universität wien

DISSERTATION | DOCTORAL THESIS

Titel | Title

Conic approaches to mixed-integer QPs with complementarity constraints and applications to sparse machine learning models

verfasst von | submitted by

Bo Peng

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 794 370 403
Dissertationsgebiet lt. Studienblatt | Field of study as it appears on the student record sheet:	Business Analytics, Logistics and Operations Research
Betreut von | Supervisor:	Univ.-Prof. i.R. Mag. Dr. Immanuel Bomze

Acknowledgements

My PhD journey started in October 2020 at the University of Vienna where I was given an opportunity to continue my passion for doing research. These past years have been a time of learning, personal growth, and professional development, and I am grateful to the many people who have supported and guided me along the way.

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Immanuel Bomze, whose expertise, patience, and guidance have been invaluable throughout my research. When I encountered difficulties in my journey, he always encouraged me and provided me with professional and thoughtful suggestions, which I found out later, were the best solutions. He never lost his patience in teaching and guiding me to be a qualified PhD candidate from every perspective, and I could not have asked for a more dedicated and encouraging supervisor.

I would also like to extend my thanks to my colleagues and collaborators. Working and studying alongside them has been a source of inspiration and motivation. In particular, I would like to deliver my thanks to the speaker, the invited lecturers, and colleagues from the Vienna Doctoral School on Computational Optimization (VGSCO), funded by FWF (Austrian Science Fund), Project W1260-N35: Professor Radu Ioan Bot, Professor Francesco Rinaldi, Professor Yurii Malitskyi, Professor Gabriele Eichfelder, Dr. Markus Gabl, Mikhail Karapetyants, Bettina Hiebl. Moreover, I would also like to thank Professor E. Alper Yıldırım and Professor Laura Palagi, who were the hosts during my research stays at the University of Edinburgh and the University of Rome for their attentive assistance with my personal life and thoughtful discussions on the joint projects. Finally, I would like to extend my special thanks to Dr. Yuzhou Qiu, a collaborator and one of my closest friends now. The experience of working with him during my stay in Edinburgh made me more determined to continue on the academic path, as I admire his passion and dedication to academia.

To my dear friends, I extend my gratitude for your encouragement and belief in me, even in moments when I doubted myself, which helped me maintain a balanced perspective. I would like to offer special thanks to Yang Zhang, Yiwen Yang, Guojun Lai, Shiyang Zhou, and Zhijin Xue, whose presence brought color and joy to the monotonous PhD journey. Finally, I am especially grateful to Claudya (Yun) Hu, who accompanied me during the thesis writing process. The time spent together not only lifted my spirits but also provided me with inspiration and courage.

Last but not least, I would like to express my deepest appreciation to my mother Shu-ying Shu, and my father Cheng-long Peng. Without your unconditional love and support, I would not have been able to complete this PhD journey. As an only child, I am deeply aware of the emotional hardship my parents have suffered, particularly during holidays when families traditionally come together. I hope to make up one day for the moments we have missed.

Thank you all for being part of this journey with me.

Abstract

The thesis addresses mixed-integer (nonconvex) quadratic optimization problems (QPs) with complementarity constraints (CC), a crucial subclass of quadratically constrained quadratic optimization problems (QCQPs). These problems serve as powerful modeling tools in various fields, including interpretable machine learning, mathematical finance, and selection in biosciences. Given the challenge of solving these problems globally, rather than solving the problems directly, the focus is placed on convexification techniques, such as linear conic relaxations. As demonstrated in the thesis, a tight yet tractable convex relaxation not only provides a robust lower bound but also yields a high-quality feasible point through heuristics based on the optimal solution of the relaxation. These relaxations can be further integrated into the branch-and-bound framework, offering a comprehensive approach for globally solving the original problems.

This cumulative thesis is primarily organized as a collection of papers already published or submitted to internationally renowned journals [Pen22, BP23, BDPP24, BPQY24, BPQY25] and a nearly complete work, and the main results can be divided into two parts. The first part consists of two chapters: In Chapter 2, I establish several completely positive (CP) reformulations for the QPCCs under new conditions that are significantly less restrictive than those in the current literature. Essentially, CP reformulations package the nonconvexity into the completely positive cone, a matrix cone that has received increasing attention in the optimization domain recently. Additionally, novel and compact CP reformulations are proposed for the (nonconvex) QPs involving a cardinality term in Chapter 3, which is based on a work nearing completion to be ready for submission.

The second part (Chapters 4 and 5) focuses on tractable convex relaxations based on the CP reformulations, which are of particular interest in real-world applications. For sparse standard quadratic optimization problems (StQPs), an analytical comparison is made among these tractable relaxations, resulting in more compact hybrid relaxations. Furthermore, novel and computationally scalable relaxations are proposed for feature selection in support vector machines (SVMs), a widely used tool in supervised machine learning. In this context, the cardinality constraint enhances the interpretability of the resulting AI models. Extensive numerical experiments demonstrate the effectiveness and efficiency of these innovative relaxations.

Kurzfassung

Die Dissertation befasst sich mit gemischt-ganzzahligen (nichtkonvexen) quadratischen Optimierungsproblemen (QPs) mit Komplementaritätsbedingungen (CC), einer wichtigen Unterklasse von quadratischen Optimierungsproblemen unter (linearen und) quadratischen Nebenbedingungen (QCQPs). Diese Probleme dienen als leistungsstarke Modellierungswerkzeuge in verschiedenen Bereichen, darunter interpretierbares maschinelles Lernen, mathematische Finanzen und die Selektion in den Biowissenschaften. Angesichts der Herausforderung, diese Probleme global zu lösen, wird der Fokus nicht auf die direkte Lösung der Probleme gelegt, sondern auf Konvexifizierungstechniken wie lineare konische Relaxationen. Wie in der Dissertation gezeigt wird, liefert eine starke, aber dennoch handhabbare konvexe Relaxation nicht nur eine robuste untere Schranke, sondern ermöglicht auch Heuristiken, die auf der optimalen Lösung der Relaxation basieren, und die einen qualitativ hochwertigen zulässigen Punkt liefern. Diese Relaxationen können zudem in den Branch-and-Bound-Ansatz integriert werden, was einen umfassenden Ansatz zur globalen Lösung der ursprünglichen Probleme bietet.

Diese kumulative Dissertation ist in erster Linie als eine Sammlung von (bereits publizierten oder bei international angesehenen Zeitschriften eingereichten) Aufsätzen [Pen22, BP23, BDPP24, BPQY24, BPQY25] und einem fast fertiggestellten Artikel organisiert. Die Hauptresultate lassen sich in zwei Teile gliedern. Der erste Teil besteht aus zwei Kapiteln: In Kapitel 2 stelle ich mehrerecompletely positive (CP) Reformulierungen für die QPCCs unter neuen Bedingungen vor, die deutlich weniger restriktiv sind als die in der aktuellen Literatur. Im Wesentlichen codieren die CP-Reformulierungen die Nichtkonvexität mit Hilfe des completely positive cone, eines Matrixkegels, der in der Optimierungsforschung in letzter Zeit zunehmend Beachtung findet. Darüber hinaus werden im Kapitel 3 neuartige und kompakte CP-Reformulierungen für die (nichtkonvexen) QPs mit Kardinalitätsbedingungen vorgeschlagen, basierend auf einer Arbeit, die beinahe zur Einreichung bereit ist.

Der zweite Teil (Kapitel 4 und Kapitel 5) konzentriert sich auf handhabbare konvexe Relaxationen, die auf den CP-Reformulierungen basieren, was besonders relevant in realen Anwendungen ist. Für Standard-Quadratische-Optimierungsprobleme (StQPs) unter Kardinalitätsbedingungen wird ein analytischer Vergleich dieser handhabbaren Relaxationen vorgenommen, was zu kompakteren hybriden Relaxationen führt. Des Weiteren werden neuartige und rechnerisch skalierbare Relaxationen für die Merkmalsauswahl in Support-Vektor-Maschinen (SVMs) vorgeschlagen, ein weit verbreitetes Werkzeug im überwachten maschinellen Lernen. In diesem Zusammenhang verbessert die Kardinalitätsbedingung die Interpretierbarkeit der resultierenden KI-Modelle. Umfangreiche numerische Experimente belegen die Effektivität und Effizienz dieser innovativen Relaxationen.

Contents

List of Tables

List of Figures

1. Introduction

In this section, we provide a brief introduction to the background and motivation of this thesis. Additionally, we present an overview of the primary technique employed: copositive optimization.

1.1. Sparsity in machine learning

Sparsity is becoming an increasingly important concept in machine learning, especially in the era of high-dimensional and high-frequency data. According to Albert Einstein, "The supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible", this principle of simplicity is highly relevant to machine learning as well. A term that represents this idea is data sparsity, which describes a situation where a substantial proportion of data points are either missing or set to zero. Mathematically, we use the notation $\|\cdot\|_0$ to capture the sparsity of a given data (vector). For a vector $\mathsf{x} \in \mathbb{R}^n$, by

$$\|\mathsf{x}\|_0 := |\{i \in [1\!:\!n] : x_i \neq 0\}|\,,$$

where by $[a\!:\!b]$, for any integers a, b, we denote the set of all integers k satisfying $a \leq k \leq b$.

we count the number of non-zero elements of x. Consider a minimization optimization problem over the variable x. By adding the term $\|\mathsf{x}\|_0$ to the objective function with a positive coefficient λ, it is evident that the solution tends to exhibit sparsity. In an extreme case where λ approaches infinity, the optimization problem effectively seeks the sparsest feasible point of the original problem. Alternatively, we can directly limit the number of nonzero elements in the variable x using a cardinality constraint

$$\|\mathsf{x}\|_0 \leq k,$$

where k is a non-negative integer. In this case, x is said to be k-sparse if the constraint is satisfied. Incorporating the sparse term $\|\cdot\|_0$ into many machine learning tasks promotes sparsity in the solutions, leading to significant advantages, as demonstrated in the following subsections.

Sparse principal component analysis

Sparse principal component analysis (sparse PCA) is a widely used data processing and reduction technique that seeks to find a low-dimensional representation of the data while enforcing sparsity on the principal components. For a given data $\mathsf{X} \in \mathbb{R}^{n\times p}$ of n observations on p variables, PCA is a process of reconstructing the empirical covariance matrix of data by linear combinations of the original p variables. This leads to the following optimization problem [Jol02]:

$$\begin{aligned} \max_{\mathsf{v}\in\mathbb{R}^p} \quad & \mathsf{v}^\top \Sigma \mathsf{v} \\ \text{s.t.} \quad & \mathsf{v}^\top \mathsf{v} = 1\,, \end{aligned}$$

where Σ is the empirical covariance matrix of data X and we call the optimal solution v^* principal component. The principal component is typically dense, e.g., a linear combination of all original variables. As a result, interpreting the principal component can be difficult, especially when

you have a large number of variables. In other words, the individual component can be too abstract to be related back to the original data. To overcome the weakness of traditional PCA, the following sparse PCA is proposed

$$\begin{aligned} \max_{\mathsf{v}\in\mathbb{R}^p} \quad & \mathsf{v}^\top \Sigma \mathsf{v} \\ \text{s.t.} \quad & \mathsf{v}^\top \mathsf{v} = 1 \\ & \|\mathsf{v}\|_0 \le k\,, \end{aligned}$$

where we restrict the number of variables involved in each principal component. Since in each principal component, only a small subset of variables are involved, this gives more direct insight into which specific variables (features) are responsible for the most significant patterns. Therefore it makes the results more interpretable and meaningful, especially in domains like biology, finance, or social sciences, where understanding the influence of individual variables is crucial.

Compressed sensing

Compressed sensing is one of the significant breakthroughs in signal processing, which allows the recovery of signals via much fewer samples than required by the Nyquist–Shannon sampling theorem [Sha48] under the condition:

- The signal is sparse in some domains (e.g., Fourier, wavelet);
- The measurement matrix A must satisfy the restricted isometry property (RIP) [CRT06, CT06]. RIP ensures that the lengths of sparse signals are approximately preserved when projected onto the lower-dimensional measurement space:

$$(1-\delta_k)\|\mathsf{x}\|_2^2 \le \|\mathsf{A}\mathsf{x}\|_2^2 \le (1+\delta_k)\|\mathsf{x}\|_2^2$$

.

Mathematically, compressed sensing turns out to find a sparse solution from an underdetermined linear system:

$$\mathsf{A}\mathsf{x} = \mathsf{b},$$

where $\mathsf{A} \in \mathbb{R}^{m\times n}$ $(m \ll n)$ is the measurement matrix and $\mathsf{b} \in \mathbb{R}^m$ is m observations. The related optimization problem reads as follows:

$$\begin{aligned} \max_{\mathsf{x}\in\mathbb{R}^n} \quad & \|\mathsf{x}\|_0 \\ \text{s.t.} \quad & \mathsf{A}\mathsf{x} = \mathsf{b}\,. \end{aligned}$$

Sparse dictionary learning

Sparse Dictionary Learning is a technique in machine learning and signal processing that seeks to represent data as a linear combination of a few elements from a larger set, known as the dictionary. The dictionary $\mathsf{D} \in \mathbb{R}^{n\times m}$ is a set of vectors that forms the building blocks for reconstructing data $\mathsf{x}_i \in \mathbb{R}^n$ for $i \in I$. Our goal is to find a set of sparse vector $\alpha_i \in \mathbb{R}^m$ and the dictionary D, such that

$$\mathsf{x}_i \approx \mathsf{D}\alpha\,.$$

The corresponding optimization problem is usually of the forms:

$$\begin{aligned} \min_{\mathsf{D},\alpha_i} \quad & \sum_{i\in I} \|\mathsf{x}_i - \mathsf{D}\alpha_i\|_2^2 \\ \text{s.t.} \quad & \|\alpha_i\|_0 \le k\,, \end{aligned}$$

or

$$\min_{\mathsf{D},\alpha_i} \quad \sum_{i\in I} \|\mathsf{x} - \mathsf{D}\alpha_i\|_2^2 + \lambda\|\alpha_i\|_0\,,$$

where the first problem directly limits the variables α_i to be k-sparse, and the second problem price the sparsity of the variables α_i in the objective by multiplying a user-defined parameter λ.

Sparse support vector machine (controlled feature selection)

Support vector machine (SVM) is a supervised learning technique mainly used for classification tasks. The key idea behind SVM is to find the optimal hyperplane that separates data points of different classes with the maximum margin. Sparse SVM extends traditional SVM by introducing the sparsity constraint in the feature space, which is particularly useful for feature selection, improving interpretability, reducing overfitting, and enhancing computational efficiency, especially in high-dimensional datasets. Here we take linear SVM with the soft margin as an example:

$$\begin{aligned} \min_{\mathsf{w},b,\xi} \quad & \frac{1}{2}\|\mathsf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y^i(\mathsf{w}^\top \mathsf{x}^i + b) \ge 1 - \xi_i\,, \quad i\in[1\!:\!m] \\ & \|\mathsf{w}\|_0 \le B \\ & \xi \ge \mathsf{o}\,, \end{aligned}$$

where the user-defined parameter $B \in \mathbb{Z}_+$ constitutes a budget on the number of features used, indicating that the number of features involved in the classifier is at most B.

After showing a few examples where we employ $\|\cdot\|_0$ for data sparsity, We close the subsection by listing the benefits of enforcing sparsity in machine learning:

- Sparsity allows models to be more interpretable because fewer variables are involved in the final model. Irrelevant or redundant features are assigned a coefficient of zero, making it clear which features are driving the model's predictions.
- In high-dimensional datasets, the number of features can be much larger than the number of observations. In these cases, enforcing sparsity in machine learning models automatically performs feature selection by excluding irrelevant variables, which reduces the risk of overfitting and simplifies the model.
- Sparse models require fewer computations because many features are excluded from the model. This reduces memory usage and computational time, especially important in large-scale applications such as natural language processing or computer vision.

1.2. Complementarity constraint

Complementarity constraints are widely used to model various combinatorial structures, such as binary constraints, cardinality constraints, logical relations, and disjunctive conditions. As a

result, they are attracting increased attention in interpretable machine learning where we focus on explainable and understandable machine learning models by using logical building blocks rather than relying on black-box approaches.

A complementarity constraint typically has the following form:

$$x \geq 0, \quad y \geq 0, \quad xy = 0\,.$$

In the following, we will demonstrate the strong modeling ability of the complementarity constraint with some examples.

Binary constraint

The binary constraint simply means a variable only takes two values. e.g., $x \in \{0, 1\}$. This can be modeled with complementarity constraint by introducing an auxiliary variable:

$$x + y = 1, \quad xy = 0\,.$$

It is important to note that mixed-integer optimization problems represent a special case of optimization problems with complementarity constraints. As a result, the latter problem class contains at least the former, which includes a wide range of problems in combinatorial optimization.

Cardinality constraint

The cardinality constraint refers to limiting the number of non-zero elements of a variable $\mathsf{x} \in \mathbb{R}^n_+$. For instance:

$$\|\mathsf{x}\|_0 \leq k\,.$$

The importance of this type of constraint has been fully addressed in the last subsection. In fact, after introducing auxiliary variables, it can be modeled with complementarity constraints as:

$$\mathsf{o} \leq \mathsf{u} \leq \mathsf{e}, \quad \mathsf{e}^\top \mathsf{u} \geq n - k, \quad \mathsf{x}^\top \mathsf{u} = 0$$

where e is the all-ones vector. Indeed, the first two constraints imply that at least $n - k$ entries of u are positive. Due to the complementarity constraint between x and u, it follows immediately that x has at least $n - k$ zero entries.

Logical If-Then constraint

Logical If-Then constraints are used to model relationships where the satisfaction of one condition (the "if" part) implies the satisfaction of another condition (the "then" part). Given two constraints of the form:

$$f(x) > 0, \quad g(x) \geq 0\,,$$

consider the logical condition:

$$\text{If} \quad f(x) > 0, \quad \text{then} \quad g(x) \geq 0\,.$$

By introducing auxiliary variables, we arrive at the following formulation with complementarity constraint:

$$\begin{aligned} f(x) + s_1 - s_2 = 0, \quad g(x) + s_3 - s_4 = 0, \\ s_1 s_2 + s_3 s_4 + s_2 s_3 = 0, \quad s_1, s_2, s_3, s_4 \geq 0\,. \end{aligned}$$

Note that s_1 and s_3 are negative parts of $f(x)$ and $g(x)$ respectively. Similarly, s_2 and s_4 are positive parts of $f(x)$ and $g(x)$ respectively. Indeed, $f(x) > 0$ implies $s_2 > 0$. Together with the constraint $s_2 s_3 = 0$, we have $s_3 = 0$ and therefore $g(x) = s_4 \geq 0$. As we may know, the If-Then constraint can be modeled by the so-called big-M formulation, where we introduce a sufficiently large constant M and binary variables. However, the existence and the choice of such M are unclear in many cases, which may introduce numerical instability, especially when it is fed to a branch and bound method. Large values of M can cause the optimization problem to become ill-conditioned, leading to slower convergence or inaccurate solutions. Instead, the formulation with complementarity constraint does not involve any big M and binary variables, and the difficulty is only packaged in the complementarity constraint.

Logical Either-Or constraint

Logical Either-Or constraint refers to the situation where exactly one of two mutually exclusive conditions must hold, simplifying the optimization problem to choosing between two constraints. For example, consider a production planning problem in which a machine can either produce Product A or Product B, but not both at the same time. Let

- $f(x)$ represents the production level of Product A;
- $g(x)$ represents the production level of Product B.

The Either-Or constraint is:

$$\text{Either} \quad f(\mathsf{x}) \geq 0 \quad \text{or} \quad g(\mathsf{x}) \geq 0.$$

This can be formulated as:

$$\begin{aligned} & f(\mathsf{x}) + s_1 \geq 0, \quad g(\mathsf{x}) + s_2 \geq 0, \\ & s_1 s_2 = 0, \quad s_1 + s_2 \geq \epsilon\,, \\ & s_1, s_2 \geq 0\,, \end{aligned}$$

where $\epsilon > 0$ is a small enough positive parameter. Due to the constraints on s_1 and s_2, we have either $s_1 = 0$ and $s_2 > 0$, or $s_1 > 0$ and $s_2 = 0$. Together with the constraints $f(\mathsf{x}) + s_1 \geq 0$ and $g(\mathsf{x}) + s_2 \geq 0$, it follows that either $f(\mathsf{x}) \geq 0$ or $g(\mathsf{x}) \geq 0$, but not both simultaneously.

Disjunctive constraint

The disjunctive constraint can be viewed as an extension of the logical Either-Or constraint and refers to the feasible region described as a union of a finite collection of disjoint sets S_i, $i \in [1\!:\!k]$. A disjunctive constraint has the form:

$$\mathsf{x} \in \cup_{i=1}^{k} S_i\,,$$

A piecewise defined function consists of different functional expressions on different parts, therefore it can be naturally represented using disjunctive constraints. We take the well-known rectified linear unit (ReLU) function as an example:

$$\mathrm{ReLU}(x) = \max(0, x)\,.$$

The function returns 0 if $x < 0$ and returns x if $x \geq 0$. Again, employing a complementarity constraint and auxiliary variables $s_i \geq 0$, the ReLU can be equivalently expressed as:

$$\begin{aligned} & x + s_1 - s_2 = 0, \quad s_1 s_2 + \mathrm{ReLU}(x) s_1 = 0, \\ & \mathrm{ReLU}(x) \leq s_2, \quad \mathrm{ReLU}(x) \geq s_2 - s_1, \\ & s_1, s_2, \mathrm{ReLU}(x) \geq 0. \end{aligned}$$

If $x < 0$, due to the nonnegativity of s_1 and s_2, and the complementarity between them, we have $s_1 > 0$ and $s_2 = 0$. Together with the constraint $\mathrm{ReLU}(x)s_1 = 0$, it follows immediately that $\mathrm{ReLU}(x) = 0$; if $x \geq 0$, due to the nonnegativity of s_1 and s_2, and the complementarity between them, we have $s_1 = 0$ and $s_2 \geq 0$. Since $s_2 - s_1 \leq \mathrm{ReLU}(x) \leq s_2$, we have $\mathrm{ReLU}(x) = x$.

As shown above, the complementarity constraint is a versatile and efficient tool for a wide range of optimization and decision-making problems. It has a strong modeling ability arising from its flexibility in handling logical relationships, piecewise functions, non-convexities, binarity, and more compact formulations for many mixed-integer optimization problems.

However, an optimization problem involving complementarity constraints is generally difficult to solve due to its non-smoothness and non-convexity. Even the simplest one, a linear optimization problem with complementarity constraints, is NP-hard [CPS92]. Nevertheless, these constraints have two distinguished properties: they are homogeneous, and the variables involved are nonnegative. Based on these two key features, we will introduce a tool that has the potential to handle this type of problem in the next section.

1.3. Copositive optimization

Copositive optimization is a branch of mathematical optimization that focuses on problems involving copositive matrices, where a matrix is said to be copositive if its quadratic form takes non-negative values for all non-negative vectors. The field has gained attention due to its ability to reformulate many hard combinatorial optimization problems, and the interested readers are referred to [Bom12, BSU12, Bur15, DR21] and references therein.

Copositive optimization is closely related to semidefinite programming (SDP), but it goes beyond SDP by encompassing problems that are generally more difficult to solve due to the structure of the copositive cone. Despite this, it has the advantage of transforming certain non-convex problems into a convex framework, offering theoretical and computational benefits.

The concept of copositive matrices was introduced in the 1950s in the context of quadratic forms and matrix theory [Mot52]. A matrix $\mathsf{Q} \in \mathcal{S}^n$ is said to be copositive if:

$$\mathsf{x}^\top \mathsf{Q}\mathsf{x} \geq 0, \quad \forall \mathsf{x} \in \mathbb{R}^n_+ .$$

The set of copositive matrices is known as the copositive cone, denoted by $\mathcal{COP}^n$, which is a convex, closed, pointed, and full-dimensional cone. For a given matrix $\mathsf{Q} \in \mathcal{S}^n$, the membership problem to decide whether or not $\mathsf{Q} \in \mathcal{COP}^n$ holds is co-NP-complete [MK87]. It is equivalent to determining whether or not $v(\mathsf{Q}) \geq 0$ holds, for

$$v(\mathsf{Q}) := \min_{\mathsf{x} \in \Delta_n} \mathsf{x}^\top \mathsf{Q}\mathsf{x} ,$$

where $\Delta_n := \{\mathsf{x} \in \mathbb{R}^n_+ : \|\mathsf{x}\|_1 = 1\}$ denotes the standard simplex. The above standard quadratic problem is NP-hard, since the maximum clique problem can be reduced to it [MS65].

For an arbitrary cone $\mathcal{K} \in \mathcal{S}$, the dual cone $\mathcal{K}^*$ is defined as:

$$\mathcal{K}^* := \{\mathsf{A} \in \mathcal{S} : \langle \mathsf{A}, \mathsf{B} \rangle \geq 0, \quad \text{for all} \quad \mathsf{B} \in \mathcal{K}\} .$$

Therefore, the dual cone of the copositive cone equals

$$\mathcal{CP}^n = \left\{ \sum_{i \in I} a_i a_i^\top : a_i \in \mathbb{R}^n_+ \right\} ,$$

which is known as the completely positive (CP) cone.

The idea of using copositive matrices in optimization started gaining attraction in the 1990s, particularly in the field of quadratic programming and combinatorial optimization. Researchers discovered that certain non-convex quadratic optimization problems could be reformulated as convex optimization problems over the copositive cone[QdKRT98].

A copositive optimization problem refers to a linear optimization problem over the completely positive cone and has the following form:

$$\begin{aligned} \min_{\mathsf{X}\in\mathcal{S}^n} \quad & \langle \mathsf{C}, \mathsf{X}\rangle \\ \text{s.t.} \quad & \mathcal{A}\mathsf{X} = \mathsf{b} \\ & \mathsf{X} \in \mathcal{CP}^n\,, \end{aligned} \tag{1.1}$$

where $\mathcal{A} : \mathcal{S}^n \to \mathbb{R}^m$ is a linear operator.

The Lagrangian dual of (1.1) is a linear conic optimization problem over the copositive cone:

$$\begin{aligned} \max_{\mathsf{y}\in\mathbb{R}^m_+} \quad & \mathsf{b}^\top \mathsf{y} \\ \text{s.t.} \quad & \mathcal{A}^*\mathsf{y} + \mathsf{Z} = \mathsf{C} \\ & \mathsf{Z} \in \mathcal{COP}^n\,, \end{aligned} \tag{1.2}$$

where $\mathcal{A}^* : \mathbb{R}^m \to \mathcal{S}^n$ is the adjoint operator of $\mathcal{A}$. For historical reasons and simplicity, both problems (sometimes the primal-dual pair of conic problems) are frequently addressed as copositive optimization problems.

