



# Enhancements of discretization approaches for non-convex mixed-integer quadratically constrained quadratic programming: part II

Benjamin Beach<sup>1</sup> · Robert Burlacu<sup>2</sup> · Andreas Bäermann<sup>3</sup> · Lukas Hager<sup>3</sup>  · Robert Hildebrand<sup>1</sup>

Received: 2 February 2023 / Accepted: 16 January 2024 / Published online: 18 March 2024  
© The Author(s) 2024

## Abstract

This is Part II of a study on mixed-integer programming (MIP) relaxation techniques for the solution of non-convex mixed-integer quadratically constrained quadratic programs (MIQCQPs). We set the focus on MIP relaxation methods for non-convex continuous variable products where both variables are bounded and extend the well-known MIP relaxation *normalized multiparametric disaggregation technique* (NMDT), applying a sophisticated discretization to both variables. We refer to this approach as *doubly discretized normalized multiparametric disaggregation technique* (D-NMDT). In a comprehensive theoretical analysis, we underline the theoretical advantages of the enhanced method D-NMDT compared to NMDT. Furthermore, we perform a broad computational study to demonstrate its effectiveness in

---

B. Beach and R. Hildebrand are supported by AFOSR grant FA9550-21-0107. R. Hildebrand was also partially supported by ONR Grant N00014-20-1-2156. Furthermore, we acknowledge 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 Bäermann  
andreas.baermann@math.uni-erlangen.de

Robert Hildebrand  
rhil@vt.edu

<sup>1</sup> Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA, USA

<sup>2</sup> Fraunhofer Institute for Integrated Circuits IIS, 90411 Nürnberg, Germany

<sup>3</sup> Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany

terms of producing tight dual bounds for MIQCQPs. Finally, we compare D-NMDT to the separable MIP relaxations from Part I and a state-of-the-art MIQCQP solver.

**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.

We enhance the *normalized multiparametric disaggregation technique* (NMDT) introduced in [7]. NMDT is a *McCormick relaxation* based MIP relaxation approach, which is applied to form relaxations of the quadratic equations  $z = x^2$  and  $z = xy$ . The McCormick relaxation is a set of four inequalities that describe the convex hull of the feasible points of the equation  $z = xy$  in the satisfying finite lower and upper bounds on  $x$  and  $y$ , see [16]. We extend NMDT by applying a discretization to both variables. We refer to the latter as *doubly discretized* NMDT (D-NMDT). Both MIP formulations, NMDT and D-NMDT, can be applied to MIQCQPs to form an MIP relaxation by introducing auxiliary variables and one such quadratic equation for each quadratic term in the MIQCQP. Such an MIP relaxation can then be solved with a standard MIP solver. We analyze these MIP relaxation approaches theoretically and computationally with respect to the quality of the dual bound they deliver for MIQCQPs.

For a thorough discussion of background on discretization and piecewise linear techniques in MIQCQPs, please refer to Part I [3].

**Contribution** We extend NMDT by a discretization of both variables, called D-NMDT. We analyze both MIP relaxations in terms of the dual bound they impose for non-convex MIQCQPs. In a theoretical analysis, we show that D-NMDT requires fewer binary variables and yields better *linear programming* (LP) relaxations at identical relaxation errors compared to NMDT. Finally, we perform an extensive numerical study where we use NMDT and D-NMDT to generate MIP relaxations of non-convex MIQCQPs. We show that D-NMDT has clear advantages, such as tighter dual bounds, shorter runtimes, and it finds more feasible solutions to the original MIQCQPs when combined with a callback function that uses the non-linear programming (NLP) solver IPOPT [19]. These effects become even more apparent in dense instances with many variable products. Moreover, we combine NMDT and D-NMDT with the *tighten sawtooth epigraph relaxation* from Part I [3] to obtain even tighter relaxations for  $z = x^2$  terms in MIQCQPs. This tightening leads to improved results in the computational study.

**Outline** In Sect. 2.1 and Sect. 2.2 we review several useful concepts, notations, and core formulations from Part I [3]. In Sect. 3, we recall the NMDT MIP relaxation and introduce the new MIP relaxation D-NMDT. In Sect. 4, we prove various properties about the strengths of the MIP relaxations focusing on volume, sharpness, and optimal choice of breakpoints. In Sect. 5, we present our computational study.

## 2 Preliminaries

### 2.1 MIP formulations

We follow Part I [3] for notation used in this work. We provide this section here for the completeness of this article. We study relaxations of general mixed-integer quadratically constrained quadratic programs (MIQCQPs), which are defined as

$$\begin{aligned}
 &\min x'Q_0x + c'_0x + d'_0y, \\
 &\text{s.t. } x'Q_jx + c'_jx + d'_jy + b_j \leq 0 \quad j \in 1, \dots, m, \\
 &\quad x_i \in [x_i, \bar{x}_i] \quad i \in 1, \dots, n, \\
 &\quad y \in \{0, 1\}^k,
 \end{aligned} \tag{1}$$

for  $Q_0, Q_j \in \mathbb{R}^{n \times n}, c_0, c_j \in \mathbb{R}^n, d_0, d_j \in \mathbb{R}^k$  and  $b_j \in \mathbb{R}, j = 1, \dots, m$ . Throughout this article, we use the following convenient notation: for any two integers  $i \leq j$ , we define  $\llbracket i, j \rrbracket := \{i, i + 1, \dots, j\}$ , and for an integer  $i \geq 1$  we define  $\llbracket i \rrbracket := \llbracket 1, i \rrbracket$ . 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 lower case letters and vectors of variables using bold face. For a vector  $\mathbf{u} = (u_1, \dots, u_n)$  and some index set  $I \subseteq \llbracket n \rrbracket$ , we write  $\mathbf{u}_I := (u_i)_{i \in I}$ . Thus, e.g.  $\mathbf{u}_{\llbracket i \rrbracket} = (u_1, \dots, u_i)$ . Furthermore, we introduce the following notation: for a function  $F: X \rightarrow \mathbb{R}$  and a subset  $B \subseteq X$ , let  $\text{gra}_B(F)$ ,  $\text{epi}_B(F)$  and  $\text{hyp}_B(F)$  denote the *graph*, *epigraph* and *hypograph* of the function  $F$  over the set  $B$ , respectively. That is,

$$\begin{aligned}
 \text{gra}_B(F) &:= \{(\mathbf{u}, z) \in B \times \mathbb{R} : z = F(\mathbf{u})\}, \\
 \text{epi}_B(F) &:= \{(\mathbf{u}, z) \in B \times \mathbb{R} : z \geq F(\mathbf{u})\}, \\
 \text{hyp}_B(F) &:= \{(\mathbf{u}, z) \in B \times \mathbb{R} : z \leq F(\mathbf{u})\}.
 \end{aligned}$$

In the following, we introduce MIP formulations as we will use them to represent these sets as well as the different notions of the strength of an MIP formulation explored in this work.

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

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

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

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

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

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

The corresponding *projected linear relaxation*  $\text{proj}_{\mathbf{u}}(P^{\text{LP}})$  onto the  $\mathbf{u}$ -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 [5, 14]. 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. In the following  $\text{conv}(S)$  denotes the convex hull of a set  $S$ .

**Definition 1** We say that the MIP formulation  $P^{\text{IP}}$  is *sharp* if

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

holds.

Sharpness expresses a tightness at the root node of 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  $P^{\text{IP}}$  to form relaxations of  $U$  in the projected space, as defined in the following.

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

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

Given a function  $F : [0, 1]^d \rightarrow \mathbb{R}$ , 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** For an MIP relaxation  $P^{\text{IP}}$  of a set  $U \subseteq \mathbb{R}^{d+1}$ , let  $\bar{\mathbf{u}} \in \text{proj}_{\mathbf{u}}(P^{\text{IP}})$ . We then define the *pointwise error* of  $\bar{\mathbf{u}}$  as

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

This enables us to define the following two error measures for  $P^{\text{IP}}$  w.r.t.  $U$ :

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

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

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

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

Via integral calculus, the second, volume-based error measure can be interpreted as the average pointwise error width of all points  $\mathbf{u} \in \text{proj}_{\mathbf{u}}(P^{\text{IP}})$ . Note that whenever the volume of  $U$  is zero (i.e. it is a lower-dimensional set), the average error width just reduces to the volume of  $\text{proj}_{\mathbf{u}}(P^{\text{IP}})$ . Both of the defined error quantities for an MIP relaxation  $P^{\text{IP}}$  can also be used to measure the tightness of the corresponding LP relaxation  $P^{\text{LP}}$ . The volume of LP relaxation as a measure of a MIP relaxation strength was previously used in [2].

### 2.2 Core relaxations

In the definition of the MIP relaxations studied in this work, we will frequently consider equations 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 will often use the function  $F : D \rightarrow \mathbb{R}$ ,  $F(x, y) = xy$ ,  $D := D_x \times D_y$ , and refer to the set of feasible solutions to the equation  $z = xy$  via the graph of  $F$ , i.e.  $\text{gra}_D(F) = \{(x, y, z) \in D \times \mathbb{R} : z = xy\}$ . In order to simplify the exposition, we will, for example, often write  $\text{gra}_D(xy)$  or refer to a relaxation of the equation  $z = xy$  instead of  $\text{gra}_D(F)$ . We will do this similarly for the univariate function  $f : D_x \rightarrow \mathbb{R}$ ,  $f(x) = x^2$  and equations of the form  $z = x^2$ , for example. For inequalities, like  $z \geq xy$  or  $z = x^2$ , we can use the epigraph.

Furthermore, we repeatedly make use of several “core” formulations for specific sets of feasible points. They are introduced in the following.

#### 2.2.1 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 [16]:

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

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

When one of the variables, here  $\beta$ , is binary, the McCormick envelope of  $z = x\beta$  simplifies to

$$\mathcal{M}(x, \beta) = \{(x, \beta, z) \in [\underline{x}, \bar{x}] \times [0, 1] \times \mathbb{R} : (6)\}. \tag{5}$$

$$\begin{aligned} \underline{x} \cdot \beta \leq z \leq \bar{x} \cdot \beta, \\ x - \bar{x} \cdot (1 - \beta) \leq z \leq x - \underline{x} \cdot (1 - \beta). \end{aligned} \tag{6}$$

For univariate continuous quadratic equations  $z = x^2$ , it simplifies to

$$\mathcal{M}(x, x) = \{(x, z) \in [\underline{x}, \bar{x}] \times \mathbb{R} : (8)\}. \tag{7}$$

$$\begin{aligned} z &\geq 2\underline{x} \cdot x - \underline{x}^2, \\ z &\geq 2\bar{x} \cdot x - \bar{x}^2, \\ z &\leq x(\bar{x} + \underline{x}) - \bar{x} \cdot \underline{x}. \end{aligned} \tag{8}$$

### 2.2.2 Sawtooth-based MIP formulations

Next, we state an MIP relaxation for equations of the form  $z \geq x^2$  that requires only logarithmically-many auxiliary variables and constraints in the number of linear segments. It makes use of an elegant piecewise linear (pwl) formulation for  $\text{gra}_{[0,1]}(x^2)$  from [20] using the recursively defined *sawtooth* function presented in [18] to formulate the approximation of  $\text{gra}_{[0,1]}(x^2)$ , as described in [5]. We will use this formulation to further strengthen the relaxation of  $z = x^2$  by NMDT or D-NMDT. To this end, we define a formulation parameterized by the depth  $L \in \mathbb{N}$ :

$$S^L := \{(x, \mathbf{g}, \boldsymbol{\alpha}) \in [0, 1] \times [0, 1]^{L+1} \times \{0, 1\}^L : (10)\} \tag{9}$$

$$\begin{aligned} g_0 &= x \\ 2(g_{j-1} - \alpha_j) &\leq g_j \leq 2g_{j-1} & j &= 1, \dots, L, \\ 2(\alpha_j - g_{j-1}) &\leq g_j \leq 2(1 - g_{j-1}) & j &= 1, \dots, L. \end{aligned} \tag{10}$$

Note that, by construction in [5, 20],  $S^L$  is defined such that 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  $j = 1, \dots, L$ , which means that it is given by the “tooth” function  $G: [0, 1] \rightarrow [0, 1]$ ,  $G(x) = \min\{2x, 2(1 - x)\}$ . Therefore, each  $g_j$  represents the output of a “sawtooth” function of  $x$ , as described in [18, 20], i.e. when  $\boldsymbol{\alpha} \in \{0, 1\}^L$ , we have

$$g_j = G^j(x) \quad \text{for } G^j := \underbrace{G \circ G \circ \dots \circ G}_j. \tag{11}$$

Now, we define the function  $F^L : [0, 1] \rightarrow [0, 1]$ ,

$$F^L(x) := x - \sum_{j=1}^L 2^{-2j} G^j(x), \tag{12}$$

which is a close approximation to  $x^2$ .

Using the relationships (11) and (12) between  $x$  and  $\mathbf{g}$ , any constraint of the form  $z = x^2$  can be approximated via the function

$$f^L : [0, 1] \times [0, 1]^{L+1} \rightarrow [0, 1],$$

$$f^L(x, \mathbf{g}) = x - \sum_{j=1}^L 2^{-2j} g_j, \quad \text{for an integer } L \geq 0. \tag{13}$$

Now, we consider the LP relaxation of  $S^L$ , where each variable  $\alpha_j$  is relaxed to the interval  $[0, 1]$ . Then, via the constraints (9), 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  $\alpha$ , the LP relaxation of  $S^L$  in terms of just  $x$  and  $\mathbf{g}$  can be stated as

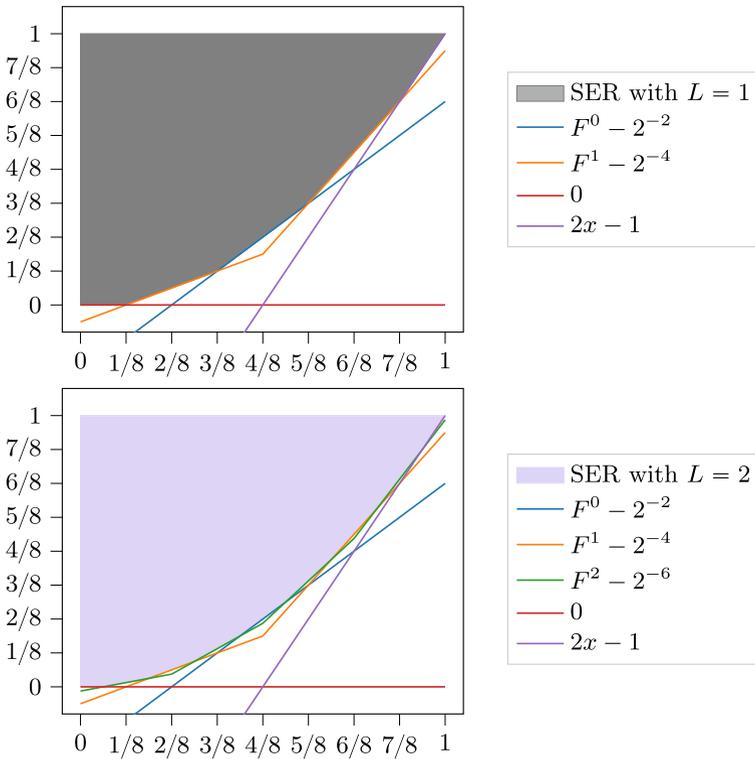
$$T^L = \left\{ (x, \mathbf{g}) \in [0, 1] \times [0, 1]^{L+1} : \begin{array}{ll} g_0 = x & \\ g_j \leq 2(1 - g_{j-1}) & j = 1, \dots, L \\ g_j \leq 2g_{j-1} & j = 1, \dots, L. \end{array} \right\} \tag{14}$$

The LP relaxation  $T^L$  is sharp by [3, Theorem 1]. Thus,  $T^L$  yields the same lower bound on  $z$  as the MIP formulation  $S^L$  due to sharpness and the convexity of  $F^L$ . This allows us to define an LP outer approximation for inequalities of the form  $z \geq x^2$ :

**Definition 4** (*Sawtooth Epigraph Relaxation, SER*) Given some  $L \in \mathbb{N}$ , the *depth- $L$  sawtooth epigraph relaxation* for  $z \geq x^2$  on the interval  $x \in [0, 1]$  is given by

$$Q^L := \left\{ (x, z) \in [0, 1] \times \mathbb{R} : \exists \mathbf{g} \in [0, 1]^{L+1} : \begin{array}{l} (16) \\ z \geq f^j(x, \mathbf{g}) - 2^{-2j-2} \quad j = 0, \dots, L \\ z \geq 0, \quad z \geq 2x - 1 \\ (x, \mathbf{g}) \in T^L. \end{array} \right\} \tag{15}$$

In [3] it is shown that that the maximum error for the sawtooth epigraph relaxation is  $2^{-2L-4}$  (Fig. 1).



