldsymbol{g} \in \check{P}^{\text{LP},\boldsymbol{\alpha}}_{L,L}|_{x=\Phi_x(\tilde{x})}} (\check{f}(\mathring{x}, \boldsymbol{g})),$$

such that (51) holds.

Proof of Claim 1 Let $\tilde{x} \in \tilde{\check{X}}^{\text{LP}}$ and $\tilde{z}^* := \min\{z : (z, \boldsymbol{g}) \in \text{proj}_{z,\boldsymbol{g}}(\check{P}^{\text{LP},\tilde{\boldsymbol{\alpha}}}_{L-1,L-1}|_{x=\tilde{x}})\}$, and let $\tilde{\boldsymbol{g}}^*$ be the optimizing solution from Lemma 2. For convenience, let $\mathring{x} := \Phi_x(\tilde{x})$ and $\boldsymbol{g}^* := \Phi_{\boldsymbol{g}}(\tilde{\boldsymbol{g}}^*)$. Then $\boldsymbol{g}^*$ takes on the optimal form from Lemma 2, with $\tilde{z}^* = \tilde{\check{f}}(\tilde{x}, \tilde{\boldsymbol{g}}^*)$, yielding

$$z^* := \Phi_z(\tilde{z}^*) = \Phi_z(\tilde{\check{f}}(\tilde{x}, \tilde{\boldsymbol{g}}^*)) = \check{f}(\mathring{x}, \boldsymbol{g}^*) = \min\{z : (z, \boldsymbol{g}) \in \text{proj}_{z,\boldsymbol{g}}(\check{P}^{\text{LP},\boldsymbol{\alpha}}_{L,L}|_{x=\mathring{x}})\},$$

such that $(\frac{1}{4}\tilde{z}^*, \Phi_{\boldsymbol{g}}(\tilde{\boldsymbol{g}}^*)) \in \text{argmin}\{z : (z, \boldsymbol{g}) \in \text{proj}_{z,\boldsymbol{g}}(\check{P}^{\text{LP},\boldsymbol{\alpha}}_{L,L}|_{x=\hat{x}})\}$, as required. As a corollary, observing that $\tilde{\check{f}}(\tilde{x}) = \min\{z : (z, \boldsymbol{g}) \in \text{proj}_{z,\boldsymbol{g}}(P^{\text{LP}}_{-}|_{x=\mathring{x}})\}$, and likewise for $\check{f}(\mathring{x})$, we have that $\Phi_z(\tilde{\check{f}}(\tilde{x})) = \frac{1}{4}\tilde{\check{f}}(\tilde{x}) = \check{f}(\mathring{x})$. □

Proof of Claim 2. In order to show $\tilde{z} = \tilde{\check{F}}_{\check{X}^{\text{IP}}}(\tilde{x})$ if and only if $\Phi_z(\tilde{y}) = \check{F}_{\check{X}^{\text{IP}}}(\Phi_x(\tilde{x}))$, we observe that, for any $\tilde{x} \in \tilde{X}$, we have $\Phi_x(\tilde{x}) \in X$, and therefore

$$\Phi_x(\tilde{\check{F}}_{\check{X}^{\text{IP}}}(\tilde{x})) = \Phi_x(\tilde{\check{f}}(\tilde{x})) = \check{f}(\Phi_x(\tilde{x})) = \check{F}_{\check{X}^{\text{IP}}}(\Phi_x(\tilde{x})).$$

Consequently, $\Phi_z(\tilde{\check{F}}_{\check{X}^{\text{IP}}}(\tilde{x})) = \check{F}_{\check{X}^{\text{IP}}}(\Phi_x(\tilde{x}))$ holds on $\tilde{X}$. Now, by Lemma 4, across any gap $\tilde{x}_-, \tilde{x}_+ \in \tilde{X}$ for which $(\tilde{x}_-, \tilde{x}_+) \cap \tilde{X} = \emptyset$ and $\tilde{x} \in [\tilde{x}_-, \tilde{x}_+]$, we have that $\tilde{\check{F}}_{\check{X}^{\text{IP}}}(\tilde{x})$ is on the line between the points $(\tilde{x}_-, \tilde{\check{f}}(\tilde{x}_-))$ and $(\tilde{x}_+, \tilde{\check{f}}(\tilde{x}_+))$. Thus, since

$\mathring{x}{:=}\Phi_x(\tilde{x})$, and since Φ is linear in x and z, $\check{f}(\mathring{x})$ lies on the line between the points $(\Phi_x(\tilde{x}_-), \tilde{\check{f}}(\tilde{x}_-)))$ and $\Phi_x((\tilde{x}_+), \tilde{\check{f}}(\tilde{x}_+))$.