In the early 2000s, researchers [BDDK$^+$00] recognized the significant potential of copositive optimization when it was discovered that many NP-hard combinatorial problems could be reformulated as linear optimization problems over the copositive cone. A seminal contribution in this field was made by Burer [Bur09], who demonstrated that the quadratic optimization problem with binary variables could be exactly reformulated as a copositive optimization problem. Another important work [PVZ15] extends the scope of applying copositive optimization to polynomial optimization problems, presenting a general characterization of the class of problems that can be formulated as a conic program over the cone of completely positive tensors. Essentially, the reformulation effectively packages the complexity of the problem within the CP cone, enabling tighter bounds and novel approximation techniques based on the CP cone.

Despite the compact form of CP reformulation, their computational intractability arises from the presence of the CP cone. In practice, outer approximations of the CP reformulation are typically considered, yielding lower bounds that are significantly tighter than those obtained through SDP relaxations. It is important to note that any feasible solution to (1.2) gives, by weak duality, a rigorous lower bound to the optimal value of (1.1) which in turn is, by relaxation, a lower bound to many NP-hard problems (sometimes the same, for CP reformulations). Moreover, (1.2)-feasibility requires copositivity testing for Z, which underlines the importance of the next section.

1.4. Copositivity tests: comparison of two recent proposals

As demonstrated in the previous section, the primary challenge of a copositive optimization problem lies in the cone constraint. In this section, we focus on methods for testing the copositivity

of a given matrix and provide a numerical comparison of these methods. Additionally, a modified variant is introduced, which offers improved efficiency in certain scenarios. This section is mainly based on a peer-reviewed and published short communication [Pen22].

1.4.1. Copositivity test based on almost copositivity

A matrix $\mathsf{Q} \in \mathcal{S}^n$ is said almost copositive if either $n = 1$ or every $n-1 \times n-1$ principal submatrix of Q is copositive. Based upon the theory of almost copositivity (see, e.g. [Dic19]), the approach of [Ans21] formulates an MILP as follows:

$$\begin{aligned} l_1(\mathsf{Q}) := \max \quad & \lambda \\ s.t. \quad & \mathsf{q}_i^\top \mathsf{x} \leq -\lambda + m_i(1-y_i), \quad i \in [1\!:\!n], \\ & \lambda \geq 0,\, \mathsf{o} \leq \mathsf{x} \leq \mathsf{y}, \\ & \mathsf{y} \in \{0,1\}^n, \mathsf{e}^\top \mathsf{y} \geq \theta, \end{aligned} \tag{COP1}$$

where $m_i = 1 + \sum_{j \neq i} \max\{q_{ij}, 0\}$ for all $i \in [1\!:\!n]$ and $\theta > 0$ is a parameter.

Theorem 1. *[Ans21] Let $\theta = 1$. Then Q is copositive if and only if the optimal value of* (COP1) *is zero.*

Essentially, the optimization problem (COP1) identifies the existence of a copositive matrix among the principal submatrices of Q. According to [Dic19], if an almost copositive submatrix of Q exists, the objective value of (COP1) must be strictly greater than zero; otherwise, Q is copositive.

1.4.2. Copositivity test based on KKT conditions

Recall that a matrix $\mathsf{Q} \in \mathcal{S}^n$ is copositive if and only if

$$\mathsf{x}^\top \mathsf{Q} \mathsf{x} \geq 0, \quad \forall \mathsf{x} \in \mathbb{R}^n_+ ,$$

which is equivalent to $v(\mathsf{Q}) \geq 0$ and it is defined as the optimal value of the optimization problem:

$$\begin{aligned} v(\mathsf{Q}) := \min_{\mathsf{x} \in \mathbb{R}^n_+} \quad & \mathsf{x}^\top \mathsf{Q} \mathsf{x} \\ \text{s.t.} \quad & \mathsf{e}^\top \mathsf{x} = 1 . \end{aligned} \tag{1.3}$$

Due to the constraints of (1.3) satisfies linear independence constraint qualification at every feasible point, for any (1.3)-optimal solution, there exists $\lambda \in \mathbb{R}$ and $\mathsf{s} \in \mathbb{R}^n$ such that:

$$\begin{aligned} \mathsf{e}^\top \mathsf{x} &= 1 \\ \mathsf{Q}\mathsf{x} - (\lambda \mathsf{e} + s) &= \mathsf{o} \\ \mathsf{x}^\top \mathsf{s} &= 0 \\ \mathsf{x} \geq \mathsf{o}, \quad \mathsf{s} &\geq \mathsf{o}. \end{aligned}$$

Considering the first two constraints, we have $\lambda = \mathsf{x}^\top \mathsf{Q}\mathsf{x}$. Therefore, the problem (1.3) is equivalent to the following problem:

$$\begin{aligned} \min_{\mathsf{x},\mathsf{s},\lambda} \quad & \lambda \\ \text{s.t.} \quad & \mathsf{e}^\top \mathsf{x} = 1 \\ & \mathsf{Q}\mathsf{x} - (\lambda \mathsf{e} + s) = \mathsf{o} \\ & \mathsf{x}^\top \mathsf{s} = 0 \\ & \mathsf{x} \geq \mathsf{o}, \quad \mathsf{s} \geq \mathsf{o}\,. \end{aligned}$$

By a standard linearization procedure for the complementarity constraint, [GY21] proposed the following equivalent MILP:

$$\begin{aligned} l_2(\mathsf{Q}) := \min \quad & \lambda \\ s.t. \quad & \mathsf{Q}\mathsf{x} - \lambda \mathsf{e} - \mathsf{s} = 0 \\ & \mathsf{e}^\top \mathsf{x} = 1 \\ & x_i \leq y_i,\ i \in [1\!:\!n] \\ & s_i \leq M_i(1 - y_i),\ i \in [1\!:\!n] \\ & \mathsf{s}, \mathsf{x} \geq \mathsf{o} \\ & \mathsf{y} \in \{0,1\}^n\,. \end{aligned} \tag{COP2}$$

We choose the parameter M_i according to the following considerations: by the first two constraints in (COP2), it follows that

$$s_i = \mathsf{q}_i^\top \mathsf{x} - \lambda \leq \max_{j \in [1:n]} q_{ij} - \lambda\,, \tag{1.4}$$

where q_i and q_{ij} are the i-th column and ij-th entry of Q, respectively. Thus, we only need to estimate the lower bound of λ. Likewise, from the complementary slackness conditions $\mathsf{s}^\top \mathsf{x} = 0$ and (1.4) it follows that $\lambda = \mathsf{x}^\top \mathsf{Q}\mathsf{x} \geq v(\mathsf{Q})$. Therefore, lower bounds on $v(\mathsf{Q})$ can be used to bound λ from below. We tried two bounds on $v(\mathsf{Q})$ from [BLT08]:

$$\lambda \geq v(\mathsf{Q}) \geq v_1(\mathsf{Q}) := \min_{1 \leq i \leq j \leq n} q_{ij} + \frac{1}{\sum_{k=1}^n (1/(q_{kk} - \min_{1 \leq i \leq j \leq n} q_{ij}))}$$

or a tighter conic bound:

$$\lambda \geq v(\mathsf{Q}) \geq v_2(\mathsf{Q}) := \min \left\{ \langle \mathsf{Q}, \mathsf{X} \rangle : \langle \mathsf{E}, \mathsf{X} \rangle = 1, \mathsf{X} \in \mathcal{S}_+^n \cap \mathcal{N}^n \right\}.$$

Despite much higher computational cost for $v_2(\mathsf{Q})$, we found in a preliminary limited study no significant improvement of $v_2(\mathsf{Q})$ over $v_1(\mathsf{Q})$, so we decided to use $v_1(\mathsf{Q})$ throughout our experiments. Summarizing, we chose $M_i = \max_{j \in [1:n]} q_{ij} - v_1(\mathsf{Q})$ for use in the comparison described in the following.

1.4.3. Numerical comparison on two copositivity tests

In this section, we conduct a numerical comparison of the two copositivity tests using constructed instances. The experiment aims to demonstrate the strengths and weaknesses of each test when applied to matrices with different characteristics.

1.4.4. Instance generation

Based on the preliminary numerical study on (COP1) and (COP2), it is noteworthy that both methods perform poorly on matrices perturbed by the positive semidefinite cone. Motivated by this observation, we introduce a matrix parameter to quantify the distance between a matrix and the positive semidefinite cone, which can also be used to predict the difficulty of the copositivity test. For a given matrix $\mathsf{Q} \in \mathcal{S}^n$, by $\{\lambda_i : i \in [1\!:\!n]\}$ we denote the set of its eigenvalues. By I_+ and I_- we denote the index set of positive eigenvalues and nonpositive eigenvalues, respectively. By $\mathcal{S}_1^n$ we denote a subset of $\mathcal{S}^n$, which contains matrices with at least one positive eigenvalue. We then define a function $h : \mathcal{S}_1^n \to \mathbb{R}_+$ by

$$h(\mathsf{Q}) := \frac{\sum_{i \in I_-} |\lambda_i|}{\sum_{i \in I_+} \lambda_i}\,.$$

By $\mathfrak{A}_{c,n}$ we denote the test set

$$\mathfrak{A}_{c,n} := \{\mathsf{Q} \in \mathcal{S}_1^n : h(\mathsf{Q}) = c\}\,,$$

where c is a nonnegative parameter. Note that the extreme cases $c \searrow 0$ correspond to the easy case $\mathsf{Q} \in \mathcal{S}_+^n$ and likewise $c \nearrow \infty$ to the easy case $-\mathsf{Q} \in \mathcal{S}_+^n$, if $\sum_{i=1}^{n} |\lambda_i|$ remains bounded.

To generate $\mathsf{Q} \in \mathfrak{A}_{c,n}$, we use the eigendecomposition $\mathsf{Q} = \mathsf{P}^\top \Lambda \mathsf{P}$ where P is a randomly generated orthogonal matrix and Λ is a constructed diagonal matrix. Note that $\mathfrak{A}_{c,n}$ includes not only copositive matrices but also non-copositive ones. We separate the problem set $\mathfrak{A}_{c,n}$ according to copositivity: let $\mathfrak{A}_{c,n}^+ := \mathfrak{A}_{c,n} \cap \mathcal{COP}^n$ and $\mathfrak{A}_{c,n}^- := \mathfrak{A}_{c,n} \setminus \mathcal{COP}^n$. In fact, we cannot generate $\mathfrak{A}_{c,n}^+$ and $\mathfrak{A}_{c,n}^-$ directly. However, we are able to analyse the numerical results separately to avoid the bias towards hardness of copositive instances, as compared to non-copositive ones.

We next adopt the method [HP73] in which a procedure of constructing an extreme copositive matrix with entries from $\{-1, 0, 1\}$ was proposed. We call a copositive matrix $\mathsf{Q} \in \mathcal{COP}^n$ exceptional if $\mathsf{Q} \notin \mathcal{SPN}^n$. Exceptional copositive matrices are always of great interest from the aspect of copositivity test since there are polynomial-time approaches for testing membership of $\mathcal{SPN}^n$, unlike the membership of $\mathcal{COP}^n \setminus \mathcal{SPN}^n$. We know that the extreme copositive matrices are either exceptional or trivial (e.g. lying in $\mathcal{S}_+^n \cup \mathcal{N}^n$) and therefore filter the non-trivial matrices as instances.

Theorem 2. *[HP73] A given symmetric matrix* B *with entries from* $\{-1, 0, 1\}$ *is on an extreme ray of* $\mathcal{COP}^n$ *if and only if the following conditions hold*

- *the diagonal entries of* B *are all equal to 1,*
- $G_{-1}(\mathsf{B})$ *is connected and contains no triangle,*
- $G_1(\mathsf{B})$ *contains precisely those edges* (i, j) *where* i *and* j *are at distance 2 in* $G_{-1}(\mathsf{B})$*,*
- $L^*(\mathsf{B})$ *has no isolated vertices and no bipartite graph for a connected component,*

where $G_{-1}(\mathsf{B})$*,* $G_0(\mathsf{B})$*,* $G_1(\mathsf{B})$ *and* $L^*(\mathsf{B})$ *are four undirected graphs associated with* B*. In particular,* $G_{-1}(\mathsf{B})$ *(*$G_0(\mathsf{B})$*,* $G_1(\mathsf{B})$*) is a graph on* n *vertices with edges* (i, j) *if* $\mathsf{B}_{ij} = -1$ *(*$\mathsf{B}_{ij} = 0$*,* $\mathsf{B}_{ij} = 1$*) and* $L^*(\mathsf{B})$ *is a graph whose vertices are the edges of* $G_0(\mathsf{B})$ *with two vertices adjacent if the corresponding edges in* $G_0(\mathsf{B})$ *have a common vertex and the third edge of the associated triangle is in* $G_{-1}(\mathsf{B})$*.*

By $\mathfrak{B}_n$ we denote the test set of all $n \times n$ symmetric matrices satisfying the conditions of Theorem 2. Note that, by construction, $\mathfrak{B}_n$ contains copositive matrices only, and that the famous Horn and Hoffman-Pereira matrices [HP73] are all included in $\mathfrak{B}_n$.

1.4.5. Computational results

In the first experiment, we test (COP1) and (COP2) on the test sets $\mathfrak{A}^+_{0.02i,50}$ and $\mathfrak{A}^-_{0.02i,50}$. For each $i\in[1:10]$, we generate 500 instances in each test set. The box plots, bar charts and curves in Figure 1.1 below present the results on copositive part $\mathfrak{A}^+_{0.02i,50}$ for $i\in[1:10]$ while Figure 1.2 summarizes results from the non-copositive part.

In the copositive case, a result immediately from Figures 1.1a and 1.1c is that in general, the performance of both methods worsens when c decreases. From Figure 1.1a we notice that several instances cannot be solved within 200 seconds by (COP1) in sets 1-4. We count solved instances taking more than 50 seconds as hard as well for the percentages shown in Figure 1.1b. Approximately 16% of the cases in the set 1 are hard for method (COP1), meaning they need even more time by (COP1) than the worst case for method (COP2).

The situation improves quickly with increasing c. A similar pattern can also be seen for method (COP2) in Figure 1.1c, however, among the hardest test sets 1-4, all of the instances are resolved within 40 seconds, which is much better than (COP1). This effect is also confirmed by the average CPU time reported in Figure 1.1d. As c increases, the times for each instance become similar for both methods (COP1) and (COP2), almost coinciding when $c \geq 0.14$.

For $\mathfrak{A}^-_{0.02i,50}$, the behavior of (COP1) and (COP2) follows a pattern similar to that for $\mathfrak{A}^+_{0.02i,50}$, computational cost for both methods decreases with increasing c. However, it seems that (COP1) is more efficient to solve non-copositive instances as with copositive cases, taking less time across all sets. Moreover, we note that (COP1) takes less time than (COP2) if $i \geq 2$, as shown in Figure 1.2d. This effect is not seen in the copositive case.

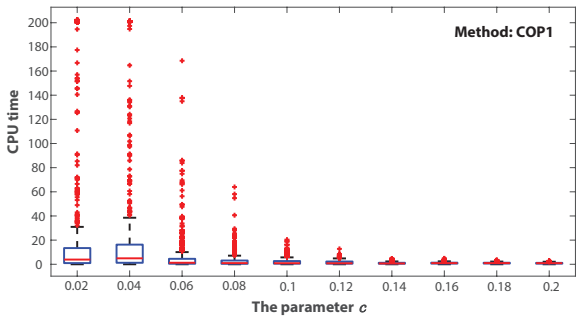
We conclude from the first experiment:

- Both methods perform worse for smaller values of c.
- For the copositive case, the performance of the method (COP2) is better than the method (COP1), while the gap reduces with increasing c.
- For the non-copositive case, performance of (COP1) seems better than that of (COP2) if $c \geq 0.04$.

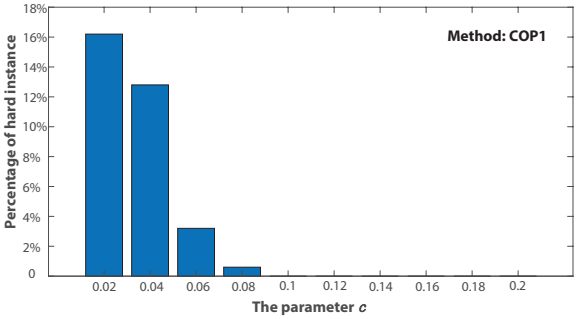
In the second experiment, we generate 500 test instances from $\mathcal{SPN}^n+\mathfrak{B}_n$ for $n \in \{50, 100, 150, 200\}$. For each generated instance Q we verified that $h(\mathsf{Q}) \geq 0.14$, which indicates that the hardness of the copositivity test from a close distance to the positive semidefinite cone is removed. Note that $\mathcal{SPN}^n + \mathfrak{B}_n \subseteq \mathcal{COP}^n$. In both Figure 1.3$a$ and Figure 1.3b, the horizontal axis represents the dimension of different problem sets while the vertical axis shows the distribution of CPU time in each set. Overall, a clearly discernible trend is that the CPU time taken by (COP2) grows nearly exponentially with the problem size. In contrast, (COP1) is not sensitive to problem dimension, taking slightly more time solving high dimension (200) instances compared to lower dimension instances. Average CPU times reported in Figure 1.3c seem to support this observation.

We conclude from the second experiment:

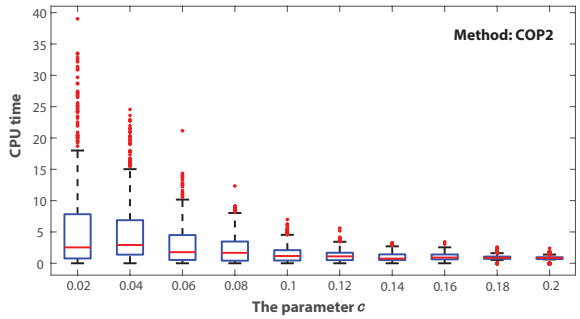
- Apart from hard instances, (COP1) performs much better than (COP2) when dealing with high dimension problems.
- Computational complexity of (COP1) grows slowly when the dimension of problems increases.
- Computational complexity of (COP2) is highly related to the dimension of problems.



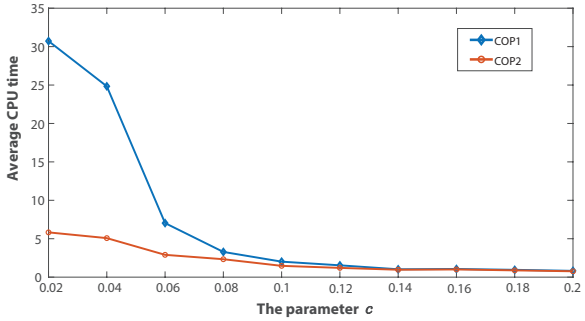
(a) CPU time by COP1



(b) Percentage of hard instances for COP1



(c) CPU timeby COP2



(d) CPU time taken by COP1

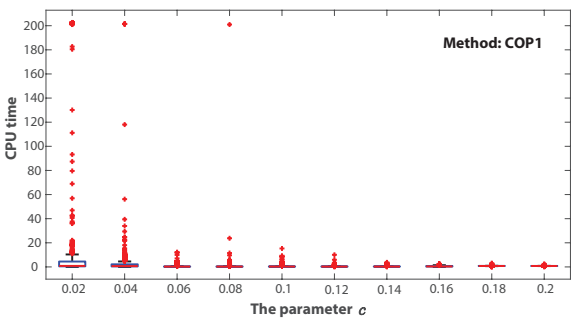
Figure 1.1.: Results on the copositive instances in $\mathfrak{A}^{+}_{0.02i,50}$

1.5. An improved variant of one copositivity test

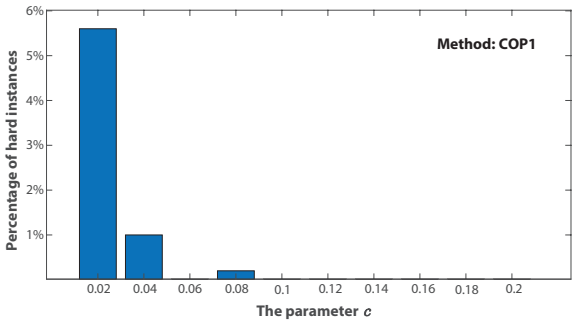
This section also contains material published in [Pen22]. As presented in the previous section, (COP1) can perform poorly in the cases when $h(\mathsf{Q}) \leq 0.04$, and this can be explained by the nature of the method (COP1). To check copositivity, (COP1) iterates over all principal submatrices of the targeted matrix Q (done under the binary variables y) until an almost copositive matrix [Dic19] is found. Given the extreme case that $\mathsf{Q} \in \mathcal{S}^n_+$, all of the principal submatrices of Q need to be checked, which is of exponential complexity. When the matrix parameter c is small, the distance between Q and $\mathcal{S}^n_+$ is small in the sense of the Frobenius norm. Therefore, by continuity of the condition for almost copositivity, we can infer that (COP1) needs to undergo a large number of principal submatrix checks to determine copositivity. However, compared with (COP2), (COP1) is much more competitive when tackling high-dimensional cases. To retain this nice property and to improve the poor performance of (COP1) when $h(\mathsf{Q})$ is small, we propose the following model:

$$
\begin{aligned}
l_3(\mathsf{Q}) := \min \quad & \lambda \\
s.t. \quad & \mathsf{q}_i^\top \mathsf{x} \leq \lambda + m_i(1-y_i), \quad i \in [1\!:\!n], \\
& \mathsf{q}_i^\top \mathsf{x} \geq \lambda, \quad i \in [1\!:\!n], \\
& \mathsf{y} \in \{0,1\}^n,\ \mathsf{o} \leq \mathsf{x} \leq \mathsf{y}, \\
& \mathsf{e}^\top \mathsf{y} \geq 1\,.
\end{aligned}
\tag{COP1$_+$}
$$

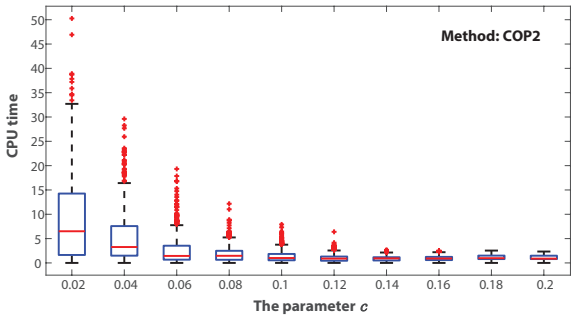
Note that the only difference between (COP1$_+$) and (COP1) lies in the additional inequalities $\mathsf{q}_i^\top \mathsf{x} \geq \lambda$, $i \in [1\!:\!n]$. The model (COP1$_+$) now is actually a MILP formulation for the Lemma 3.2(4)



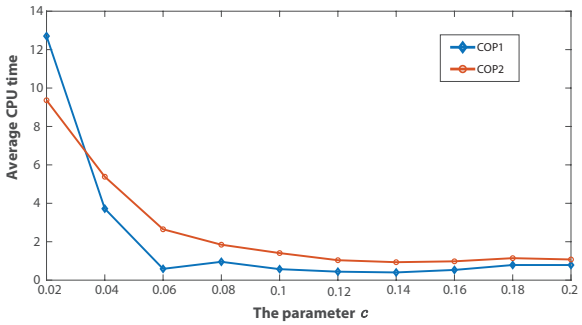
(a) CPU time by COP1



(b) Percentage of hard instances for COP1



(c) CPU time by COP2



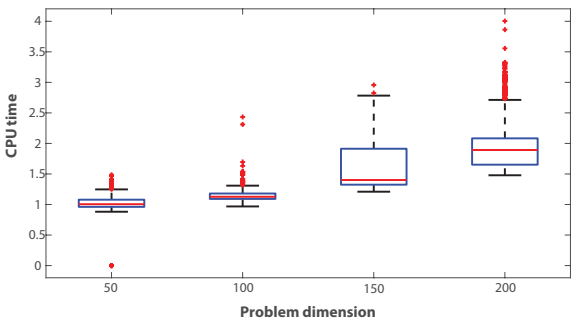
(d) Comparison on average CPU time

Figure 1.2.: Results on the non-copositive instances in $\mathfrak{A}^{-}_{0.02i,50}$

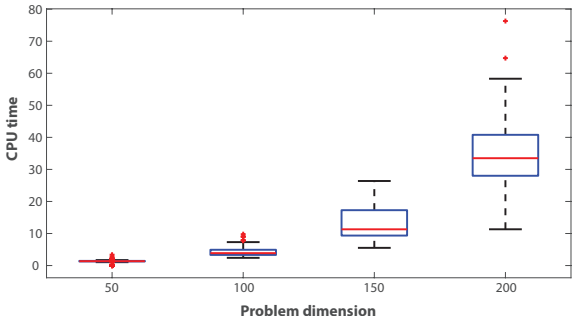
in [Dic19]. With these additional constraints, fewer principal submatrices satisfy the stronger condition for almost copositivity in the process of searching almost copositive matrices, which results in a reduction in the nodes used in solving MILP. The following Table 1.1 also confirms our speculation with roughly 80% reduction in CPU time for the test set when $c = 0.01$. Here $\mathfrak{A}_{0.01,50}$ and $\mathfrak{A}_{0.02,50}$ are two randomly generated test sets containing 500 instances respectively. It is worth noting that there are many instances that cannot be resolved within 200 seconds by (COP1) while (COP1$_+$) solved all of the instances within the time limit. One may also notice that the gap between average times comes closer as the parameter c increases from 0.01 to 0.02. The reason for it is that the test set contains more easy instances(noncopositive cases) when we increase the parameter c. However, as long as Q is copositive, the method (COP1) still needs much computational effort to determine its copositivity.

Table 1.1.: Comparison between (COP1) and (COP1$_+$) on the test sets with small c

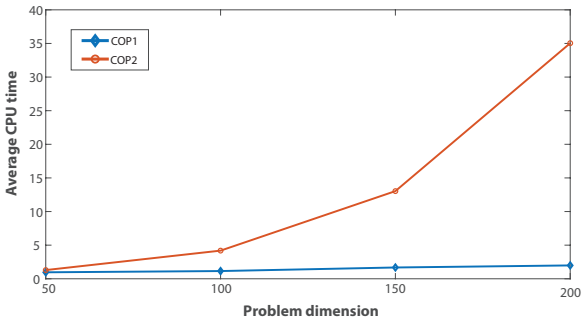
test sets	methods	average time	# hard instances	# unsolved instances
$\mathfrak{A}_{0.01,50}$	COP1	101.2	376	212
	COP1$_+$	21.2	83	0
$\mathfrak{A}_{0.02,50}$	COP1	30.3	89	62
	COP1$_+$	15.3	72	0



(a) CPU time by COP1



(b) CPU time by COP2



(c) Comparison on average CPU time

Figure 1.3.: Results of the second experiment

2. CP reformulation for QPCCs

In this chapter, we study (nonconvex) quadratic optimization problems (QPs) with complementarity constraints, establishing an exact completely positive reformulation under — apparently new — mild conditions involving only the constraints, not the objective. Moreover, we also give the conditions for strong conic duality between the obtained completely positive problem and its dual. The content of this chapter is mainly based on the published paper [BP23].