**Fig. 1** The sawtooth epigraph relaxations  $Q^L$  for  $L = 1$  and  $L = 2$ . By increasing  $L$ , 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

### 3 MIP relaxations for non-convex MIQCQPs

In this section, we present MIP relaxations for bivariate equations of the form  $z = xy$  and univariate equations of the form  $z = x^2$ . 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.

We proceed as follows. First, we recall the well-known MIP relaxation technique NMDT. Then, we introduce an enhanced version of it, called D-NMDT, which is designed to reduce the number of binary variables required to reach the same level of approximation accuracy compared to NMDT for completely dense MIQCQPs. Finally, we define the two *tightened* variants of NMDT and D-NMDT, for which we also incorporate the sawtooth epigraph relaxation (15) for all  $z = x_i^2$  terms. We call these methods T-NMDT and T-D-NMDT, respectively. We will mention the corresponding maximum errors of the presented MIP relaxations and derive them in detail in Sect. 4.1.

### 3.1 Base-2 NMDT

The Normalized Multiparametric Disaggregation Technique (NMDT) was introduced by Castro [7]. Later it was used in [4, 5] along with its univariate form (see [5, Appendix A]). While in [7] a base of 10 was chosen for the discretization, in [4, 5] NMDT is described with a base of 2. We use the latter here and provide both the bivariate and univariate definition of base-2 NMDT according to [5] here.

In NMDT, the key idea for relaxing  $z = xy$  is to discretize one variable, e.g.  $x$ , using binary variables  $\beta \in \{0, 1\}^L$  and a residual term  $\Delta_x^L$  and then relaxing the resulting products  $\beta_i y$  and  $\Delta_x^L y$  using McCormick envelopes. The following derivation of NMDT can be transferred one-to-one to bases different to 2. We start with the base-2 discretization of the variable  $x$ :

$$x = \sum_{j=1}^L 2^{-j} \beta_j + \Delta_x^L.$$

Then we multiply by  $y$  to obtain the exact representation

$$x = \sum_{j=1}^L 2^{-j} \beta_j + \Delta_x^L, \quad z = \sum_{j=1}^L 2^{-j} \beta_j y + \Delta_x^L y \tag{17}$$

$$\Delta_x^L \in [0, 2^{-L}], \quad \beta \in \{0, 1\}^L.$$

Next, we use McCormick envelopes to model all remaining product terms,  $\beta_j y$  and  $\Delta_x^L \cdot y$ , to obtain the final formulation.

**Definition 5** (NMDT, [7]) The MIP relaxation *NMDT* of  $z = xy$  with  $x \in [0, 1]$ ,  $y \in [0, 1]$  and a depth of  $L \in \mathbb{N}$  is defined as follows:

$$x = \sum_{j=1}^L 2^{-j} \beta_j + \Delta_x^L$$

$$z = \sum_{j=1}^L 2^{-j} u_j + \Delta_z^L \tag{18}$$

$$(y, \beta_j, u_j) \in \mathcal{M}(y, \beta_j) \quad j = 1, \dots, L$$

$$(\Delta_x^L, y, \Delta_z^L) \in \mathcal{M}(\Delta_x^L, y)$$

$$\Delta_x^L \in [0, 2^{-L}], \quad y \in [0, 1], \quad \beta \in \{0, 1\}^L.$$

Since McCormick envelopes are exact reformulations of the variable products if at least one of the variables is required to be binary, the maximum error of NMDT with respect to  $z = xy$  is purely due to the McCormick relaxation of  $\Delta_z^L = \Delta_x^L \cdot y$ , with a value of  $2^{-L-2}$ .

An advantage of the NMDT approach compared to the separable formulations from Part I is that it requires fewer binary variables to reach the desired level of

accuracy for *bipartite* MIQCQPs, for which the quadratic part in each constraint is of the form  $\mathbf{x}^T Q \mathbf{y}$ . This is due to the fact that one has only to discretize either  $\mathbf{x} \in \mathbb{R}^n$  or  $\mathbf{y} \in \mathbb{R}^m$ . Thus, to reach a maximum error of  $2^{-2L-2}$  for each bilinear term, NMDT requires only  $2L \min\{m, n\}$  binary variables instead of the  $L(m+n)$  variables required by the approaches D-NMDT (see Sect. 3.2) or HybS (from Part I). In contrast, NMDT requires twice the number of binary variables to reach the same level of accuracy if all quadratic terms  $x_i x_k$  and  $x_l^2$  with  $k = 1, \dots, n$  and  $l = 1, \dots, m$  must be modelled, for example if  $Q$  is dense, see Table 1.

Next, we show how to model univariate quadratic equations  $z = x^2$  with the NMDT technique:

**Definition 6 (Univariate NMDT ([7]))** The MIP relaxation *NMDT* of  $z = x^2$  with  $x \in [0, 1]$  and a depth of  $L \in \mathbb{N}$  is defined as follows:

$$\begin{aligned}
 x &= \sum_{j=1}^L 2^{-j} \beta_j + \Delta_x^L \\
 z &= \sum_{j=1}^L 2^{-j} u_j + \Delta_z^L \tag{19} \\
 (x, \beta_j, u_j) &\in \mathcal{M}(x, \beta_j) & j = 1, \dots, L \\
 (\Delta_x^L, x, \Delta_z^L) &\in \mathcal{M}(\Delta_x^L, x) \\
 \Delta_x^L &\in [0, 2^{-L}], \quad x \in [0, 1], \quad \beta \in \{0, 1\}^L.
 \end{aligned}$$

Note that for any depth  $L$ , the univariate formulation NMDT yields a maximum error of slightly less than  $2^{-L-2}$  instead of the  $2^{-2L-2}$  in the sawtooth relaxation from [3]. Further, the formulation NMDT is not sharp. For example at  $x = \frac{1}{2}$ , its LP relaxation admits the solution  $\beta_j = \frac{1}{2}$  for all  $j \in \llbracket L \rrbracket$ ,  $\Delta_x^L = 2^{-L-1}$ ,  $u_j = 0$  for all  $j \in \llbracket L \rrbracket$ ,  $\Delta_z^L = 0$  and  $z = 0$ , which is not in the convex hull of the MIP formulation Univariate NMDT stated in (19).

However, we can tighten the lower bound on  $z$  in (19) by adding the sawtooth epigraph relaxation (15) of depth  $L_1$  (with  $L_1 \geq L$ ), i.e.  $(x, z) \in Q^{L_1}$ . We refer to NMDT with this lower-bound tightening for univariate quadratic terms as *T-NMDT*. Note that univariate T-NMDT is a sharp MIP formulation, which we discuss in Sect. 4.3.

**Definition 7 (Univariate T-NMDT)** The MIP relaxation *T-NMDT* of  $z = x^2$  with  $x \in [0, 1]$  and a depth of  $L, L_1 \in \mathbb{N}$  with  $L_1 \geq L$  is defined as follows:

$$\begin{aligned}
 (x, \Delta_x^L, z, \Delta_z^L, \mathbf{u}, \beta) &\text{ satisfy (19)} \\
 (x, z) &\in Q^{L_1}. \tag{20}
 \end{aligned}$$

### 3.2 Doubly discretized NMDT

The key idea behind the novel MIP relaxation *Doubly Discretized NMDT* (D-NMDT) for  $z = xy$  is to further increase the accuracy of NMDT by discretizing the second variable  $y$  as well, which leads to a *double* NMDT substitution, namely in the  $\Delta_x^L y$ -term. In this way, for problems where NMDT would require discretizing all  $x_i$ -variables, e.g. if we have some dense constraint, we can double the accuracy of the relaxation for the equations  $z_{ij} = x_i x_j$  without adding additional binary variables by taking advantage of the fact that both variables are discretized anyway. In NMDT, we could choose to discretize either  $x$  or  $y$  for each equation of the form  $z = xy$ . For D-NMDT, we consider both options of discretization, and then, by introducing a parameter  $\lambda \in [0, 1]$ , we can model a hybrid version of the two resulting MIP relaxations. Namely, we write

$$xy = \lambda xy + (1 - \lambda)xy,$$

then discretize  $y$  first in the relaxation of  $\lambda xy$  and  $x$  first in the relaxation of  $(1 - \lambda)xy$ . Finally, the complete MIP relaxation D-NMDT is obtained by relaxing the resulting products via McCormick envelopes (see Appendix A for the detailed derivation).

**Definition 8** (*D-NMDT*) The MIP relaxation *D-NMDT* of  $z = xy$  with  $x, y \in [0, 1]$ , a depth of  $L \in \mathbb{N}$  and the parameter  $\lambda \in [0, 1]$  is defined as follows:

$$\begin{aligned}
 x &= \sum_{j=1}^L 2^{-j} \beta_j^x + \Delta_x^L, & y &= \sum_{j=1}^L 2^{-j} \beta_j^y + \Delta_y^L \\
 z &= \sum_{j=1}^L 2^{-j} (u_j + v_j) + \Delta_z^L \\
 \left( \lambda \Delta_y^L + (1 - \lambda)y, \beta_j^x, u_j \right) &\in \mathcal{M} \left( \lambda \Delta_y^L + (1 - \lambda)y, \beta_j^x \right) & j &= 1, \dots, L \\
 \left( (1 - \lambda) \Delta_x^L + \lambda x, \beta_j^y, v_j \right) &\in \mathcal{M} \left( (1 - \lambda) \Delta_x^L + \lambda x, \beta_j^y \right) & j &= 1, \dots, L \\
 \left( \Delta_x^L, \Delta_y^L, \Delta_z^L \right) &\in \mathcal{M} \left( \Delta_x^L, \Delta_y^L \right) \\
 \Delta_x^L, \Delta_y^L &\in [0, 2^{-L}], & x, y &\in [0, 1], \quad \beta^x, \beta^y \in \{0, 1\}^L.
 \end{aligned}
 \tag{21}$$

As McCormick envelopes are exact reformulations of bilinear products if one of the variables is binary, we only make an error in the relaxation of the continuous variable product  $\Delta_x^L \Delta_y^L$ . This yields a maximum error of  $2^{-2L-2}$  for D-NMDT. For bounds on the terms  $(1 - \lambda) \Delta_x^L + \lambda x$  and  $\lambda \Delta_y^L + (1 - \lambda)y$ , see Appendix B.

**Remark 1** For our implementation of the D-NMDT technique used in Sect. 5, we set  $\lambda = \frac{1}{2}$  for the sake of formulation symmetry in  $x$  and  $y$ .

To model the univariate quadratic terms with this method, we set  $y = x$  in  $z = xy$  and get an MIP relaxation for  $z = x^2$ . The resulting MIP relaxation is stronger than the univariate NMDT approach from Definition 6, which we will prove later.

**Definition 9** (*Univariate D-NMDT*) The MIP relaxation *D-NMDT* of  $z = x^2$  with  $x \in [0, 1]$  and a depth of  $L \in \mathbb{N}$  is defined as follows:

$$\begin{aligned}
 x &= \sum_{j=1}^L 2^{-j} \beta_j + \Delta_x^L \\
 z &= \sum_{j=1}^L 2^{-j} u_j + \Delta_z^L \\
 (\Delta_x^L + x, \beta_j, u_j) &\in \mathcal{M}(\Delta_x^L + x, \beta_j) & j = 1, \dots, L \\
 (\Delta_x^L, \Delta_z^L) &\in \mathcal{M}(\Delta_x^L, \Delta_z^L) \\
 \Delta_x^L &\in [0, 2^{-L}], \quad x \in [0, 1], \quad \beta \in \{0, 1\}^L.
 \end{aligned}
 \tag{22}$$

Again, as McCormick envelopes are exact reformulations of bilinear products if one of the variables is required to be binary, we only make an error in the relaxation of the continuous variable product  $\Delta_x^L \Delta_z^L$ . This yields a maximum error of  $2^{-2L-2}$  for univariate D-NMDT. Note that the upper bound of this formulation is formed by exactly the same pwl approximation for  $z = x^2$  as the sawtooth formulations. Unfortunately, the univariate D-NMDT is not sharp; for example, at  $x = \frac{1}{2}$ , its LP relaxation admits the solution  $\beta_j = \frac{1}{2}$  for all  $j \in \llbracket L \rrbracket$ ,  $\Delta_x^L = 2^{-L-1}$ ,  $\Delta_z^L = 0$ ,  $u_j = 0$  for all  $j \in \llbracket L \rrbracket$  and  $z = 0$ , which is not in the convex hull of  $\text{gra}_{[0,1]}(x^2)$ .

To formulate a tightened version of D-NMDT, we tighten the lower bound on  $z$  in (22), by removing all McCormick lower bounds and adding the sawtooth epigraph relaxation (15) of depth  $L_1$  (with  $L_1 \geq L$ ). Note that univariate T-D-NMDT is a sharp MIP formulation, which we will prove in Sect. 4.3.

**Definition 10** (*Univariate T-D-NMDT*) The MIP relaxation *T-D-NMDT* of  $z = x^2$  with  $x \in [0, 1]$  and depths  $L, L_1 \in \mathbb{N}$  with  $L_1 \geq L$  is defined as follows:

$$\begin{aligned}
 (x, \Delta_x^L, z, \Delta_z^L, \mathbf{u}, \boldsymbol{\beta}) &\text{ satisfy (22)} \\
 (x, z) &\in Q^{L_1}.
 \end{aligned}
 \tag{23}$$

In Table 1 in Sect. 4, we give a summary of the number of binary variables and constraints as well as the accuracy of each MIP relaxation when applied to a dense MIQCQP of the form (1).

**Remark 2** (Binary Variables and Dense MIQCQPs) When modelling Problem (1) using the MIP relaxations NMDT and D-NMDT, for each variable  $x_i$ , we will need a discretization of the form  $x_i = \sum_{j=1}^L 2^{-j} \beta_j + \Delta_{x_i}^L$  with  $\beta \in \{0, 1\}^L$ . Thus, both of these formulations use  $nL$  binary variables in the case of a dense MIQCQP. However, the improved binarizations in D-NMDT reduces the errors exponentially compared to NMDT.