Now, observe that, since $\Phi_x(\tilde{\check{X}}) = X \cap [0, \frac{1}{2}]$, we have that $(x_-, x_+){:=}(\Phi_x(\tilde{x}_-), \Phi_x(\tilde{x}_+))$ is a gap in X, with $x_-, x_+ \in X$ and $(x_-, x_+) \cap X = \emptyset$. Furthermore, as $x_+, x_- \in X$, we have that $\check{F}_{\check{X}^{\mathrm{IP}}}(\hat{x}) = \Phi_x(\tilde{\check{F}}_{\check{X}^{\mathrm{IP}}}(\tilde{x}_-))$, and similarly for x_+. Then, by Lemma 4, we have for $x \in (x_+, x_-)$ that $\check{F}_{\check{X}^{\mathrm{IP}}}(\Phi_x(\tilde{x})) = \check{F}_{\check{X}^{\mathrm{IP}}}(x) = \Phi_x(\check{F}_{\check{X}^{\mathrm{IP}}}(\tilde{x}))$, as required.

Case II.A.2 : $[\mathring{x}_-, \mathring{x}_+] \subseteq \mathrm{conv}(\check{X}^{\mathrm{IP}} \cap [1/2, 1])$. Applying Lemma 3 to $\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}$, we immediately recover sharpness on $1 - \Phi_x(\tilde{\check{X}}^{\mathrm{LP}}) = \mathrm{conv}(\check{X}^{\mathrm{IP}} \cap [\frac{1}{2}, 1])$. To see this, let $x \in \Phi_x(\tilde{\check{X}}^{\mathrm{LP}})$. Then, via Lemma 3, we obtain exactly the same feasible regions for $\boldsymbol{g}, \boldsymbol{\alpha}$ with $x = 1 - \mathring{x}$ as with $x = \mathring{x}$, i.e. $\mathrm{proj}_{\boldsymbol{g},\boldsymbol{\alpha}_{[\![2,L]\!]}}(\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}|_{x=\mathring{x}}) = \mathrm{proj}_{\boldsymbol{g},\boldsymbol{\alpha}_{[\![2,L]\!]}}(\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}|_{x=1-\mathring{x}})$, and moreover, similar to Lemma 3, it is not hard to show that we have $\mathring{x}^2 - \check{F}(\hat{x}) = (1-\mathring{x})^2 - \check{F}(1-\mathring{x})$. Thus, we have that both $\check{F}(1-\mathring{x})$ and $\min_{\boldsymbol{g} \in \check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}_{L,L}|_{x=\mathring{x}}}(\check{f}(1-\mathring{x}, \boldsymbol{g}))$ maintain the same distance below $(1-\mathring{x})^2$ as $\check{F}_{\check{X}^{\mathrm{IP}}}(\mathring{x})$ and $\min_{\boldsymbol{g} \in \check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}_{L,L}|_{x=\mathring{x}}}(\check{f}(\mathring{x}, \boldsymbol{g}))$, respectively. Since the second pair coincides, so must the first pair, such that

$$\check{F}_{\check{X}^{\mathrm{IP}}}(1-\mathring{x}) = \min_{\boldsymbol{g} \in \check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}|_{x=\mathring{x}}} (\check{f}(1-\mathring{x}, \boldsymbol{g})),$$

and therefore sharpness holds on $1 - \Phi_x(\tilde{\check{X}}^{\mathrm{LP}})$. □

Case II.A.3 : $\frac{1}{2} \in [\mathring{x}_-, \mathring{x}_+]$.