2.1. The problem

Many real-world applications can be modeled by (nonconvex) quadratic optimization problems with complementarity constraints (QPCCs) [LPR96]. Due to nonconvexity of the objective function and the complementarity constraints, QPCCs form an NP-hard class of problems. We study in this chapter the following problem where $\mathsf{Q} \in \mathbb{R}^{n\times n}$, $\mathsf{c} \in \mathbb{R}^n$, $\mathsf{x} \in \mathbb{R}^n$, $\mathsf{A} \in \mathbb{R}^{m\times n}$, $\mathsf{b} \in \mathbb{R}^m$ and $\mathsf{F} \in \mathbb{R}^{n\times n}_+$:

$$\begin{aligned} \min_{\mathsf{x}\in\mathbb{R}^n_+} \quad & \mathsf{x}^\top \mathsf{Q}\mathsf{x} + 2\mathsf{c}^\top \mathsf{x} \\ \text{s.t.} \quad & \mathsf{A}\mathsf{x} = \mathsf{b}\,,\ \mathsf{x}^\top \mathsf{F}\mathsf{x} = 0\,. \end{aligned} \tag{2.1}$$

Since all $F_{ij} \geq 0$, observe that $F_{ii} > 0$ would entail $x_i^2 = 0$ for all feasible x, so that this variable can be dropped. Hence we may and do assume without loss of generality that all $F_{ii} = 0$; likewise we may and do assume symmetry: $\mathsf{F}^\top = \mathsf{F}$ and $\mathsf{Q}^\top = \mathsf{Q}$ as well. Finally, we assume $\mathsf{F} \neq \mathsf{O}$.

Problem (2.1) actually covers a wide range of relevant problems, including, for instance, all mixed-binary quadratic optimization problems which are hard as well [DDM17, AOHE19]:

$$\begin{aligned} \min_{\mathsf{x}\in\mathbb{R}^n_+} \quad & \mathsf{x}^\top \mathsf{Q}\mathsf{x} + 2\mathsf{c}^\top \mathsf{x} \\ \text{s.t.} \quad & \mathsf{A}\mathsf{x} = \mathsf{b}\,,\ x_i \in \{0,1\}\,,\ \text{all } i \in I\,, \end{aligned}$$

with I a suitable subset of all coordinate indices.

One of the possible approaches to tackle problem (2.1) is conic relaxation. In particular, [Bur09] proposed a conic relaxation model, which links QPCCs to conic linear optimization problems over the completely positive cone, and which is exact (i.e., a conic reformulation) under some conditions. However, these conditions are not necessarily satisfied in many real-world applications including quadratic optimization problems with a true sparsity term $\|\mathsf{x}\|_0$ (see below). The purpose of this work is to generalize the results of [Bur09] such that the exact relaxation technique is applicable, for instance, to all sparse QPs.

2.2. CP reformulation for QPCCs

2.2.1. Conic formulation: Burer's key condition

As announced, we focus on QPCCs of the form

$$\begin{aligned} z_P^* := \inf_{\mathsf{x}\in\mathbb{R}^n_+} \quad & f(\mathsf{x}) := \mathsf{x}^\top \mathsf{Q}\mathsf{x} + 2\mathsf{c}^\top \mathsf{x} \\ s.t. \quad & \mathsf{a}_i^\top \mathsf{x} = b_i\,, \quad i\in[1:m]\,, \\ & \mathsf{x}^\top \mathsf{F}\mathsf{x} = 0\,, \end{aligned} \tag{P}$$

where $\mathsf{Q}\in\mathcal{S}^n$, $\mathsf{F}\in\mathcal{N}^n\setminus\{\mathsf{O}\}$ with zero diagonal, $\{\mathsf{c},\mathsf{a}_i\}\subset\mathbb{R}^n$ and $\mathsf{b}=[b_1,\ldots,b_m]^\top\in\mathbb{R}^m$. We collect $\mathsf{A}^\top=[\mathsf{a}_1,\ldots,\mathsf{a}_m]$ to get an $m\times n$ matrix A and represent the linear constraints in the form

$$\mathsf{x}\in L := \{\mathsf{x}\in\mathbb{R}^n_+ : \mathsf{A}\mathsf{x}=\mathsf{b}\}\,.$$

Problem (P) is an all-quadratic problem of special structure; remark that the case where $\{\mathsf{Q},\mathsf{F}\}\subset\mathcal{S}^n_+$ reduces to a traditional convex QP of the form

$$\inf_{\mathsf{x}\in\mathbb{R}^n_+}\{f(\mathsf{x}) : \mathsf{A}\mathsf{x}=\mathsf{b},\ \mathsf{F}\mathsf{x}=\mathsf{o}\}\ ,$$

which can be solved in polynomial time to arbitrary accuracy. In all other cases, this problem class includes NP-hard nonconvex subclasses. Problems of form (P) have been addressed in the seminal paper by Burer [Bur09] who related (P) to the following conic optimization problem:

$$\begin{aligned} z_C^* := \min \quad & g(\mathsf{Y}) := \langle \mathsf{Q},\mathsf{X}\rangle + 2\mathsf{c}^\top\mathsf{x} \\ s.t. \quad & \mathsf{a}_i^\top\mathsf{x} = b_i, \quad i\in[1:m]\,, \\ & \mathsf{a}_i^\top\mathsf{X}\mathsf{a}_i = b_i^2\,, \quad i\in[1:m]\,, \\ & \langle \mathsf{F},\mathsf{X}\rangle = 0\,, \\ & \mathsf{Y} := \begin{pmatrix} 1 & \mathsf{x}^\top \\ \mathsf{x} & \mathsf{X}\end{pmatrix} \in \mathcal{CP}^{n+1}\,. \end{aligned} \tag{C}$$

The following horizon cone of L plays a crucial role in this analysis:

$$L_\infty := \left\{\mathsf{x}\in\mathbb{R}^n_+ : \mathsf{A}\mathsf{x}=\mathsf{o}\right\}\ .$$

The set of indices belonging to variables that remain bounded over L,

$$B := \{i\in[1:n]:\ \text{there is an } M_i>0 \text{ such that } x_i\le M_i \text{ if } \mathsf{x}\in L\}\ , \tag{2.2}$$

is related to L_∞ by the implication $\mathsf{z}\in L_\infty \Longrightarrow \mathsf{z}_B=\mathsf{o}$. We need one more index set, namely

$$S_\mathsf{F} := \{i\in[1:n] : F_{ij}>0 \text{ for some } j\in[1:n]\}\ ,$$

the projection of the support of F,

$$\mathcal{E} := \{\{i,j\} : F_{ij}>0\}\ .$$

We will view the sparsity pattern of F as an undirected graph $\mathcal{G}_\mathsf{F} := (V,\mathcal{E})$ on the vertex set $V=[1:n]$ and edge set $\mathcal{E}$. Since by assumption $\mathsf{F}\in\mathcal{N}^n$, we have $\mathsf{x}^\top\mathsf{F}\mathsf{x}=0$ if and only if $x_ix_j=0$

for all $\{i,j\} \in \mathcal{E}$, so $\mathcal{G}_{\mathsf{F}}$ comprises all information on F relevant to (P). Thus we can always replace F by $\mathsf{E} := \sum_{\{i,j\}\in\mathcal{E}} \mathsf{e}_i \mathsf{e}_j^\top$ without changing the problem.

With these notational preliminaries, we easily can formulate the so-called key condition put forward by Burer [Bur09] in the analysis of QPCCs:

$$S_{\mathsf{F}} \subseteq B\,, \tag{K}$$

in other words, any variable x_j actually involved in the complementarity constraints, must remain bounded over the linear portion L of the feasible set. Under (K), Burer proved that the conic relaxation (C) of (P) is exact.

2.2.2. Relaxing the key condition: vertex covers

Unfortunately, condition (K) is not necessarily met in some relevant applications, and in other applications with a convex objective, the key condition is superfluous (even if the problem is not convex as may happen for $\mathsf{F} \notin \mathcal{S}^n_+$). In this section, we will propose less restrictive conditions ensuring exactness which can be successfully applied to several optimization problems, including sparse least-squares regression with soft or hard sparsity constraints, a hot topic in data science.

Given a graph $\mathcal{G} = (V,\mathcal{E})$, recall that a subset $C \subseteq V$ is said to be a vertex cover for $\mathcal{G}$ if $\{i,j\} \cap C \neq \emptyset$ for all $\{i,j\} \in \mathcal{E}$ (this holds in particular if $\{C, V \setminus C\}$ is a bipartition of $\mathcal{G}$ in the sense that $(i,j) \in C \times (V \setminus C) \cup (V \setminus C) \times C$ for any $\{i,j\} \in \mathcal{E}$, but more general vertex covers C may also have insider edges, $C \times C \cap \mathcal{E} \neq \emptyset$). Obviously, V itself is a vertex cover but it may be too large for our purposes, as we aim to replace Burer's S_{F} with a (typically smaller) vertex cover C for $\mathcal{G}_{\mathsf{F}}$.

The *relaxed key condition* reads as follows, using notation (2.2):

$$\text{There is a vertex cover } C \subseteq B \text{ for } \mathcal{G}_{\mathsf{F}} \text{ and a point } \hat{\mathsf{x}} \in L \text{ with } \hat{\mathsf{x}}_C = \mathsf{o}\,, \tag{R}$$

or equivalently, there is a vertex cover C for $\mathcal{G}_{\mathsf{F}}$ and some (P)-feasible $\hat{\mathsf{x}}$ such that for all $j \in C$ there exists an $M_j > 0$ such that

$$\hat{x}_j = 0 \quad \text{and} \quad x_j \leq M_j \text{ for all } \mathsf{x} \in L\,.$$

Obviously, the conditions in (R) are met more easily if C is smaller, so the system $\mathfrak{C}(\mathsf{F})$ in which we collect all minimal vertex covers[1] for $\mathcal{G}_{\mathsf{F}}$ is of particular relevance. But to keep our approach more transparent, we will focus on one vertex cover C first (which is sufficient for application to sparse regression) and later on, extend the following conditions for future applications to more general QPCCs. First let us note that indeed (R) relaxes (K):

Proposition 3. *Assume that (P) is feasible, so that there exists a point $\hat{\mathsf{x}} \in L$ such that $\hat{\mathsf{x}}^\top \mathsf{F} \hat{\mathsf{x}} = 0$. Then* (K) *implies* (R).

Proof. We show that $C := \{j \in S_{\mathsf{F}} : \hat{x}_j = 0\}$ and $\hat{\mathsf{x}} \in L$ satisfy (R) under condition (K). Indeed, by definition $\hat{\mathsf{x}}_C = \mathsf{o}$ and $C \subseteq S_{\mathsf{F}}$. Furthermore, $\hat{\mathsf{x}}^\top \mathsf{F} \hat{\mathsf{x}} = 0$ implies $\hat{x}_i \hat{x}_j = 0$ for all $\{i,j\} \in \mathcal{E}$, so that $\{i,j\} \cap C \neq \emptyset$, hence C is a vertex cover for $\mathcal{G}_{\mathsf{F}}$. □

[1] note that we do not suggest to determine $\mathfrak{C}(\mathsf{F})$ which may be exponential in size (as with the maximum-clique problem) but rather propose to exploit structural knowledge of the problem at hand.

In the case of a convex objective, condition (K) and even condition (R) are unnecessary. Actually, we only need the conclusion that Q is copositive on the cone

$$L_\infty \cap F := \left\{ \mathsf{z} \in \mathbb{R}^n_+ : \mathsf{Az} = \mathsf{o} \text{ and } \mathsf{z}^\top \mathsf{Fz} = 0 \right\},$$

the intersection of the horizon cone of L with the complementarity set

$$F := \left\{ \mathsf{x} \in \mathbb{R}^n_+ : \mathsf{x}^\top \mathsf{Fx} = 0 \right\}.$$

Condition (R) guarantees that $L_\infty \cap F = L_\infty$ and the mentioned copositivity property of Q, which is also implied by the convexity of the objective. Essential arguments are summarized in the following.

Lemma 4. *Assume $z^*_P > -\infty$, so that problem* (P) *is bounded, and that* (R) *holds. Then $L_\infty \subseteq F$, and for any $\mathsf{z} \in L_\infty$ we have $\mathsf{z}^\top \mathsf{Qz} \geq 0$. The latter property holds if $\mathsf{Q} \in \mathcal{COP}^n$ (e.g., because Q is positive-semidefinite).*

Proof. If condition (R) holds for some vertex cover $C \subseteq V$ of $\mathcal{G}_\mathsf{F}$, pick an arbitrary $\mathsf{z} \in L_\infty$ and select $\hat{\mathsf{x}} \in L$ with $\hat{\mathsf{x}}_C = \mathsf{o}$. Then for all $t \geq 0$, we have $\mathsf{x}(t) = \hat{\mathsf{x}} + t\,\mathsf{z} \in L$ which implies, by $C \subseteq B$ from (R), that $\mathsf{z}_C = \mathsf{o}$, which in turn implies $\mathsf{x}(t)_C = \mathsf{o}$ for all $t \geq 0$. So $x_i(t)x_j(t) = 0$ for all $j \in C$ (and all $i \in V$) which yields $\mathsf{x}(t)^\top \mathsf{Fx}(t) = 0$ since C is a vertex cover for $\mathcal{G}_\mathsf{F}$. Hence $\mathsf{x}(t) \in F$ and in a similar fashion, we derive $\mathsf{z} \in F$. We conclude $L_\infty \subseteq F$ and furthermore that $\mathsf{x}(t)$ is (P)-feasible, hence

$$-\infty < z^*_P \leq f(\mathsf{x}(t)) = f(\hat{\mathsf{x}}) + 2t\,\mathsf{z}^\top(\mathsf{Q}\hat{\mathsf{x}} + \mathsf{c}) + t^2\,\mathsf{z}^\top \mathsf{Qz} \quad \text{for all } t \geq 0\,,$$

which implies $\mathsf{z}^\top \mathsf{Qz} \geq 0$. This implication is trivial if $\mathsf{Q} \in \mathcal{COP}^n$, since $L_\infty \subseteq \mathbb{R}^n_+$ by definition. □

Note: if $\mathsf{z}^\top \mathsf{Qz} = 0$, we could infer, by the same reasoning, that $\mathsf{z}^\top \nabla f(\hat{\mathsf{x}}) = 2\mathsf{z}^\top(\mathsf{Q}\hat{\mathsf{x}} + \mathsf{c}) \geq 0$, i.e., that z could not be a strict descent direction for f at $\hat{\mathsf{x}}$. But we will not explore this first-order information further here.

Now we formulate a variant applicable to other QPCCs:

Lemma 5. *Assume $z^*_P > -\infty$, so that problem* (P) *is bounded. Suppose that for any $C \in \mathfrak{C}(\mathsf{F})$ with nontrivial cone $\{\mathsf{x} \in L_\infty : \mathsf{x}_C = \mathsf{o}\}$, there exists a point $\hat{\mathsf{x}} \in L$ such that $\hat{\mathsf{x}}_C = \mathsf{o}$. Then Q is $(L_\infty \cap F)$-copositive.*

Proof. First observe that

$$F \subseteq \bigcup_{C \in \mathfrak{C}(\mathsf{F})} \{\mathsf{z} \in \mathbb{R}^n_+ : \mathsf{z}_C = \mathsf{o}\}\,.$$

Indeed, for any $\mathsf{z} \in F$ consider $C_\mathsf{z} = \{i \in V : z_i = 0\} \neq \emptyset$ (due to $\mathsf{F} \neq \mathsf{O}$), which obviously is a vertex cover for $\mathcal{G}_\mathsf{F}$. Clearly there is an $C \in \mathfrak{C}(\mathsf{F})$ with $C \subseteq C_\mathsf{z}$, so $\mathsf{z}_C = \mathsf{o}$. Then the result follows by arguments similar to those employed in the previous proof. □

The following examples shall demonstrate wider applicability of our proposed conditions. But for the readers' convenience let us start with a counterexample already presented in [BMP16] which shows that some conditions are needed to avoid a positive, or even infinite, conic relaxation gap:

Example 1.

$$\begin{aligned} &\min -x_2^2 \\ s.t. \quad & x_1 + x_2 - x_3 = 3\,, \\ & x_1 - x_4 = 2\,, \\ & x_1 x_2 = 0\,, \\ & \mathsf{x} \in \mathbb{R}^4_+\,. \end{aligned} \tag{2.3}$$

As argued in [BMP16, Ex.1], the relaxation (C) is unbounded below while the optimal value of (P) is zero. As can be seen easily, conditions (K) and (R) are violated, and as well the one of Lemma 5: all $\mathsf{x} \in L$ *must have* $x_1 \geq 2 > 0$*, so* $\mathsf{x}_C \neq \mathsf{o}$ *for* $C = \{1\} \in \mathfrak{C}(\mathsf{F})$ *while* L_∞ *contains, e.g.,* $(0,1,1,0)^\top$*.*

Example 2.

$$\begin{aligned} x_1 - x_2 - x_3 + x_4 &= 1\,, \\ x_1 + x_2 &= 1\,, \\ x_1 x_2 + x_1 x_3 &= 0\,, \\ \mathsf{x} &\in \mathbb{R}^4_+\,. \end{aligned} \tag{2.4}$$

Note that the second constraint implies boundedness of x_1 *and* x_2*, so we have* $B = \{1,2\}$*. However, since the set* S *of variables involved in the complementary constraint is* $S = \{1,2,3\} \supset \{1,2\}$*, condition* (K) *is violated. On the other hand, the class of minimal vertex covers of* (2.4) *is* $\mathfrak{C}(\mathsf{F}) = \{\{1\},\{2,3\}\}$*, with a vertex cover* $C = \{1\} \subset B$*. Moreover,* $(0,1,1,3)^\top$ *is* (2.4)*-feasible and therefore condition* (R) *is met. Also, the condition of Lemma 5 holds, as* $C = \{2,3\}$ *yields a trivial cone* $\{\mathsf{z} \in L_\infty : \mathsf{z}_C = \mathsf{o}\}$*.*

Example 3.

$$\begin{aligned} x_1 - x_2 - x_3 + x_4 - x_5 &= 1\,, \\ x_1 + x_2 &= 1\,, \\ x_2 x_3 + x_3 x_4 &= 0\,, \\ \mathsf{x} &\in \mathbb{R}^5_+\,. \end{aligned} \tag{2.5}$$

As in (2.4)*, we have* $B = \{1,2\}$*. Again, condition* (K) *is violated since* $S = \{2,3,4\}$ *is not contained in* B*. The class of minimal vertex covers of* (2.5) *is* $\mathfrak{C}(\mathsf{F}) = \{\{3\},\{2,4\}\}$*. None of them is contained in* B*, so condition* (R) *is violated. For* $C = \{3\}$*, the cone*

$$\{\mathsf{x} \in L_\infty : \mathsf{x}_C = \mathsf{o}\} = \{\mathsf{x} \in \mathbb{R}^5_+ : x_4 = x_5, x_1 = x_2 = x_3 = 0\}$$

is nontrivial, and the point $\hat{\mathsf{x}} = (1,0,0,0,0)^\top \in L$ *satisfies* $\hat{\mathsf{x}}_C = \mathsf{o}$*. But for* $C = \{2,4\}$*, the cone*

$$\{\mathsf{x} \in L_\infty : \mathsf{x}_C = \mathsf{o}\} = \{\mathsf{x} \in \mathbb{R}^5_+ : x_3 + x_5 = 0, x_1 = x_2 = x_4 = 0\} = \{\mathsf{o}\}\,.$$

Therefore the system satisfies the condition in Lemma 5.

Example 4.

$$\begin{aligned} x_1 + x_4 - x_5 + x_6 &= 5\,, \\ x_2 + x_3 - x_5 + x_6 &= 1\,, \\ x_1 + x_2 &= 1\,, \\ x_1 x_3 + x_1 x_4 &= 0\,, \\ \mathsf{x} &\in \mathbb{R}^6_+\,. \end{aligned} \tag{2.6}$$

As in (2.4)*, we have* $B = \{1, 2\}$*. Again, condition* (K) *is violated since* $S = \{1, 3, 4\}$ *is not contained in* B*. On the other hand, the class of minimal vertex covers of* (2.6) *is* $\mathfrak{C}(\mathsf{F}) = \{\{1\}, \{3, 4\}\}$*, with a vertex cover* $C = \{1\} \subset B$*. Moreover,* $(0, 1, 0, 5, 2, 2)^\top$ *is* (2.6)*-feasible and therefore condition* (R) *is met. For* $C = \{3, 4\}$*, the cone*

$$\{\mathsf{x} \in L_\infty : \mathsf{x}_C = \mathsf{o}\} = \{\mathsf{x} \in \mathbb{R}^6_+ : x_5 = x_6, x_1 = x_2 = x_3 = x_4 = 0\}$$

is nontrivial. However, the set $\{\mathsf{x} \in L : \mathsf{x}_C = \mathsf{o}\}$ *is empty now, because* $\mathsf{x}_C = \mathsf{o}$ *implies a contradiction, looking at the sum of the first two constraints and the third one, in conjunction with the sign constraints on* x_1 *and* x_2*. Therefore, the condition of Lemma 5 is violated.*

While (K) implies (R), it does not imply the condition of Lemma 5; we give the following example:

Example 5.

$$\begin{aligned} x_3 - x_4 + x_5 &= 2\,, \\ x_2 - x_4 + x_5 &= 1\,, \\ x_1 + x_2 + x_3 &= 2\,, \\ x_1x_2 + x_1x_3 &= 0\,, \\ \mathsf{x} &\in \mathbb{R}^5_+\,. \end{aligned} \tag{2.7}$$

Since we have $S = \{1, 2, 3\} = B$*, condition* (K) *holds. On the other hand, the class of minimal vertex covers of* (2.7) *is* $\mathfrak{C}(\mathsf{F}) = \{\{1\}, \{2, 3\}\}$*. For* $C = \{2, 3\}$*, the cone*

$$\{\mathsf{x} \in L_\infty : \mathsf{x}_C = \mathsf{o}\} = \{\mathsf{x} \in \mathbb{R}^5_+ : x_4 = x_5, x_1 = x_2 = x_3 = 0\}$$

is nontrivial. However, the set $\{\mathsf{x} \in L : \mathsf{x}_C = \mathsf{o}\}$ *is empty now because of the contradiction between the first two constraints. Therefore, the condition of Lemma 5 is violated.*

As a final note, it is easy to construct examples for which both (K) and the condition of Lemma 5 are satisfied [BMP16].

Now we are in a position to state and prove the following result which generalizes the one in [Bur09]. Although the main arguments are well known by now, we repeat them here for the readers' convenience. Note that even without any assumptions, the first part of the following proof also shows that (P) is feasible if (C) is feasible.

Theorem 6. *Suppose that* Q *is* $(L_\infty \cap F)$*-copositive in the sense of Lemmas 4, 5. Then the QPCC problem (P) is equivalent to the conic problem* (C)*.*

Proof. It is easy to see that any (P)-feasible $\mathsf{x} \in \mathbb{R}^n_+$ gives rise to a (C)-feasible point $\mathsf{Y} = \mathsf{u}\mathsf{u}^\top \in \mathcal{CP}^{n+1}$ of rank one with $\mathsf{u}^\top = [1, \mathsf{x}^\top]$ and same objective value $f(\mathsf{x}) = \langle \mathsf{Q}, \mathsf{x}\mathsf{x}^\top \rangle + 2\mathsf{c}^\top\mathsf{x} = g(\mathsf{Y})$. Therefore (C) is a conic relaxation of (P) and $z^*_C \le z^*_P$. We proceed to show equivalence by establishing the reverse inequality. To this end, take a (C)-feasible and optimal Y (or approximately optimal, in case of non-attainability for either of the problems (P) or (C); we will address primal attainability later). Then consider the following decomposition for this completely positive matrix:

$$\mathsf{Y} = \begin{pmatrix} 1 & \mathsf{x}^\top \\ \mathsf{x} & \mathsf{X} \end{pmatrix} = \sum_{k \in K} \begin{pmatrix} \lambda_k \\ \mathsf{z}_k \end{pmatrix} \begin{pmatrix} \lambda_k \\ \mathsf{z}_k \end{pmatrix}^\top, \quad \lambda_k \in \mathbb{R}_+,\ \mathsf{z}_k \in \mathbb{R}^n_+,\ k \in K\,, \tag{2.8}$$

with $g(\mathsf{Y}) = \langle \mathsf{Q}, \mathsf{X}\rangle + 2\mathsf{c}^\top \mathsf{x} \approx z_C^*$. Denote by $K_0 := \{k \in K : \lambda_k = 0\}$ and $K_+ := \{k \in K : \lambda_k > 0\}$ respectively. The following equalities hold directly by the decomposition (2.8) and the constraints in (C):

$$1 = \sum_{k\in K} \lambda_k^2\,, \quad (\text{ so } K_+ \neq \emptyset)\,, \tag{2.9}$$

$$b_i = a_i^\top \mathsf{x} = \sum_{k\in K} \lambda_k \mathsf{a}_i^\top \mathsf{z}_k, \quad i\in[1\!:\!m]\,, \tag{2.10}$$

$$b_i^2 = \mathsf{a}_i^\top \mathsf{X}\mathsf{a}_i = \sum_{k\in K} (\mathsf{a}_i^\top \mathsf{z}_k)^2, \quad i\in[1\!:\!m]\,. \tag{2.11}$$

Combining the equalities (2.9), (2.10) and (2.11), we have for any fixed $i\in[1\!:\!m]$,

$$\sum_{k\in K} \lambda_k^2 \cdot \sum_{k\in K} (\mathsf{a}_i^\top \mathsf{z}_k)^2 = 1 \cdot b_i^2 = b_i^2 = \left(\sum_{k\in K} \lambda_k \mathsf{a}_i^\top \mathsf{z}_k\right)^2 .$$

By the condition of equality case of the Cauchy-Schwarz inequality, there exists $\beta_i \in \mathbb{R}$ such that

$$\mathsf{a}_i^\top \mathsf{z}_k = \beta_i \lambda_k \quad \text{for all } k \in K\,. \tag{2.12}$$

Multiplying by λ_k and summing relations (2.12) across k and by (2.9) and (2.10), we obtain

$$\beta_i = b_i \quad \text{for all } i\in[1\!:\!m]\,. \tag{2.13}$$

Moreover, by (C)-feasibility of Y we have

$$0 = \langle \mathsf{F}, \mathsf{X}\rangle = \sum_{k\in K} \mathsf{z}_k^\top \mathsf{F}\mathsf{z}_k \Longrightarrow \mathsf{z}_k^\top \mathsf{F}\mathsf{z}_k = 0\,, \text{ for all } k \in K\,.$$

Therefore, by taking $\mathsf{x}_k := \frac{1}{\lambda_k}\mathsf{z}_k$ if $\lambda_k > 0$, we get $\mathsf{x}_k \in L \cap F$ for all $k \in K_+$ while for $\lambda_k = 0$ we obtain $\mathsf{A}\mathsf{z}_k = \mathsf{o}$, i.e., $\mathsf{z}_k \in L_\infty \cap F$ and therefore $\mathsf{z}_k^\top \mathsf{Q}\mathsf{z}_k \geq 0$ for all $k \in K_0$. We obtain that all x_k are (P)-feasible, so that

$$z_P^* \leq z_+^* := \min\{f(\mathsf{x}_k) : k \in K_+\} \leq z_C^*\,,$$

where the last inequality follows from

$$\begin{aligned} z_+^* &= \sum_{k\in K_+} (\lambda_k)^2 z_+^* \leq \sum_{k\in K_+} (\lambda_k)^2 f(\mathsf{x}_k) \\ &= \sum_{k\in K_+} \left[\mathsf{z}_k^\top \mathsf{Q}\mathsf{z}_k + \lambda_k \mathsf{c}^\top \mathsf{z}_k\right] \leq \sum_{k\in K_+} \left[\mathsf{z}_k^\top \mathsf{Q}\mathsf{z}_k + \lambda_k \mathsf{c}^\top \mathsf{z}_k\right] + \sum_{k\in K_0} \mathsf{z}_k^\top \mathsf{Q}\mathsf{z}_k \\ &= \sum_{k\in K} \left[\mathsf{z}_k^\top \mathsf{Q}\mathsf{z}_k + \lambda_k \mathsf{c}^\top \mathsf{z}_k\right] = g(\mathsf{Y}) \approx z_C^*\,, \end{aligned}$$

which establishes the desired inequality $z_P^* \leq z_C^*$ and moreover that any x^k realizing the minimum in z_+^* provides an (approximate) optimal solution to (P). □

Note: since the feasible set of (P) is a finite union of polyhedra (if not empty), the Frank-Wolfe theorem guarantees attainability of z_P^* if this value is finite. Slater's condition (strict feasibility of the conic dual below) would ensure attainability of z_C^*, but above equivalence ensures this automatically: indeed, $\mathsf{Y}^* = \mathsf{u}^*(\mathsf{u}^*)^\top$ solves (C) if $(\mathsf{u}^*)^\top = [1, (\mathsf{x}^*)^\top]$ and x^* solves (P).