**Table 1** A summary of characteristics of the different MIP relaxations for  $z = xy$

| MIP relax. | # Bin. var.             | # Constraints                               | Max. err.   | Avg. err. width      |
|------------|-------------------------|---|-------------|----------------------|
| NMDT       | $nL$                    | $n(\frac{1}{2}(5n + 7) + 2(n + 1)L)$        | $2^{-L-2}$  | $\frac{1}{6}2^{-L}$  |
| D-NMDT     | $nL$                    | $n(\frac{1}{2}(5n + 5) + 4nL)$              | $2^{-2L-2}$ | $\frac{1}{6}2^{-2L}$ |
| 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       | $\frac{1}{2}(n^2 + 1)L$ | $n(\frac{1}{2}(3n - 1) + (n + 1)(L + L_1))$ | $2^{-2L-1}$ | $\frac{1}{2}2^{-2L}$ |

Binary variables and constraints are given in the worst-case, in which every possible quadratic term is modelled, for example if some matrix  $Q_i$  is dense. The average errors for HybS, Bin2 and Bin3 with respect to  $\text{gra}_{[0,1]^2}(xy)$  are calculated for  $L_1 \rightarrow \infty$  and without the McCormick envelopes added. Finally, the average errors for Bin2 and Bin3 apply only to  $L \geq 1$ ; the corresponding volumes are  $\frac{7}{12}$  for  $L = 0$ . Finite  $L_1$  leads to slightly increased error bounds for the methods Bin2, Bin3 and HybS

Note that it is possible that some preprocessing or reformulation, such as via a convex quadratic reformulation may improve the number of binary variables needed. We do not use such reformulations in this work, but just focus on applying our MIP relaxations as is.

### 4 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. In particular, we analyze their maximum error, average errors, formulation strengths, i.e. sharpness, as well as the optimal placement of breakpoints to minimize average errors. Our results are summarized in Table 1, which also includes the results for the separable methods HybS, Bin2, and Bin3 from Part I [3].

#### 4.1 Maximum error

We start by discussing the maximum errors. We will derive the maximum errors of the NMDT-based formulations by reducing the error calculations to the error of a single McCormick relaxation per grid piece. In general, for the equation  $z = xy$  over a grid piece  $[\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}]$ , 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. [15, page23].

For NMDT, to show that the maximum error can be computed from a single McCormick relaxation, we fix  $\beta \in \{0, 1\}^L$  in (18) and observe two facts: (1) we get  $x = k2^{-L} + \Delta_x^L$  for some integer  $k$  and therefore  $x$  varies only with  $\Delta_x^L \in [0, 2^{-L}]$ , and (2) the McCormick relaxation  $(y, \beta_i, u_i) \in \mathcal{M}(y, \beta_i)$  is exact for each  $i = 1, \dots, L$ , i.e. , the relaxation equals  $u_i = y\beta_i$ . These two facts imply that the only error incurred on this small interval stems from the single McCormick relaxation  $(\Delta_x^L, y, \Delta_z^L) \in \mathcal{M}(\Delta_x^L, y)$  over regions of the form  $(\Delta_x^L, y) \in [0, 2^{-L}] \times [0, 1]$ .

This yields a maximum error of  $\frac{1}{4}(2^{-L} \cdot 1) = 2^{-L-2}$ . Similarly, for D-NMDT and univariate NMDT and D-NMDT, one can also show that all errors come from the McCormick relaxations of the continuous error terms. The maximum errors of the different MIP relaxations are listed in the following propositions.

**Proposition 1** *The maximum error in the NMDT MIP relaxation for  $z = xy$  with  $x, y \in [0, 1]$  is  $\frac{1}{4}(2^{-L} \cdot 1) = 2^{-L-2}$ .*

Likewise, for D-NMDT, the maximum error in  $z = xy$  is purely in the McCormick relaxation of the term  $(\Delta_x^L, \Delta_y^L, \Delta_z^L) \in \mathcal{M}(\Delta_x^L, \Delta_y^L)$  over the region  $(\Delta_x^L, \Delta_y^L) \in [0, 2^{-L}] \times [0, 2^{-L}]$ , yielding a maximum error of  $\frac{1}{4}(2^{-L} \cdot 2^{-L}) = 2^{-2L-2}$ .

**Proposition 2** *The maximum error in the D-NMDT MIP relaxation for  $z = xy$  with  $x, y \in [0, 1]$  is  $\frac{1}{4}(2^{-L} \cdot 2^{-L}) = 2^{-2L-2}$ .*

For univariate D-NMDT, the maximum error in  $z = x^2$  arises from the McCormick relaxation  $(\Delta_x^L, \Delta_z^L) \in \mathcal{M}(\Delta_x^L, \Delta_x^L)$  over the interval  $\Delta_x^L \in [0, 2^{-L}]$ , yielding a maximum error of  $2^{-2L-2}$ .

**Proposition 3** *The maximum error in the univariate D-NMDT MIP relaxation for  $z = x^2$  with  $x, y \in [0, 1]$  is  $2^{-2L-2}$ .*

Finally, for univariate NMDT, the error is incurred by the McCormick relaxation  $(\Delta_x^L, x, \Delta_z^L) \in \mathcal{M}(\Delta_x^L, x)$  over the box  $(\Delta_x^L, x) \in [0, 2^{-L}] \times [0, 1]$  with  $x = k2^{-L} + \Delta_x^L$  for some  $k \in \{0, \dots, 2^L - 1\}$ . Over this box, the error-maximizing point  $(x, \Delta_x^L) = (\frac{1}{2}, 2^{-L-1})$  derived in [15] is not feasible, as  $x = \frac{1}{2}$  implies  $\Delta_x^L = 0$ . In fact, we can show that the maximum error is slightly less than the expected  $2^{-L-2}$ .

To prove this, we focus on the maximum error of the underestimating part of the McCormick envelope with respect to  $x\Delta_x^L$  and skip the overestimating part as it works analogously. By (4), the McCormick relaxation underestimator over the box  $(\Delta_x^L, x) \in [0, 2^{-L}] \times [0, 1]$  is given as

$$\max_{\substack{\Delta_x^L \in [0, 2^{-L}], \\ k \in \{0, \dots, 2^L - 1\}}} \{0, \Delta_x^L - 2^{-L}(1 - x)|x = k2^{-L} + \Delta_x^L\}.$$

The underestimator is zero at points in the domain where

$$\Delta_x^L \leq -2^{-L}x + 2^{-L} = 2^{-L}(1 - 2^{-L}k - \Delta_x^L) \tag{24}$$

holds and  $\Delta_x^L - 2^{-L}(1 - 2^{-L}k - \Delta_x^L)$  at the rest of the domain. The maximum error of the McCormick underestimation is

$$\begin{aligned} & \max_{\substack{\Delta_x^L \in [0, 2^{-L}], \\ k \in \{0, \dots, 2^L - 1\}}} \{x\Delta_x^L - \max\{0, \Delta_x^L - 2^{-L}(1 - x)\}|x = k2^{-L} + \Delta_x^L\} \\ &= \max_{\substack{\Delta_x^L \in [0, 2^{-L}], \\ k \in \{0, \dots, 2^L - 1\}}} \{2^{-L}k\Delta_x^L + (\Delta_x^L)^2 - \max\{0, \Delta_x^L - 2^{-L}(1 - \Delta_x^L - k2^{-L})\}\}. \end{aligned}$$

First, we determine the maximum error on the piece where the McCormick underestimator is the zero function. In the  $(\Delta_x^L, k)$  space the region described by the inequality (24) equals  $\Delta_x^L \leq \frac{2^L - k}{2^L + 4^L}$ . Now suppose we are at some point in this region, then we can increase the error function  $2^{-L}k\Delta_x^L + (\Delta_x^L)^2 - 0$  by increasing either  $k$  or  $\Delta_x^L$ . Consequently, the maximum error is attained if  $\Delta_x^L = \frac{2^L - k}{2^L + 4^L}$ . The error at these points can be purely expressed as a quadratic function in  $k$ :

$$x\Delta_x^L - 0 = (2^{-L}k + \Delta_x^L)\Delta_x^L = \left(2^{-L}k + \frac{2^L - k}{2^L + 4^L}\right) \left(\frac{2^L - k}{2^L + 4^L}\right).$$

It is maximized and symmetric at  $k^* = \frac{1}{2}(2^L - 1) = 2^{L-1} - \frac{1}{2}$ . Since  $k^* \notin \mathbb{N}$  for any  $L \geq 1$ , the maximum error is attained at  $k_1 = 2^{L-1} - 1$  and  $k_2 = 2^{L-1}$ . It has a value of  $2^{-L-2} - 2^{-3L-2}(1 + 2^{-L})^{-2}$ . We can use the same reasoning for the region  $\Delta_x^L \geq \frac{2^L - k}{2^L + 4^L}$  and the increase in the error function by decreasing either  $k$  or  $\Delta_x^L$  and obtaining the same maximum error at the same points. The values  $k_1$  and  $k_2$  correspond to

$$(\Delta_x^L, x) = \left(\frac{1}{2(2^L+1)}, \frac{1}{2} \pm \frac{1}{2(2^L+1)}\right).$$

The maximum overestimation error with the McCormick envelope, where the proof works very similarly, is obtained at  $(\Delta_x^L, x) = (\frac{1}{4}, \frac{1}{4})$  and  $(\Delta_x^L, x) = (\frac{1}{4}, \frac{3}{4})$  with a value of  $2^{-4}$  if  $L = 1$ . However, for  $L \geq 2$  the value is somewhat lower, namely  $2^{-L-2} - 2^{-3L-2}(1 - 2^{-L})^{-2}$  attained at

$$(\Delta_x^L, x) = \left(\frac{1}{2(2^L-1)}, \frac{1}{2} \pm \frac{1}{2(2^L-1)}\right) \text{ if } L \geq 2.$$

The maximum error is therefore set by the underestimation. We summarize these findings in the following proposition.

**Proposition 4** *The maximum error in the univariate NMDT relaxation for  $z = x^2$  with  $x, y \in [0, 1]$  is  $2^{-L-2} - 2^{-3L-2}(1 + 2^{-L})^{-2}$ .*

A summary of the maximum error analysis results can be found in Table 1. It should be noted that for a fixed depth  $L$ , HybS and D-NMDT provide the smallest maximum errors among the considered MIP relaxations in our study.

### 4.2 Average error and minimizing the average error

In this section, we will study the average error of the considered MIP relaxation. In Definition 3 the average error is defined as the volume enclosed by the projected MIP relaxation. We consider it to be an additional measure of the quality of a MIP relaxation besides the maximum error.

For equations of the form  $z = x^2$ , univariate D-NMDT gives piecewise McCormick relaxations. In [5, Proposition5], it is shown that uniform discretization is optimal for

fixed numbers of breakpoints. However, for univariate NMDT the calculation of the volume is much more complicated, so we omit it here.

Next, we compute the average errors of NMDT and D-NMDT for the equation  $z = xy$ . Then we prove that the uniform discretizations, which are used in the definition of NMDT and D-NMDT, are indeed optimal in terms of the 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).

**Proposition 5** *Let  $P_{\text{NMDT}}^{\text{IP}}$  and  $P_{\text{D-NMDT}}^{\text{IP}}$  be the MIP relaxations of NMDT and D-NMDT for  $z = xy$  for some  $L \geq 0$  as defined in (18) and (21), respectively. Their respective average errors are*

$$\mathcal{E}^{\text{avg}}(P_{\text{NMDT}}^{\text{IP}}, \text{gra}_{[0,1]^2}(xy)) = \frac{1}{6}2^{-L-2}$$

and

$$\mathcal{E}^{\text{avg}}(P_{\text{D-NMDT}}^{\text{IP}}, \text{gra}_{[0,1]^2}(xy)) = \frac{1}{6}2^{-2L-2}.$$

**Proof** Note that the discretization in NMDT and D-NMDT yields piecewise McCormick relaxations over a uniformly spaced grid, where each grid piece corresponds to some fixed integer solution  $\beta^x, \beta^y \in \{0, 1\}^L, \Delta_x^L, \Delta_y^L \in [0, 2^{-L}]$ . The volume of the McCormick envelope over a single grid piece is  $\frac{1}{6}w_x^2w_y^2$ , where  $w_x$  is its  $x$ -width and  $w_y$  is its  $y$ -width (see e.g. [15, page 22]). The average error is then the sum over all grid piece volumes. Now, for NMDT we have  $2^L$  grid pieces with  $w_y = 1$  and  $w_x = 2^{-L}$ , yielding a volume per grid piece of  $\frac{1}{6}2^{-2L}$  and thus a total volume of  $\frac{1}{6}2^{-L}$ . Similarly, for D-NMDT we have  $2^{2L}$  grid pieces with  $w_x = w_y = 2^{-L}$ , which yields a volume per grid piece of  $\frac{1}{6}2^{-4L}$  and thus a total volume of  $\frac{1}{6}2^{-2L}$ .  $\square$

When applied to  $\text{gra}_{[0,1]^2}(xy)$ , NMDT and D-NMDT are both *piecewise McCormick relaxations*, defined as

$$\bigcup_{i \in \llbracket n \rrbracket, j \in \llbracket m \rrbracket} \mathcal{M}([x_{i-1}, x_i], [y_{j-1}, y_j]),$$

where we use the notation  $\mathcal{M}([x_{i-1}, x_i], [y_{j-1}, y_j])$  to mean the McCormick envelope  $\mathcal{M}(x, y)$  with  $x \in [x_{i-1}, x_i]$  and  $y \in [y_{j-1}, y_j]$ , for  $0 = x_0 < x_1 < \dots < x_n = 1$  and  $0 = y_0 < y_1 < \dots < y_m = 1$ .

We now prove that a uniform placement of breakpoints minimizes the average error in a piecewise McCormick relaxation. For  $n = 2^L$  and  $m = 1$ , this yields precisely the NMDT relaxation of depth  $L$ , and if  $n = m = 2^L$ , then this yields precisely the D-NMDT relaxation of depth  $L$ . Hence, they are optimal discretizations. The average error in NMDT is  $\frac{1}{6n} = \frac{1}{6}2^{-L}$ , and  $\frac{1}{6n^2} = \frac{1}{6}2^{-2L}$  in D-NMDT. This follows from the proof below.

**Theorem 1** *Let  $0 = x_0 < x_1 < \dots < x_n = 1$  and  $0 = y_0 < y_1 < \dots < y_m = 1$  be sets of breakpoints. Then a uniform spacing of these breakpoints minimizes the average error over all piecewise McCormick relaxations of  $\text{gra}_{[0,1]^2}(xy)$ .*

**Proof** Let  $w_{x_i} := x_i - x_{i-1}$  and  $w_{y_j} := y_j - y_{j-1}$  with  $i \in \llbracket n \rrbracket$  and  $j \in \llbracket m \rrbracket$  be the widths of the grid pieces  $[x_{i-1}, x_i] \times [y_{j-1}, y_j]$ . The volume of the McCormick envelope  $\mathcal{M}([x_{i-1}, x_i], [y_{j-1}, y_j])$  over a single grid piece is  $\frac{1}{6} w_{x_i}^2 w_{y_j}^2$ , see [15, page 22]. Therefore, the problem of minimizing the average error of a piecewise McCormick relaxation can be formulated as

$$\begin{aligned} & \frac{1}{6} \min \sum_{i=1}^n \sum_{j=1}^m w_{x_i}^2 w_{y_j}^2 \\ & \text{s.t. } \sum_{i=1}^n w_{x_i} = 1 \\ & \quad \sum_{j=1}^m w_{y_j} = 1 \\ & \quad w_{x_i} \geq 0 \quad i = 1, \dots, n \\ & \quad w_{y_j} \geq 0 \quad j = 1, \dots, m. \end{aligned} \tag{25}$$

The objective function in (25) sums the average errors over the single grid pieces while the constraints ensure that all single grid widths sum up to 1 and are greater than or equal to 0. Rewriting it to

$$\begin{aligned} & \frac{1}{6} \min (\sum_{i=1}^n w_{x_i}^2) \cdot (\sum_{j=1}^m w_{y_j}^2) \\ & \text{s.t. } \sum_{i=1}^n w_{x_i} = 1 \\ & \quad \sum_{j=1}^m w_{y_j} = 1 \\ & \quad w_{x_i} \geq 0 \quad i = 1, \dots, n \\ & \quad w_{y_j} \geq 0 \quad j = 1, \dots, m. \end{aligned} \tag{26}$$

lets (26) decompose into the two independent convex subproblems

$$\begin{aligned} & \frac{1}{6} \min \sum_{i=1}^n w_{x_i}^2 \\ & \text{s.t. } \sum_{i=1}^n w_{x_i} = 1 \\ & \quad w_{x_i} \geq 0 \quad i = 1, \dots, n, \end{aligned} \tag{27}$$

$$\begin{aligned} & \frac{1}{6} \min \sum_{j=1}^m w_{y_j}^2 \\ & \text{s.t. } \sum_{j=1}^m w_{y_j} = 1 \\ & \quad w_{y_j} \geq 0 \quad j = 1, \dots, m. \end{aligned} \tag{28}$$

Applying the KKT conditions to (27) and (28), which are sufficient for global optimality here, directly shows that a uniform placement of the breakpoints with  $w_{x_i} = \frac{1}{n}$  and  $w_{y_j} = \frac{1}{m}$  is optimal for (25). The total average error is then  $\frac{1}{6nm}$ .  $\square$

**Corollary 1** Let  $0 = x_0 < x_1 < \dots < x_n = 1$  and  $0 = y_0 < y_1 < \dots < y_m = 1$  be sets of breakpoints with  $n = 2^L$  and  $P_L^{\text{IP}}$  a depth- $L$  NMDT MIP relaxation of  $\text{gra}_{[0,1]^2}(xy)$  from (18). Then  $P_L^{\text{IP}}$  is an optimal piecewise McCormick relaxation with an average error of  $\mathcal{E}^{\text{avg}}(P_L^{\text{IP}}, \text{gra}_{[0,1]^2}(xy)) = \frac{1}{6} 2^{-2L}$ .