Since we showed sharpness on both $\mathrm{conv}(\check{X}^{\mathrm{IP}} \cap [0, \frac{1}{2}])$ and $\mathrm{conv}(\check{X}^{\mathrm{IP}} \cap [\frac{1}{2}, 1])$, we only have to show sharpness on the gap $(\mathring{x}_-, \mathring{x}_+)$ in $\check{X}^{\mathrm{IP}}$. Note, in this case, $\frac{1}{2} \notin \check{X}^{\mathrm{IP}}$. We wish to show that $\min_{\boldsymbol{g} \in \check{P}^{\mathrm{LP},\underline{\boldsymbol{\alpha}}}|_{x=\mathring{x}}}(\check{f}(\mathring{x}, \boldsymbol{g}))$ coincides with the line between $(\mathring{x}_-, \check{f}(\mathring{x}_-))$ and $(\mathring{x}_+, \check{f}(\mathring{x}_+))$.

To show this, we first note that both endpoints coincide with $\check{f}^{j_{\max}}(x, \boldsymbol{g}^*)$ for some $j_{\max}$, and by Lemma 3, both this value of j and the corresponding solution $\boldsymbol{g}^*$ must be the same for both gap endpoints. Further, since $\mathring{x}_-, \mathring{x}_+$ are the endpoints of a gap, we have that $\check{f}(\mathring{x}_-) = \mathring{x}_-^2$ and $\check{f}(\mathring{x}_+) = \mathring{x}_+^2$. This can be seen as follows: first, by [7, Lemma 6], we have that each $\check{f}^j$, $j \geqslant 0$, is incident with x^2 exactly at the points $x = \frac{k}{2^j} + \frac{1}{2^{j+1}}$, $k = 0, \ldots, 2^j - 1$. Furthermore, the points at which the $\boldsymbol{\alpha}$-vector changes, and thus the possible gaps in $\check{X}^{\mathrm{IP}}$, are exactly the points $\mathring{x} = k2^{-L}$, which must take the form above for some $j \in [\![0, L-1]\!]$, so that $\check{F}^{j-1}(x) = x^2$ for $x \in \{\mathring{x}_-, \mathring{x}_+\}$. Since each other $\check{f}^j(x) \leqslant x^2$ at these points, this yields $\check{f}(x) = x^2$ for $x \in \{\mathring{x}_-, \mathring{x}_+\}$.

Now, let $[a_1, b_1]$ be the bounds on g_1 from Lemma 1. Then we have $g_1^* = b_1$: through the mapping Φ, we have $g_1^* = \tilde{x} = \tilde{b}_0$ at both $\mathring{x}_-$ and $\mathring{x}_+$, where $\tilde{b}_0$ is defined in the manner of Lemma 1. Thus, since g_1 is subject to every constraint in $\check{P}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}$ that $\tilde{x}$ is in $\tilde{\check{P}}^{\mathrm{IP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}$, we have that $b_1 \leqslant \tilde{b}_0 = g_1^* \leqslant b_1$, such that $g_1^* = b_1$.

Furthermore, by the convexity of $\text{proj}_{x,\boldsymbol{g}}\left(\check{P}^{\text{LP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}\right)$, since $(\mathring{x}_-, \boldsymbol{g}^*), (\mathring{x}_+, \boldsymbol{g}^*) \in \text{proj}_{x,\boldsymbol{g}}\left(\check{P}^{\text{LP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}\right)$, we have that $(\mathring{x}, \boldsymbol{g}^*) \in \text{proj}_{x,\boldsymbol{g}}\left(\check{P}^{\text{LP},\underline{\boldsymbol{\alpha}}}_{(x,\boldsymbol{g},\boldsymbol{\alpha})}\right)$ for all $\hat{x} \in (\mathring{x}_-, \mathring{x}_+)$. Thus, we have for any such $\mathring{x}$ that

$$g_1^* = b_1 \geqslant \min(2\mathring{x}, 2(1-\mathring{x}), b_1) \geqslant g_1^*,$$

yielding by Lemma 2 that $\boldsymbol{g}^* \in \text{argmin}\{z : (z, \boldsymbol{g}) \in \text{proj}_{z,\boldsymbol{g}}(\check{P}^{\text{LP},\underline{\boldsymbol{\alpha}}}_{L,L})|_{x=\mathring{x}}\}$. Thus, we have

$$\check{f}(\mathring{x}, \boldsymbol{g}^*) = \min_{\boldsymbol{g} \in \check{P}^{\text{LP},\underline{\boldsymbol{\alpha}}}_{L,L}|_{x=\mathring{x}}} (\check{f}(\mathring{x}, \boldsymbol{g})) = \check{f}(\mathring{x})$$