2.2.3. A simplification of (C) and its dual

We show a possible simplification of (C) which is initially proposed in [Bur09], reducing the cone in the feasible set from $\mathcal{CP}^{n+1}$ to $\mathcal{CP}^n$, and which works under

Condition 1. *There exists* $\mathsf{y} \in \mathbb{R}^m_+$ *such that*

$$\mathsf{d} := \mathsf{A}^\top \mathsf{y} = \sum_{i=1}^m y_i \mathsf{a}_i \in \mathbb{R}^n_+ \quad \textit{and} \quad \mathsf{b}^\top \mathsf{y} = 1\,.$$

Condition 1 is certainly met if $\mathsf{a}_i \in \mathbb{R}^m_+$ and $b_i > 0$ for at least one $i \in [1\!:\!m]$, taking $\mathsf{y} = \frac{1}{b_i}\mathsf{e}_i \in \mathbb{R}^m_+$.

It is clear that the constraints $\mathsf{Ax} = \mathsf{b}$ imply $\mathsf{d}^\top \mathsf{x} = \mathsf{y}^\top \mathsf{b} = 1$. From the decomposition (2.8) and as in the proof of Theorem 6 we have $\mathsf{d}^\top \mathsf{z}_k = \mathsf{y}^\top \mathsf{A} \mathsf{z}_k = \mathsf{y}^\top \mathsf{o} = 0$ for all $k \in K_0$ and

$$\mathsf{x} = \sum_{k \in K_+} \lambda_k \mathsf{z}_k = \sum_{k \in K_+} \lambda_k^2 \mathsf{x}_k$$

with $\mathsf{A}\mathsf{x}_k = \mathsf{b}$, entailing $\mathsf{d}^\top \mathsf{x}_k = 1$ for all $k \in K_+$, yielding

$$\mathsf{X}\mathsf{d} = \sum_{k \in K} \mathsf{z}_k \mathsf{z}_k^\top \mathsf{d} = \sum_{k \in K_+} \lambda_k^2 \mathsf{x}_k \mathsf{x}_k^\top \mathsf{d} + \sum_{k \in K_0} \mathsf{z}_k \mathsf{z}_k^\top \mathsf{d} = \sum_{k \in K_+} \lambda_k^2 \mathsf{x}_k + \mathsf{o} = \mathsf{x}\,,$$

which allows us to rewrite (C) as:

$$\begin{array}{ll} \inf & \langle \mathsf{Q} + (\mathsf{d}\mathsf{c}^\top + \mathsf{c}\mathsf{d}^\top), \mathsf{X} \rangle = \langle \mathsf{Q}, \mathsf{X} \rangle + 2\mathsf{c}^\top \mathsf{X}\mathsf{d} \\ s.t. & \langle \mathsf{d}\mathsf{a}_i^\top + \mathsf{a}_i \mathsf{d}^\top, \mathsf{X} \rangle = 2\mathsf{a}_i^\top \mathsf{X}\mathsf{d} = 2b_i\,, \quad i \in [1\!:\!m]\,, \\ & \langle \mathsf{a}_i \mathsf{a}_i^\top, \mathsf{X} \rangle = \mathsf{a}_i^\top \mathsf{X} \mathsf{a}_i = b_i^2\,, \quad i \in [1\!:\!m]\,, \\ & \langle \mathsf{F}, \mathsf{X} \rangle = 0\,, \\ & \mathsf{X} \in \mathcal{CP}^n\,. \end{array} \tag{C_{sim}}$$

We abbreviate $\mathsf{G} := \mathsf{Q} + (\mathsf{d}\mathsf{c}^\top + \mathsf{c}\mathsf{d}^\top)$ and $\mathsf{A}_i := \frac{1}{2}(\mathsf{d}\mathsf{a}_i^\top + \mathsf{a}_i \mathsf{d}^\top)$ as well as $\bar{\mathsf{A}}_i := \mathsf{a}_i \mathsf{a}_i^\top$ for all $i \in [1\!:\!m]$ to rewrite (C_{sim}) in a simpler form, recalling that $\langle \mathsf{F}, \mathsf{X} \rangle = 0 \Leftrightarrow \langle \mathsf{E}, \mathsf{X} \rangle = 0$:

$$\begin{array}{ll} \inf & \langle \mathsf{G}, \mathsf{X} \rangle \\ s.t. & \langle \mathsf{A}_i, \mathsf{X} \rangle = b_i\,, \quad i \in [1\!:\!m]\,, \\ & \langle \bar{\mathsf{A}}_i, \mathsf{X} \rangle = b_i^2\,, \quad i \in [1\!:\!m]\,, \\ & \langle \mathsf{E}, \mathsf{X} \rangle = 0\,, \\ & \mathsf{X} \in \mathcal{CP}^n\,. \end{array} \tag{CP}$$

By introducing dual variables $\mathsf{r} \in \mathbb{R}^m$ for constraints $\langle \mathsf{A}_i, \mathsf{X} \rangle = b_i$, $\mathsf{t} \in \mathbb{R}^m$ for constraints $\langle \bar{\mathsf{A}}_i, \mathsf{X} \rangle = b_i^2$ and $h \in \mathbb{R}$ for the constraint $\langle \mathsf{E}, \mathsf{X} \rangle = 0$, the dual problem of (CP) can be obtained:

$$\begin{array}{ll} \sup & \displaystyle\sum_{i=1}^m (r_i b_i + t_i b_i^2) \\ s.t. & \mathsf{S}(\mathsf{r}, \mathsf{t}, h) := \mathsf{G} - \displaystyle\sum_{i=1}^m (r_i \mathsf{A}_i + t_i \bar{\mathsf{A}}_i) - h\mathsf{E} \in \mathcal{COP}^n \\ & (\mathsf{r}, \mathsf{t}, h) \in \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}\,. \end{array} \tag{CD}$$

While we have, by preceding arguments under the assumption of $(L_\infty \cap F)$-copositivity of Q and $z_P^* \in \mathbb{R}$,

$$z_P^* = z_C^* = \inf\{\langle \mathsf{G}, \mathsf{X}\rangle : \mathsf{X} \text{ is } \text{ (CP)-feasible}\}$$

as the optimal value for (CP), we denote by

$$z_D^* := \sup\left\{\sum_{i=1}^m (r_i b_i + t_i b_i^2) : \mathsf{S}(\mathsf{r}, \mathsf{t}, h) \in \mathcal{COP}^n\right\}$$

the optimal value for (CD).

We will now establish sufficient conditions for strict feasibility of (CD), which means that Slater's condition holds for (CD). The condition seems quite natural as it is a sharpening of the conclusion of Lemma 4. It is met if Q is positive-definite, and for general Q if L is bounded (implying $L_\infty = \{\mathsf{o}\}$), by default:

Theorem 7. *Suppose that $z_P^* \in \mathbb{R}$ and that Q is strictly L_∞-copositive: $\mathsf{z}^\top \mathsf{Q}\mathsf{z} > 0$ if $\mathsf{z} \in L_\infty \setminus \{\mathsf{o}\}$. Then*

$$z_C^* = z_D^* \quad \text{and the optimal value } z_C^* \text{ is attained.}$$

Proof. First observe that by definition of $\mathsf{d} = \mathsf{A}^\top \mathsf{y}$, we have $\mathsf{z}^\top \mathsf{d} = (\mathsf{A}\mathsf{z})^\top \mathsf{d} = 0$ for all $\mathsf{z} \in L_\infty$, hence $\mathsf{z}^\top \mathsf{G}\mathsf{z} = \mathsf{z}^\top \mathsf{Q}\mathsf{z}$ on L_∞. Now, under the assumptions, there is a constant $\sigma > 0$ such that

$$\mathsf{z}^\top \mathsf{G}\mathsf{z} = \mathsf{z}^\top \mathsf{Q}\mathsf{z} \geq 2\sigma \quad \text{for all } \mathsf{z} \in \Delta_\infty := L_\infty \cap \Delta^n\,.$$

By continuity, there is a $\delta > 0$ such that also in a small relative neighbourhood

$$U_\delta := \{\mathsf{x} \in \Delta^n : \text{dist}\,(\mathsf{x}, \Delta_\infty) < \delta\}$$

we have $\mathsf{z}^\top \mathsf{G}\mathsf{z} \geq \sigma > 0$. The continuous function $\|\mathsf{A}\mathsf{x}\|^2$ takes its positive minimum $\mu > 0$ on the compact set $\Delta^n \setminus U_\delta$ which is disjoint from L_∞; furthermore, we have that $\gamma := \min_{\mathsf{x} \in \Delta^n} \mathsf{z}^\top \mathsf{G}\mathsf{z} \in \mathbb{R}$. Choosing $\alpha := \max\left\{0, \frac{\sigma - \gamma}{\mu}\right\}$ gives therefore a strictly copositive matrix $\mathsf{S}(\mathsf{o}, -\alpha\,\mathsf{e}, 0)$ with a quadratic form $\mathsf{z}^\top \mathsf{G}\mathsf{z} + \alpha\,\|\mathsf{A}\mathsf{z}\|^2$ taking values not smaller than σ over U_δ but also, by choice of α, not smaller than σ over $\Delta^n \setminus U_\delta$. Hence Slater's condition holds for (CD) and the claim follows. □

2.3. QPCC and CP formulations for quadratic problems under sparsity constraints

In this section, we discuss two models for pursuing sparsity in quadratic optimization problems with the direct use of $\|\cdot\|_0$ put forward by [BKS16, FMP+18]. Both are based on the following idea: negating the binary variables $|\text{sign}\,(y_i)| \in \{0, 1\}$, we arrive again at binary variables $u_i := 1 - |\text{sign}\,(y_i)| \in \{0, 1\}$ which satisfy $u_i |y_i| = 0$ for all i. As no negative variables appear in this complementarity relation, we can proceed towards a QPCC with a single homogeneous bilinear equality constraint. An additional interesting feature of this approach is that binarity of u_i is not needed here explicitly as a constraint, only the upper bounds $u_i \leq 1$ are already sufficient, as will be argued below. As already noted in the introduction, a naive approach via Burer's result [Bur09] would result in additional constraints in the conic formulation which may be detrimental to algorithmic performance. Anyhow, some of them may be added in the implementation of specially structured problems.

2.3.1. Soft sparsity: regularization by zero-norm

No sign constraints on the variables

We will first treat sparse quadratic problems under linear equality constraints, but no sign constraints on the variables. At first sight, this may seem a restriction of models but actually any sign constraint on some variables would reduce dimensionality rather than increase complexity, as will be clear in the next subsection.

So consider the following soft sparsity model:

$$\min_{\mathsf{y}\in\mathbb{R}^n} \left\{ \mathsf{y}^\top \mathsf{Q}_0 \mathsf{y} + 2\mathsf{c}_0^\top \mathsf{y} + \rho\|\mathsf{y}\|_0 : \mathsf{c}_i^\top \mathsf{y} = b_i \,, i\in[1\!:\!m] \right\} , \tag{P_{S_0}}$$

where $\rho > 0$ is a regularization parameter for balancing accuracy against sparsity: the larger ρ, the more sparse the solution obtained from (P_{S_0}) will be.

By introduction of new variables $\mathsf{u} \in \mathbb{R}^n_+$ and decomposition of $\mathsf{y} = \mathsf{v} - \mathsf{w}$, where $\{\mathsf{v}, \mathsf{w}\} \subset \mathbb{R}^n_+$, problem (P_{S_0}) can be exactly expressed by a quadratic problem:

$$\begin{array}{ll} \min\limits_{\{\mathsf{v},\mathsf{w},\mathsf{u}\}\subset\mathbb{R}^n_+} & (\mathsf{v}-\mathsf{w})^\top \mathsf{Q}_0 (\mathsf{v}-\mathsf{w}) + 2\mathsf{c}_0^\top (\mathsf{v}-\mathsf{w}) + \rho n - \rho\, \mathsf{e}^\top \mathsf{u} \\ s.t. & \mathsf{c}_i^\top (\mathsf{v}-\mathsf{w}) = b_i \,, \quad i\in[1\!:\!m] \,, \\ & \mathsf{u}^\top (\mathsf{v}+\mathsf{w}) = 0 \,, \\ & \mathsf{u} \le \mathsf{e} \,. \end{array} \tag{$\mathrm{P}_{S'}$}$$

Because of nonnegativity of $(\mathsf{u}, \mathsf{v}, \mathsf{w})$, the homogeneous bilinear equality constraint $\mathsf{u}^\top(\mathsf{v}+\mathsf{w}) = 0$ is equivalent to the equality system $u_i(v_i + w_i) = 0$, $i\in[1\!:\!n]$.

Theorem 8. *Problems* (P_{S_0}) *and* $(\mathrm{P}_{S'})$ *are equivalent.*

Proof. This was proved in [FMP+18], but we provide a short argument for the readers' convenience here. For any $\alpha \in \mathbb{R}$, let $\alpha_+ := \max\{\alpha, 0\}$ and $\alpha_- = \alpha_+ - \alpha$; then $\alpha_\pm \ge 0$ and $|\alpha| = \alpha_+ + \alpha_-$ as well as $\alpha = \alpha_+ - \alpha_-$. If y is feasible to (P_{S_0}), then putting $v_i = (y_i)_+$ and $w_i = (y_i)_-$ as well as $u_i = 1 - |\mathrm{sign}\ y_i| \in \{0, 1\}$ one gets a feasible solution to $(\mathrm{P}_{S'})$ with same objective value, so $opt(\mathrm{P}_{S'}) \le opt(\mathrm{P}_{S_0})$. Conversely, let

$$F' := \{(\mathsf{v}, \mathsf{w}, \mathsf{u}) \in \mathbb{R}^{3n}_+ : \mathsf{u}^\top(\mathsf{v}+\mathsf{w}) = 0,\ \mathsf{u} \le \mathsf{e}\} \,.$$

For $(\mathsf{v}, \mathsf{w}, \mathsf{u}) \in F'$ feasible to $(\mathrm{P}_{S'})$, put $\mathsf{y} := \mathsf{v} - \mathsf{w} \in \mathbb{R}^n$ and consider $v'_i := (y_i)_+$, $w'_i := (y_i)_-$ and $u'_i := \mathrm{sign}\ u_i$. Then also $(\mathsf{v}', \mathsf{w}', \mathsf{u}') \in F'$ is $(\mathrm{P}_{S'})$-feasible with lower or equal objective value due to $\mathsf{e}^\top \mathsf{u} \le \mathsf{e}^\top \mathsf{u}'$ and $\mathsf{v}' - \mathsf{w}' = \mathsf{v} - \mathsf{w}$, and again, this value coincides with the objective value of y in (P_{S_0}). Thus $opt(\mathrm{P}_{S_0}) \le opt(\mathrm{P}_{S'})$. □

Given $(\mathsf{v}, \mathsf{w}, \mathsf{u}) \in F'$, we introduce slack variables $\mathsf{s} := \mathsf{e} - \mathsf{u} \in \mathbb{R}^n_+$ if $\mathsf{u} \le \mathsf{e}$ and stack all variables to $\mathsf{x} := [\mathsf{v}^\top, \mathsf{w}^\top, \mathsf{u}^\top, \mathsf{s}^\top]^\top \in \mathbb{R}^{4n}_+$. Then problem $(\mathrm{P}_{S'})$ can be rewritten, dropping the additive constant ρn:

$$\begin{array}{ll} \min\limits_{\mathsf{x}\in\mathbb{R}^{4n}_+} & \mathsf{x}^\top \mathsf{Q} \mathsf{x} + 2\mathsf{c}^\top \mathsf{x} \\ s.t. & \mathsf{a}_i^\top \mathsf{x} = b_i \,, \quad i\in[1\!:\!m+n] \,, \\ & \displaystyle\sum_{(i,j)\in\mathcal{E}} x_i x_j = 0 \,, \end{array} \tag{P_S}$$

where

$$\mathsf{Q} = \begin{pmatrix} \mathsf{Q}_0 & -\mathsf{Q}_0 & \mathsf{O} \\ -\mathsf{Q}_0 & \mathsf{Q}_0 & \mathsf{O} \\ \mathsf{O} & \mathsf{O} & \mathsf{O} \end{pmatrix} \in \mathcal{S}^{4n} \text{ and } \quad \mathsf{c} = [\mathsf{c}_0^\top, -\mathsf{c}_0^\top, -\frac{\rho}{2}\mathsf{e}^\top, \mathsf{o}^\top]^\top \in \mathbb{R}^{4n},$$

as well as the linear constraints, putting $b_i = 1$ for all $i \in [m+1:m+n]$ while keeping all b_i unchanged for $i \in [1:m]$:

$$\mathsf{a}_i = \begin{cases} [\mathsf{c}_i^\top, -\mathsf{c}_i^\top, \mathsf{o}^\top]^\top \in \mathbb{R}^{4n}, & \text{if } i \in [1:m], \\ \mathsf{e}_{i-m+2n} + \mathsf{e}_{i-m+3n} \in \mathbb{R}^{4n}, & \text{if } i \in [m+1:m+n]. \end{cases}$$

Furthermore, we use the graph $\mathcal{G}_\mathsf{F} = (V, \mathcal{E})$ with

$$V = [1:4n] \text{ and edge set } \mathcal{E} = \{\{i+2n, j\} : i \in [1:n], j \in \{i, i+n\}\}$$

to express the bilinear constraints $\mathsf{u}^\top \mathsf{v} + \mathsf{u}^\top \mathsf{w} = 0$ modelling $u_i|y_i| = 0$.

The first $2n$ variables (v, w) are not necessarily bounded but nevertheless are involved in the complementarity constraint $\mathsf{u}^\top(\mathsf{v} + \mathsf{w}) = 0$, so the classical key condition (K) fails. Fortunately, by construction we find a vertex cover $C := [2n+1:3n]$ of $\mathcal{G}_\mathsf{F}$ such that any $x_j = u_{j-2n} \leq 1$ for all $j \in C$. Moreover, for any feasible solution $\hat{\mathsf{y}} \in \mathbb{R}^n$ to the original QP (i.e., such that $\mathsf{C}\hat{\mathsf{y}} = \mathsf{b}$),

$$\hat{\mathsf{x}}^\top = [(\hat{y}_1)^+, \ldots, (\hat{y}_n)^+, (\hat{y}_1)^-, \ldots, (\hat{y}_n)^-, \mathsf{o}^\top, \mathsf{e}^\top]$$

satisfies $\hat{\mathsf{x}}_C = \mathsf{o}$, so that condition (R) is met if the original QP is feasible. Lemma 4 can be applied, and yields $L_\infty \subseteq F$ for this class of problems. Note that [BMP16, Thm.4] requires a priori knowledge of copositivity of the Hessian of the objective while our approach yields it automatically only by property of the constraints alone, and the fact that the optimal objective value is finite.

Therefore the conic relaxation is exact whenever the QP is feasible and bounded, which automatically would be the case for positive-semidefinite Q_0, given that the linear system $\mathsf{c}_i^\top \mathsf{y} = b_i$, $i \in [1:m]$, has a solution. Furthermore, the slackness constraints $\mathsf{u} + \mathsf{s} = \mathsf{e}$ provide some rows $\mathsf{a}_i \in \mathbb{R}^{4n}_+$ with $b_i = 1 > 0$, so that also the reduction step as in Section 2.2.3 can be performed.

Unfortunately, the constructed Q is never strictly L_∞-copositive even if Q_0 is positive-definite, hence we cannot establish strong duality along the general lines in Section 2.2.3. Note that in the present case, by construction of A,

$$L_\infty = \left\{ [\mathsf{v}^\top, \mathsf{w}^\top, \mathsf{o}^\top]^\top \in \mathbb{R}^{4n}_+ : \mathsf{c}_i^\top(\mathsf{v} - \mathsf{w}) = 0\,,\ i \in [1:m] \right\}, \tag{2.14}$$

which includes all vectors of the form $[\mathsf{v}^\top, \mathsf{v}^\top, \mathsf{o}^\top]^\top \in \mathbb{R}^{4n}_+$ for which any slack matrix $\mathsf{S}(\mathsf{r}, \mathsf{t}, h)$ gives a zero quadratic form.

However, even if Q_0 is merely positive-semidefinite, adding an ℓ^2 term $\tau \sum_{i=1}^{n} (v_i^2 + w_i^2)$ with small $\tau > 0$ to the objective provides a regularization which ensures strong conic duality. See Section 2.3.3 for more details.

Sparsity in QPs with sign constraints on variables

Additional linear inequality constraints can also be incorporated by introducing additional slack variables. We leave the obvious adaptations to the interested readers and concentrate for ease of presentation on the following model:

$$\min_{\mathsf{y}\in\mathbb{R}^n_+}\left\{\mathsf{y}^\top\mathsf{Q}_0\mathsf{y}+2\mathsf{c}_0^\top\mathsf{y}+\rho\|\mathsf{y}\|_0:\ \mathsf{c}_i^\top\mathsf{y}=b_i\,,i\in[1:m]\right\}\,. \tag{P_S^+}$$

Here we need not split $\mathsf{y}=\mathsf{v}-\mathsf{w}$ into two nonnegative vectors any more, but we will keep u and the slack variables as in the previous subsection. So, here $\mathsf{x}^\top=[\mathsf{y}^\top,\mathsf{u}^\top,\mathsf{s}^\top]$ with $\mathsf{x}\in\mathbb{R}^{3n}_+$ now, and we end up with a QPCC formulation very similar to (P_S), but with reduced dimensionality $3n$ and smaller primitive data: $\mathsf{x}^\top\mathsf{Q}\mathsf{x}+2\mathsf{c}^\top\mathsf{x}=\mathsf{y}^\top\mathsf{Q}_0\mathsf{y}+2\mathsf{c}_0^\top\mathsf{y}$ for

$$\mathsf{Q}=\begin{pmatrix}\mathsf{Q}_0 & \mathsf{O}\\ \mathsf{O} & \mathsf{O}\end{pmatrix}\in\mathcal{S}^{3n}\ \text{ and }\ \ \mathsf{c}=[\mathsf{c}_0^\top,-\frac{\rho}{2}\mathsf{e}^\top,\mathsf{o}^\top]^\top\in\mathbb{R}^{3n}\,,$$

The linear constraints read now, putting $b_i=1$ for all $i\in[m+1:m+n]$ while keeping b_i unchanged for all $i\in[1:m]$:

$$\mathsf{a}_i=\begin{cases}[\mathsf{c}_i^\top,\mathsf{o}^\top]^\top\in\mathbb{R}^{3n}\,, & \text{if } i\in[1:m]\,,\\ \mathsf{e}_{i-m+n}+\mathsf{e}_{i-m+2n}\in\mathbb{R}^{3n}\,, & \text{if } i\in[m+1:m+n]\,.\end{cases}$$

Furthermore, we use the edge set $\mathcal{E}=\{\{i,i+n\}:i\in[1:n]\}$ to express the bilinear constraints $\mathsf{u}^\top\mathsf{y}=0$ modelling $u_iy_i=0$. Here $C=[n+1:2n]$ is a vertex cover for $\mathcal{G}_\mathsf{F}$. Again, similar to the case with no sign constraints, Lemma 4 can be applied and yields $L_\infty\subseteq F$ also for this problem class. Indeed, for any feasible $\hat{\mathsf{y}}$ here $\hat{\mathsf{x}}:=[\hat{\mathsf{y}},\mathsf{o}^\top,\mathsf{e}^\top]^\top\in L$ satisfies $\hat{\mathsf{x}}_C=\mathsf{o}$.

The two resulting, already simplified conic reformulation problems read exactly as in (CP) and (CD) for their duals, with n replaced by $3n$ and m replaced by $m+n$.