**Corollary 2** Let  $0 = x_0 < x_1 < \dots < x_n = 1$  and  $0 = y_0 < y_1 < \dots < y_m = 1$  be sets of breakpoints with  $n = m = 2^L$  and  $P_L^{\text{IP}}$  a depth- $L$  D-NMDT MIP relaxation of  $\text{gra}_{[0,1]^2}(xy)$  from (21). Then  $P_L^{\text{IP}}$  is an optimal piecewise McCormick relaxation with an average error of  $\mathcal{E}^{\text{avg}}(P_L^{\text{IP}}, \text{gra}_{[0,1]^2}(xy)) = \frac{1}{6} 2^{-2L}$ .

We summarize the key results of Sect. 4.2 in the remark below and in Table 1.

**Remark 3** (Tightness of MIP Relaxations) For an equation  $z = x^2$  and a fixed depth  $L$ , the tightened sawtooth relaxation [3, Definition 7], and the separable formulations from Part I that employ it, have the smallest volume in the projected MIP relaxation among all studied formulations: they are equivalent in upper bound, with a tightened lower bound, compared to univariate NMDT and D-NMDT. For  $z = xy$ , D-NMDT is the tightest formulation, as it yields the convex hull of  $\text{gra}_D(xy)$  on each grid piece  $D = [k^x 2^{-L}, (k^x + 1)2^{-L}] \times [k^y 2^{-L}, (k^y + 1)2^{-L}]$ ,  $k^x, k^y \in \llbracket 0, 2^L - 1 \rrbracket$ . Combining these facts, T-D-NMDT is the tightest relaxation presented for the full MIQCQP.  $\diamond$

### 4.3 Formulation strength: sharpness and LP relaxations

In the previous section, we discussed maximum error and average errors incurred from using certain discretizations. We will now consider the strength of the resulting MIP relaxations by analyzing their LP relaxation, i.e. we will check for sharpness. Sharpness means that the projected LP relaxation equals the convex hull of the MIP relaxation.

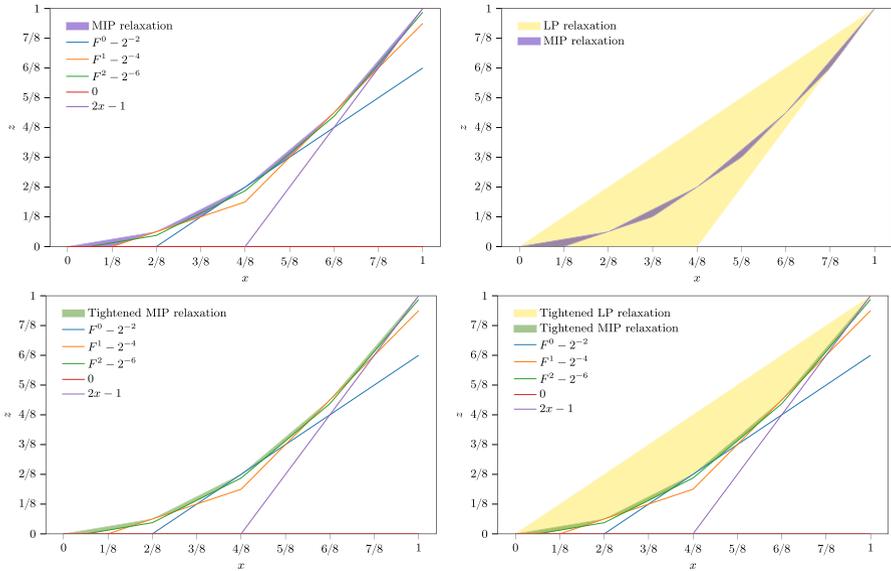
We start with the core formulations from Sect. 2.2. It is well known that the McCormick relaxation yields the convex hull of the feasible set of  $z = xy$  over box domains  $D = [\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}]$ . Therefore, it is obviously sharp. The volume is  $1/6(\bar{x} - \underline{x})(\bar{y} - \underline{y})$ . In [3] it is further shown that the sawtooth epigraph relaxation is also sharp. Since the epigraph of  $f$  is an unbounded set, we do not discuss volume here.

Next, we look at the formulations from Sect. 3. The LP relaxations of NMDT and D-NMDT for  $z = xy$  yield the McCormick envelope over  $D$ , and thus they are sharp. The LP relaxation volumes of NMDT and D-NMDT for  $z = xy$  is thus  $1/6(\bar{x} - \underline{x})(\bar{y} - \underline{y})$  and independent of the choice of  $L$ . In Sect. 3 we proved that univariate NMDT as well as univariate D-NMDT are not sharp by giving points that are feasible for the LP relaxation but are not in the convex hull of the MIP relaxations. Finally, we consider the two tightened formulations univariate T-NMDT and univariate T-D-NMDT for  $z = x^2$ . We show that both formulations are sharp for any  $L_1$  with  $L_1 \geq L$ . A graphical illustration of how tightening leads to sharp MIP formulations in the univariate cases can be seen in Fig. 2 for D-NMDT and Fig. 3 for NMDT. We begin with a lemma about the structure of (non-tightened) univariate D-NMDT MIP relaxations.

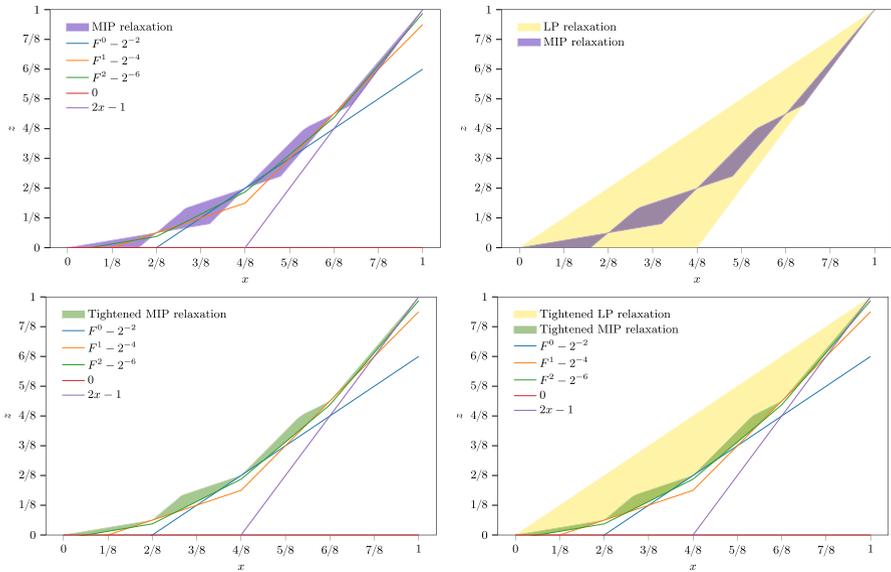
**Lemma 1** *Let  $P_L^{\text{IP}}$  be the univariate D-NMDT MIP relaxation with depth  $L \geq 1$  for  $z = x^2$  as defined in (22). Then the projection of  $P_L^{\text{IP}}$  in the  $(x, z)$ -space gives a series of McCormick envelopes, i.e.*

$$\text{proj}_{(x,z)}(P_L^{\text{IP}}) = \bigcup_{i \in \llbracket 0, 2^L - 1 \rrbracket} \mathcal{M}_i \tag{29}$$

where  $\mathcal{M}_i$  is the McCormick relaxation  $\mathcal{M}(x, x)$  with  $x \in [i2^{-L}, (i + 1)2^{-L}]$  and  $i = 0, \dots, 2^L - 1$ .



**Fig. 2** Feasible set of the univariate MIP relaxation D-NMDT and its LP relaxation with  $L = 2$ . In three of the plots, we display the lower bounds obtained from tightening to show how this affects the MIP relaxation



**Fig. 3** Feasible set of the univariate MIP relaxation NMDT and its LP relaxation with  $L = 2$ . In three of the plots we display the lower bounds obtained from tightening to show how this affects the MIP relaxation

The proof of Lemma 1 is stated in Appendix “Piecewise McCormick relaxations of univariate DNMDT”. We use Lemma 1 to prove sharpness of the tightened version univariate T-D-NMDT.

**Theorem 2** *The univariate T-D-NMDT MIP relaxation for  $z = x^2$  is sharp.*

**Proof** Let  $P_{L,L_1}^{IP}$  be a univariate T-D-NMDT MIP relaxation with  $L_1 \geq L$  and let  $P_{L,L_1}^{LP}$  be the corresponding LP relaxation. To prove sharpness, we have to show that

$$\text{proj}_{(x,z)}(P_{L,L_1}^{LP}) = \text{conv}(\text{proj}_{(x,z)}(P_{L,L_1}^{IP})).$$

As,  $\text{proj}_{(x,z)}(P_{L,L_1}^{LP}) \supseteq \text{conv}(\text{proj}_{(x,z)}(P_{L,L_1}^{IP}))$  is obvious for any LP relaxation, we have to show

$$\text{proj}_{(x,z)}(P_{L,L_1}^{LP}) \subseteq \text{conv}(\text{proj}_{(x,z)}(P_{L,L_1}^{IP})).$$

To do that we analyze the minimum and maximum values of  $z$  in  $\text{proj}_{(x,z)}(P_{L,L_1}^{LP})$  and  $\text{conv}(\text{proj}_{(x,z)}(P_{L,L_1}^{IP}))$ .

In Lemma 1, we showed that the projected MIP relaxation of univariate D-NMDT is a series of small McCormick relaxations. As the MIP relaxation contains the points  $(0, 0)$  and  $(1, 1)$  it follows that its convex hull contains the line connecting these points and thus

$$\max_{z \in [0,1]} \text{conv}(\text{proj}_{(x,z)}(P_{L,L_1}^{IP} |_{x=\tilde{x}})) = \tilde{x}, \forall \tilde{x} \in [0, 1]$$

holds.

Next, we show that the same inequality bounds the maximum value of  $z$  in the LP relaxation, i.e. we prove that

$$\max_{z \in [0,1]} \text{proj}_{(x,z)}(P_{L,L_1}^{LP} |_{x=\tilde{x}}) \leq \tilde{x}, \forall \tilde{x} \in [0, 1].$$

The McCormick relaxations in (22) give the following inequalities

$$\Delta_z^L \stackrel{(*)}{\leq} 2^{-L} \Delta_x^L, \quad u_j \stackrel{(**)}{\leq} x + \Delta_x^L, \quad \text{and} \quad u_j \stackrel{(***)}{\leq} (1 + 2^{-L})\beta_j \quad \forall j = 1, \dots, 2^L - 1.$$

Thus in the LP relaxation,  $z$  is bounded as follows

$$\begin{aligned} z &= \sum_{j=1}^L 2^{-j} u_j + \Delta_z^L \stackrel{(*),(**)}{\leq} (x + \Delta_x^L) \left( \sum_{j=1}^L 2^{-j} \beta_j \right) + 2^{-L} \Delta_x^L \\ &\leq (x + \Delta_x^L)(1 - 2^{-L}) + 2^{-L} \Delta_x^L = x + \Delta_x^L - 2^{-L} x. \end{aligned}$$

and

$$\begin{aligned} z &= \sum_{j=1}^L 2^{-j} u_j + \Delta_z^L \stackrel{(*),(***)}{\leq} (1 + 2^{-L}) \left( \sum_{j=1}^L 2^{-j} \beta_j \right) + 2^{-L} \Delta_x^L \\ &= (1 + 2^{-L})(x - \Delta_x^L) + 2^{-L} \Delta_x^L = x + 2^{-L} x - \Delta_x^L. \end{aligned}$$

Combining both inequalities we have

$$z \leq \begin{cases} x + (\Delta_x^L - 2^{-L}x), \\ x + (2^{-L}x - \Delta_x^L), \end{cases}$$

which implies  $z \leq x$  and therefore

$$\max_{z \in [0,1]} \text{proj}_{(x,z)}(P_{L,L_1}^{\text{LP}}|_{x=\tilde{x}}) \leq \tilde{x}, \forall \tilde{x} \in [0, 1].$$

It remains to show that

$$\min_{z \in [0,1]} \text{proj}_{(x,z)}(P_{L,L_1}^{\text{LP}}|_{x=\tilde{x}}) \geq \min_{z \in [0,1]} \text{conv}(\text{proj}_{(x,z)}(P_{L,L_1}^{\text{IP}}|_{x=\tilde{x}})), \forall \tilde{x} \in [0, 1]. \tag{30}$$

We start with the linear cuts given by the tightened sawtooth relaxation. The sawtooth relaxation is bounded from below by the recursively defined function  $F^L$ . From [3, Proposition 1] in Part I it follows that for each  $L$  and every breakpoint  $x_i := \frac{i}{2^L}$  with  $i = 1, \dots, 2^L - 1$ , there is a function  $F^j - 2^{-2j-2}$  with  $j < L$  such that  $F^j$  lies tangent to  $x^2$  at  $x_i$ . These cuts are exactly the McCormick underestimators of the MIP relaxation derived in Lemma 1,  $z \geq 2x(i2^{-L}) - (i2^{-L})^2$  at  $x_i := \frac{i}{2^L}$  for  $i = 1, \dots, 2^L - 1$ . As the additional sawtooth cuts for  $L_1 \geq L$  tighten both the MIP and LP relaxations, we have

$$\min_{z \in [0,1]} \text{proj}_{(x,z)}(P_{L,L_1}^{\text{LP}}|_{x=\tilde{x}}) \geq \min_{z \in [0,1]} \text{conv}(\text{proj}_{(x,z)}(P_{L,L_1}^{\text{IP}}|_{x=\tilde{x}})), \forall \tilde{x} \in [2^{-L-1}, 1 - 2^{-L-1}].$$

Finally, consider the boundary regions  $[0, 2^{-L-1}]$  and  $[1 - 2^{-L-1}, 1]$ . We show that both the MIP and LP relaxations yield the same minimum value. As established in Lemma 1, D-NMDT provides a piecewise McCormick relaxation. Consequently, for  $x \in [0, 2^{-L-1}]$ , the minimum value is  $z = 0$ . Meanwhile, for  $x \in [1 - 2^{-L-1}, 1]$ , the minimum value is  $z = 2x - 1$ . We further assert that  $z \geq 0$  and  $z \geq 2x - 1$  are valid cuts for the LP relaxation. Based on (22), we can deduce the following McCormick cuts for the LP relaxation:

$$\begin{aligned} u_j &\stackrel{(*)}{\geq} x + \Delta_x^L - (1 + 2^{-L})(1 - \beta_j), & u_j &\stackrel{(**)}{\geq} 0 \quad \forall j = 1, \dots, L \\ \Delta_z^L &\stackrel{(\dagger)}{\geq} 2 \cdot 2^{-L} \Delta_x^L - 2^{-2L}, & \Delta_z^L &\stackrel{(\ddagger)}{\geq} 0. \end{aligned}$$

From the above, we can estimate  $z = \sum_{j=1}^L 2^{-j} u_j + \Delta_z^{L, (**), (')} \geq 0$  and

$$z = \sum_{j=1}^L 2^{-j} u_j \stackrel{(*)}{\geq} (1 - 2^{-L})(x + \Delta_x^L - (1 + 2^{-L})) + (1 + 2^{-L}) \left( \sum_{j=1}^L 2^{-j} \beta_j \right) + \Delta_z^L$$

$$\stackrel{(\dagger)}{\geq} 2x - 2 \cdot 2^{-L} \Delta_x^L - 1 + 2^{-2L} + 2 \cdot 2^{-L} \Delta_x^L - 2^{-2L} = 2x - 1.$$

This concludes our proof. □

Lastly, we state that the univariate T-NMDT is also sharp.