is linear in $\mathring{x}$ across the gap $[\mathring{x}_-, \mathring{x}_+]$ and coincides with $\check{f}(\mathring{x})$ at the endpoints, as required. Therefore, we have that $\check{P}^{\text{LP},\underline{\boldsymbol{\alpha}}}_{L,L}$ is sharp across the gap. We have now established sharpness of $\check{P}^{\text{LP},\underline{\boldsymbol{\alpha}}}_{L,L}$ over all of $\text{conv}(P_{(x,\boldsymbol{g},\boldsymbol{\alpha})})$, and thus the proof is complete for $1 \notin I$.

Case II.B: $1 \in I$. Finally, to recover sharpness if $1 \in I$, we only have to observe that inserting 1 into I, thereby restricting $\alpha_1 = 1$ or $\alpha_1 = 0$, simply restricts $\check{P}^{\text{IP},\underline{\boldsymbol{\alpha}}}_{L,L}$ to either $x \in \Phi_x(\check{X}^{\text{IP}})$ or $x \in 1 - \Phi_x(\check{X}^{\text{IP}})$, on which sharpness holds exactly as the sharpness result on the image of Φ (or its reflection) with $1 \in I$, with one difference: we define Φ so that $\alpha_1 = \hat{\alpha}_1$. However, this difference has no effect on the z-minimal solutions for g_1^* within $\check{X}^{\text{LP}}$, and thus no effect on sharpness. □

C Auxiliary results and proofs

In this section of the appendix, we give the proofs of Lemma 4 and Proposition 7 which we have moved here for better readability.

C.1 Epigraphs over non-contiguous domains

Here we present the proof of Lemma 4.

Proof of Lemma 4 We first note that we have $F_X(x) \geqslant F(x)$ for all $x \in \text{conv}(X)$: for all $x \in \text{conv}(X)$, we have that either $F_X(x) = F(x)$ or that $F_X(x)$ is the line between two points on the graph of f, which must lie above the graph of f by the convexity of f. Further, we have that F_X is convex, as it is a maximum between the convex function F and some of its secant lines, which are also convex.

Now, trivially, by the convexity of F_X, we have

$$\text{conv}(\text{epi}_X(F)) = \text{conv}(\text{epi}_X(F_X)) \subseteq \text{conv}(\text{epi}_{\text{conv}(X)}(F_X)) = \text{epi}_{\text{conv}(X)}(F_X)$$

To show that $\text{epi}_{\text{conv}(X)}(F_X) \subseteq \text{conv}(\text{epi}_X(F))$, let $(x, y) \in \text{epi}_{\text{conv}(X)}(F_X)$. Then if $x \in X$, $y \geqslant F_X(x) = F(x)$, such that $(x, y) \in \text{epi}_X(F) \subseteq \text{conv}(\text{epi}_X(F))$. On

the other hand, if $x \in \text{conv}(X) \setminus X$, then by definition of F_X we have that there exist some $\lambda \in [0, 1]$ and $x_1, x_2 \in X$ such that $x = \lambda x_1 + (1 - \lambda)x_2$ and $F_X(x) = \lambda F(x_1) + (1-\lambda) F(x_2)$. Then we have that (x, y) is a convex combination of the points $(x_1, f(x_1) + (y - F_X(x)))$ and $(x_2, F(x_2) + (y - F_X(x)))$, which are in $\text{epi}_X(F)$ (since $y - F_X(x) \geqslant 0$), yielding $(x, y) \in \text{conv}(\text{epi}_X(F))$ as required. □

C.2 Volume proof for Bin2 and Bin3

Now we prove Proposition 7.