Apart from employing a conic representation with reduced dimensionality, we also can infer strong duality for the conic primal-dual pair by imposing conditions which in a realistic setting are met more frequently, compared to the setting with no sign constraints on the primitive variables. Indeed, as explained in the remark after the following proof, if the primitive polyhedron is bounded, then the strict copositivity condition holds by default.

Theorem 9. *In model* (P_S^+)*, assume that* Q_0 *is strictly* L_∞^C*-copositive with*

$$L_\infty^\mathsf{C}:=\left\{\mathsf{y}\in\mathbb{R}^n_+:\mathsf{C}\mathsf{y}=\mathsf{o}\right\}\quad \textit{where}\ \mathsf{C}^\top=[\mathsf{c}_1,\ldots,\mathsf{c}_m]\,,$$

i.e., $\mathsf{y}^\top\mathsf{Q}_0\mathsf{y}>0$ *for all* $\mathsf{y}\in\mathbb{R}^n_+\setminus\{\mathsf{o}\}$ *with* $\mathsf{C}\mathsf{y}=\mathsf{o}$*. Then*

- *any feasible problem instance is bounded below;*
- *the conic representation is exact;*
- *the primal-dual conic pair has zero duality gap.*

Proof. Observe that in our case $L_\infty=L_\infty^\mathsf{C}\times\{\mathsf{o}\}\subset\mathbb{R}^{3n}_+$, so that $\mathsf{z}\in L_\infty\setminus\{\mathsf{o}\}$ if and only if $z^\top=[\mathsf{y}^\top,\mathsf{o}^\top]$ with $\mathsf{y}\in L_\infty^\mathsf{C}\setminus\{\mathsf{o}\}$, and then $\mathsf{z}^\top\mathsf{Q}\mathsf{z}=\mathsf{y}^\top\mathsf{Q}_0\mathsf{y}>0$ by assumption. The claims now follows by application of Theorem 7. □

Note that the assumption that Q_0 be strictly L_∞^C-copositive is trivially satisfied if the primitive feasible polyhedron $M=\left\{\mathsf{y}\in\mathbb{R}^n_+:\mathsf{C}\mathsf{y}=\mathsf{b}\right\}$ is bounded, because then $L_\infty^\mathsf{C}=\{\mathsf{o}\}$.

2.3.2. Hard sparsity constraints

Sometimes we want to force the solution of a quadratic problem to be k-sparse, where the integer $k > 0$ is a hard parameter for sparsity. To this end, we consider another sparsity model for constrained quadratic problems:

$$\min_{\mathsf{y}\in\mathbb{R}^n} \left\{ \mathsf{y}^\top \mathsf{Q}_0 \mathsf{y} + 2\mathsf{c}_0^\top \mathsf{y} : \mathsf{c}_i^\top \mathsf{y} = b_i \,, i\in[1\!:\!m] \,, \quad \|\mathsf{y}\|_0 \leq k \right\} . \tag{P_{H_0}}$$

Instead of using the soft parameter ρ to pursuit sparsity, alternatively, we employ $\|\cdot\|_0$ directly in a hard constraint. In analogy to the formulation of the soft sparsity problem (P_{S_0}), we can also exactly express the hard sparsity problem (P_{H_0}) as follows:

$$\begin{aligned} \min_{\{\mathsf{v},\mathsf{w},\mathsf{u}\}\subset\mathbb{R}^n_+} \quad & (\mathsf{v}-\mathsf{w})^\top \mathsf{Q}_0 (\mathsf{v}-\mathsf{w}) + 2\mathsf{c}_0^\top (\mathsf{v}-\mathsf{w}) \\ s.t. \quad & \mathsf{c}_i^\top (\mathsf{v}-\mathsf{w}) = b_i \,, \quad i\in[1\!:\!m] \,, \\ & \mathsf{u}^\top (\mathsf{v}+\mathsf{w}) = 0 \,, \\ & -\mathsf{e}^\top \mathsf{u} \leq k-n \,, \\ & \mathsf{u} \leq \mathsf{e} \,. \end{aligned} \tag{$\mathrm{P}_{H'}$}$$

Theorem 10. *Problems* (P_{H_0}) *and* ($\mathrm{P}_{H'}$) *are equivalent.*

Proof. The proof was presented in [BKS16] and is very similar to that of Theorem 8. □

Similarly, with the help of slack variables $\mathsf{s} := \mathsf{e} - \mathsf{u} \in \mathbb{R}^n_+$ and $\sigma := \mathsf{e}^\top \mathsf{u} + k - n \geq 0$ and after stacking all variables to $\mathsf{x} := [\mathsf{v}^\top, \mathsf{w}^\top, \mathsf{u}^\top, \mathsf{s}^\top, \sigma]^\top \in \mathbb{R}^{4n+1}_+$, the problem ($\mathrm{P}_{H'}$) can be rewritten:

$$\begin{aligned} \min_{\mathsf{x}\in\mathbb{R}^{4n+1}_+} \quad & \mathsf{x}^\top \mathsf{Q} \mathsf{x} + 2\mathsf{c}^\top \mathsf{x} \\ s.t. \quad & \mathsf{a}_i^\top \mathsf{x} = b_i \,, \quad i\in[1\!:\!m+n+1] \,, \\ & \sum_{(i,j)\in\mathcal{E}} x_i x_j = 0 \,, \end{aligned} \tag{P_H}$$

where

$$\mathsf{Q} = \begin{pmatrix} \mathsf{Q}_0 & -\mathsf{Q}_0 & \mathsf{O} \\ -\mathsf{Q}_0 & \mathsf{Q}_0 & \mathsf{O} \\ \mathsf{O} & \mathsf{O} & \mathsf{O} \end{pmatrix} \in \mathcal{S}^{4n+1} \text{ and } \quad \mathsf{c} = [\mathsf{c}_0^\top, -\mathsf{c}_0^\top, \mathsf{o}^\top]^\top \in \mathbb{R}^{4n+1} \,,$$

as well as the linear constraints, putting $b_i = 1$ for all $i\in[m+1\!:\!m+n]$ and $b_i = k-n$ for $i = m+n+1$:

$$\mathsf{a}_i = \begin{cases} [\mathsf{c}_i^\top, -\mathsf{c}_i^\top, \mathsf{o}^\top]^\top \in \mathbb{R}^{4n+1} \,, & \text{if } i\in[1\!:\!m] \,, \\ \mathsf{e}_{i-m+2n} + \mathsf{e}_{i-m+3n} \in \mathbb{R}^{4n+1} \,, & \text{if } i\in[m+1\!:\!m+n] \,, \\ [\mathsf{o}^\top, \mathsf{o}^\top, -\mathsf{e}^\top, \mathsf{o}^\top, 1]^\top \in \mathbb{R}^{4n+1} \,, & \text{if } i = m+n+1 \,. \end{cases}$$

Similarly to the soft sparsity case, the model with sign constraints on the variables can also be treated, leading to

$$\min_{\mathsf{y}\in\mathbb{R}^n_+} \left\{ \mathsf{y}^\top \mathsf{Q}_0 \mathsf{y} + 2\mathsf{c}_0^\top \mathsf{y} : \mathsf{c}_i^\top \mathsf{y} = b_i \,, i\in[1\!:\!m] \,, \quad \|\mathsf{y}\|_0 \leq k \right\} . \tag{P_H^+}$$

Again, we need not split $\mathsf{y} = \mathsf{v} - \mathsf{w}$ into two nonnegative vectors anymore, but we will keep u and the slack variables as in the previous subsection. So, here $\mathsf{x}^\top = [\mathsf{y}^\top, \mathsf{u}^\top, \mathsf{s}^\top, \sigma]$ with $\mathsf{x} \in \mathbb{R}^{3n+1}_+$ now, and again $\mathsf{x}^\top \mathsf{Q} \mathsf{x} + 2\mathsf{c}^\top \mathsf{x} = \mathsf{y}^\top \mathsf{Q}_0 \mathsf{y} + 2\mathsf{c}_0^\top \mathsf{y}$. The QPCC formulation is similar to (P_H), using the same

$$\mathsf{Q} = \begin{pmatrix} \mathsf{Q}_0 & \mathsf{O} \\ \mathsf{O} & \mathsf{O} \end{pmatrix} \in \mathcal{S}^{3n+1} \quad \text{and} \quad \mathsf{c} = [\mathsf{c}_0^\top, \mathsf{o}^\top]^\top \in \mathbb{R}^{3n+1}\,,$$

adding a linear constraint with $b_{m+n+1} = k - n$ while keeping b_i unchanged for all $i \in [1\!:\!m+n]$:

$$\mathsf{a}_i = \begin{cases} [\mathsf{c}_i^\top, \mathsf{o}^\top]^\top \in \mathbb{R}^{3n+1}\,, & \text{if } i \in [1\!:\!m]\,, \\ \mathsf{e}_{i-m+n} + \mathsf{e}_{i-m+2n} \in \mathbb{R}^{3n+1}\,, & \text{if } i \in [m+1\!:\!m+n] \\ [\mathsf{o}^\top, -\mathsf{e}^\top, \mathsf{o}^\top, 1]^\top \in \mathbb{R}^{3n+1}\,, & \text{if } i = m+n+1\,. \end{cases}$$

The conic reformulation problems read exactly as in (CP) and (CD) for their duals, with n replaced by $3n+1$ and m replaced by $m+n+1$.

Theorem 11. *In model* (P_H^+)*, assume that* Q_0 *is strictly* L_∞^{C}*-copositive with*

$$L_\infty^{\mathsf{C}} := \big\{\mathsf{y} \in \mathbb{R}^n_+ : \mathsf{C}\mathsf{y} = \mathsf{o}\big\} \quad \textit{where } \mathsf{C}^\top = [\mathsf{c}_1, \ldots, \mathsf{c}_m]\,,$$

i.e., $\mathsf{y}^\top \mathsf{Q}_0 \mathsf{y} > 0$ *for all* $\mathsf{y} \in \mathbb{R}^n_+ \setminus \{\mathsf{o}\}$ *with* $\mathsf{C}\mathsf{y} = \mathsf{o}$*. Then*

- *any feasible problem instance is bounded below;*
- *the conic representation is exact;*
- *the primal-dual conic pair has zero duality gap.*

Proof. Employ the same arguments as for Theorem 9. □

Again, the assumption that Q_0 be strictly L_∞^{C}-copositive is trivially satisfied if the primitive feasible polyhedron $M = \big\{\mathsf{y} \in \mathbb{R}^n_+ : \mathsf{C}\mathsf{y} = \mathsf{b}\big\}$ is bounded, because then $L_\infty^{\mathsf{C}} = \{\mathsf{o}\}$. It is also satisfied if Q_0 is positive-definite as in the following section which presents an application that motivated this study.

2.3.3. Sparse linear regression: conic reformulation

In a linear least-squares/regression context, $\mathsf{Q}_0 = \mathsf{A}_0^\top \mathsf{A}_0$ with A_0 an $m \times n$ matrix with $m < n$, so Q_0 is always positive-semidefinite but will have some zero eigenvalues. Hence strict copositivity of Q_0 cannot be ensured by spectral properties of Q_0 alone; one possibility is the condition that A_0 contains some strictly positive rows which renders $\mathsf{Q}_0 = \mathsf{P} + \mathsf{N}$ with $\mathsf{P} \in \mathcal{S}^n_+$ and $\mathsf{N} \in \mathcal{N}^n$. This is enough to ensure strict $\mathbb{R}^n_+$-copositivity of Q_0 which establishes all favorable properties of our conic reformulation by application of Theorems 9 and 11. Another class of instances is that with bounded primitive feasible polyhedron $M = \big\{\mathsf{y} \in \mathbb{R}^n_+ : \mathsf{C}\mathsf{y} = \mathsf{b}\big\}$, as explained in the note after these results.

In absence of sign constraints, another option already hinted at in Section 2.3.1 would consist in adding a regularization term $\tau \sum_{i=1}^{n} (v_i^2 + w_i^2)$, with small $\tau > 0$:

Proposition 12. *Consider a regularized sparse linear regression problem with a small* $\tau > 0$ *where, in the notation of* (P_{S_0}) *or* (P_{H_0})*,*

$$\mathsf{y}^\top \mathsf{Q}_0 \mathsf{y} = \|\mathsf{A}_0 \mathsf{y}\|^2 + \tau \|\mathsf{y}\|^2\,.$$

Then

- *any feasible problem instance is bounded below;*
- *the conic representation is exact;*
- *the primal-dual conic pair has zero duality gap.*

Proof. By reformulation as before, we only need to add $\tau\left(\|\mathsf{v}\|^2 + \|\mathsf{w}\|^2\right)$ to the original objective of (P_{S_0}) or (P_{H_0}) instead of $\tau\|\mathsf{y}\|^2 = \tau\|\mathsf{v} - \mathsf{w}\|^2$; indeed, in the optimum we have $\|\mathsf{v}\|^2 + \|\mathsf{w}\|^2 = \|\mathsf{v} - \mathsf{w}\|^2$ as optimality already implies $v_i = y_i^+$ and $w_i = y_i^-$ and therefore $\mathsf{v}^\top\mathsf{w} = 0$, as argued before. Thus we can regularize by a positive-definite term in (v, w) which renders Q strictly L_∞-copositive where L_∞ is given as in (2.14) (replace $4n$ with $4n + 1$ when dealing with hard sparsity constraints). The assertions then follow by application of Theorems 6 and 7. □

Thus, for a large class of sparse regression problems, be they constrained or not, our theory applies in full strength. Apart from that, even if the last line of arguments cannot be carried through for all instances with no sign constraints on the variables, we still can use weak conic duality to obtain strong dual bounds, as our experiments show.

2.4. An iterative algorithm for dual problems (CD)

For the convenience of expression, we write (CD) in the following form:

$$\max_{\mathsf{x}\in\mathbb{R}^m} \left\{ \mathsf{a}^\top \mathsf{x} : g(\mathsf{x}) \in \mathcal{COP}^n \right\}, \tag{2.15}$$

where $g : \mathbb{R}^m \to \mathcal{S}^n$ is a linear-affine function in x and $\mathsf{a} \in \mathbb{R}^m$. By convexity of $\mathcal{COP}^n$, it is obvious that the set $\{\mathsf{x} : g(\mathsf{x}) \in \mathcal{COP}^n\}$ is convex too. Therefore it is possible to solve the convex problem (2.15) via a directional search scheme, provided a feasible initial guess. Recall that any feasible solution to (2.15) already provides a rigorous lower bound $\mathsf{a}^\top \mathsf{x}$ for (P).

Now we introduce a local search algorithm for (2.15) where a subscript i denotes the i-th entry of a vector and a superscript k stands for the iteration counter.

Data: a problem of the form (2.15), a feasible initial guess $\mathsf{x}^0 \in \mathbb{R}^m$, adaptive parameter $0 < \eta < 1$, stopping tolerance $\varepsilon > 0$, initial step size $\lambda^0 \geq \varepsilon\,\mathsf{e}$
Result: a solution x^* to (2.15)
while $\|\lambda^k\|_\infty \geq \varepsilon$ **do**
  Generate a permutation $\mathsf{p}^k \in \mathbb{R}^m$ of set $[1\!:\!d]$ such that the following inequalities hold:
  $\lambda^k_{p_i^k}\text{sign}\,(a_{p_i^k}) \leq \lambda^k_{p_j^k}\text{sign}\,(a_{p_j^k})$ if $i < j$
  $\mathsf{y}^0 \Leftarrow \mathsf{x}^k, \quad \beta^0 \Leftarrow \lambda^k, \quad \mathsf{p} \Leftarrow \mathsf{p}^k$
  while $i \in [1\!:\!m]$ **do**
    $\bar{\mathsf{y}}^i \Leftarrow \mathsf{y}^{i-1},\ \beta^i \Leftarrow \beta^{i-1}$
    $\bar{y}^i_{p_i} \Leftarrow y^{i-1}_{p_i} + \beta^{i-1}_{p_i} a_{p_i}$
    if $g(\bar{\mathsf{y}}^i) \in \mathcal{COP}^n$ **then**
      $\mathsf{y}^i \Leftarrow \bar{\mathsf{y}}^i,\ \beta^i_{p_i} \Leftarrow \frac{1}{\eta}\beta^{i-1}_{p_i}$
    else
      $\mathsf{y}^i \Leftarrow \mathsf{y}^{i-1},\ \beta^i_{p_i} \Leftarrow \eta\beta^{i-1}_{p_i}$
    end
    $i \Leftarrow i + 1$
  end
  $\mathsf{x}^{k+1} \Leftarrow \mathsf{y}^m, \quad \lambda^{k+1} \Leftarrow \beta^m, \quad k \Leftarrow k + 1$
end

Algorithm 1: Greedy coordinate descent algorithm

As for the copositivity detection involved in the Algorithm 1, one of possible methods is [Ans21], which amounts to solve the following MILP:

$$\begin{aligned} \max_{\mathsf{x},\mathsf{y}} \quad & \gamma \\ s.t. \quad & \mathsf{s}_i^\top \mathsf{x} \leq -\gamma + m_i(1 - y_i)\,, \quad i \in [1\!:\!n], \\ & \gamma \geq 0\,, \mathsf{o} \leq \mathsf{x} \leq \mathsf{y}, \\ & \mathsf{y} \in \{0,1\}^n\,, \mathsf{e}^\top \mathsf{y} \geq 1\,, \end{aligned} \tag{2.16}$$

where s_i is the i-th column of targeted matrix S and the parameter m_i is typically chosen $m_i = 1 + \sum_{i\neq j}\{s_{ij} : s_{ij} > 0\}$ as in [Ans21].

Theorem 13. *[Ans21] Let $\mathsf{S} \in \mathcal{S}^n$. Then $\mathsf{S} \in \mathcal{COP}^n$ if and only if the optimal value of (2.16) is zero.*

2.5. Conclusion

In this chapter, we propose an exact completely copositive reformulation of QPCCs under mild conditions on the constraints, which extend previously known ones. An application on pursuing sparse solutions of (nonconvex) quadratic optimization is shown to satisfy our proposed conditions and therefore an equivalent completely positive optimization problem is obtained. Moreover, the condition for strong conic duality is also studied. To solve the conic optimization problem from the dual side, we introduce a direct search method with embedded MILP-based copositivity detection.

3. CP reformulation for sparse QPs

In this chapter, we focus on the sparse quadratic optimization problems and propose novel and more compact completely positive reformulations for them. This part is based on a nearly complete joint work with Immanuel Bomze and Markus Gabl, which hopefully will be ready for submission soon.

3.1. The problem

In this chapter, we aim to obtain sparse solutions to quadratic optimization problems (QPs). We consider the QP in standard form:

$$\begin{aligned} z^* := \min_{\mathsf{x}\in\mathbb{R}^n_+} \quad & \mathsf{x}^\top \mathsf{Q}\mathsf{x} + 2\mathsf{c}^\top \mathsf{x} \\ \text{s.t.} \quad & \mathsf{A}\mathsf{x} = \mathsf{b}\,, \end{aligned} \tag{QP}$$

where Q is a (possibly indefinite) symmetry real matrix and the linear operator $\mathsf{A} : \mathbb{R}^n \to \mathbb{R}^m$, represented by an $m \times n$ matrix. Throughout the chapter, we assume the (non-sparse) problem (QP) is feasible and bounded below,

$$-\infty < \min\left\{\mathsf{x}^\top \mathsf{Q}\mathsf{x} + 2\mathsf{c}^\top \mathsf{x} : \mathsf{A}\mathsf{x} = \mathsf{b}\,,\, \mathsf{x} \in \mathbb{R}^n_+\right\} < +\infty\,. \tag{3.1}$$

Incorporating the ℓ_0-pseudonorm

$$\|\mathsf{x}\|_0 := |\{i\in[1:n] : x_i \neq 0\}|,$$

term into an optimization problem, one can obtain sparse solutions to this problem.

Generally, there are two ways to exploit the ℓ_0-norm to boost the sparsity of the optimal solution. By adding the term to the objective, we consider the problem of the form, which is referred to as soft sparse QP:

$$\begin{aligned} \min_{\mathsf{x}\in\mathbb{R}^n_+} \quad & \mathsf{x}^\top \mathsf{Q}\mathsf{x} + 2\mathsf{c}^\top \mathsf{x} + \gamma\|\mathsf{x}\|_0 \\ \text{s.t.} \quad & \mathsf{A}\mathsf{x} = \mathsf{b}\,, \end{aligned} \tag{Soft-QP}$$

where $\gamma > 0$ is a trade-off parameter to balance between the sparsity of the optimal solution and the optimality of the original QP. It is worth noticing that the optimal solution of (Soft-QP) is in fact always attained. Indeed, considering all possible supports of x^*, the problem (Soft-QP) reduces to a finite collection of 2^n QPs (some of them may be infeasible). To this end, the optimal solution to the problem (Soft-QP) can be interpreted as the best choice from a candidate pool consisting of optimal solutions to a finite collection of QPs and therefore the attainability of the optimal solution follows from the celebrated Frank/Wolfe Theorem [FW56].

Alternatively, we can also limit the number of nonzero elements of variables directly by a so-called cardinality constraint and arrive at the problem, which is referred to as hard sparse QP:

$$\begin{aligned} \min_{\mathsf{x}\in\mathbb{R}^n_+} \quad & \mathsf{x}^\top \mathsf{Q}\mathsf{x} + 2\mathsf{c}^\top \mathsf{x} \\ \text{s.t.} \quad & \mathsf{A}\mathsf{x} = \mathsf{b} \\ & \|\mathsf{x}\|_0 \le \rho\,, \end{aligned} \tag{Hard-QP}$$

where $\rho \in [0, n]$ is a user-defined parameter.[1] Note that this problem is feasible if and only if $\rho \ge \underline{\rho} := \min_{\mathsf{x}\in\mathbb{R}^n_+}\{\|\mathsf{x}\|_0 : \mathsf{A}\mathsf{x} = \mathsf{b}\}$, see the discussion before Proposition 21 below. Then the optimal solution of (Hard-QP) is attainable for the same reason as in the soft sparse QP case, for a finite collection of classical QPs.

It is known that both (Soft-QP) and (Hard-QP) are in the NP-hard problem class even if the term $\|\cdot\|_0$ is absent. In this paper, we develop the completely positive (CP) technique following important works [Bur09] and propose a novel, hopefully more efficient and effective, CP reformulation of these two problems. CP reformulation is a convexification technique for some particular subsets of quadratic-constrained quadratic optimization problems (QCQPs), packaging the nonconvexity of the original problem into the completely positive cone (or its dual cone: copositive cone), and ends up with a linear conic optimization problem. For more excellent works and literature in this field, we refer interested readers to [Bom12, BDDK$^+$00, Bur15, DR21, PVZ15] and references therein.

3.2. CP reformulation for soft sparse QPs

In this section, we propose two compact CP relaxations of (Soft-QP) and prove the exactness of the relaxations.

3.2.1. Burer's representation theorem and its extension

In this subsection, we review Burer's CP reformulation for quadratic optimization problems with complementarity constraints (QPCCs), which is mainly based on the seminal paper [Bur09]. For two vectors $\mathsf{a} \in \mathbb{R}^n$ and $\mathsf{b} \in \mathbb{R}^n$, we denote their Hadamard product by $\mathsf{a} \circ \mathsf{b} := [a_i b_i]_{i\in[1:n]} \in \mathbb{R}^n$.

First of all, we consider an optimization problem of the form (QP), and linearize the objective function by replacing $\mathsf{x}\mathsf{x}^\top$ by X. To this end, we introduced a new variable X and the objective function becomes a linear function in X and x, e.g., $\langle \mathsf{X}, \mathsf{Q}\rangle + 2\mathsf{c}^\top\mathsf{x}$, which is known as the lifting technique. It is worth noticing that the lifted objective is always a lower bound of the original objective, as for any feasible point x, the point $(\mathsf{x}, \mathsf{x}\mathsf{x}^\top)$ is also feasible to the lifted problem and gives the same objective value. Hence, in order to have the lifted problem as tight as the original problem, some extra constraints are usually posed to the variable (x, X). Now consider the following lifted set, to shed more light on the effect of squaring a linear constraint:

Lemma 14. *If the point* (x, X) *is an element of the set*

$$\left\{(\mathsf{x}, \mathsf{X}) \in \mathbb{R}^n_+ \times \mathcal{S}^n : \begin{array}{l} \mathsf{A}\mathsf{x} = \mathsf{b} \\ \operatorname{diag}(\mathsf{A}\mathsf{X}\mathsf{A}^\top) = \mathsf{b}\circ\mathsf{b} \end{array}, \begin{pmatrix} 1 & \mathsf{x}^\top \\ \mathsf{x} & \mathsf{X} \end{pmatrix} \in \mathcal{CP}^{n+1}\right\}, \tag{3.2}$$

[1] not necessarily integer.

it has the following decomposition:

$$\begin{pmatrix} 1 & \mathsf{x}^\top \\ \mathsf{x} & \mathsf{X} \end{pmatrix} = \sum_{k\in K_+} \beta_k \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \end{pmatrix}^\top + \sum_{k\in K_0} \begin{pmatrix} 0 \\ \mathsf{x}^k \end{pmatrix} \begin{pmatrix} 0 \\ \mathsf{x}^k \end{pmatrix}^\top,$$

where $\beta_k = \lambda_k^2 > 0$ *for all* $k \in K_+$ *with* $\sum_{k\in K_+} \beta_k = 1$, $\mathsf{x}^k \in \mathbb{R}^n_+$ *for all* $k \in K_+ \cup K_0$, *and the following equations hold:*

$$\frac{\mathsf{x}^k}{\lambda_k} \in L := \{\mathsf{x} \in \mathbb{R}^n_+ : \mathsf{A}\mathsf{x} = \mathsf{b}\}, \quad \forall k \in K_+ \,,$$

and

$$\mathsf{x}^k \in L_0 := \{\mathsf{x} \in \mathbb{R}^n_+ : \mathsf{A}\mathsf{x} = \mathsf{o}\}, \quad \forall k \in K_0 \,.$$

Proof. See in the paper [Bur09]. □

Note that the lifted variable X consists of two parts, namely,

$$\mathsf{X} = \sum_{k\in K_+} \beta_k \frac{\mathsf{x}^k}{\lambda_k} \left(\frac{\mathsf{x}^k}{\lambda_k}\right)^\top + \sum_{k\in K_0} \mathsf{x}^k (\mathsf{x}^k)^\top,$$

where x^k, β_k, and λ_k satisfy the property presented in Lemma 14. Now we consider the objective in the lifted problem:

$$\begin{aligned} & \langle \mathsf{Q}, \mathsf{X} \rangle + 2\mathsf{c}^\top \mathsf{x} \\ = & \sum_{k\in K_+} \beta_k \left(\frac{\mathsf{x}^k}{\lambda_k}\right)^\top \mathsf{Q} \frac{\mathsf{x}^k}{\lambda_k} + \sum_{k\in K_0} (\mathsf{x}^k)^\top \mathsf{Q}\mathsf{x}^k + 2\mathsf{c}^\top \mathsf{x} \\ \geq & z^* + \sum_{k\in K_0} (\mathsf{x}^k)^\top \mathsf{Q}\mathsf{x}^k \,, \end{aligned}$$

where z^* is the optimal value to the (QP).