**Theorem 3** *The univariate T-NMDT MIP relaxation is sharp.*

The proof of Theorem 3 works by showing that the projection of D-NMDT is a subset of the projection of NMDT and is stated in Appendix “Sharpness of univariate NMDT”.

## 5 Computational results

In order to test the MIP relaxations from Sect. 3 with respect to their ability to determine dual bounds, we now perform an indicative computational study. More precisely, we will derive MIP relaxations of non-convex MIQCQP instances. The MIP relaxations are then solved using Gurobi [13] as an MIP solver to determine dual bounds and a callback function that uses the non-linear programming (NLP) solver IPOPT [19] to find a feasible solution for the MIQCQP. The MIP relaxation methods are tested for several discretization depths. To compare the considered methods to state-of-the-art spatial branching based solvers, we also run Gurobi as an MIQCQP solver.

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.

### 5.1 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 [5, 8, 11] and earlier works, 20 AC optimal power flow (ACOPF) instances from the NESTA benchmark set (v0.7.0) (see [9]), previously used in [1], and 20 MIQCQP instances from the QPLIB [12]. In Appendix D you will find links that contain download options and detailed descriptions of the instances. For an overview

**Table 2** In the study, we consider the parameters cuts, depth, and formulation to create MIP relaxations for 60 MIQCQP instances

| Depth   | Formulation | Instances            |
|---|-------------|----------------------|
| $L = 1, 2, 4, 6$  | HybS        | boxQP (20 instances) |
| $L_1 = L$   | NMDT        | ACOPF (20 instances) |
| Tightened:<br>$L = 1, 2, 4, 6$<br>$L_1 = \max\{2, 1.5L\}$ | D-NMDT      | QPLIB (20 instances) |

of the IDs of all instances, see Table 9. The benchmark set is equally divided into 30 sparse and 30 dense instances. We refer to dense instances if either the objective function and/or at least one quadratic function in the constraint set is of the form  $x^T Qx$ , 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. 3 using various approximation depths of  $L \in \{1, 2, 4, 6\}$  and a time limit of 8 h. All MIP relaxations are solved twice. Once in the standard versions from Sect. 3 and once with a tightened underestimator version for univariate quadratic terms where  $L_1 = \max\{2, 1.5L\}$ . Note that the tightened MIP relaxations T-NMDT and T-D-NMDT are equivalent to the non-tightened MIP relaxations NMDT and D-NMDT when applied to bilinear terms of the form  $z = xy$ . However, they differ from them in that all lower bounding McCormick constraints in the univariate quadratic terms of the form  $z = x^2$  are replaced by a tighter sawtooth epigraph relaxation (15) as described in Sects. 3.1 and 3.2. Furthermore, we include HybS, the most performant separable MIP relaxation from Part I, in the study. However, we do not apply tightening to HybS, as it was shown in Part I that this does not result in computational improvements.

In Table 2, one can see an overview of the different parameters in our study. In total, we have 24 parameter configurations for 60 original problems. However, as we do not apply tightening to HybS we end up with 1200 MIP instances. For the comparison with Gurobi as a state-of-the-art MIQCQP solver, we solve an additional 480 MIP instances and 120 MIQCQP instances. These additional MIP instances arise from disabling the cuts in Gurobi for the winner of the NMDT-based methods and HybS. The 120 MIQCQP instances are built by solving all 60 benchmark problems once with cuts enabled and once with cuts disabled.

See Sect. 5.2.2 for more details on the latter.

**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 NLP locally via IPOPT in an attempt to find a feasible solution for the original MIQCQP problem.

## 5.2 Results

In the following, we present the results of our study. In particular, we aim to answer the following questions regarding dual bounds:

- Is our enhanced method D-NMDT computationally superior to its predecessors NMDT?
- Is it beneficial to use tightened versions of the NMDT and D-NMDT, i.e., to choose  $L_1 > L$ ?
- How do the studied methods compare to the state-of-the-art MIQCQP solver Gurobi?

We provide performance profile plots as proposed by Dolan and More [10] to illustrate the results of the computational study regarding the dual bounds, see Fig. 4, 5, 6, 7, 8 and 9. The performance profiles work as follows: Let  $d_{p,s}$  be the best dual bound obtained by MIP relaxation or MIQCQP solver  $s$  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. [17]. The plots are divided into two blocks, one for NMDT-based methods and one for the comparison against HybS and Gurobi as an MIQCQP solver. 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.

Although the main criterion of the study is the dual bound, we also discuss run times. 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  $t_1, \dots, t_n$  with shift  $s$  is defined as  $(\prod_{i=1}^n (t_i + s))^{1/n} - s$ . It has the advantage that it is neither affected by very large outliers (in contrast to the arithmetic mean) nor by very small 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.

Finally, we will highlight some important results regarding primal bounds in the comparison of our methods with Gurobi [13] as an MIQCQP solver.

### 5.2.1 NMDT-based MIP relaxations

We start our analysis of the results by looking at the dual bounds, run times and feasible solution of the NMDT-based MIP relaxations.

**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 set, where  $d_{p,s}$  is the corresponding best dual bound found by

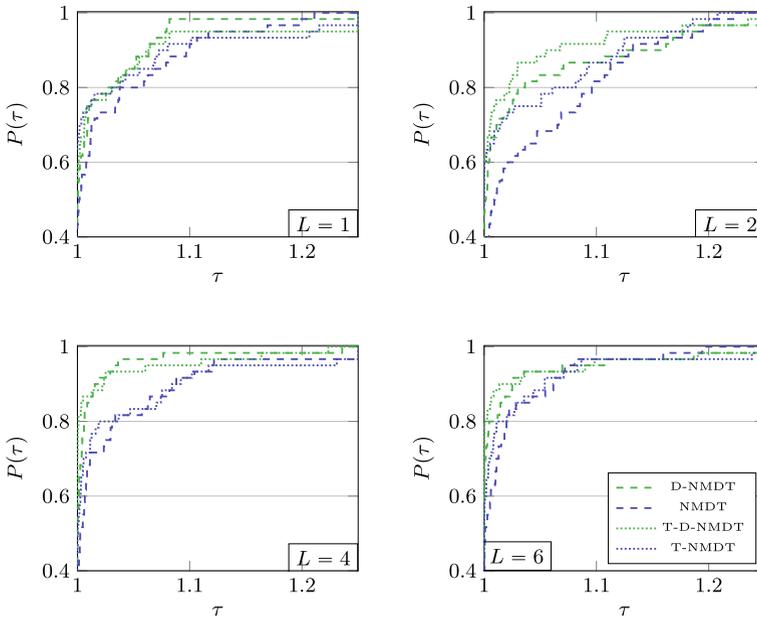
**Table 3** Arithmetic (left) and geometric (right) mean of relative optimality gaps (in %) on all instances for NMDT-based methods

|         | NMDT               | T-NMDT             | DNMDT              | T-DNMDT                     |
|---------|--------------------|--------------------|--------------------|-----------------------------|
| all     |                    |                    |                    |                             |
| $L = 1$ | 31.33/3.60         | <b>29.85</b> /3.37 | 36.41/3.14         | 35.39/ <b>3.00</b>          |
| $L = 2$ | 24.57/3.41         | <b>23.89</b> /3.05 | 25.58/2.30         | 24.72/ <b>2.20</b>          |
| $L = 4$ | 32.76/1.69         | 33.12/1.26         | <b>28.45</b> /1.32 | 30.18/ <b>1.03</b>          |
| $L = 6$ | <b>32.34</b> /1.08 | 37.45/0.90         | 38.55/0.85         | 37.92/ <b>0.70</b>          |
| sparse  |                    |                    |                    |                             |
| $L = 1$ | 19.18/11.48        | 17.91/10.84        | 18.66/10.71        | <b>18.57</b> / <b>10.66</b> |
| $L = 2$ | 15.79/7.16         | 14.02/5.81         | 15.70/5.65         | <b>15.50</b> / <b>5.51</b>  |
| $L = 4$ | 11.47/1.49         | 11.23/0.85         | 11.03/1.21         | <b>11.00</b> / <b>0.78</b>  |
| $L = 6$ | 11.17/0.56         | 10.93/0.40         | 10.72/0.44         | <b>10.56</b> / <b>0.35</b>  |
| dense   |                    |                    |                    |                             |
| $L = 1$ | <b>43.47</b> /1.13 | 41.80/1.05         | 54.16/0.92         | 52.22/ <b>0.85</b>          |
| $L = 2$ | <b>33.36</b> /1.63 | 33.77/1.60         | 35.45/0.94         | 33.94/ <b>0.88</b>          |
| $L = 4$ | 54.06/1.92         | 55.00/1.86         | <b>45.87</b> /1.44 | 49.37/ <b>1.37</b>          |
| $L = 6$ | <b>53.51</b> /2.10 | 63.97/2.04         | 66.38/1.66         | 65.28/ <b>1.42</b>          |

method  $s$  and  $b_p$  is the best known primal bound for instance  $p$ . Looking at Table 3, which displays the arithmetic (left) and geometric (right) means of the relative optimality gaps for all 60 instances of the benchmark set based on NMDT methods, several observations can be made. Across all instances, the gap generally decreases with increasing  $L$  values, although there are exceptions. While the outcomes for the arithmetic mean, in which outliers play a greater role, the outcomes for the geometric mean are clear. Here, T-D-NMDT is the winner exhibiting the best geometric means for all  $L$  values.

Additionally, in Fig. 4 we show performance profiles for the dual bounds that are obtained by the different NMDT-based MIP relaxations. Starting from  $L = 2$ , we can see that both D-NMDT and T-D-NMDT deliver notably tighter bounds within the run time limit of 8 h. The largest difference is at  $L = 4$ , where D-NMDT and T-D-NMDT are able to find dual bounds that are within a factor 1.05 of the overall best bounds for nearly all instances. In contrast, NMDT and T-NMDT require a corresponding factor of more than 1.1. In addition, the tightened versions perform somewhat better than the corresponding counterparts, especially for  $L = 4$ .

To gain a deeper insight into the benefits of D-NMDT and the tightening of NMDT-based relaxations, we divide the benchmark set into sparse and dense instances. For sparse instances, the advantage of the new methods in the performance profiles is rather small; see the performance profiles in Fig. 6. Here, T-D-NMDT provides marginally better bounds than the other methods in case of  $L = 4$  and  $L = 6$ . For  $L = 1$  and  $L = 2$ , however, T-NMDT dominates all other approaches. Moreover, the tightened versions outperform their counterparts for all depths  $L$ . Regarding the arithmetic and geometric means, the gaps consistently decrease with increasing  $L$



**Fig. 4** Performance profiles to dual bounds of NMDT-based methods on all instances

values for all methods, indicating improved performance. T-NMDT and T-D-NMDT are generally competitive, with T-D-NMDT having the best geometric mean in all cases.

For dense instances, D-NMDT and T-D-NMDT are clearly superior to NMDT and T-NMDT; see Fig. 6. Regardless of the relaxation depth, the new methods yield the tightest dual bounds, with T-D-NMDT being superior to D-NMDT only in case of  $L = 2$ , where the tightened version T-D-NMDT is able to find the best dual bound for roughly 10% more instances than D-NMDT. Tightening the NMDT method does not deliver better bounds, in fact, T-NMDT is surpassed by NMDT for  $L = 1$ . Regarding the arithmetic and geometric means, there's no clear trend of improvement with increasing  $L$  values. NMDT overall has the best arithmetic means for  $L = 1$ ,  $L = 2$  and  $L = 4$ , while other methods shine at different  $L$  values. In summary, the performance of the NMDT-based methods varies depending on the  $L$  value and dataset density. For sparse instances, the performance seems to consistently improve with increasing  $L$  values, while for dense instances, no clear trend is discernible. Overall, T-D-NMDT showcases the performance in most scenarios. Tightening leads to an improvement in dual bounds across all instances, but this is mainly due to the sparse instances. In dense instances, we assume that the large number of additional cuts in the tightened variants leads to slower computations and thus worse bounds (Fig. 5).

**Run times** Table 4 showcases the shifted geometric mean for run times. Throughout all instances, T-D-NMDT consistently outperforms the other methods, indicating its efficiency, especially at  $L = 1$  and  $L = 2$ . As  $L$  increases, the run times generally rise

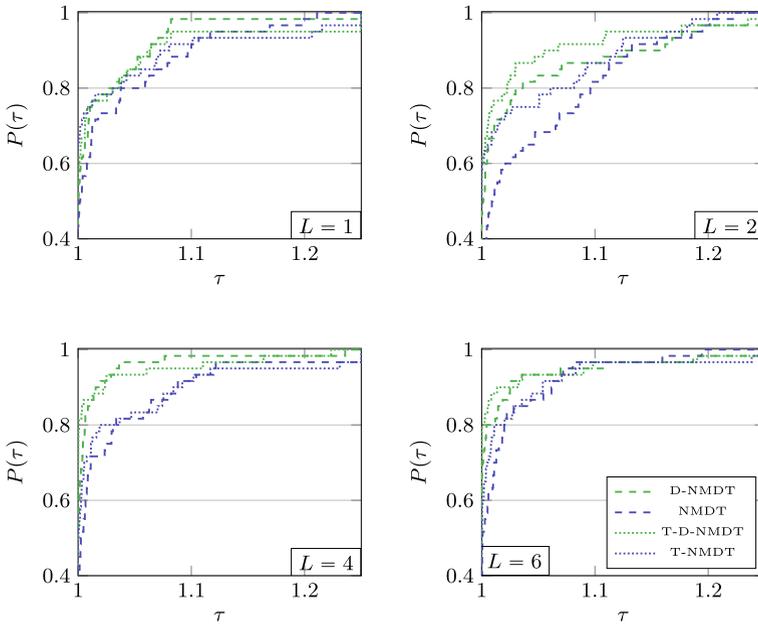


Fig. 5 Performance profiles to dual bounds of NMDT-based methods on sparse instances