Proof of Proposition 7 Let P^{IP}_{L,L_1} be the MIP relaxation Bin2, where F^L is the sawtooth approximation of $z_x = x^2$ and $z_y = y^2$ that consists of secant lines to x^2 between consecutive breakpoints $x_k = k2^{-L}$ and $y_k = k2^{-L}$ for $k \in [\![0, 2^L]\!]$. Further, for $L_1 \to \infty$ we have

$$\begin{aligned}&\lim_{L_1 \to \infty} \{(p, z_p) \in [0, 1] \times \mathbb{R} : (p, z_p) \in Q^{L_1}\} \\ &\quad = \{(p, z_p) \in [0, 1] \times \mathbb{R} : (p, z_p) \in \text{epi}_{[0,1]}(p^2)\}\end{aligned}$$

under Hausdorff distance. As a result, we obtain

$$\begin{aligned}\lim_{L, L_1 \to \infty} (\text{proj}_{x,y,z}(P^{\text{IP}}_{L,L_1})) = \Big\{&(x, y, z) \in [0, 1]^2 \times \mathbb{R} : \\ &\tfrac{1}{2}\left((x + y)^2 - F^L(x) - F^L(y)\right) \leqslant z \leqslant \tfrac{1}{2}\left(4F^L\left(\tfrac{x+y}{2}\right) - x^2 - y^2\right)\Big\}.\end{aligned}$$

Now let and $w_x = w_y = 2^{-(L-1)}$ be the distance between any two consecutive breakpoints x_k, x_{k-1} and y_k, y_{k-1}, respectively, and consider the volume of $\text{proj}_{x,y,z}(P^{\text{IP}}_{L,L_1})$ over the grid piece $[x_{k-1}, x_k] \times [y_{k-1}, y_k]$:

$$\begin{aligned}&\lim_{L, L_1 \to \infty} \text{vol}(\text{proj}_{x,y,z}(P^{\text{IP}}_{L,L_1})) \\ &= \tfrac{1}{2}\int_{x_{k-1}}^{x_k}\int_{y_{k-1}}^{y_k}\left(4F^L\left(\tfrac{x+y}{2}\right) - x^2 - y^2 - \left((x + y)^2 - F^L(x) - F^L(y)\right)\right) dydx \\ &= \tfrac{1}{2}\int_{x_{k-1}}^{x_k}\int_{y_{k-1}}^{y_k}\left(\left(4F^L\left(\tfrac{x+y}{2}\right) - (x + y)^2\right) + (F^L(x) - x^2) + (F^L(y) - y^2)\right) dydx \\ &= \tfrac{w_y}{2}\int_{x_{k-1}}^{x_k}(F^L(x) - x^2)\, dx + \tfrac{w_x}{2}\int_{y_{k-1}}^{y_k}(F^L(y) - y^2)\, dy \\ &\quad + 2\int_{x_{k-1}}^{x_k}\int_{y_{k-1}}^{y_k}\left(F^L\left(\tfrac{x+y}{2}\right) - (\tfrac{x+y}{2})^2\right) dydx.\end{aligned}$$

The first two integrals are each the overapproximation volumes for the sawtooth approximation over two consecutive univariate domain segments, each of which has an area of $\frac{1}{6}2^{-3L}$, see [7, Appendix A]. Thus, since $w_x = w_y = 2 * 2^{-L}$, we have that the first two integrals add up to $\frac{2}{3}2^{-4L}$.

To process the third integral, we apply the two substitutions $u = \frac{(x-x_{k-1})+(y-y_{k-1})}{2}$ and $v = \frac{(x-x_{k-1})-(y-y_{k-1})}{2}$. The integral then becomes

$$
\begin{aligned}
&2\int_{x_{k-1}}^{x_k}\int_{y_{k-1}}^{y_k}\left(F^L\left(\tfrac{x+y}{2}\right)-(\tfrac{x+y}{2})^2\right)dydx \\
&\quad= 2\int_0^{2^{-L}}(F^L\left(u+\tfrac{x_{k-1}+y_{k-1}}{2}\right)-(u+\tfrac{x_{k-1}+y_{k-1}}{2})^2)\int_{-u}^{u}1\,dvdu \\
&\qquad+ 2\int_{2^{-L}}^{2\cdot 2^{-L}}(F^L\left(u+\tfrac{x_{k-1}+y_{k-1}}{2}\right)-(u+\tfrac{x_{k-1}+y_{k-1}}{2})^2)\int_{-(2\cdot 2^{-L}-u)}^{2\cdot 2^{-L}-u}1\,dvdu \\
&\quad= 4\int_0^{2^{-L}}u(F^L\left(u+\tfrac{x_{k-1}+y_{k-1}}{2}\right)-(u+\tfrac{x_{k-1}+y_{k-1}}{2})^2)\,du \\
&\qquad+ 4\int_{2^{-L}}^{2\cdot 2^{-L}}(2\cdot 2^{-L}-u)(F^L\left(u+\tfrac{x_{k-1}+y_{k-1}}{2}\right)-(u+\tfrac{x_{k-1}+y_{k-1}}{2})^2)\,du \\
&\quad= 8\int_0^{2^{-L}}u(F^L\left(u+\tfrac{x_{k-1}+y_{k-1}}{2}\right)-(u+\tfrac{x_{k-1}+y_{k-1}}{2})^2)\,du \qquad \text{(J1)} \\
&\quad= 8\int_0^{2^{-L}}u(u(2^{-L}-u))\,du \qquad \text{(J2)} \\
&\quad= 8\int_0^{2^{-L}}(2^{-L}u^2-u^3)\,du \\
&\quad= 8(\tfrac{1}{3}2^{-4L}-\tfrac{1}{4}2^{-4L}) = \tfrac{2}{3}2^{-4L}.
\end{aligned}
$$