We notice that if the quadratic form $\mathsf{x}^\top \mathsf{Q}\mathsf{x}$ takes no negative values over the set $\mathsf{x} \in L_0$, then at least one optimal solution in the above lifted set (3.2) can be expressed as the convex hull of the set of dyadic matrices $\mathsf{x}^k(\mathsf{x}^k)^\top$ based upon elements x^k from the feasible region of (QP), which leads to an exact CP relaxation.

Following a similar idea, Burer proposed a so-called key condition[2] ensuring the exactness of CP relaxation for QPCC. Here is the form of QPCC considered in [Bur09]:

$$\begin{aligned} \min_{\mathsf{x}\in\mathbb{R}^n_+} \quad & \mathsf{x}^\top \mathsf{Q}\mathsf{x} + 2\mathsf{c}^\top \mathsf{x} \\ \text{s.t.} \quad & \mathsf{A}\mathsf{x} = \mathsf{b} \,, \\ & x_i x_j = 0, \quad \forall (i,j) \in B_1 \times B_2 \subseteq \in [1\!:\!n] \times \in [1\!:\!n] \,. \end{aligned} \tag{QPCC}$$

Under the key condition:

$$\mathsf{x} \in L \implies \exists u \geq 0 \text{ such that } 0 \leq x_j \leq u, \quad \forall j \in B_1 \cup B_2 \,,$$

[2] this condition ensuring exactness of the CP relaxation for QPCCs can be significantly relaxed [BP23].

(QPCC) is equivalent to the following CP relaxation:

$$\begin{aligned}
\min_{\mathsf{x}\in\mathbb{R}^n_+} \quad & \langle \mathsf{Q}, \mathsf{X}\rangle + 2\mathsf{c}^\top \mathsf{x} \\
\text{s.t.} \quad & \mathsf{A}\mathsf{x} = \mathsf{b}\,, \\
& \operatorname{diag}\left(\mathsf{A}\mathsf{X}\mathsf{A}^\top\right) = \mathsf{b}\circ\mathsf{b}\,, \\
& X_{ij} = 0, \quad \forall (i,j) \in B_1 \times B_2\,, \\
& \begin{pmatrix} 1 & \mathsf{x}^\top \\ \mathsf{x} & \mathsf{X} \end{pmatrix} \in \mathcal{CP}^{n+1}\,.
\end{aligned}$$

In the next sections, we will see that both soft sparse QP and hard sparse QP can be expressed in the form of (QPCC) and therefore Burer's CP relaxation results are applicable. However, the key condition is not necessarily satisfied if the feasible region is unbounded. Even for the bounded (binary) variables, we need to introduce the slack variables to indicate the boundedness from the perspective of set L, which leads to an increase in the dimension of problem. To overcome this difficulty, we propose more compact CP relaxations and prove their exactness for both soft and hard sparse QP cases in a direct way.

3.2.2. New and more compact CP reformulation for soft sparse QPs

In this subsection, we propose new and compact CP reformulations for soft sparse QPs.

Recall that the soft sparse QP is of the form:

$$\begin{aligned}
\min_{\mathsf{x}\in\mathbb{R}^n_+} \quad & \mathsf{x}^\top \mathsf{Q}\mathsf{x} + 2\mathsf{c}^\top \mathsf{x} + \gamma\|\mathsf{x}\|_0 \\
\text{s.t.} \quad & \mathsf{A}\mathsf{x} = \mathsf{b}\,.
\end{aligned} \tag{Soft-QP}$$

To express the term $\|\mathsf{x}\|_0$ in a more tractable way, we introduce auxiliary binary variables and employ a complementarity constraint, arriving at a mixed-integer reformulation of (Soft-QP):

$$\begin{aligned}
\zeta^*_\gamma := \min_{\mathsf{x}\in\mathbb{R}^n_+} \quad & \mathsf{x}^\top \mathsf{Q}\mathsf{x} + 2\mathsf{c}^\top \mathsf{x} - \gamma \mathsf{e}^\top \mathsf{u} \\
\text{s.t.} \quad & \mathsf{A}\mathsf{x} = \mathsf{b}\,, \\
& \mathsf{x}^\top \mathsf{u} = 0\,, \\
& \mathsf{u} \in \{0,1\}^n\,,
\end{aligned} \tag{MIQP-soft(γ)}$$

where u serves as an indicator variable. It is clear that, if x_i is positive for some i, the corresponding u_i has to be 0 due to the complementarity constraint, whereas, u_i has to be 1 if x_i equals 0 since we maximize u_i in the objective. To this end, the sparse term $\|\mathsf{x}\|_0$ is captured by $n - \mathsf{e}^\top \mathsf{u}$. After eliminating the constant term in the objective, we arrive at (MIQP-soft(γ)).

Note that the resulting problem is seemingly difficult due to the presence of two types of non-convexity, coming from the binarity and the complementarity of constraints. The most common way in the current literature to proceed with the problem is to linearize the complementarity constraint, involving bounds (big-M) on all variables. However, the consequence of the above procedure can be highly dependent on the estimation of big-M, which is impractical in many real-world applications.

Inspired by the previous work [Bur09, BKS16, BP23], we propose the following CP relaxation of the problem (MIQP-soft(γ)):

$$
\begin{aligned}
z_\gamma^* = \inf \quad & z_\gamma(\mathsf{Y}) := \langle \mathsf{Q}, \mathsf{Y}^{\mathsf{xx}} \rangle + 2\mathsf{c}^\top \mathsf{x} - \gamma \mathsf{e}^\top \mathsf{u} \\
\text{s.t.} \quad & \mathsf{A}\mathsf{x} = \mathsf{b}\,, \\
& \operatorname{diag}\left(\mathsf{A}\mathsf{Y}^{\mathsf{xx}}\mathsf{A}^\top\right) = \mathsf{b} \circ \mathsf{b}\,, \\
& \mathsf{e}^\top \operatorname{diag}\left(\mathsf{Y}^{\mathsf{xu}}\right) = 0\,, \\
& \operatorname{diag}\left(\mathsf{Y}^{\mathsf{uu}}\right) = \mathsf{u}\,, \\
& \mathsf{Y} := \begin{pmatrix} 1 & \mathsf{x}^\top & \mathsf{u}^\top \\ \mathsf{x} & \mathsf{Y}^{\mathsf{xx}} & \mathsf{Y}^{\mathsf{xu}} \\ \mathsf{u} & (\mathsf{Y}^{\mathsf{xu}})^\top & \mathsf{Y}^{\mathsf{uu}} \end{pmatrix} \in \mathcal{CP}^{2n+1}\,.
\end{aligned}
\tag{CP-soft(γ)}
$$

In the following, we will show that the problems (MIQP-soft(γ)) and (CP-soft(γ)) are equivalent. For the sake of compact expression, we abbreviate the above matrix relation by $\mathsf{Y} = \mathcal{Y}(\mathsf{x}, \mathsf{Y}^{\mathsf{xx}}, \mathsf{u}, \mathsf{Y}^{\mathsf{xu}}, \mathsf{Y}^{\mathsf{uu}})$.

Lemma 15. *Given any $\mathsf{p} \in \mathbb{R}^d_+$ with finite d, consider the following maximization problem:*

$$
\sup_{\mathsf{v}, s} \left\{ \mathsf{p}^\top \mathsf{v} : \mathsf{p}^\top \mathsf{v} = \|\mathsf{v}\|^2 + s, (s, \mathsf{v}) \in \mathbb{R}^{d+1}_+ \right\}.
$$

This problem has the unique optimal solution $[\mathsf{v}^{*\top}, s^*] = [\mathsf{p}^\top, 0]$.

Proof. It is clear that over the feasible set we have $\|\mathsf{v}\|^2 = \mathsf{p}^\top \mathsf{v} - s \le \|\mathsf{p}\|\|\mathsf{v}\| - s$, which implies that $\|\mathsf{v}\| \le \|\mathsf{p}\|$ and the objective function $\mathsf{p}^\top \mathsf{v} \le \|\mathsf{p}\|\|\mathsf{v}\| \le \|\mathsf{p}\|^2$. On the other hand, it is clear that $\mathsf{v}^* = \mathsf{p}$ is the only feasible point that attains the upper bound, namely, $\mathsf{p}^\top \mathsf{v}^* = \|\mathsf{p}\|^2$. In this case, $s^* = 0$. □

Lemma 16. *For a given* (CP-soft(γ))*-feasible point Y_0, there exists an improving* (CP-soft(γ))*-feasible point Y (i.e., $z_\gamma(\mathsf{Y}) \le z_\gamma(\mathsf{Y}_0)$ holds) which has the decomposition*

$$
\mathsf{Y} = \begin{pmatrix} 1 & \mathsf{x}^\top & \mathsf{u}^\top \\ \mathsf{x} & \mathsf{Y}^{\mathsf{xx}} & \mathsf{Y}^{\mathsf{xu}} \\ \mathsf{u} & (\mathsf{Y}^{\mathsf{xu}})^\top & \mathsf{Y}^{\mathsf{uu}} \end{pmatrix} = \sum_{k \in K_+} \beta_k \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \\ \frac{\mathsf{u}^k}{\lambda_k} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \\ \frac{\mathsf{u}^k}{\lambda_k} \end{pmatrix}^\top + \sum_{k \in K_0} \begin{pmatrix} 0 \\ \mathsf{x}^k \\ \mathsf{u}^k \end{pmatrix} \begin{pmatrix} 0 \\ \mathsf{x}^k \\ \mathsf{u}^k \end{pmatrix}^\top, \tag{3.3}
$$

where $\beta_k = \lambda_k^2 > 0$ for all $k \in K_+$ with $\sum_{k \in K_+} \beta_k = 1$ and $\{\mathsf{x}^k, \mathsf{u}^k\} \subset \mathbb{R}^n_+$ for all $k \in K := K_+ \cup K_0$. Moreover, $\frac{\mathsf{u}^k}{\lambda_k} \in \{0,1\}^n$ and $\mathsf{A}\mathsf{x}^k = \lambda_k \mathsf{b}$ for all $k \in K_+$, while $\mathsf{u}^k = \mathsf{o}$ and $\mathsf{A}\mathsf{x}^k = \mathsf{o}$ for all $k \in K_0$.

Proof. Let $\mathsf{Y}_0 = \mathcal{Y}(\mathsf{x}_0, \mathsf{Y}_0^{\mathsf{xx}}, ...)$ be (CP-soft(γ))-feasible. We construct another (CP-soft(γ))-feasible point $\mathsf{Y} = \mathcal{Y}(\mathsf{x}, \mathsf{Y}^{\mathsf{xx}}, \mathsf{u}, \mathsf{Y}^{\mathsf{xu}}, \mathsf{Y}^{\mathsf{uu}})$ with $\mathsf{x} = \mathsf{x}_0$ and $\mathsf{Y}^{\mathsf{xx}} = \mathsf{Y}_0^{\mathsf{xx}}$. Hence $\mathsf{A}\mathsf{x} = \mathsf{b}$ and $\operatorname{diag}\left(\mathsf{A}\mathsf{Y}^{\mathsf{xx}}\mathsf{A}^\top\right) = \mathsf{b} \circ \mathsf{b}$ are granted. It remains to show $z_\gamma(\mathsf{Y}) \le z_\gamma(\mathsf{Y}_0)$ and the property: $\frac{\mathsf{u}^k}{\lambda_k} \in \{0,1\}^n$ for all $k \in K_+$ and $\mathsf{u}^k = \mathsf{o}$ for all $k \in K_0$, which entails (CP-soft(γ))-feasibility of Y.

We start with the CP decomposition of the north-west principal submatrix of Y (which coincides with that of Y_0) as follows, resulting from Lemma 14:

$$
\begin{pmatrix} 1 & \mathsf{x}^\top \\ \mathsf{x} & \mathsf{Y}^{\mathsf{xx}} \end{pmatrix} = \begin{pmatrix} 1 & \mathsf{x}_0^\top \\ \mathsf{x}_0 & \mathsf{Y}_0^{\mathsf{xx}} \end{pmatrix} = \sum_{k \in K_+} \beta_k \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \end{pmatrix}^\top + \sum_{k \in K_0} \begin{pmatrix} 0 \\ \mathsf{x}^k \end{pmatrix} \begin{pmatrix} 0 \\ \mathsf{x}^k \end{pmatrix}^\top,
$$

where $\mathsf{A}\mathsf{x}^k = \lambda_k \mathsf{b}$ and $\beta_k = \lambda_k^2 > 0$ for all $k \in K_+$ with $\sum_{k\in K_+} \beta_k = 1$ while $\mathsf{A}\mathsf{x}^k = \mathsf{o}$ for all $k \in K_0$ and $\mathsf{x}^k \in \mathbb{R}^n_+$ for all $k \in K := K_+ \cup K_0$. With $(\mathsf{x}, \mathsf{Y}^{\mathsf{xx}})$ and the above CP decomposition fixed, we aim to find a set of $\mathsf{u}^k \in \mathbb{R}^n_+$ for all $k \in K$ such that the extended point $\mathsf{Y} = \mathcal{Y}(\mathsf{x}, \mathsf{Y}^{\mathsf{xx}}, \mathsf{u}, \mathsf{Y}^{\mathsf{xu}}, \mathsf{Y}^{\mathsf{uu}})$ defined by the CP decomposition (3.3) satisfies the stated property.

Regarding the constraint $\mathsf{e}^\top \mathrm{diag}\,(\mathsf{Y}^{\mathsf{xu}}) = 0$, we notice that this is equivalent to stating

$$x_i^k u_i^k = 0 \quad \text{for all } k \in K \text{ and } i \in [1\!:\!n]\,, \tag{3.4}$$

where the subscript index i denotes the i-th coordinate of the underlying vector.

Focusing on the i-th coordinate of x^k across all $k \in K_+$, we split the index set $K_+ = P_i \cup N_i$ into two parts according to positivity. i.e.,

$$P_i = \{k \in K_+ : x_i^k > 0\} \quad \text{and} \quad N_i = K_+ \setminus P_i = \{k \in K_+ : x_i^k = 0\}\,.$$

To this end, to obtain (3.4), we let

$$u_i^k := 0 \quad \text{for all } i \in [1\!:\!n] \quad \text{and all } k \in P_i\,. \tag{3.5}$$

Moreover, we let the remaining part of u^k be the optimal solution to the following problem:

$$\begin{aligned} \max_{\mathsf{u}^k \in \mathbb{R}^n_+, k\in K} \quad & \sum_{k\in K_+} \lambda_k \mathsf{e}^\top \mathsf{u}^k \\ \text{s.t.} \quad & \sum_{k\in N_i} (u_i^k)^2 + \sum_{k\in P_i} (u_i^k)^2 + \sum_{k\in K_0} (u_i^k)^2 = \sum_{k\in K_+} \lambda_k u_i^k, \quad \forall i \in [1\!:\!n]\,, \\ & x_i^k u_i^k = 0, \quad \forall k \in K, \forall i \in [1\!:\!n]\,. \end{aligned} \tag{3.6}$$

It follows immediately from the objective of (CP-soft(γ)) that then $z_\gamma(\mathsf{Y}) \le z_\gamma(\mathsf{Y}_0)$.

Notice that, in the optimization problem (3.6), there is no constraint involving two variables from different coordinates and the objective function is separable across the coordinates. Hence, the problem (3.6) can be equivalently represented by a set of mini-problems indexed by $i \in [1\!:\!n]$:

$$\begin{aligned} \max_{u_i^k \in \mathbb{R}_+, k\in K} \quad & \sum_{k\in K_+} \lambda_k u_i^k \\ \text{s.t.} \quad & \sum_{k\in N_i} (u_i^k)^2 + \sum_{k\in P_i} (u_i^k)^2 + \sum_{k\in K_0} (u_i^k)^2 = \sum_{k\in K_+} \lambda_k u_i^k\,, \\ & x_i^k u_i^k = 0, \quad \forall k \in K\,. \end{aligned}$$

Taking the fact (3.5) into account, we arrive at the following problems across all $i \in [1\!:\!n]$:

$$\begin{aligned} \max_{s_i, u_i^k \in \mathbb{R}_+, k\in N_i} \quad & \sum_{k\in N_i} \lambda_k u_i^k \\ \text{s.t.} \quad & \sum_{k\in N_i} (u_i^k)^2 + s_i = \sum_{k\in N_i} \lambda_k u_i^k\,, \end{aligned} \tag{3.7}$$

where $s_i = \sum_{k\in K_0} (u_i^k)^2$.

Applying Lemma 15 to the problem (3.7) with a fixed coordinate i, we specify p by $[\lambda_k]_{k\in N_i}$ and v by $[u_i^k]_{k\in N_i}$. It follows immediately that $u_i^k = \lambda_k$ for all $k \in N_i$, which implies that

$$\frac{u_i^k}{\lambda_k} = 1, \quad \forall k \in N_i.$$

Moreover, due to $s_i^* = 0$, we have

$$u_i^k = 0, \quad \forall k \in K_0.$$

Together with the fact (3.5) and applying Lemma 15 to the problems (3.7) across all coordinates $i \in [1:n]$, we complete the proof. □

Theorem 17. *Assume that* (3.1) *holds. Then, for all* $\gamma \in \mathbb{R}$*, the conic relaxation* (CP-soft(γ)) *is equivalent to the problem* (MIQP-soft(γ)) *in the sense of objective value.*

Proof. By Lemma 14, for any (CP-soft(γ))-feasible Y, we have a CP decomposition

$$\begin{aligned} \mathsf{Y} = \begin{pmatrix} 1 & \mathsf{x}^\top & \mathsf{u}^\top \\ \mathsf{x} & \mathsf{Y}^{\mathsf{xx}} & \mathsf{Y}^{\mathsf{xu}} \\ \mathsf{u} & (\mathsf{Y}^{\mathsf{xu}})^\top & \mathsf{Y}^{\mathsf{uu}} \end{pmatrix} = \sum_{k \in K} \begin{pmatrix} \lambda_k \\ \mathsf{x}^k \\ \mathsf{u}^k \end{pmatrix} \begin{pmatrix} \lambda_k \\ \mathsf{x}^k \\ \mathsf{u}^k \end{pmatrix}^\top = \mathsf{M}_1 + \mathsf{M}_2\,, \quad \text{with} \\ \mathsf{M}_1 := \sum_{k \in K_+} \beta_k \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \\ \frac{\mathsf{u}^k}{\lambda_k} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \\ \frac{\mathsf{u}^k}{\lambda_k} \end{pmatrix}^\top \quad \text{and} \quad \mathsf{M}_2 := \sum_{k \in K_0} \begin{pmatrix} 0 \\ \mathsf{x}^k \\ \mathsf{u}^k \end{pmatrix} \begin{pmatrix} 0 \\ \mathsf{x}^k \\ \mathsf{u}^k \end{pmatrix}^\top \end{aligned} \tag{3.8}$$

where $\lambda_k^2 = \beta_k$ for all $k \in K_+$ and $\sum_{k \in K_+} \beta_k = 1$. Moreover, it holds that $\mathsf{A}\mathsf{x}^k = \lambda_k \mathsf{b}$ for all $k \in K_+$ while $\mathsf{A}\mathsf{x}^k = \mathsf{o}$ for all $k \in K_0$.

Now consider a (CP-soft(γ))-feasible Y^* which is (nearly)[3] optimal, $z_\gamma(\mathsf{Y}^*) \approx z_\gamma^*$.

If $K_0 = \emptyset$, then by default $\mathsf{M}_2 = \mathsf{O}$, so that it is easy to see the equivalence. Indeed, Lemma 16 gives

$$z_\gamma(\mathsf{Y}^*) = z_\gamma(\mathsf{M}_1) = \sum_{k \in K_+} \beta_k \big(\frac{(\mathsf{x}^k)^\top}{\lambda_k} \mathsf{Q} \frac{\mathsf{x}^k}{\lambda_k} + 2\mathsf{c}^\top \frac{\mathsf{x}^k}{\lambda_k} - \gamma \mathsf{e}^\top \frac{\mathsf{u}^k}{\lambda_k} \big)$$

with $\left(\frac{\mathsf{x}^k}{\lambda_k}\right)^\top \frac{\mathsf{u}^k}{\lambda_k} = 0$ and $\frac{\mathsf{u}^k}{\lambda_k} \in \{0,1\}^n$ for all $k \in K_+$, hence all $\left(\frac{\mathsf{x}^k}{\lambda_k}, \frac{\mathsf{u}^k}{\lambda_k}\right)$ are (MIQP-soft(γ))-feasible, so

$$z_\gamma(\mathsf{Y}^*) \geq \min_{k \in K_+} \left[\frac{(\mathsf{x}^k)^\top}{\lambda_k} \mathsf{Q} \frac{\mathsf{x}^k}{\lambda_k} + 2\mathsf{c}^\top \frac{\mathsf{x}^k}{\lambda_k} - \lambda \mathsf{e}^\top \frac{\mathsf{u}^k}{\lambda_k} \right] \geq \zeta_\gamma^*\,,$$

and we get $\zeta_\gamma^* \leq z_\gamma^*$; the reverse inequality $\zeta_\gamma^* \geq z_\gamma^*$ follows by relaxation.

If $K_0 \neq \emptyset$ for $\mathsf{Y} = \mathsf{Y}^*$, then Lemma 16 ensures $\mathsf{u}^k = \mathsf{o}$ and $\mathsf{A}\mathsf{x}^k = \mathsf{o}$ for all $k \in K_0$. It follows that $\mathsf{x}(t) := \mathsf{x} + t\mathsf{x}^k$ is (MIQP-soft(γ))-feasible for all $t > 0$ and all $k \in K_0$, hence Assumption (3.1) guarantees $(\mathsf{x}^k)^\top \mathsf{Q} \mathsf{x}^k \geq 0$ for all $k \in K_0$. Recalling that

$$\mathsf{Y}^* = \sum_{k \in K_+} \beta_k \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \\ \frac{\mathsf{u}^k}{\lambda_k} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \\ \frac{\mathsf{u}^k}{\lambda_k} \end{pmatrix}^\top + \sum_{k \in K_0} \begin{pmatrix} 0 \\ \mathsf{x}^k \\ \mathsf{o} \end{pmatrix} \begin{pmatrix} 0 \\ \mathsf{x}^k \\ \mathsf{o} \end{pmatrix}^\top,$$

this implies

$$z_\gamma^* \approx z_\gamma(\mathsf{Y}^*) \geq z_\gamma(\bar{\mathsf{Y}}) \quad \text{with } \bar{\mathsf{Y}} = \sum_{k \in K_+} \beta_k \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \\ \frac{\mathsf{u}^k}{\lambda_k} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \\ \frac{\mathsf{u}^k}{\lambda_k} \end{pmatrix}^\top,$$

and $z_\gamma(\bar{\mathsf{Y}}) \geq \zeta_\gamma^*$ as discussed for the case $K_0 = \emptyset$ above. This completes the proof. □

[3] We need to argue by approximate optimality here since attainability of the optimal value of conic problems is not guaranteed *a priori*. But see Corollary 18 below.

Corollary 18. *The optimal value z_γ^* of* (CP-soft(γ)) *is attained at an optimal solution* Y^* *of the form*

$$\mathsf{Y}^* = \sum_{k\in K_+} \beta_k \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \\ \frac{\mathsf{u}^k}{\lambda_k} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \\ \frac{\mathsf{u}^k}{\lambda_k} \end{pmatrix}^\top,$$

where $\beta_k = \lambda_k^2 > 0$ *and* $\sum_{k\in K_+} \beta_k = 1$. *Moreover, we have*

$$\frac{\mathsf{u}^k}{\lambda_k} \in \{0,1\}^n,$$

and $u_i^k x_i^k = 0$ *for all* $i\in[1\!:\!n]$ *and* $k \in K_+$.

Proof. As argued in Theorem 17, there exists a nearly (CP-soft(γ))-optimal solution $\bar{\mathsf{Y}}$ of the form:

$$\bar{\mathsf{Y}} = \sum_{k\in K_+} \beta_k \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \\ \frac{\mathsf{u}^k}{\lambda_k} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \\ \frac{\mathsf{u}^k}{\lambda_k} \end{pmatrix}^\top,$$

where $\beta_k = \lambda_k^2 > 0$ and $\sum_{k\in K_+} \beta_k = 1$. Moreover, we have

$$\frac{\mathsf{u}^k}{\lambda_k} \in \{0,1\}^n,$$

and $u_i^k x_i^k = 0$ for all $i\in[1\!:\!n]$ and $k \in K_+$. In each component of the decomposition of $\bar{\mathsf{Y}}$, if the support of u_k is given, to find the corresponding x_k reduces to find an optimal solution of a QP, which is attainable. On the other hand, the number of supports of u^k is finite ($\leq 2^n$), which implies that the system of possible sets K_+ is finite.[4] Therefore, the claim follows. □

3.2.3. The decomposed CP reformulation

In this subsection, we propose another exact CP relaxation for (MIQP-hard) based on the sparse pattern of the problem.

We notice that the problem (CP-soft(γ)) has a sparse structure, namely, only the submatrices of the underlying variable Y are involved in both the objective function and the constraints. To be more precise, the specification graph of the variable Y involved in the problem (CP-soft(γ)) is chordal. It is known that a partial SDP matrix is SDP-completable if its specification graph is chordal, and a partial CP matrix is CP-completable if this graph is block-clique [SMB21]. Inspired by this fact, we consider the following decomposed CP/DNN relaxation:

$$\begin{aligned} \hat{z}_\gamma^* = \inf_{\mathsf{X},\mathsf{x},\mathsf{u}} \quad & \hat{z}_\gamma(\mathsf{X},\mathsf{x},\mathsf{u}) := \langle \mathsf{Q}, \mathsf{X}\rangle + 2\mathsf{c}^\top\mathsf{x} - \gamma \mathsf{e}^\top \mathsf{u} \\ \text{s.t.} \quad & \mathsf{A}\mathsf{x} = \mathsf{b}\,, \\ & \mathrm{diag}\,(\mathsf{A}\mathsf{X}\mathsf{A}^\top) = \mathsf{b}\circ\mathsf{b}\,, \\ & \begin{pmatrix} 1 & x_i & u_i \\ x_i & X_{ii} & 0 \\ u_i & 0 & u_i \end{pmatrix} \in \mathcal{DNN}^3, \quad \forall i\in[1\!:\!n], \\ & \begin{pmatrix} 1 & \mathsf{x}^\top \\ \mathsf{x} & \mathsf{X} \end{pmatrix} \in \mathcal{CP}^{n+1}. \end{aligned} \qquad \text{(CP-soft-decomposed}(\gamma)\text{)}$$

[4] Note that $|K_+|$ may exceed the CP-rank of $\bar{\mathsf{Y}}$ so that we cannot claim a tighter bound on $|K_+|$ like those in [BSU15]. In other words, decomposition (3.8) may not be a minimal CP decomposition of Y^*, so we cannot even apply Caratheodory's theorem. But fortunately only finiteness matters here.