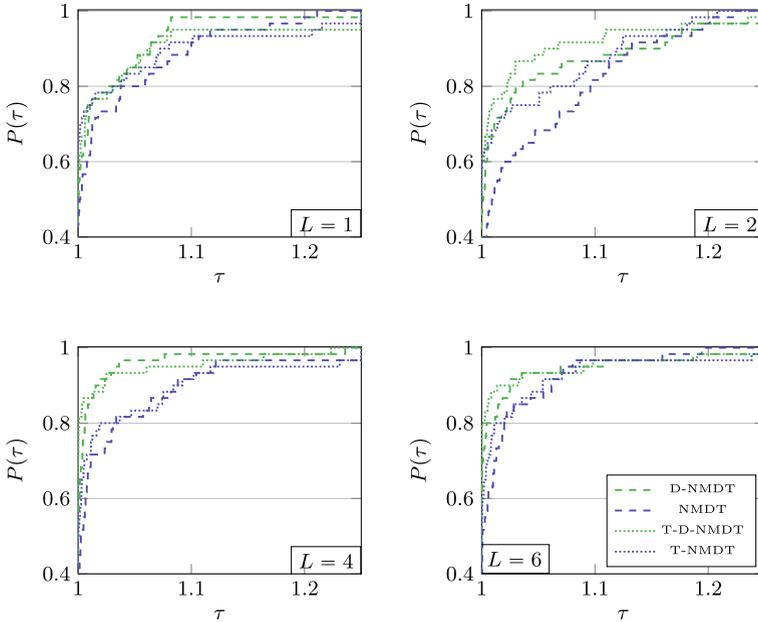


Fig. 6 Performance profiles to dual bounds of NMDT-based methods on dense instances

**Table 4** Shifted geometric mean for run times on all 60 instances in NMDT-based MIP relaxations

|        | NMDT    | T-NMDT | D-NMDT        | T-D-NMDT      |
|--------|---------|--------|---------------|---------------|
| all    |         |        |               |               |
| L=1    | 68.88   | 69.81  | 57.71         | <b>42.13</b>  |
| L=2    | 155.40  | 152.57 | 52.85         | <b>47.18</b>  |
| L=4    | 347.69  | 294.51 | 138.44        | <b>135.16</b> |
| L=6    | 630.35  | 489.71 | 288.87        | <b>272.09</b> |
| sparse |         |        |               |               |
| L=1    | 70.78   | 70.11  | 75.97         | <b>47.28</b>  |
| L=2    | 141.66  | 131.06 | 44.86         | <b>41.27</b>  |
| L=4    | 531.43  | 351.58 | 151.42        | <b>132.42</b> |
| L=6    | 1364.11 | 695.06 | 409.26        | <b>291.78</b> |
| dense  |         |        |               |               |
| L=1    | 65.98   | 69.35  | 36.44         | <b>34.92</b>  |
| L=2    | 178.87  | 191.98 | 67.39         | <b>57.55</b>  |
| L=4    | 187.84  | 228.25 | <b>121.69</b> | 139.17        |
| L=6    | 199.78  | 292.14 | <b>172.23</b> | 245.60        |

**Table 5** Number of instances with feasible solutions found with different relative optimality gaps.

|     | NMDT             | T-NMDT            | D-NMDT                    | T-D-NMDT                          |
|-----|------------------|-------------------|---------------------------|-----------------------------------|
| L=1 | <b>32</b> /34/40 | 31/ <b>35</b> /41 | 29/33/ <b>42</b>          | 29/33/40                          |
| L=2 | 32/37/45         | <b>34</b> /37/43  | <b>34</b> /38/42          | <b>34</b> /37/42                  |
| L=4 | 42/44/48         | 39/44/48          | 37/42/49                  | <b>45</b> / <b>47</b> / <b>51</b> |
| L=6 | 43/45/48         | 42/43/47          | 44/ <b>47</b> / <b>50</b> | <b>46</b> / <b>47</b> / <b>50</b> |

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 finding a feasible solution

for all methods, but T-D-NMDT remains the fastest. In sparse instances, T-D-NMDT retains its edge in efficiency, especially evident at  $L = 6$ . However, the run times of other methods, particularly NMDT, escalate significantly. For dense instances, T-D-NMDT is the fastest for small values while D-NMDT takes the lead at  $L = 4$  and  $L = 6$ . This observation is in line with increasing gaps for tightened variants. Despite these fluctuations, T-D-NMDT remains the most efficient methods across most scenarios.

**Feasible solutions** In Table 5, we can see that the QP heuristic (IPOPT) we mentioned at the beginning of this section delivers high-quality feasible solutions for the original (MIQC-)QP instances. With increasing  $L$  values, IPOPT is able to find more feasible solutions with all NMDT-based methods quite similarly. For  $L = 6$ , T-D-NMDT combined with IPOPT yields feasible solutions for 50 out of 60 benchmark instances, 47 of which have a relative optimality gap below 1% and 46 of which are even globally optimal, i.e., which have a gap below 0.01%.

In summary, both T-D-NMDT and D-NMDT are clearly superior to the previously known NMDT approach. The double discretization and the associated reduction in the

number of binary variables while maintaining the same relaxation error are most likely the reason for this. Surprisingly, the tightening of the lower bounds in the univariate quadratic terms and the resulting introduction of new constraints does not lead to higher run times. Thus, the latter is recommended. Moreover, T-D-NMDT is slightly ahead of the other methods in computing good solutions for the MIP relaxations that are used by the NLP solver IPOPT to find feasible solutions for the original MIQCQP instances. Altogether, we consider T-D-NMDT to be the winner among the NMDT-based methods.

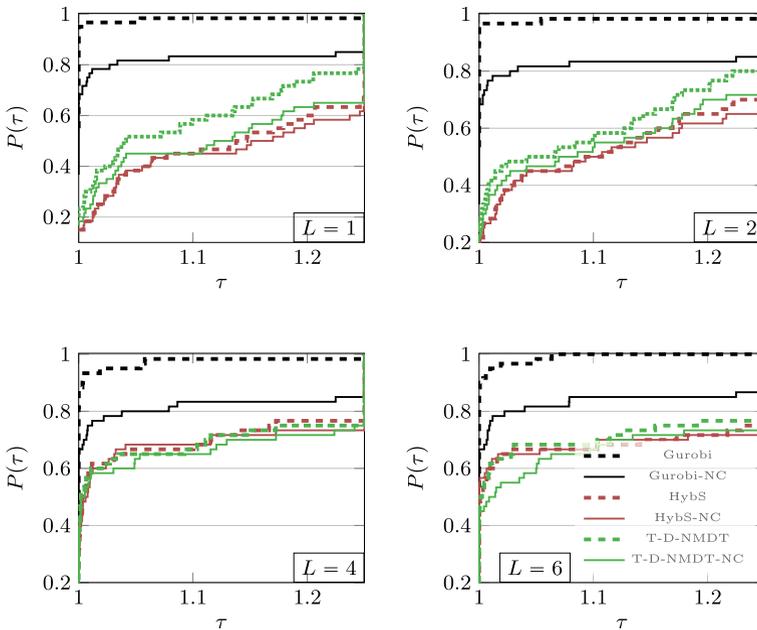
## 5.2.2 Comparison with state-of-the-art MIQCQP Solver Gurobi

Finally, we compare the two winners T-D-NMDT and HybS of the NMDT-based and separable Methods (Part I) with the state-of-the-art MIQCQP solver Gurobi 9.5.1. We perform the comparison in two ways. Firstly, with Gurobi's default settings, and secondly, with cuts disabled, i.e., we set the parameter "*Cuts = 0*". The reason for running Gurobi again with cuts turned off is that cuts are one of the most important components of MIQCQP/MIP solvers that rely on the structure of the problem. While constructing the MIP relaxations with T-D-NMDT and HybS, the original problem is transformed in such a way that Gurobi can no longer recognize the original quadratic structure of the problem. However, many cuts would still be valid and applicable in the MIP relaxations, for instance, RLT and PSD cuts.

**Dual bounds** We start our comparison with showing performance profiles for Gurobi, T-D-NMDT, HybS, and their variants without cuts ("-NC") on all instances in Fig. 7. As expected, Gurobi performs best for all  $L$  values, followed by its variant without cuts in second place. However, as the depth  $L$  increases, the MIP relaxations provide gradually tighter dual bounds. For  $L = 6$ , T-D-NMDT and HybS are able to find the best dual bounds for more than 50% of the cases, while Gurobi delivers the best bounds for roughly 90% and its variant without cuts for about 70% of the cases. Surprisingly, in contrast to T-D-NMDT, disabling cuts in case of HybS has little effect on the quality of the dual bounds. In Table 6 we display the arithmetic and geometric means of the relative optimality gaps for various methods and their "no cuts" (NC) versions. Here, the findings are in line with those from the performance plots.

As before, we divide the benchmark set into sparse and dense instances. For sparse instances, the dual bounds computed by T-D-NMDT and HybS become progressively tighter with increasing  $L$ ; see Fig. 8. For  $L = 4$  and  $L = 6$ , T-D-NMDT and HybS are able to find the best dual bounds in about 60% of the instances, while Gurobi delivers the best bounds for roughly 80%. Compared to Gurobi-NC, our new methods T-D-NMDT, HybS, and most notably HybS-NC perform almost equally well. Regarding the means in Table 6, the cuts contribute minimally to improved gaps. The differences between the versions with and without cuts are marginal. For HybS, the results even indicate that the cuts can be counterproductive.

In the case of dense instances, a different picture emerges, see Fig. 9. Again, Gurobi and also Gurobi-NC are dominant for all approximation depths and thanks to the cuts, Gurobi can solve all instances to a gap of 0%, see Table 6. However, for  $L = 1$ , T-D-NMDT delivers dual bounds that are within a factor 1.1 of the dual bounds provided by the variant of Gurobi without cuts. With higher  $L$  values, T-D-NMDT, HybS, and



**Fig. 7** Performance profiles on dual bounds of best MIP relaxation compared to Gurobi as MIQCQP solver, with and without cuts, on all 60 instances

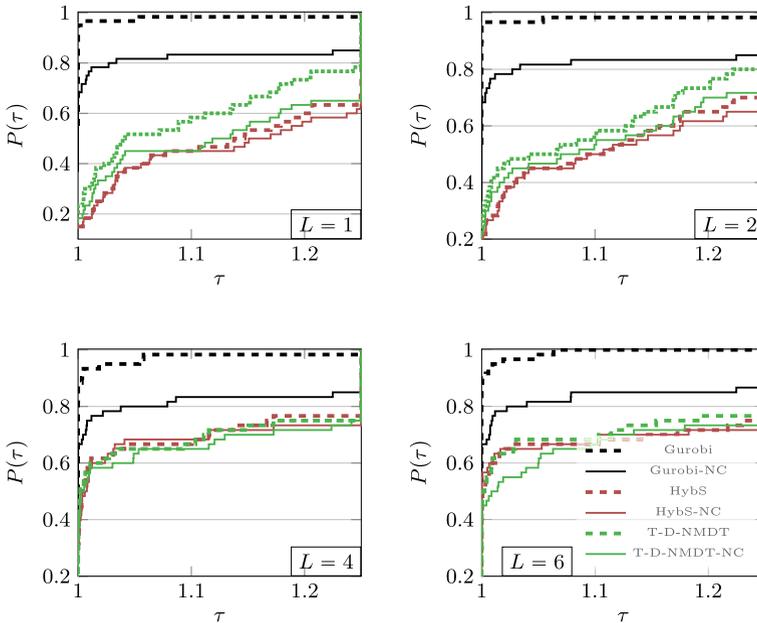
**Table 6** Arithmetic (left) and geometric (right) mean of relative optimality gaps (in %) on all instances

|        | HybS        | HybS-NC     | T-D-NMDT    | T-D-NMDT-NC | Gurobi           | Gurobi-NC         |
|--------|-------------|-------------|-------------|-------------|------------------|-------------------|
| all    |             |             |             |             |                  |                   |
| L1     | 46.13/7.94  | 51.36/8.78  | 29.85/3.37  | 44.64/5.21  | <b>4.79/0.04</b> | 21.11/0.17        |
| L2     | 33.07/4.96  | 38.76/5.49  | 23.89/3.05  | 33.32/3.36  | <b>4.79/0.04</b> | 21.11/0.17        |
| L4     | 24.84/1.81  | 30.29/1.98  | 33.12/1.26  | 30.59/1.20  | <b>4.79/0.04</b> | 21.11/0.17        |
| L6     | 32.97/1.05  | 35.93/0.99  | 37.45/0.90  | 35.95/0.91  | <b>4.79/0.04</b> | 21.11/0.17        |
| sparse |             |             |             |             |                  |                   |
| L1     | 23.85/14.01 | 23.85/14.01 | 17.91/10.84 | 18.61/10.69 | <b>9.58/0.20</b> | 9.67/ <b>0.16</b> |
| L2     | 21.21/11.52 | 21.24/11.54 | 14.02/5.81  | 15.58/5.54  | <b>9.58/0.20</b> | 9.67/ <b>0.16</b> |
| L4     | 14.93/2.19  | 14.87/2.17  | 11.23/0.85  | 11.67/0.90  | <b>9.58/0.20</b> | 9.67/ <b>0.16</b> |
| L6     | 10.91/0.72  | 11.73/0.58  | 10.93/0.40  | 11.52/0.38  | <b>9.58/0.20</b> | 9.67/ <b>0.16</b> |
| dense  |             |             |             |             |                  |                   |
| L1     | 68.40/4.50  | 78.87/5.50  | 52.22/0.85  | 70.67/2.54  | <b>0.00/0.01</b> | 32.56/0.17        |
| L2     | 44.94/2.14  | 56.27/2.62  | 33.94/0.88  | 51.06/2.04  | <b>0.00/0.01</b> | 32.56/0.17        |
| L4     | 34.76/1.49  | 45.72/1.80  | 49.37/1.37  | 49.51/1.60  | <b>0.00/0.01</b> | 32.56/0.17        |
| L6     | 55.04/1.52  | 60.13/1.71  | 65.28/1.42  | 60.39/2.18  | <b>0.00/0.01</b> | 32.56/0.17        |

HybS-NC compute in about 40% of the cases the best bounds, while Gurobi yields the best bounds in all cases and Gurobi-NC for roughly 70% of the instances. When

**Table 7** Shifted geometric mean for run times on all instances for best MIP relaxation compared to Gurobi as MIQCQP solver with cuts and without cuts (-NC)

|    | HybS    | HybS-NC | T-D-NMDT | T-D-NMDT-NC | Gurobi       | Gurobi-NC |
|----|---------|---------|----------|-------------|--------------|-----------|
| L1 | 188.23  | 225.0   | 163.52   | 244.73      | <b>77.32</b> | 388.74    |
| L2 | 342.32  | 279.0   | 266.37   | 340.8       | <b>54.0</b>  | 307.03    |
| L4 | 1008.09 | 964.63  | 950.73   | 1012.11     | <b>25.16</b> | 193.75    |
| L6 | 2548.31 | 2315.29 | 1665.28  | 1618.91     | <b>20.75</b> | 174.54    |

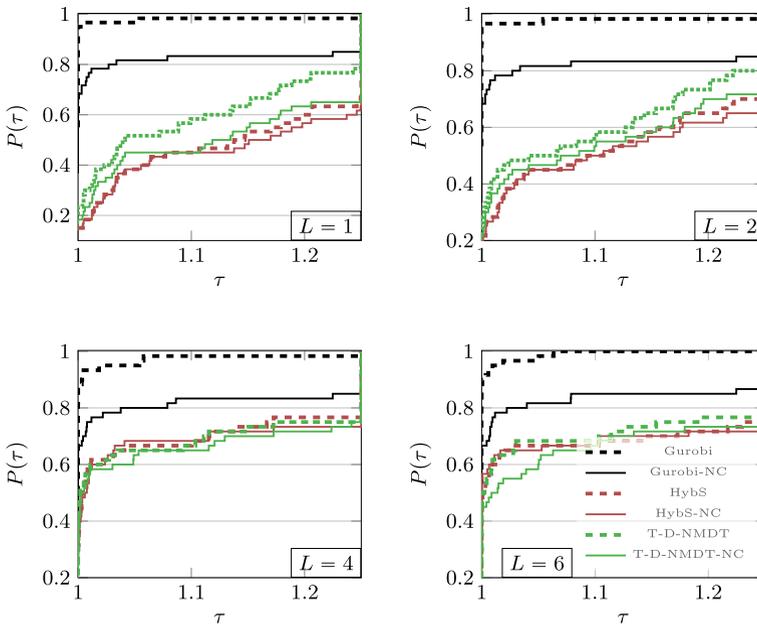


**Fig. 8** Performance profiles on dual bounds of best MIP relaxation compared to Gurobi as MIQCQP solver, with and without cuts, on sparse instances

comparing D-NMDT and HybS, D-NMDT demonstrates advantages for smaller  $L$  values. In contrast, HybS delivers better gaps for larger  $L$  values.