The steps J1 and J2 rely on the observation that F^L is the secant line to x^2 across the intervals $[\frac{x_{k-1}+y_{k-1}}{2}, \frac{x_{k-1}+y_{k-1}}{2}+2^{-2L}]$ and $[\frac{x_{k-1}+y_{k-1}}{2}+2^{-2L}, \frac{x_{k-1}+y_{k-1}}{2}+2\cdot 2^{-2L}]$, due to the positions of x_{k-1} and y_{k-1}. In addition, for some $\mathring{x} \in [x_{k-1}, x_k]$, the error between and x^2 and the secant line to x^2 at points x_{k-1} and x_k is given by $(x-x_{k-1})(x_k-x)$ - the product of distances to each endpoint. Thus, for $u \in [0, 2^{-L}]$, we have

$$F^L\left(u+\tfrac{x_{k-1}+y_{k-1}}{2}\right)-(u+\tfrac{x_{k-1}+y_{k-1}}{2})^2 = u(2^L-u),$$

yielding the validity of step J2. On the other hand, to show that step J1 is valid, we observe for $u \in [0, 2^{-L}]$ that

$$F^L\left(u+\tfrac{x_{k-1}+y_{k-1}}{2}\right)-(u+\tfrac{x_{k-1}+y_{k-1}}{2})^2 = (u-2^L)(2^{-2L}-u)$$

holds, such that the second integral becomes the first integral under the substitution $\tilde{u} = 2^{-L} - u$, since the secant-error portion of the integrand is symmetric about $u = 2^{-L}$. Thus, the volume related to the second integral is $\frac{4}{3}2^{-4L}$. The volume of P^{IP}_{L,L_1} over each grid piece converges to $2\cdot 2^{-4L}$, yielding a total volume convergence

of

$$\lim_{L_1 \to \infty} \text{vol}(\text{proj}_{x,y,z}(P^{\text{IP}}_{L,L_1})) = 2^{2(L-1)}(2 \cdot 2^{-4L}) = \tfrac{1}{2}2^{-2L}.$$

The proof for Bin3 is similar and therefore omitted here. □

D Instance set

In Table 8 we show a listing of all instances of the computational study from Sect. 6. The boxQP instances are publicly available at https://github.com/joehuchette/quadratic-relaxation-experiments. The ACOPF instances are also publicly available at https://github.com/robburlacu/acopflib. The QPLIB instances are available at https://qplib.zib.de/. In total, we have 60 instances, of which 30 are dense and 30 are sparse.

Table 8 IDs of all 60 instances used in the computational study

boxQP instances: spar				
020-100-1	**020-100-2**	**030-060-1**	**030-060-3**	**040-030-1**
040-030-2	**050-030-1**	**050-030-2**	**060-020-1**	**060-020-2**
070-025-2	**070-050-1**	**080-025-1**	**080-050-2**	**090-025-1**
090-050-2	**100-025-1**	**100-050-2**	**125-025-1**	**125-050-1**
ACOPF instances: miqcqp_ac_opf_nesta_case				
3_lmbd_api	4_gs_api	4_gs_sad	5_pjm_api	5_pjm_sad
6_c_api	6_c_sad	6_ww_sad	6_ww	9_wscc_api
9_wscc_sad	14_ieee_api	14_ieee_sad	24_ieee_rts_api	24_ieee_rts_sad
29_edin_api	29_edin_sad	30_fsr_api	30_ieee_sad	9_epri_api
QPLIB instances: QPLIB_				
0031	**0032**	**0343**	0681	0682
0684	0698	**0911**	**0975**	**1055**
1143	**1157**	**1423**	**1922**	2882
2894	2935	2958	3358	3814