Some comments related to (CP-soft-decomposed(γ)):

- (CP-soft-decomposed(γ)) is a direct consequence of a procedure that relaxes the cone constraint in (CP-soft(γ)) to its specified principal submatrices, which is known as chordal decomposition. Generally speaking, after the chordal decomposition for the CP constraint, the two relaxations are not necessarily equivalent in the sense of the objective value. However, we will show the equivalence in our cases.
- It is worth noticing that when $n \leq 4$, we have $\mathcal{DNN}^n = \mathcal{CP}^n$ [SMB21], hence we replace $\mathcal{CP}^3$ by $\mathcal{DNN}^3$ in the formulation. Note that this condition is equivalent to the two conditions $u_i \geq 0$ and $X_{ii}(1-u_i) \geq x_i^2$, given that $X_{ii} \geq 0$.
- In the decomposed CP relaxation, we notice that the dimension of the cone constraint reduces dramatically from $2n+1$ to $n+1$, meanwhile the complexity of the problem decreases from $O(2^{2n+1})$ to $O(2^{n+1})$.

In the following, we will show that the relaxation (CP-soft-decomposed(γ)) is also exact.

Theorem 19. *The CP relaxation* (CP-soft-decomposed(γ)) *is exact.*

Proof. In analogy to Lemma 16, we study the necessary condition of a point $(\mathsf{X}, \mathsf{x}, \mathsf{u})$ being (CP-soft-decomposed(γ))-optimal. First, similar to the proof of Lemma 16, for a given point $(\mathsf{X}, \mathsf{x}, \mathsf{u})$ which is (CP-soft-decomposed(γ))-feasible with

$$\begin{pmatrix} 1 & \mathsf{x} \\ \mathsf{x} & \mathsf{X} \end{pmatrix} = \sum_{k \in K_+} \beta_k \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \end{pmatrix}^\top + \sum_{k \in K_0} \begin{pmatrix} 0 \\ \mathsf{x}^k \end{pmatrix} \begin{pmatrix} 0 \\ \mathsf{x}^k \end{pmatrix}^\top,$$

we construct another (CP-soft-decomposed(γ))-feasible point $(\mathsf{X}, \mathsf{x}, \hat{\mathsf{u}})$ with

$$\hat{z}_\gamma(\mathsf{X}, \mathsf{x}, \hat{\mathsf{u}}) \leq \hat{z}_\gamma(\mathsf{X}, \mathsf{x}, \mathsf{u}) \tag{3.9}$$

by defining $\hat{u}_i^k := 1 - \text{sign}\,(x_i^k) \in \{0, 1\}$. Since $\hat{u}_i = \sum_{k \in K_+} \beta_k \hat{u}_i^k \geq u_i$, the claimed relation (3.9) follows. Furthermore, the following CP decomposition across all $i \in [1\!:\!n]$ holds:

$$\begin{pmatrix} 1 & x_i & \hat{u}_i \\ x_i & X_{ii} & 0 \\ \hat{u}_i & 0 & \hat{u}_i \end{pmatrix} = \sum_{k \in K_+} \beta_k \begin{pmatrix} 1 \\ \frac{x_i^k}{\lambda_k} \\ \hat{u}_i^k \end{pmatrix} \begin{pmatrix} 1 \\ \frac{x_i^k}{\lambda_k} \\ \hat{u}_i^k \end{pmatrix}^\top.$$

Here $\mathsf{x}^k \in \mathbb{R}^n_+$ for all $k \in K_+ \cup K_0$, which is a point of the projection of the feasible region of (CP-soft(γ)) onto $(\mathsf{X}, \mathsf{x}, \mathsf{u})$, which implies that $\hat{z}_\gamma^* \geq z_\gamma^*$. On the other hand, it is clear that the feasible region of (CP-soft-decomposed(γ)) contains the corresponding projected set of the feasible region of (CP-soft(γ)). Therefore, together with Theorem 17 we have

$$\hat{z}_\gamma^* = z_\gamma^* = \zeta_\gamma,$$

with which we complete the proof. □

Corollary 20. *The optimal value $\hat{z}_\gamma^*$ of* (CP-soft(γ)) *is attained at an optimal solution $(\mathsf{X}^*, \mathsf{x}^*, \mathsf{u}^*)$ of the form*

$$\begin{pmatrix} 1 & x_i & u_i \\ x_i & X_{ii} & 0 \\ u_i & 0 & u_i \end{pmatrix} = \sum_{k \in K_+} \beta_k \begin{pmatrix} 1 \\ \frac{x_i^k}{\lambda_k} \\ u_i^k \end{pmatrix} \begin{pmatrix} 1 \\ \frac{x_i^k}{\lambda_k} \\ u_i^k \end{pmatrix}^\top, \quad \forall i \in \in [1\!:\!n],$$

where $\beta_k = \lambda_k^2 > 0$ *and* $\sum_{k\in K_+} \beta_k = 1$. *Moreover, we have*

$$\mathsf{u}^k \in \{0,1\}^n,$$

and $u_i^k x_i^k = 0$ *for all* $i \in [1\!:\!n]$ *and* $k \in K_+$.

Proof. Directly from Theorem 19, there exists a nearly (CP-soft-decomposed(γ))-optimal solution having the properties stated above. As the rest is quite identical to the proof of Corollary 18, we omit it for the sake of brevity. □

3.3. CP reformulation for hard sparse QPs

In this section, we propose two compact CP relaxations of (MIQP-hard) and prove the exactness of the relaxations.

3.3.1. New and more compact CP reformulation for hard sparse QPs

Recall the hard sparse QP is of the form:

$$\begin{aligned} \ell_\rho^* := \min_{\mathsf{x}\in\mathbb{R}^n_+} \quad & \mathsf{x}^\top \mathsf{Q}\mathsf{x} + 2\mathsf{c}^\top \mathsf{x} \\ \text{s.t.} \quad & \mathsf{A}\mathsf{x} = \mathsf{b}\,, \\ & \|\mathsf{x}\|_0 \le \rho\,, \end{aligned} \tag{Hard-QP}$$

where $\rho \in [0, n]$ is a parameter restricting the number of nonzero elements of x. Notice that we only assume the feasible region of (QP) is nonempty as shown in (3.1), therefore it may happen for small ρ, the problem (Hard-QP) is infeasible. Indeed, if the feasible region of (Hard-QP) is nonempty for $\rho = \mu \in [0, n]$, then the set remains nonempty for $\rho \in [\mu, n]$. To accommodate the above case, by $\underline{\rho} \in \in [0\!:\!n]$ we define the l_0-pseudonorm of the sparsest feasible point of the problem (QP). It follows immediately that the problem (Hard-QP) is feasible for $\rho \in [\underline{\rho}, n]$. As explained above, for all these ρ, the optimal solution of (Hard-QP) is attainable due to the Frank/Wolfe Theorem.

Proposition 21. *For a given optimization problem* (Hard-QP)*, we have* $\ell^*_{\rho_1} \ge \ell^*_{\rho_2}$ *if* $\rho_1 \le \rho_2$.

Proof. The claim follows since the feasible region of (Hard-QP) enlarges when the parameter ρ increases. By default, if an optimization problem is infeasible, the optimal value is $+\infty$. So the relation remains true even if $\rho_1 < \underline{\rho}$. □

Before loading to the CP reformulation, we present a useful property on the optimal value of (Hard-QP) when the sparsity parameter ρ varies: the objective value is not only decreasing, but even convex decreasing in ρ.

Lemma 22. *The problem* (Hard-QP) *has the property:*

$$\alpha\ell^*_\rho + (1-\alpha)\ell^*_{\rho'} \ge \ell^*_{\alpha\rho+(1-\alpha)\rho'} \quad \textit{for all } \{\rho,\rho'\} \subset [0,n] \textit{ and } \alpha \in [0,1]\,.$$

Proof. In the case of $\min\{\rho,\rho'\} < \underline{\rho}$, it is clear that the $\max\{\ell^*_\rho, \ell^*_{\rho'}\} = +\infty$, therefore

$$\alpha\ell^*_\rho + (1-\alpha)\ell^*_{\rho'} \ge \ell^*_{\alpha\rho+(1-\alpha)\rho'}$$

holds for all $\alpha \in [0,1]$.

In the following, we show the property when $\min\{\rho, \rho'\} \geq \underline{\rho}$. First, by x^*_ρ we denote an optimal solution to the problem (Hard-QP) and consider any strictly positive vector $\mathsf{d} \in int\ \mathbb{R}^n_+$. Letting

$$h := \max_{\rho \in [\underline{\rho}:n]} \mathsf{d}^\top \mathsf{x}^*_\rho = \max_{\rho \in [\underline{\rho},n]} \mathsf{d}^\top \mathsf{x}^*_\rho \,,$$

we add the redundant constraint $\mathsf{d}^\top \mathsf{x} + t = h$ to (Hard-QP):

$$\begin{array}{rll} \ell^*_\rho = \min\limits_{\mathsf{x} \in \mathbb{R}^n_+, t \in \mathbb{R}_+} & \mathsf{x}^\top \mathsf{Q} \mathsf{x} + 2\mathsf{c}^\top \mathsf{x} & \\ \text{s.t.} & \mathsf{A}\mathsf{x} = \mathsf{b}\,, & \\ & \mathsf{d}^\top \mathsf{x} + t = h\,, & \\ & \|\mathsf{x}\|_0 \leq \rho\,. & \end{array} \tag{3.10}$$

Due to the construction of h, it is clear that the problem (3.10) is equivalent to the problem (Hard-QP) for all $\rho \in [\underline{\rho}, n]$. By introducing auxiliary variables and slack variables $\{\mathsf{u}, \mathsf{s}\} \subset \mathbb{R}^n_+$, the problem (3.10) can be rewritten:

$$\begin{array}{rll} \ell^*_\rho = \min\limits_{\{\mathsf{x},\mathsf{u},\mathsf{s}\} \subset \mathbb{R}^n_+, t \in \mathbb{R}_+} & \mathsf{x}^\top \mathsf{Q} \mathsf{x} + 2\mathsf{c}^\top \mathsf{x} & \\ \text{s.t.} & \mathsf{A}\mathsf{x} = \mathsf{b}\,, & \\ & \mathsf{d}^\top \mathsf{x} + t = h\,, & \\ & \mathsf{e}^\top \mathsf{u} = n - \rho\,, & \\ & \mathsf{u} + \mathsf{s} = \mathsf{e}\,, & \\ & \mathsf{x}^\top \mathsf{u} = \mathsf{o}\,. & \end{array} \tag{3.11}$$

Since the feasible region of problem (3.11) is bounded, it satisfies the condition ensuring exactness of the CP relaxation proposed in [Bur09], it admits an exact CP reformulation of the form

$$\begin{array}{rll} \ell^*_\rho = \inf\limits_{\mathsf{Y}} & \langle \mathsf{Q}, \mathsf{Y}^{\mathsf{xx}} \rangle + \mathsf{c}^\top \mathsf{x} & \\ \text{s.t.} & \mathsf{Y}^{00} = 1\,, & [y_1] \\ & \mathsf{A}\mathsf{x} = \mathsf{b}\,, & [\mathsf{z}_1] \\ & \mathsf{d}^\top \mathsf{x} + t = h\,, & [y_2] \\ & \mathsf{e}^\top \mathsf{u} = n - \rho\,, & [y_3] \\ & \mathsf{u} + \mathsf{s} = \mathsf{e}\,, & [\mathsf{z}_2] \\ & \text{diag}\,(\mathsf{A}\mathsf{Y}^{\mathsf{xx}}\mathsf{A}^\top) = \mathsf{b} \circ \mathsf{b}\,, & [\mathsf{z}_3] \\ & \langle \mathsf{d}\mathsf{d}^T, \mathsf{Y}^{\mathsf{xx}} \rangle + 2\mathsf{d}^\top \mathsf{Y}^{\mathsf{x}t} + \mathsf{Y}^{tt} = h^2\,, & [y_4] \\ & \text{diag}\,(\mathsf{Y}^{\mathsf{uu}} + 2\mathsf{Y}^{\mathsf{us}} + \mathsf{Y}^{\mathsf{ss}}) = \mathsf{e}\,, & [\mathsf{z}_4] \\ & \mathsf{e}^\top \text{diag}\,(\mathsf{Y}^{\mathsf{xu}}) = 0\,, & [y_5] \\ & \langle \mathsf{E}, \mathsf{Y}^{\mathsf{uu}} \rangle = (n - \rho)^2\,, & [y_6] \\ & \mathsf{Y} := \begin{pmatrix} \mathsf{Y}^{00} & \mathsf{x}^\top & \mathsf{u}^\top & \mathsf{s}^\top & t \\ \mathsf{x} & \mathsf{Y}^{\mathsf{xx}} & \mathsf{Y}^{\mathsf{xu}} & \mathsf{Y}^{\mathsf{xs}} & \mathsf{Y}^{\mathsf{x}t} \\ \mathsf{u} & (\mathsf{Y}^{\mathsf{xu}})^\top & \mathsf{Y}^{\mathsf{uu}} & \mathsf{Y}^{\mathsf{us}} & \mathsf{Y}^{\mathsf{u}t} \\ \mathsf{s} & (\mathsf{Y}^{\mathsf{xs}})^\top & (\mathsf{Y}^{\mathsf{us}})^\top & \mathsf{Y}^{\mathsf{ss}} & \mathsf{Y}^{\mathsf{s}t} \\ t & (\mathsf{Y}^{\mathsf{x}t})^\top & (\mathsf{Y}^{\mathsf{u}t})^\top & (\mathsf{Y}^{\mathsf{s}t})^\top & \mathsf{Y}^{tt} \end{pmatrix} \in \mathcal{CP}^{3n+2}\,, & \end{array} \tag{3.12}$$

in which the associated dual variable for each constraint is introduced on the right. Notice that $\mathsf{z}_i \in \mathbb{R}^n$ for all $i \in [1\!:\!4]$ and $y_j \in \mathbb{R}$ for all $j \in [1\!:\!6]$. Further observe that the equality constraint $\langle \mathsf{E}, \mathsf{Y}^{\mathsf{uu}} \rangle = (n-\rho)^2$ can be equivalently relaxed to the inequality $\langle \mathsf{E}, \mathsf{Y}^{\mathsf{uu}} \rangle \leq (n-\rho)^2$. To show this, for any (3.12)-feasible Y^{uu}, we notice that $\mathsf{Y}^{\mathsf{uu}} \succeq \mathsf{u}\mathsf{u}^\top$ and $\mathsf{e}^\top \mathsf{u} = n - \rho$, which implies that $\langle \mathsf{E}, \mathsf{Y}^{\mathsf{uu}} \rangle \geq (n-\rho)^2$. Therefore, such replacement does not change the feasible set of (3.12). An immediate implication is that the dual variable y_6 has to be non-negative.

The conic dual of (3.12) reads as

$$\begin{array}{rll} d_\rho^* := \sup\limits_{\mathsf{S}} & (n-\rho) y_3 + (n-\rho)^2 y_6 + f(y_i, \mathsf{z}_j) & \\ s.t. & y_6 \geq 0 & \\ & \mathsf{S}(y_i, \mathsf{z}_j) \in \mathcal{COP}^{3n+2}\,, & \end{array} \tag{3.13}$$

where $f(y_i, \mathsf{z}_j) := y_1 + \mathsf{b}^\top \mathsf{z}_1 + h y_2 + \mathsf{e}^\top \mathsf{z}_2 + (\mathsf{b} \circ \mathsf{b})^\top \mathsf{z}_3 + h^2 y_4 + \mathsf{e}^\top \mathsf{z}_4$ and the upper triangular part of the symmetric slack matrix $\mathsf{S}(y_i, \mathsf{z}_j)$ is defined as an affine function of all dual variables (y_i, z_j) as follows:

$$\begin{pmatrix} -2y_1 & (\mathsf{c} - \mathsf{A}^\top \mathsf{z}_1 - y_2 \mathsf{d})^\top & -(y_3 \mathsf{e} + \mathsf{z}_2)^\top & -\mathsf{z}_2^\top & -y_2 \\ * & 2\,\mathsf{B}(y_4, \mathsf{z}_3) & -y_5 \mathsf{I} & \mathsf{O} & -2y_4 \mathsf{d} \\ * & * & -2(\mathrm{Diag}\,(\mathsf{z}_4) + y_6 \mathsf{E}) & -\mathrm{Diag}\,(\mathsf{z}_4) & \mathsf{o} \\ * & * & * & -2\mathrm{Diag}\,(\mathsf{z}_4) & \mathsf{o} \\ * & * & * & * & -2y_4 \end{pmatrix},$$

with $\mathsf{B}(y_4, \mathsf{z}_3) := \mathsf{Q} - \mathrm{Diag}\,(\mathsf{z}_3)\mathsf{A}\mathsf{A}^\top - y_4 \mathsf{d}\mathsf{d}^\top$. Let

$$\bar{y}_2 < \min\{0\} \cup \{\frac{c_i}{d_i} : i \in [1\!:\!n]\} \quad \text{and} \quad \bar{y}_4 < \min\{0\} \cup \{\frac{Q_{ij}}{d_i d_j} : (i,j) \in \in [1\!:\!n]^2\}\,.$$

Furthermore, let $\bar{y}_1 = -0.5$, $\bar{y}_3 = \bar{y}_5 = \bar{y}_6 0$, $\bar{\mathsf{z}}_1 = \bar{\mathsf{z}}_2 = \bar{\mathsf{z}}_3 = \mathsf{o}$ while $\bar{\mathsf{z}}_4 = -\mathsf{e}$. Then the slack matrix $\bar{\mathsf{S}} := \mathsf{S}(\bar{y}_i, \bar{\mathsf{z}}_j)$ generates a quadratic form in $[r, \mathsf{x}^\top, \mathsf{u}^\top, \mathsf{s}^\top, t]^\top \in \mathbb{R}^{3n+2}_+ \setminus \{\mathsf{o}\}$ satisfying

$$[r, \mathsf{x}^\top, \mathsf{u}^\top, \mathsf{s}^\top, t] \bar{\mathsf{S}} [r, \mathsf{x}^\top, \mathsf{u}^\top, \mathsf{s}^\top, t]^\top \geq r^2 + 2\left(\mathsf{x}^\top \mathsf{B}(\bar{y}_4, \bar{\mathsf{z}}_3)\mathsf{x} + \|\mathsf{u}\|^2 + \|\mathsf{s}\|^2\right) + t^2\,,$$

which implies that $\bar{\mathsf{S}}$ is a strictly copositive matrix, since $\mathsf{B}(\bar{y}_4, \bar{\mathsf{z}}_3) = \mathsf{Q} - \bar{y}_4 \mathsf{d}\mathsf{d}^\top$ is so, having only positive entries by choice of $\bar{y}_4$. Therefore strong duality holds for the pair (3.12) and (3.13): $\ell_\rho^* = d_\rho^*$. But by definition in (3.13), the function d_ρ^* is convex in ρ as a pointwise supremum of convex functions. We conclude that ℓ_ρ^* is also convex in ρ, which completes the proof. □

Now we turn to the mixed-integer reformulation of (Hard-QP). Again, by introducing auxiliary (now binary) variables u, we arrive at the reformulation:

$$\begin{array}{rll} \ell_\rho^* = \min\limits_{\mathsf{x} \in \mathbb{R}^n_+, \mathsf{u} \in \{0,1\}^n} & \mathsf{x}^\top \mathsf{Q} \mathsf{x} + \mathsf{c}^\top \mathsf{x} & \\ \text{s.t.} & \mathsf{A}\mathsf{x} = \mathsf{b}\,, & \\ & \mathsf{e}^\top \mathsf{u} = n - \rho\,, & \\ & \mathsf{x}^\top \mathsf{u} = 0\,. & \end{array} \tag{MIQP-hard}$$

Consider the following CP relaxation of (MIQP-hard):

$$\begin{aligned}
\ell_\rho^{\mathrm{CP}} := \inf \quad & \langle \mathsf{Q}, \mathsf{Z}^{\mathsf{xx}} \rangle + \mathsf{c}^\top \mathsf{x} \\
\text{s.t.} \quad & \mathsf{A}\mathsf{x} = \mathsf{b} \\
& \mathrm{diag}\,(\mathsf{A}\mathsf{Z}^{\mathsf{xx}}\mathsf{A}^\top) = \mathsf{b} \circ \mathsf{b}\,, \\
& \mathsf{e}^\top \mathsf{u} = n - \rho \\
& \mathsf{e}^\top \mathrm{diag}\,(\mathsf{Z}^{\mathsf{xu}}) = 0 \\
& \mathrm{diag}\,(\mathsf{Z}^{\mathsf{uu}}) = \mathsf{u} \\
& \mathsf{Z} := \begin{pmatrix} 1 & \mathsf{x}^\top & \mathsf{u}^\top \\ \mathsf{x} & \mathsf{Z}^{\mathsf{xx}} & \mathsf{Z}^{\mathsf{xu}} \\ \mathsf{u} & (\mathsf{Z}^{\mathsf{xu}})^\top & \mathsf{Z}^{\mathsf{uu}} \end{pmatrix} \in \mathcal{CP}^{2n+1}.
\end{aligned} \tag{CP-hard(ρ)}$$

We notice that the relaxation has fewer variables, with only $3n+2$ linear constraints and a much smaller conic one, compared to CP reformulations in [BP23, Bur09].

In the following part, we will show, under the assumption (3.1), the equivalence between (MIQP-hard) and (CP-hard(ρ)). First of all, we establish an important result similar to Lemma 15.

Lemma 23. *For any* (CP-hard(ρ))*-feasible point* Z*, the following decomposition holds:*

$$\mathsf{Z} = \begin{pmatrix} 1 & \mathsf{x}^\top & \mathsf{u}^\top \\ \mathsf{x} & \mathsf{Z}^{\mathsf{xx}} & \mathsf{Z}^{\mathsf{xu}} \\ \mathsf{u} & (\mathsf{Z}^{\mathsf{xu}})^\top & \mathsf{Z}^{\mathsf{uu}} \end{pmatrix} = \sum_{k \in K_+} \beta_k \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \\ \frac{\mathsf{u}^k}{\lambda_k} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{\mathsf{x}^k}{\lambda_k} \\ \frac{\mathsf{u}^k}{\lambda_k} \end{pmatrix}^\top + \sum_{k \in K_0} \begin{pmatrix} 0 \\ \mathsf{x}^k \\ \mathsf{u}^k \end{pmatrix} \begin{pmatrix} 0 \\ \mathsf{x}^k \\ \mathsf{u}^k \end{pmatrix}^\top,$$

where $\beta_k = \lambda_k^2 > 0$ *for all* $k \in K_+$ *and* $\sum_{k \in K_+} \beta_k = 1$*. Furthermore, we have* $\sum_{k \in K_+} \beta_k \|\mathsf{x}^k\|_0 \le \rho$ *and* $\mathsf{u}^k = \mathsf{o}$ *for all* $k \in K_0$*.*

Proof. For a given (CP-hard(ρ))-feasible point Z, there is always a CP decomposition of the form (3.8). Therefore, we can present a feasible point either by Z or $(\lambda_k, \mathsf{x}^k, \mathsf{u}^k)_{k \in K}$. Let us fix Z^{xx} and x (or equivalently λ_k and x^k for all $k \in K$), and try to find the necessary condition on u^k for the point $(\lambda_k, \mathsf{x}^k, \mathsf{u}^k)_{k \in K}$ being (CP-hard(ρ))-feasible. Again, we need some index sets which play a significant role in the proof: for all $i \in [1\!:\!n]$, consider

$$P_i = \{k \in K_+ : x_i^k > 0\} \quad \text{and} \quad N_i = K_+ \setminus P_i = \{k \in K_+ : x_i^k = 0\}\,.$$

Now consider the following optimization problem:

$$\begin{aligned}
\ell := \max_{\mathsf{u}^k \in \mathbb{R}^n_+} \quad & \sum_{k \in K_+} \lambda_k \mathsf{e}^\top \mathsf{u}^k \\
\text{s.t.} \quad & \sum_{k \in K} (u_i^k)^2 = \sum_{k \in K_+} \lambda_k u_i^k, \quad \forall i \in [1\!:\!n]\,, \\
& u_i^k x_i^k = 0, \quad \forall i \in [1\!:\!n], \quad \forall k \in K\,.
\end{aligned} \tag{3.14}$$

It is clear that one necessary condition of the point $(\lambda_k, \mathsf{x}^k, \mathsf{u}^k)_{k \in K}$ being (CP-hard(ρ))-feasible is that the optimal value of (3.14) satisfying

$$\ell \ge n - \rho. \tag{3.15}$$

Due to the fully decomposed structure in both the objective and the constraints regarding u^k, and by the defined sets P_i and N_i, (3.14) can be equivalently splitted into a set of more simple problems with index $i\in[1:n]$:

$$\begin{aligned} \max_{u_i^k\in\mathbb{R}_+} \quad & \sum_{k\in K_+} \lambda_k u_i^k \\ \text{s.t.} \quad & \sum_{k\in N_i} (u_i^k)^2 + \sum_{k\in K_0} (u_i^k)^2 = \sum_{k\in K_+} \lambda_k u_i^k \,. \end{aligned} \tag{3.16}$$

It follows from Lemma 15 that the optimal solution to the problem (3.16) is

$$(u_i^k)^* = \begin{cases} \lambda_k, & \text{if } k \in N_i; \\ 0, & \text{else.} \end{cases}$$

The optimal value of (3.16) is

$$\ell = \sum_{k\in K_+} \beta_k \mathsf{e}^\top \left(\frac{(\mathsf{u}^k)^*}{\lambda_k} \right) = \sum_{k\in K_+} \beta_k (n - \|\mathsf{x}^k\|_0) = n - \sum_{k\in K_+} \beta_k \|\mathsf{x}^k\|_0.$$

Together with (3.15), we have $\sum_{k\in K_+} \beta_k \|\mathsf{x}^k\|_0 \le \rho$, which is exactly the first half of the statement. The second half follows immediately from the structure of the optimal solution to (3.16). i.e., $(u_i^k)^* = 0$ for all $k \in K_0$ and $i\in[1:n]$. □

Theorem 24. *Under the assumption* (3.1)*, the CP relaxation* (CP-hard(ρ)) *is equivalent to the problem* (MIQP-hard)*.*

Proof. It is clear that the problem (CP-hard(ρ)) is a relaxation of the problem (MIQP-hard), hence

$$\ell_\rho^* \ge \ell_\rho^{\mathrm{CP}} \,.$$

To show the converse, we take a (CP-hard(ρ)) feasible point Z and compare the objective values between (CP-hard(ρ)) and (MIQP-hard). Let us consider

$$\begin{aligned} & \langle \mathsf{Q}, \mathsf{Z}^{\mathsf{xx}} \rangle \\ \overset{(3.8)}{=} \quad & \sum_{k\in K_+} \beta_k (\frac{\mathsf{x}^k}{\lambda_k})^\top \mathsf{Q} \frac{\mathsf{x}^k}{\lambda_k} \\ \ge \quad & \sum_{k\in K_+} \beta_k \ell^*_{\|\mathsf{x}^k\|_0} \\ \ge \quad & \ell^*_{\sum_{k\in K_+} \beta_k \|\mathsf{x}^k\|_0} \\ \ge \quad & \ell_\rho^*, \end{aligned} \tag{3.17}$$

where the last inequality follows from Lemma 23 and Proposition 21, while the previous one follows from Lemma 22. So (3.17) holds which implies $\ell_\rho^{\mathrm{CP}} \ge \ell_\rho^*$. □

Corollary 25. *The optimal solution of* (CP-hard(ρ)) *is attained.*

Proof. Directly from Theorem 24 via the Frank/Wolfe theorem. □

3.4. Conclusion

In this chapter, we introduced new, significantly more compact CP reformulations for QPs with a sparsity term under mild conditions. By leveraging the sparse structure of these CP reformulations, we also proved the equivalence of their decomposed forms, demonstrating that the complexity of the problems increases only mildly with dimension. These compact relaxations facilitate the development of tight yet tractable relaxations in the area of sparse QPs.