**Run times** In Table 7 we show the shifted geometric mean values of the run times for solving all instances with Gurobi and the corresponding MIP relaxations constructed with T-D-NMDT and HybS. The variants of Gurobi, T-D-NMDT, and HybS without cuts are also contained. Gurobi has significantly shorter run times than all other approaches. However, with  $L = 1$  and  $L = 2$ , T-D-NMDT, HybS, T-D-NMDT-NC and HybS-NC are somewhat faster than Gurobi-NC.

**Remark 4** Note, that for calculating the shifted geometric mean only those instances are used for which at least one method computed the optimal solution within the run time limit of 8 h. Since with higher  $L$  values the complexity of the MIP relaxations increases, fewer instances are solved to optimality by T-D-NMDT and HybS. Therefore, the shifted geometric mean decreases for Gurobi and Gurobi-NC with higher  $L$  values.



**Fig. 9** Performance profiles on dual bounds of best MIP relaxation compared to Gurobi as MIQCQP solver, with and without cuts, on sparse instances. See Fig. 8 for the legend

This inherent nature of the shifted geometric mean is also the reason why we see different values in Tables 7 and 4 for the same methods.  $\diamond$

**Feasible Solutions.** In combination with IPOPT as a QP heuristic, T-D-NMDT, HybS, and their variants without cuts are competitive with Gurobi for high  $L$  values when it comes to finding feasible solutions, as Table 8 shows. HybS-NC with IPOPT is able to find feasible with a relative optimality gap below 1% for 48 out of 60 benchmark instances, while Gurobi finds 50 feasible solutions with a gap below 1%. T-D-NMDT computes 46 solutions that are globally optimal, whereas Gurobi achieves this for 50 instances. Surprisingly, the variant without cuts of HybS delivers more feasible solutions than its variant with cuts enabled. Finally, we note that some MIQCQP instances have been solved to global optimality by the MIP relaxation methods, while Gurobi reached the run time limit of 8 h. For instance, T-D-NMDT with IPOPT is able to solve the QPLIB instance “QPLIB\_0698” to global optimality for  $L \in \{2, 4, 6\}$  with a run time below 5 min, while Gurobi has a relative optimality gap of more than 5% after a run time of 8 h.

Overall, the comparison with Gurobi as a state-of-the-art MIQCQP solver has shown that the new methods T-D-NMDT and HybS can be relevant for practical applications. For sparse instances, the dual bounds provided by T-D-NMDT and HybS are of similar quality to those provided by Gurobi. In terms of MIQCQP-feasible solutions, for most instances the two methods are able to find very high quality solutions in combination with IPOPT as NLP solver.

**Table 8** Number of feasible solutions found with different relative optimality gaps

|    | HybS     | HybS-NC  | T-D-NMDT | T-D-NMDT-NC | Gurobi          | Gurobi-NC |
|----|----------|----------|----------|-------------|-----------------|-----------|
| L1 | 31/33/40 | 28/31/40 | 29/33/40 | 31/34/42    | <b>50/50/57</b> | 46/49/56  |
| L2 | 32/37/44 | 31/36/41 | 34/37/42 | 35/41/44    | <b>50/50/57</b> | 46/49/56  |
| L4 | 41/44/50 | 40/45/53 | 45/47/51 | 40/45/50    | <b>50/50/57</b> | 46/49/56  |
| L6 | 40/43/51 | 43/48/50 | 46/47/50 | 40/46/49    | <b>50/50/57</b> | 46/49/56  |

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

Moreover, there is still plenty of room for improvement. First, numerical studies have shown before that an adaptive refinement of nonlinearities drastically decreases run times for solving MINLPs by piecewise linear MIP relaxations; see [6] for example. Hence, an approach with an adaptive refinement of the approximation depth  $L$  is even more promising. Second, HybS and its variant without cuts HybS-NC have performed very similarly in our computational study. In addition, HybS-NC was relatively close to Gurobi-NC in both solution quality and dual bounds for the MIQCQPs. Since most MIQCQP-specific cuts can still be integrated into the HybS approach, we believe that HybS can be further improved by embedding it in a branch-and-cut solution framework that is able to add MIQCQP-specific cuts, such as BQP and PSD cuts, to the MIP relaxations. In this way, we obtain both tighter dual bounds and MIP relaxation solutions that are more likely to yield feasible solutions for the MIQCQP in combination with IPOPT.

## 6 Conclusion

We introduced an enhanced *mixed-integer programming* (MIP) relaxation technique for non-convex *mixed-integer quadratically constrained quadratic programs* (MIQCQP), called *doubly discretized normalized multiparametric disaggregation technique* (D-NMDT). We showed that it has clear theoretical advantages over its predecessor NMDT, i.e. it requires a significantly lower number of binary variables to achieve the same accuracy. In addition, we combined both, D-NMDT and NMDT, with the *sawtooth epigraph relaxation* from Part I [3] to further strengthen the relaxations for univariate quadratic terms.

In a two-part computational study, we first compared D-NMDT to NMDT. We showed that D-NMDT determines far better dual bounds than NMDT and also has shorter run times. Furthermore, we were able to show that our tightening in both methods led to better dual bounds while simultaneously shortening the computation time. In the second part of the computational study, we compared the *tightened D-NMDT* (T-D-NMDT) against *Hybrid Separable* (HybS), the best-performing MIP relaxation from Part I. We showed that HybS does perform slightly better in terms of dual bounds. However, both new methods were 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. Furthermore, we showed that they both method can partially compete with the state-of-the-art MIQCQP solver Gurobi.

Finally, we gave some indications on how to further improve the new approaches. Two of the most promising directions in this context are employing adaptivity and adding MIQCQP-specific cuts that are valid but not recognized by the MIP solvers. This is the subject of future work.

**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.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## Appendix A: Detailed derivation of the MIP relaxation D-NMDT

For the derivation of the MIP relaxation D-NMDT for  $\text{gra}_{[0,1]^2}(xy)$ , we first define

$$\begin{aligned}
 x &= \sum_{j=1}^L 2^{-j} \beta_j^x + \Delta_x^L, & y &= \sum_{j=1}^L 2^{-j} \beta_j^y + \Delta_y^L, \\
 \Delta_x^L &\in [0, 2^{-L}], \Delta_y^L \in [0, 2^{-L}], & \beta^x &\in \{0, 1\}^L, \beta^y \in \{0, 1\}^L.
 \end{aligned}
 \tag{31}$$

Then we use the NMDT representation (17), expand the  $\Delta_x^L y$ -term and obtain

$$\begin{aligned}
 z = xy &= y \left( \sum_{j=1}^L 2^{-j} \beta_j^x + \Delta_x^L \right) \\
 &= \sum_{j=1}^L 2^{-j} \beta_j^x y + y \Delta_x^L \\
 &= \sum_{j=1}^L 2^{-j} \beta_j^x y + \Delta_x^L \left( \sum_{j=1}^L 2^{-j} \beta_j^y + \Delta_y^L \right) \\
 &= \sum_{j=1}^L 2^{-j} (\beta_j^x y + \beta_j^y \Delta_x^L) + \Delta_x^L \Delta_y^L.
 \end{aligned}$$

Alternatively, if we discretize  $y$  first, then expand the term  $\Delta_y^L x$ , we obtain

$$z = \sum_{j=1}^L 2^{-j} (\beta_j^y x + \beta_j^x \Delta_y^L) + \Delta_x^L \Delta_y^L.$$

Finally, to balance between the two formulations, we observe for any  $\lambda \in [0, 1]$  that

$$\begin{aligned} z &= xy = \lambda xy + (1 - \lambda)xy \\ &= \lambda \left( \sum_{j=1}^L 2^{-j} (\beta_j^y x + \beta_j^x \Delta_y^L) + \Delta_x^L \Delta_y^L \right) \\ &\quad + (1 - \lambda) \left( \sum_{j=1}^L 2^{-j} (\beta_j^x y + \beta_j^y \Delta_x^L) + \Delta_x^L \Delta_y^L \right) \\ &= \sum_{j=1}^L 2^{-j} [\beta_j^y ((1 - \lambda) \Delta_x^L + \lambda x) + \beta_j^x (\lambda \Delta_y^L + (1 - \lambda)y)] + \Delta_x^L \Delta_y^L \end{aligned}$$

holds. This yields

$$\begin{aligned} x &= \sum_{j=1}^L 2^{-j} \beta_j^x + \Delta_x^L, \quad y = \sum_{j=1}^L 2^{-j} \beta_j^y + \Delta_y^L \\ z &= \sum_{j=1}^L 2^{-j} [\beta_j^y ((1 - \lambda) \Delta_x^L + \lambda x) + \beta_j^x (\lambda \Delta_y^L + (1 - \lambda)y)] + \Delta_x^L \Delta_y^L \tag{32} \\ \Delta_x^L, \Delta_y^L &\in [0, 2^{-L}], \quad x, y \in [0, 1], \quad \beta^x, \beta^y \in \{0, 1\}^L. \end{aligned}$$

Finally, we obtain the complete MIP relaxation D-NMDT stated in (21) by applying McCormick envelopes to the product terms  $\beta_j^y ((1 - \lambda) \Delta_x^L + \lambda x)$ ,  $\beta_j^x (\lambda \Delta_y^L + (1 - \lambda)y)$  and  $\Delta_x^L \Delta_y^L$ . For bounds on the terms  $((1 - \lambda) \Delta_x^L + \lambda x)$  and  $(\lambda \Delta_y^L + (1 - \lambda)y)$ , see Appendix 1.

### Appendix B: MIP relaxations on general intervals

In this section, we generalize the MIP relaxations for  $\text{gra}_{[0,1]^2}(xy)$  and  $\text{gra}_{[0,1]}^2(x^2)$  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  $y$ .

#### B.1 MIP relaxations for bivariate quadratic equations

First, we consider MIP relaxations for  $z = xy$  and give explicit models of NMDT and D-NMDT for general box domains.

Next, we consider the MIP relaxation NMDT. To derive the general formulation, we first introduce  $\hat{x} \in [0, 1]$  and define  $\hat{z} = \hat{x}y$ , then use the definitions  $x = l_x\hat{x} + \underline{x}$  and

$$z = xy = (l_x\hat{x} + \underline{x})y = l_x\hat{z} + \underline{x} \cdot y$$

to obtain

$$\begin{aligned} x &= l_x \sum_{j=1}^L 2^{-j} \beta_j + \Delta_x^L + \underline{x} \\ z &= l_x \sum_{j=1}^L 2^{-j} \beta_j y + \Delta_x^L \cdot y + \underline{x} \cdot y \\ \Delta_x^L &\in [0, 2^{-L}(\bar{x} - \underline{x})], \quad y \in [\underline{y}, \bar{y}], \quad \beta \in \{0, 1\}^L. \end{aligned} \tag{33}$$

In this way, we are able to formulate the MIP relaxation NMDT on a general box domain as follows:

$$\begin{aligned} x &= l_x \sum_{j=1}^L 2^{-i} \beta_j + \Delta_x^L + \underline{x} \\ z &= l_x \sum_{j=1}^L 2^{-j} u_j + \Delta_z^L + \underline{x} \cdot y \end{aligned} \tag{34}$$

$$\begin{aligned} (x, \alpha_j, u_j) &\in \mathcal{M}(x, \beta_j) && j \in 1, \dots, L \\ (\Delta_x^L, y, \Delta_z^L) &\in \mathcal{M}(\Delta_x^L, y) \\ \Delta_x^L &\in [0, 2^{-L}l_x], \quad y \in [\underline{y}, \bar{y}], \quad \beta \in \{0, 1\}^L \end{aligned}$$

Finally, we present the modelling of D-NMDT on general box domains. Analogously as for NMDT, we apply McCormick envelopes to model all remaining product terms  $\alpha_j y$  and  $\Delta_x^L \cdot y$ . Further, we introduce the variables  $\hat{x} \in [0, 1]$  and  $\hat{z} \in [0, 1]$  to map the domain to  $[0, 1]$  intervals by using the transformations  $x := l_x\hat{x} + \underline{x}$  and  $y := l_y\hat{z} + \underline{y}$  as well as

$$\begin{aligned} z = xy &= (l_x\hat{x} + \underline{x})(l_y\hat{z} + \underline{y}) \\ &= l_x l_y \hat{x} \hat{z} + l_x \hat{x} \underline{y} + l_y \hat{z} \underline{x} + \underline{x} \underline{y} \\ &= l_x l_y \hat{z} + l_x \hat{x} \underline{y} + l_y \hat{z} \underline{x} + \underline{x} \underline{y}. \end{aligned}$$

As in the derivation of (21), we then obtain the formulation D-NMDT by applying McCormick envelopes to the product terms  $\beta_i((1 - \lambda)\Delta_x^L + \lambda\hat{x})$ ,  $\alpha_i(\lambda\Delta_z^L + (1 - \lambda)\hat{z})$  and  $\Delta_x^L \Delta_z^L$ . As in (21), we incorporate the following bounds to construct McCormick envelopes:

$$\begin{aligned} (1 - \lambda)\Delta_x^L + \lambda\hat{x} &\in [0, (1 - \lambda)2^{-L} + \lambda] \\ \lambda\Delta_z^L + (1 - \lambda)\hat{z} &\in [0, \lambda 2^{-L} + (1 - \lambda)]. \end{aligned}$$

Altogether, we are now ready to state the MIP relaxation D-NMDT on general box domains:

$$\begin{aligned}
 x &= l_x \hat{x} + \underline{x}, & y &= l_y \hat{z} + \underline{y} \\
 z &= l_x l_y \hat{z} + l_x \hat{x} \underline{y} + l_y \hat{z} \underline{x} + \underline{x} \underline{y} \\
 \hat{x} &= \sum_{j=1}^L 2^{-j} \beta_j^x + \Delta_x^L, & \hat{z} &= \sum_{j=1}^L 2^{-j} \beta_j^y + \Delta_z^L \\
 \hat{z} &= \sum_{j=1}^L 2^{-j} (u_j + v_j) + \Delta_z^L \\
 (\lambda \Delta_z^L + (1 - \lambda) \hat{z}, \beta_j^x, u_j) &\in \mathcal{M}(\lambda \Delta_z^L + (1 - \lambda) \hat{z}, \alpha_j) & j &\in 1, \dots, L \\
 ((1 - \lambda) \Delta_x^L + \lambda \hat{x}, \beta_j^y, v_j) &\in \mathcal{M}((1 - \lambda) \Delta_x^L + \lambda \hat{x}, \beta_j) & j &\in 1, \dots, L \\
 (\Delta_x^L, \Delta_z^L, \Delta_z^L) &\in \mathcal{M}(\Delta_x^L, \Delta_z^L) \\
 \Delta_x^L, \Delta_z^L &\in [0, 2^{-L}], & \hat{x}, \hat{z} &\in [0, 1], \quad \beta^x, \beta^y \in \{0, 1\}^L
 \end{aligned}
 \tag{35}$$

### B.2 MIP relaxations for univariate quadratic equations

For NMDT and D-NMDT, we derive the general formulations by using the derivations in Sect. 1 with  $x = y$ . In the case of NMDT, where the original model is (19), this leads to

$$\begin{aligned}
 x &= l_x \sum_{i=1}^L 2^{-i} \beta_i + \Delta_x^L + \underline{x} \\
 z &= l_x \sum_{i=1}^L 2^{-i} u_i + \Delta_z^L + \underline{x} \cdot x \\
 (x, \beta_i, u_i) &\in \mathcal{M}(x, \alpha_i) & i &\in 1, \dots, L \\
 (\Delta_x^L, x, \Delta_z^L) &\in \mathcal{M}(\Delta_x^L, x) \\
 \Delta_x^L &\in [0, 2^{-L} l_x], & x &\in [\underline{x}, \bar{x}], \quad \beta \in \{0, 1\}^L.
 \end{aligned}
 \tag{36}$$

For D-NMDT, we obtain (22) for general domains as follows:

$$\begin{aligned}
 x &= l_x \sum_{i=1}^L 2^{-i} \beta_i + l_x \Delta_x^L + \underline{x} \\
 z &= l_x \sum_{i=1}^L 2^{-i} u_i + l_x^2 \Delta_z^L + \underline{x} (x + l_x \Delta_x^L) \\
 (l_x \Delta_x^L + x, \beta_i, u_i) &\in \mathcal{M}(l_x \Delta_x^L + x, \beta_i) & i &\in 1, \dots, L \\
 (\Delta_x^L, \Delta_z^L) &\in \mathcal{M}(\Delta_x^L) \\
 \Delta_x^L &\in [0, 2^{-L}], & x &\in [\underline{x}, \bar{x}], \quad \beta \in \{0, 1\}^L,
 \end{aligned}
 \tag{37}$$

with  $l_x \Delta_x^L + x \in [\underline{x}, l_x 2^{-L} + \bar{x}]$ .