In bold are the IDs of the instances that are dense

References

1. Adjiman, C.S., Dallwig, S., Floudas, C.A., Neumaier, A.: A global optimization method, αbb, for general twice-differentiable constrained NLPs—i. theoretical advances. Comput. Chem. Eng. **22**(9), 1137–1158 (1998)
2. Aigner, K.-M., Burlacu, R., Liers, F., Martin, A.: Solving ac optimal power flow with discrete decisions to global optimality. INFORMS J. Comput. **35**(2), 458–474 (2023)
3. Androulakis, I.P., Maranas, C.D., Floudas, C.A.: αbb: a global optimization method for general constrained nonconvex problems. J. Glob. Optim. **7**(4), 337–363 (1995)
4. Appa, G.M., Pitsoulis, L., Williams, H.P.: Handbook on Modelling for Discrete Optimization, vol. 88. Springer Science & Business Media, Berlin (2006)
5. Bärmann, A., Burlacu, R., Hager, L., Kleinert, T.: On piecewise linear approximations of bilinear terms: structural comparison of univariate and bivariate mixed-integer programming formulations. J. Glob. Optim. **85**(4), 789–819 (2023)
6. Beach, B., Hildebrand, R., Ellis, K., Lebreton, B.: An approximate method for the optimization of long-horizon tank blending and scheduling operations. Comput. Chem. Eng. **141**, 106839 (2020)
7. Beach, B., Hildebrand, R., Huchette, J.: Compact mixed-integer programming formulations in quadratic optimization. J. Glob. Optim. **84**(4), 869–912 (2022)
8. Belotti, P., Lee, J., Liberti, L., Margot, F., Wächter, A.: Branching and bounds tightening techniques for non-convex MINLP. Optim. Methods Softw. **24**(4–5), 597–634 (2009)
9. Billionnet, A., Elloumi, S., Lambert, A.: Extending the QCR method to general mixed-integer programs. Math. Program. **131**(1–2), 381–401 (2012)
10. Bärmann, A., Martin, A., Schneider, O.: The bipartite Boolean quadric polytope with multiple-choice constraints, 2022. Available at: arXiv:2009.11674
11. Burlacu, R., Geißler, B., Schewe, L.: Solving mixed-integer nonlinear programmes using adaptively refined mixed-integer linear programmes. Optim. Methods Softw. **35**(1), 37–64 (2020)
12. Castillo, P.A.C., Castro, P.M., Mahalec, V.: Global optimization of MIQCPs with dynamic piecewise relaxations. J. Glob. Optim. **71**(4), 691–716 (2018)
13. Castro, P.M.: Normalized multiparametric disaggregation: an efficient relaxation for mixed-integer bilinear problems. J. Glob. Optim. **64**(4), 765–784 (2015)
14. Castro, P.M.: Tightening piecewise McCormick relaxations for bilinear problems. Comput. Chem. Eng. **72**, 300–311 (2015)
15. Castro, P.M.: Source-based discrete and continuous-time formulations for the crude oil pooling problem. Comput. Chem. Eng. **93**, 382–401 (2016)
16. Castro, P.M., Liao, Q., Liang, Y.: Comparison of mixed-integer relaxations with linear and logarithmic partitioning schemes for quadratically constrained problems. Optim. Eng. **23**, 717–747 (2022)
17. Chen, J., Burer, S.: Globally solving nonconvex quadratic programming problems via completely positive programming. Math. Program. Comput. **4**(1), 33–52 (2012)
18. Coffrin, C., Gordon, D., Scott, P.: NESTA, the NICTA energy system test case archive. arXiv preprint arXiv:1411.0359 (2014)
19. Correa-Posada, C.M., Sánchez-Martín, P.: Gas network optimization: a comparison of piecewise linear models. Optimization Online (2014)
20. Dolan, E.D., Moré, J.J.: Benchmarking optimization software with performance profiles. Math. Program. **91**(2), 201–213 (2002)
21. Dong, H.: Relaxing nonconvex quadratic functions by multiple adaptive diagonal perturbations. SIAM J. Optim. **26**(3), 1962–1985 (2016)
22. Dong, H., Luo, Y.: Compact disjunctive approximations to nonconvex quadratically constrained programs. arXiv preprint: arXiv:1811.08122 (2018)
23. Faria, D.C., Bagajewicz, M.J.: Novel bound contraction procedure for global optimization of bilinear MINLP problems with applications to water management problems. Comput. Chem. Eng. **35**(3), 446–455 (2011)
24. Furini, F., Traversi, E., Belotti, P., Frangioni, A., Gleixner, A., Gould, N., Liberti, L., Lodi, A., Misener, R., Mittelmann, H., et al.: Qplib: a library of quadratic programming instances. Math. Program. Comput. **11**(2), 237–265 (2019)
25. Furini, F., Traversi, E., Belotti, P., Frangioni, A., Gleixner, A., Gould, N., Liberti, L., Lodi, A., Misener, R., Mittelmann, H., Sahinidis, N.V., Vigerske, S., Wiegele, A.: QPLIB: a library of quadratic programming instances. Math. Program. Comput. **11**(2), 237–265 (2019)