### Appendix C: Auxiliary proofs

In this section of the appendix, we give the proofs of Lemma 1 and Theorem 3 which we have moved here for better readability.

#### C.1 Piecewise McCormick relaxations of univariate DNMDT

We start with the proof of Lemma 1 which says that univariate D-NMDT gives a piecewise McCormick relaxation of  $\text{gra}(x^2)$ .

**Proof** Let  $P_L^{\text{IP}}$  be a univariate MIP relaxations as defined in (22). Consider a component of the MIP relaxation defined by fixing the variables  $\beta \in \{0, 1\}^L$ . In doing so, the condition  $(\Delta_x^L + x, \beta_j, u_j) \in \mathcal{M}(\Delta_x^L + x, \beta_j)$  becomes tight in the sense that we recover exactly  $(\Delta_x^L + x)\beta_j = u_j$ . This means that the model can reduce to this

$$x = \sum_{j=1}^L 2^{-j} \beta_j + \Delta_x^L, \quad z = \sum_{j=1}^L 2^{-j} (\Delta_x^L + x)\beta_j + \Delta_z^L$$

$$(\Delta_x^L, \Delta_z^L) \in \mathcal{M}(\Delta_x^L, \Delta_x^L), \quad \Delta_x^L \in [0, 2^{-L}], \quad x \in [0, 1].$$

The remaining McCormick containment  $(\Delta_x^L, \Delta_z^L) \in \mathcal{M}(\Delta_x^L, \Delta_x^L)$ , gives the bounds

$$\Delta_z^L \stackrel{(*)}{\geq} 0, \quad \Delta_z^L \stackrel{(**)}{\geq} 2 \cdot 2^{-L} \Delta_x^L - (2^{-L})^2, \quad \text{and} \quad \Delta_z^L \stackrel{(***)}{\leq} 2^{-L} \Delta_x^L.$$

We define  $\underline{x} := \sum_{j=1}^L \beta_j$  and  $\bar{x} := \underline{x} + 2^{-L}$ . With that  $\Delta_x^L = x - \underline{x}$  and  $\underline{x} = \bar{x} - 2^{-L}$  or equivalently  $\Delta_x^L = x - \bar{x} + 2^{-L}$  and  $\underline{x}^2 = \bar{x}^2 - 2\bar{x} \cdot 2^{-L} + (2^{-L})^2$  holds. The inequalities (\*\*) and (\*\*\*) can be rewritten as

$$\Delta_z^L \stackrel{(**)}{\geq} 2 \cdot 2^{-L} (x - \bar{x} + 2^{-L}) - (2^{-L})^2 = 2 \cdot 2^{-L} (x - \bar{x}) + (2^{-L})^2,$$

$$\Delta_z^L \stackrel{(***)}{\leq} 2^{-L} (x - \bar{x} + 2^{-L}).$$

We project out the variables  $\Delta_x^L$

$$z = \sum_{j=1}^L 2^{-j} (\Delta_x^L + x)\beta_j + \Delta_z^L = (\Delta_x^L + x) \sum_{j=1}^L 2^{-j} \beta_j + \Delta_z^L$$

$$= (\Delta_x^L + x)(\underline{x}) + \Delta_z^L = (2x - \underline{x})(\underline{x}) + \Delta_z^L = 2x\underline{x} - \underline{x}^2 + \Delta_z^L,$$

and apply the inequalities (\*), (\*\*) and (\*\*\*) to replace  $\Delta_z^L$  and get bounds on  $z$  in terms of  $x$ ,

$$\begin{aligned} z &\geq 2x\underline{x} - \underline{x}^2, \\ z &\geq 2x\underline{x} - \underline{x}^2 + 2 \cdot 2^{-L}(x - \bar{x}) + (2^{-L})^2, \\ z &\leq 2x\underline{x} - \underline{x}^2 + 2^{-L}(x - \underline{x}). \end{aligned}$$

Finally, we simplify the second and third inequality

$$\begin{aligned} z &\geq 2x(\bar{x} - 2^{-L}) - (\bar{x}^2 - 2\bar{x} \cdot 2^{-L} + (2^{-L})^2) + 2 \cdot 2^{-L}(x - \bar{x}) + (2^{-L})^2 \\ &= 2x\bar{x} - \bar{x}^2 \\ z &\leq 2x\underline{x} - \underline{x}^2 + 2^{-L}(x - \underline{x}) = x(2^{-L} + 2\underline{x}) - \underline{x}(2^{-L} - \underline{x}) = x(\bar{x} + \underline{x}) - \bar{x}\underline{x} \end{aligned}$$

and have

$$z \geq 2x\underline{x} - \underline{x}^2, \quad z \geq 2x\bar{x} - \bar{x}^2, \quad \text{and} \quad z \leq x(\bar{x} + \underline{x}) - \bar{x}\underline{x}. \tag{38}$$

Thus, on each interval the univariate D-NMDT MIP relaxation exactly recovers the McCormick envelope, i.e.  $\mathcal{M}(x, x)$  where  $x \in [\frac{i}{2^L}, \frac{i+1}{2^L}]$  with  $i = 0, \dots, 2^L - 1$ .  $\square$

### C.2 Sharpness of univariate NMDT

Next, we proof the sharpness of univariate NMDT proposed in Theorem 3.

**Proof** Let  $P_{L,L_1}^{IP}$  be the univariate T-D-NMDT with  $L_1 \geq L$  and  $P_{L,L_1}^{LP}$  the corresponding LP relaxation. We will proceed analogously to the proof of Theorem 2 and show

$$\text{proj}_{(x,z)}(P_{L,L_1}^{LP}) = \text{conv}(\text{proj}_{(x,z)}(P_{L,L_1}^{IP}))$$

by analyzing the minimum and maximum values of  $z$  in  $\text{proj}_{(x,z)}(P_{L,L_1}^{LP})$  and  $\text{conv}(\text{proj}_{(x,z)}(P_{L,L_1}^{IP}))$ . Similarly as for D-NMDT, also for NMDT

$$\max_{z \in [0,1]} \text{conv}(\text{proj}_{(x,z)}(P_{L,L_1}^{IP} |_{x=\bar{x}})) = \tilde{x}, \quad \forall \tilde{x} \in [0, 1]$$

applies. Next, we prove that

$$\max_{z \in [0,1]} \text{proj}_{(x,z)}(P_{L,L_1}^{LP} |_{x=\bar{x}}) \leq \tilde{x}, \quad \forall \tilde{x} \in [0, 1].$$

The McCormick cuts in (19) give the following upper bounds,

$$\Delta_z^L \leq \Delta_x^L, \quad \Delta_z^L \leq 2^{-L}x \text{ and } u_j \leq x, \quad u_j \leq \beta_j, \quad \forall j = 1, \dots, L.$$

This allows the following estimation for  $z$  in  $P_{L,L_1}^{LP}$ ,

$$z = \sum_{j=1}^L 2^{-j} u_j + \Delta_z^L \leq x \sum_{j=1}^L 2^{-j} + \Delta_z^L \leq 1(1 - 2^{-L}) + 2^{-L}x = x, \forall x \in [0, 1].$$

Next, we analyze the minimum of  $z$  in  $P_{L,L_1}^{IP}$ . We will show that for any fixed  $x \in [\underline{x}, \bar{x}]$ , with  $\underline{x} := i2^{-L}$  and  $\bar{x} := (i + 1)2^{-L}$  for  $i = 0, 1, \dots, 2^L$ , the lower bound on  $z$  in  $P_{L,L_1}^{IP}$  satisfies

$$z \geq \max\{x\underline{x}, (1 + \bar{x})x - \bar{x}\}. \tag{39}$$

We recall the lower bound for univariate D-NMDT in (38)

$$z \geq \max\{2x\underline{x} - \underline{x}^2, 2x\bar{x} - \bar{x}^2\},$$

and can easily show that the lower bound in univariate NMDT is weaker,

$$2x\underline{x} - \underline{x}^2 \geq 2x\underline{x} \geq x\underline{x}, 2x\bar{x} - \bar{x}^2 \geq 2x\bar{x} \geq x\bar{x} \geq x\bar{x} - (\bar{x} - x) = (1 + \bar{x})x - \bar{x}.$$

Hence, it follows from the same proof logic of Theorem 2 that the tightening inequalities from the sawtooth relaxation are stronger than the MIP relaxation lower bound of univariate NMDT, and hence tightening, in this case, recovers sharpness of the formulation as well.

Proof of (39): Fix some  $x \in [0, 1]$ . If  $x = i2^{-L}$  for some  $i$ , then the formulation is tight to  $z = x^2$  and there is nothing to show.

Otherwise, there is a unique choice of  $\beta \in \{0, 1\}^n$  and  $\Delta_x^L \in [0, 2^{-L}]$  such that

$$x = \sum_{j=1}^L 2^{-j} \beta_j + \Delta_x^L.$$

So consider the MIP relaxation and assume that  $\beta$  and  $\Delta_x^L$  are fixed.

Hence, the McCormick envelope  $(x, \beta_j, u_j) \in \mathcal{M}(x, \beta_j)$  is tight, meaning that  $x\beta_j = u_j$ .

We will decompose  $z$  and observe that

$$z = \sum_{j=1}^L 2^{-j} u_j + \Delta_z^L = x \sum_{j=1}^L 2^{-j} \beta_j + \Delta_z^L = x\underline{x} + \Delta_z^L = x(\bar{x} - 2^{-L}) + \Delta_z^L. \tag{40}$$

Next, we reformulate the lower McCormick inequalities on  $\Delta_z^L$  in (19),

$$\Delta_z^L \geq \max\{0, 2^{-L}x + \Delta_x^L - 2^{-L}\} = \max\{0, (1 + 2^{-L})x - \bar{x}\}. \tag{41}$$

The final inequality above is reached by the substitution  $\Delta_x^L = x - \bar{x} + 2^{-L}$ . If we incorporate (41) into (40) we get exactly the bounds proposed in (39). Since these are the only lower bounds on  $\Delta_z^L$ , the bounds in (39) are attained when minimizing over  $z$ , which finishes the proof.  $\square$

### Appendix D: Instance set

In Table 9 we show a listing of all instances of the computational study from Sect. 5. The boxQP instances are publicly available at <https://github.com/joe-huchette/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 9** IDs of all 60 instances used in the computational study.

|   |                  |                  |                  |                  |
|---|------------------|------------------|------------------|------------------|
| boxQP instances: spar                     |                  |                  |                  |                  |
| <b>020-100-1</b>                          | <b>020-100-2</b> | <b>030-060-1</b> | <b>030-060-3</b> | <b>040-030-1</b> |
| <b>040-030-2</b>                          | <b>050-030-1</b> | <b>050-030-2</b> | <b>060-020-1</b> | <b>060-020-2</b> |
| 070-025-2                                 | <b>070-050-1</b> | <b>080-025-1</b> | <b>080-050-2</b> | <b>090-025-1</b> |
| <b>090-050-2</b>                          | <b>100-025-1</b> | <b>100-050-2</b> | <b>125-025-1</b> | <b>125-050-1</b> |
| 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_wscs_api       |
| 9_wscs_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_                   |                  |                  |                  |                  |
| <b>0031</b>                               | <b>0032</b>      | <b>0343</b>      | 0681             | 0682             |
| 0684                                      | 0698             | <b>0911</b>      | <b>0975</b>      | <b>1055</b>      |
| <b>1143</b>                               | <b>1157</b>      | <b>1423</b>      | <b>1922</b>      | 2882             |
| 2894                                      | 2935             | 2958             | 3358             | 3814             |

In bold are the IDs of the instances that are dense

## References

1. Aigner, K.-M., Burlacu, R., Liers, F., Martin, A.: Solving an optimal power flow with discrete decisions to global optimality. *INFORMS J. Comput.* **35**(2), 458–474 (2023)
2. 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. Global Optim.* **1–31**, 85 (2022)
3. Beach, B., Burlacu, R., Bärmann, A., Hager, L., Hildebrand, R.: Enhancements of discretization approaches for non-convex mixed-integer quadratically constrained quadratic programming: Part I. arXiv preprint [arXiv:2211.00876](https://arxiv.org/abs/2211.00876) (2022)
4. 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)
5. Beach, B., Hildebrand, R., Huchette, J.: Compact mixed-integer programming formulations in quadratic optimization. *J. Global Optim.* (2022)
6. 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)
7. Castro, P.M.: Normalized multiparametric disaggregation: an efficient relaxation for mixed-integer bilinear problems. *J. Global Optim.* **64**(4), 765–784 (2015)
8. Chen, J., Burer, S.: Globally solving nonconvex quadratic programming problems via completely positive programming. *Math. Program. Comput.* **4**(1), 33–52 (2012)
9. Coffrin, C., Gordon, D., Scott, P.: NESTA, the NICTA energy system test case archive. arXiv preprint [arXiv:1411.0359](https://arxiv.org/abs/1411.0359) (2014)
10. Dolan, E.D., Moré, J.J.: Benchmarking optimization software with performance profiles. *Math. Program.* **91**(2), 201–213 (2002)
11. Dong, H., Luo, Y.: Compact disjunctive approximations to nonconvex quadratically constrained programs (2018)
12. 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)
13. Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual (2022)
14. Huchette, J.A.: Advanced mixed-integer programming formulations: methodology, computation, and application. PhD thesis, Massachusetts Institute of Technology (2018)
15. Linderoth, J.: A simplicial branch-and-bound algorithm for solving quadratically constrained quadratic programs. *Math. Program.* **103**(2), 251–282 (2005)
16. McCormick, G.P.: Computability of global solutions to factorable nonconvex programs: Part I—convex underestimating problems. *Math. Program.* **10**(1), 147–175 (1976)
17. 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 (2016)
18. Telgarsky, M.: Representation benefits of deep feedforward networks. [arxiv:https://arxiv.org/abs/1509.08101](https://arxiv.org/abs/1509.08101) (2015)
19. Wachter, A.: An interior point algorithm for large-scale nonlinear optimization with applications in process engineering. PhD thesis, Carnegie Mellon University (2002)
20. Yarotsky, D.: Error bounds for approximations with deep relu networks. *Neural Netw.* **94**, 103–114 (2017)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.