26. Galli, L., Letchford, A.N.: A compact variant of the QCR method for quadratically constrained quadratic 0–1 programs. Optim. Lett. **8**(4), 1213–1224 (2014)
27. Geißler, B., Martin, A., Morsi, A., Schewe, L.: Using piecewise linear functions for solving MINLPs. In: Mixed Integer Nonlinear Programming, pp. 287–314. Springer (2012)
28. Gurobi Optimization, LLC: Gurobi Optimizer Reference Manual (2022)
29. Huchette, J.A.: Advanced mixed-integer programming formulations: methodology, computation, and application. PhD thesis, Massachusetts Institute of Technology (2018)
30. Joly, M., Pinto, J.M.: Mixed-integer programming techniques for the scheduling of fuel oil and asphalt production. Chem. Eng. Res. Des. **81**(4), 427–447 (2003)
31. Kolodziej, S.P., Grossmann, I.E., Furman, K.C., Sawaya, N.W.: A discretization-based approach for the optimization of the multiperiod blend scheduling problem. Comput. Chem. Eng. **53**, 122–142 (2013)
32. Kutzer, K.: Using piecewise linear approximation techniques to handle bilinear constraints. PhD thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg (2020)
33. Linderoth, J.: A simplicial branch-and-bound algorithm for solving quadratically constrained quadratic programs. Math. Program. **103**(2), 251–282 (2005)
34. McCormick, G.P.: Computability of global solutions to factorable nonconvex programs: part I—convex underestimating problems. Math. Program. **10**(1), 147–175 (1976)
35. Misener, R., Floudas, C.A.: Global optimization of mixed-integer quadratically-constrained quadratic programs (MIQCQP) through piecewise-linear and edge-concave relaxations. Math. Program. **136**(1), 155–182 (2012)
36. Nagarajan, H., Mowen, L., Wang, S., Bent, R., Sundar, K.: An adaptive, multivariate partitioning algorithm for global optimization of nonconvex programs. J. Global Optim. **74**, 639–675 (2019)
37. Phan-huy-Hao, E.: Quadratically constrained quadratic programming: some applications and a method for solution. Z. Oper. Res. **26**(1), 105–119 (1982)
38. Siqueira, A.S., da Silva, R.C., Santos, L.-R.: Perprof-py: a python package for performance profile of mathematical optimization software. J. Open Res. Softw. **4**(1), e12–e12 (2016)
39. Telgarsky, M.: Representation benefits of deep feedforward networks. arXiv:1509.08101 (2015)
40. Vielma, J.P., Ahmed, S., Nemhauser, G.: Mixed-integer models for nonseparable piecewise-linear optimization: unifying framework and extensions. Oper. Res. **58**(2), 303–315 (2010)
41. Wachter, A.: An interior point algorithm for large-scale nonlinear optimization with applications in process engineering. PhD thesis, Carnegie Mellon University (2002)
42. Yarotsky, D.: Error bounds for approximations with deep ReLU networks. Neural Netw. **94**, 103–114 (2017)