4. Conic relaxations for sparse StQPs

In this chapter, we consider standard quadratic optimization problems (StQPs) under a hard sparsity constraint and propose novel relaxations based on the corresponding CP reformulations with significant dimensional reduction, which is essential for the tractability of these relaxations when the problem size grows. Moreover, an instance-generation procedure of StQPs is proposed, which systematically avoids too easy instances. Extensive computational results illustrate the high quality of the relaxation bounds in a significant number of instances. The content of this chapter is based on the submitted paper [BPQY24]. A precursor of the material presented in the previous chapter, which also complements the study in this chapter, is published as [BPQY25].

4.1. The problem

An StQP consists of minimizing a (not necessarily convex) quadratic form over the standard simplex:

$$\ell_n(\mathsf{Q}) := \min_{\mathsf{x}\in\mathbb{R}^n} \left\{ \mathsf{x}^\top \mathsf{Q}\mathsf{x} : \mathsf{x} \in F_n \right\}, \tag{StQP}$$

where

$$F_n := \left\{ \mathsf{x} \in \mathbb{R}^n_+ : \mathsf{e}^\top \mathsf{x} = 1 \right\} \tag{4.1}$$

is the standard simplex, the simplest polytope in the n-dimensional Euclidean space $\mathbb{R}^n$. Here, $\mathsf{e} \in \mathbb{R}^n$ denotes the vector of all ones.

Despite its simplicity, this problem class serves to model manifold real-world applications, ranging from the analysis of social networks (community detection via dominant-set-clustering) [BKL21, MS65] through biology, game theory to economy and finance [Bom02, Mar52], to cite just a few. For a survey on mixed-integer convex quadratic optimization approaches to portfolio selection, see, e.g., [MD19].

One motivation to introduce sparsity constraints comes from high-dimensional portfolio selection. Recently, the significant benefits of controlling the cardinality of a portfolio have been widely recognized in the literature, e.g., [DGW23], [HV19].

Therefore, we propose to study the StQP with an exact, hard sparsity constraint, with the aim of developing tight yet tractable relaxations. Such a problem can be expressed as follows:

$$\ell_\rho(\mathsf{Q}) := \min_{\mathsf{x}\in\mathbb{R}^n} \left\{ \mathsf{x}^\top \mathsf{Q}\mathsf{x} : \mathsf{x} \in F_\rho \right\}, \tag{StQP(ρ)}$$

where

$$F_\rho := \{\mathsf{x} \in F_n : \|\mathsf{x}\|_0 \le \rho\} = \left\{ \mathsf{x} \in \mathbb{R}^n_+ : \mathsf{e}^\top \mathsf{x} = 1, \quad \|\mathsf{x}\|_0 \le \rho \right\}. \tag{4.2}$$

The sparse StQP problem appears in numerous applications, notably in portfolio selection with asset limits [Bie96, Per84, XTYP24]. `MATLAB`'s latest release [The24] includes a MILP framework to solve MIQP formulations of convex portfolio problems, handling lower and upper cardinality bounds and semi-continuous variables. However, tests show that even for $n = 30$ and extreme sparsity $\rho = 1$, solution time was around 40 seconds, and no solution was found for $\rho = 2$ within

two hours. Additional applications include selecting regression variables [Mil84] and compressive sampling [CW08].

The sparse StQP problem has been studied extensively and can be categorized into two main approaches. The first addresses sparsity constraints indirectly through regularization methods such as the ℓ_1-norm or the ℓ_p-pseudonorms with $p \in (0,1)$ [BDDM$^+$09, CPZ13]. The second approach directly tackles cardinality constraints using various methods. Semidefinite programming-based reformulations for generating perspective cuts have been explored in [FG07, ZSL14], while nonconvex reformulations using regularization methods are presented in [BKS16]. Heuristic methods based on semidefinite programming that provide upper bounds for the sparse StQP problem are proposed in [BM05], and tight semidefinite relaxations under convexity assumptions are discussed in [WZ21]. Additionally, mixed-integer formulations are solved through discrete optimization techniques, either to optimality or as approximations [BS09, Bie96, DLLR$^+$12, MS12, RTGMS10, SZL13, ZSLS14]. In this work, our focus is on finding tight but tractable relaxations that scale well with the problem dimension.

4.2. Two exact completely positive formulations

In this section, we present two exact formulations of the sparse StQP problem (StQP(ρ)) as mixed-binary quadratic optimization problems, a direct one covered by Section 4.2.1, whereas Section 4.2.2 is devoted to a formulation involving a complementarity constraint.

4.2.1. Conic formulation of direct MIQP

By introducing binary variables, the sparse StQP can be reformulated as the following mixed-binary quadratic optimization problem:

$$\begin{aligned} \ell_\rho(\mathsf{Q}) = \min_{(\mathsf{x},\mathsf{u})\in\mathbb{R}^n\times\mathbb{R}^n} \quad & \mathsf{x}^\top\mathsf{Q}\mathsf{x} \\ \text{s.t.} \quad & \mathsf{e}^\top\mathsf{x} = 1 \\ & \mathsf{e}^\top\mathsf{u} = \rho \\ & \mathsf{x} \le \mathsf{u} \\ & \mathsf{u} \in \{0,1\}^n \\ & \mathsf{x} \ge \mathsf{o}\,. \end{aligned} \tag{P1(ρ)}$$

By introducing the redundant constraints $\mathsf{u} \le \mathsf{e}$, slack variables $\mathsf{y} \ge \mathsf{o}$ and $\mathsf{v} \ge \mathsf{o}$ so that $\mathsf{x}+\mathsf{y} = \mathsf{u}$ and $\mathsf{u}+\mathsf{v} = \mathsf{e}$, we arrive at the following reformulation of (P1(ρ)):

$$\begin{aligned} \ell_\rho(\mathsf{Q}) = \min_{(\mathsf{x},\mathsf{u},\mathsf{v},\mathsf{y})\in\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{R}^n} \quad & \mathsf{x}^\top\mathsf{Q}\mathsf{x} \\ \text{s.t.} \quad & \mathsf{e}^\top\mathsf{x} = 1 \\ & \mathsf{e}^\top\mathsf{u} = \rho \\ & \mathsf{x}+\mathsf{y} = \mathsf{u} \\ & \mathsf{u}+\mathsf{v} = \mathsf{e} \\ & \mathsf{u} \in \{0,1\}^n \\ & \mathsf{x},\mathsf{y},\mathsf{z} \ge \mathsf{o} \end{aligned} \tag{P1A(ρ)}$$

By [Bur09], (P1A(ρ)) admits an equivalent reformulation as the following conic optimization problem:

$$
\begin{aligned}
\ell_\rho(\mathsf{Q}) = \min_{\mathsf{Z}\in\mathcal{S}^{4n+1}} \quad & \langle \mathsf{Q}, \mathsf{Z}^{\mathsf{xx}}\rangle \\
\text{s.t.} \quad & \mathsf{e}^\top \mathsf{x} = 1 \\
& \mathsf{e}^\top \mathsf{u} = \rho \\
& \mathsf{x} + \mathsf{y} = \mathsf{u} \\
& \mathsf{u} + \mathsf{v} = \mathsf{e} \\
& \langle \mathsf{E}, \mathsf{Z}^{\mathsf{xx}}\rangle = 1 \\
& \langle \mathsf{E}, \mathsf{Z}^{\mathsf{uu}}\rangle = \rho^2 \\
& \operatorname{diag}(\mathsf{Z}^{\mathsf{uu}}) = \mathsf{u} \\
& \operatorname{diag}(\mathsf{Z}^{\mathsf{xx}}) + \operatorname{diag}(\mathsf{Z}^{\mathsf{yy}}) + \operatorname{diag}(\mathsf{Z}^{\mathsf{uu}}) + 2\left[\operatorname{diag}(\mathsf{Z}^{\mathsf{xy}}) - \operatorname{diag}(\mathsf{Z}^{\mathsf{xu}}) - \operatorname{diag}(\mathsf{Z}^{\mathsf{uy}})\right] = \mathsf{o} \\
& \operatorname{diag}(\mathsf{Z}^{\mathsf{uu}}) + 2\operatorname{diag}(\mathsf{Z}^{\mathsf{uv}}) + \operatorname{diag}(\mathsf{Z}^{\mathsf{vv}}) = \mathsf{e} \\
& \mathsf{Z} := \begin{pmatrix} 1 & \mathsf{x}^\top & \mathsf{u}^\top & \mathsf{v}^\top & \mathsf{y}^\top \\ \mathsf{x} & \mathsf{Z}^{\mathsf{xx}} & \mathsf{Z}^{\mathsf{xu}} & \mathsf{Z}^{\mathsf{xv}} & \mathsf{Z}^{\mathsf{xy}} \\ \mathsf{u} & (\mathsf{Z}^{\mathsf{xu}})^\top & \mathsf{Z}^{\mathsf{uu}} & \mathsf{Z}^{\mathsf{uv}} & \mathsf{Z}^{\mathsf{uy}} \\ \mathsf{v} & (\mathsf{Z}^{\mathsf{xv}})^\top & (\mathsf{Z}^{\mathsf{uv}})^\top & \mathsf{Z}^{\mathsf{vv}} & \mathsf{Z}^{\mathsf{vy}} \\ \mathsf{y} & (\mathsf{Z}^{\mathsf{xy}})^\top & (\mathsf{Z}^{\mathsf{uy}})^\top & (\mathsf{Z}^{\mathsf{vy}})^\top & \mathsf{Z}^{\mathsf{yy}} \end{pmatrix} \in \mathcal{CP}^{4n+1}.
\end{aligned}
\tag{CP1A(ρ)}
$$

By [Bur09], (P1A(ρ)) and (CP1A(ρ)) are equivalent in the following sense: Both of their optimal values are equal to $\ell_\rho(\mathsf{Q})$ and for any optimal solution $\mathsf{Z} \in \mathcal{S}_+^{4n+1}$ of (CP1A(ρ)), $(\mathsf{x}, \mathsf{u}, \mathsf{y}, \mathsf{v}) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$ is in the convex hull of the set of optimal solutions of (P1A(ρ)). Therefore, (CP1A(ρ)) is an exact convex reformulation of the nonconvex optimization problem (P1A(ρ)).

4.2.2. Conic formulation of MIQP with complementarity constraint

In (P1(ρ)), sparsity constraints are handled by big-M constraints. Alternatively, sparsity can be enforced by dropping the big-M constraints $\mathsf{x} \le \mathsf{u}$ and replacing them with the complementarity constraints $\mathsf{x}_j(1 - \mathsf{u}_j) = 0,\ j = 1, \ldots, n$ (see, e.g., [BKS16]). Defining $\mathsf{v} = \mathsf{e} - \mathsf{u} \in \{0,1\}^n$, we obtain the following quadratic optimization problem with complementarity constraints:

$$
\begin{aligned}
\ell_\rho(Q) = \min_{(\mathsf{x},\mathsf{v})\in\mathbb{R}^n\times\mathbb{R}^n} \quad & \mathsf{x}^\top \mathsf{Q} \mathsf{x} \\
\text{s.t.} \quad & \mathsf{e}^\top \mathsf{x} = 1 \\
& \mathsf{e}^\top \mathsf{v} = n - \rho \\
& \mathsf{x}^\top \mathsf{v} = 0 \\
& \mathsf{v} \in \{0,1\}^n \\
& \mathsf{x} \ge \mathsf{o}.
\end{aligned}
\tag{P2(ρ)}
$$

Similarly, adding the redundant constraints $\mathsf{v} \le \mathsf{e}$ and reintroducing the slack variables $\mathsf{u} \ge \mathsf{o}$

so that $\mathsf{u}+\mathsf{v}=\mathsf{e}$, (P2($\rho$)) can be reformulated as follows:

$$\begin{aligned}\ell_\rho(\mathsf{Q}) = \min_{(\mathsf{x},\mathsf{u},\mathsf{v})\in\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{R}^n} \quad & \mathsf{x}^\top\mathsf{Q}\mathsf{x} \\ \text{s.t.} \quad & \mathsf{e}^\top\mathsf{x}=1 \\ & \mathsf{e}^\top\mathsf{u}=\rho \\ & \mathsf{u}+\mathsf{v}=\mathsf{e} \\ & \mathsf{x}^\top\mathsf{v}=0 \\ & \mathsf{u}\in\{0,1\}^n \\ & \mathsf{x},\mathsf{v},\mathsf{u}\geq\mathsf{o}\,.\end{aligned} \tag{P2A(ρ)}$$

By [BP23, Bur09], (P2A(ρ)) admits an equivalent reformulation as the following copositive optimization problem:

$$\begin{aligned}\ell_\rho(\mathsf{Q}) = \min_{\mathsf{Y}\in\mathcal{S}^{3n+1}} \quad & \langle\mathsf{Q},\mathsf{Y}^{\mathsf{xx}}\rangle \\ \text{s.t.} \quad & \mathsf{e}^\top\mathsf{x}=1 \\ & \mathsf{e}^\top\mathsf{u}=\rho \\ & \mathsf{u}+\mathsf{v}=\mathsf{e} \\ & \langle\mathsf{E},\mathsf{Y}^{\mathsf{xx}}\rangle=1 \\ & \langle\mathsf{E},\mathsf{Y}^{\mathsf{uu}}\rangle=\rho^2 \\ & \operatorname{diag}(\mathsf{Y}^{\mathsf{uu}})=\mathsf{u} \\ & \operatorname{diag}(\mathsf{Y}^{\mathsf{uu}})+2\operatorname{diag}(\mathsf{Y}^{\mathsf{uv}})+\operatorname{diag}(\mathsf{Y}^{\mathsf{vv}})=\mathsf{e} \\ & \mathsf{e}^\top\operatorname{diag}(\mathsf{Y}^{\mathsf{xv}})=0 \\ & \mathsf{Y}:=\begin{pmatrix}1 & \mathsf{x}^\top & \mathsf{u}^\top & \mathsf{v}^\top \\ \mathsf{x} & \mathsf{Y}^{\mathsf{xx}} & \mathsf{Y}^{\mathsf{xu}} & \mathsf{Y}^{\mathsf{xv}} \\ \mathsf{u} & (\mathsf{Y}^{\mathsf{xu}})^\top & \mathsf{Y}^{\mathsf{uu}} & \mathsf{Y}^{\mathsf{uv}} \\ \mathsf{v} & (\mathsf{Y}^{\mathsf{xv}})^\top & (\mathsf{Y}^{\mathsf{uv}})^\top & \mathsf{Y}^{\mathsf{vv}}\end{pmatrix}\in\mathcal{CP}^{3n+1}.\end{aligned} \tag{CP2A(ρ)}$$

Once again, the equivalence between (P2A(ρ)) and (CP2A(ρ)) is understood similarly to that between (P1A(ρ)) and (CP1A(ρ)).

4.3. Two convex relaxations and their smaller equivalents

In this section, we consider two convex relaxations arising from the conic formulations (CP1A(ρ)) and (CP2A(ρ)).

Despite the fact that (CP1A(ρ)) and (CP2A(ρ)) are convex optimization problems, they are, in general, NP-hard [MK87]. A well-known tractable convex relaxation of completely positive conic problems is obtained by replacing the intractable cone $\mathcal{CP}^d$ with the larger cone $\mathcal{DNN}^d$ of doubly nonnegative (DNN) matrices, i.e., the convex cone of symmetric $d\times d$ matrices that are both PSD and componentwise nonnegative. It follows that the optimal value of the DNN relaxation is a lower bound on the optimal value of the completely positive conic optimization problem. Here, we present the DNN relaxations arising from conic optimization problems (CP1A(ρ)) and (CP2A(ρ)). Furthermore, we establish that each of the two DNN relaxations can, in fact, be reformulated in smaller dimensions without sacrificing the strength of the corresponding lower bound. Finally, we present a theoretical comparison of the strength of the two DNN relaxations.

4.3.1. DNN relaxations of the direct MIQP formulation

In this section, we consider the DNN relaxation of the problem (CP1A(ρ)). By replacing the intractable conic constraint $\mathsf{Z} \in \mathcal{CP}^{4n+1}$ in (CP1A(ρ)) by $\mathsf{Z} \in \mathcal{DNN}^{4n+1}$, where $\mathcal{DNN}^{4n+1}$ denotes the cone of doubly nonnegative (DNN) matrices in $\mathcal{S}_+^{4n+1}$, we obtain the following tractable convex relaxation of (P1A(ρ)):

$$
\begin{aligned}
\nu(\text{D1A}(\rho)) := \min_{\mathsf{Z} \in \mathcal{S}^{4n+1}} \quad & \langle \mathsf{Q}, \mathsf{Z}^{\mathsf{xx}} \rangle \\
\text{s.t.} \quad & \mathsf{e}^\top \mathsf{x} = 1 \\
& \mathsf{e}^\top \mathsf{u} = \rho \\
& \mathsf{x} + \mathsf{y} = \mathsf{u} \\
& \mathsf{u} + \mathsf{v} = \mathsf{e} \\
& \langle \mathsf{E}, \mathsf{Z}^{\mathsf{xx}} \rangle = 1 \\
& \langle \mathsf{E}, \mathsf{Z}^{\mathsf{uu}} \rangle = \rho^2 \\
& \text{diag}\,(\mathsf{Z}^{\mathsf{uu}}) = \mathsf{u} \\
& \text{diag}\,(\mathsf{Z}^{\mathsf{xx}}) + \text{diag}\,(\mathsf{Z}^{\mathsf{yy}}) + \text{diag}\,(\mathsf{Z}^{\mathsf{uu}}) + \\
& 2\,[\text{diag}\,(\mathsf{Z}^{\mathsf{xy}}) - \text{diag}\,(\mathsf{Z}^{\mathsf{xu}}) - \text{diag}\,(\mathsf{Z}^{\mathsf{uy}})] = \mathsf{o} \\
& \text{diag}\,(\mathsf{Z}^{\mathsf{uu}}) + 2\,\text{diag}\,(\mathsf{Z}^{\mathsf{uv}}) + \text{diag}\,(\mathsf{Z}^{\mathsf{vv}}) = \mathsf{e} \\
& \mathsf{Z} := \begin{pmatrix} 1 & \mathsf{x}^\top & \mathsf{u}^\top & \mathsf{v}^\top & \mathsf{y}^\top \\ \mathsf{x} & \mathsf{Z}^{\mathsf{xx}} & \mathsf{Z}^{\mathsf{xu}} & \mathsf{Z}^{\mathsf{xv}} & \mathsf{Z}^{\mathsf{xy}} \\ \mathsf{u} & (\mathsf{Z}^{\mathsf{xu}})^\top & \mathsf{Z}^{\mathsf{uu}} & \mathsf{Z}^{\mathsf{uv}} & \mathsf{Z}^{\mathsf{uy}} \\ \mathsf{v} & (\mathsf{Z}^{\mathsf{xv}})^\top & (\mathsf{Z}^{\mathsf{uv}})^\top & \mathsf{Z}^{\mathsf{vv}} & \mathsf{Z}^{\mathsf{vy}} \\ \mathsf{y} & (\mathsf{Z}^{\mathsf{xy}})^\top & (\mathsf{Z}^{\mathsf{uy}})^\top & (\mathsf{Z}^{\mathsf{vy}})^\top & \mathsf{Z}^{\mathsf{yy}} \end{pmatrix} \in \mathcal{DNN}^{4n+1}.
\end{aligned}
\tag{D1A(ρ)}
$$

Note that (D1A(ρ)) consists of $5n + 4$ linear equality constraints and a doubly nonnegative constraint in $\mathcal{S}_+^{4n+1}$. First, we make several observations about feasible solutions of (D1A(ρ)).

Lemma 26. *Let $\mathsf{Z} \in \mathcal{S}_+^{4n+1}$ be (D1A(ρ))-feasible. Then, the following relations hold:*

$$
\begin{aligned}
\mathsf{x}\mathsf{e}^\top - \mathsf{Z}^{\mathsf{xu}} &= \mathsf{Z}^{\mathsf{xv}} \geq \mathsf{O}, && (4.3)\\
-\mathsf{Z}^{\mathsf{xx}} + \mathsf{Z}^{\mathsf{xu}} &= \mathsf{Z}^{\mathsf{xy}} \geq \mathsf{O}, && (4.4)\\
\mathsf{Z}^{\mathsf{xx}} - \mathsf{Z}^{\mathsf{xu}} - (\mathsf{Z}^{\mathsf{xu}})^\top + \mathsf{Z}^{\mathsf{uu}} &= \mathsf{Z}^{\mathsf{yy}} \geq \mathsf{O}, && (4.5)\\
\mathsf{e}\mathsf{e}^\top - \mathsf{e}\mathsf{u}^\top - \mathsf{u}\mathsf{e}^\top + \mathsf{Z}^{\mathsf{uu}} &= \mathsf{Z}^{\mathsf{vv}} \geq \mathsf{O}, && (4.6)\\
-\mathsf{e}\mathsf{x}^\top + (\mathsf{Z}^{\mathsf{xu}})^\top + \mathsf{e}\mathsf{u}^\top - \mathsf{Z}^{\mathsf{uu}} &= \mathsf{Z}^{\mathsf{vy}} \geq \mathsf{O}, && (4.7)\\
\textit{diag}\,(\mathsf{Z}^{\mathsf{xu}}) &= \mathsf{x}, && (4.8)\\
\textit{diag}\,(\mathsf{Z}^{\mathsf{xv}}) &= \mathsf{o}, && (4.9)\\
\textit{diag}\,(\mathsf{Z}^{\mathsf{vy}}) &= \mathsf{o}, && (4.10)\\
\textit{diag}\,(\mathsf{Z}^{\mathsf{vv}}) &= \mathsf{v}, && (4.11)\\
\textit{diag}\,(\mathsf{Z}^{\mathsf{uv}}) &= \mathsf{o}. && (4.12)
\end{aligned}
$$

Proof. Let $\mathsf{Z} \in \mathcal{S}_+^{4n+1}$ be (D1A(ρ))-feasible. By using the Schur complementation, we obtain

$$
\begin{pmatrix} \mathsf{Z}^{\mathsf{xx}} & \mathsf{Z}^{\mathsf{xu}} & \mathsf{Z}^{\mathsf{xv}} & \mathsf{Z}^{\mathsf{xy}} \\ (\mathsf{Z}^{\mathsf{xu}})^\top & \mathsf{Z}^{\mathsf{uu}} & \mathsf{Z}^{\mathsf{uv}} & \mathsf{Z}^{\mathsf{uy}} \\ (\mathsf{Z}^{\mathsf{xv}})^\top & (\mathsf{Z}^{\mathsf{uv}})^\top & \mathsf{Z}^{\mathsf{vv}} & \mathsf{Z}^{\mathsf{vy}} \\ (\mathsf{Z}^{\mathsf{xy}})^\top & (\mathsf{Z}^{\mathsf{uy}})^\top & (\mathsf{Z}^{\mathsf{vy}})^\top & \mathsf{Z}^{\mathsf{yy}} \end{pmatrix} = \begin{pmatrix} \mathsf{x} \\ \mathsf{u} \\ \mathsf{v} \\ \mathsf{y} \end{pmatrix} \begin{pmatrix} \mathsf{x} \\ \mathsf{u} \\ \mathsf{v} \\ \mathsf{y} \end{pmatrix}^\top + \sum_{k \in K} \begin{pmatrix} \mathsf{a}^k \\ \mathsf{b}^k \\ \mathsf{c}^k \\ \mathsf{d}^k \end{pmatrix} \begin{pmatrix} \mathsf{a}^k \\ \mathsf{b}^k \\ \mathsf{c}^k \\ \mathsf{d}^k \end{pmatrix}^\top,
$$

where K is a finite set and $\{\mathsf{a}^k, \mathsf{b}^k, \mathsf{c}^k, \mathsf{d}^k\} \subset \mathbb{R}^n$ for each $k \in K$. Using $\operatorname{diag}(\mathsf{Z}^{\mathsf{uu}}) + 2\operatorname{diag}(\mathsf{Z}^{\mathsf{uv}}) + \operatorname{diag}(\mathsf{Z}^{\mathsf{vv}}) = \mathsf{e}$, we obtain

$$\begin{aligned} u_i^2 + \sum_{k\in K}\left(b_i^k\right)^2 + 2\left[u_i v_i + \sum_{k\in K}\left(b_i^k\right)\left(c_i^k\right)\right] + v_i^2 + \sum_{k\in K}\left(c_i^k\right)^2 &= (u_i+v_i)^2 + \sum_{k\in K}\left(b_i^k + c_i^k\right)^2 \\ &= 1 + \sum_{k\in K}\left(b_i^k + c_i^k\right)^2 = 1 \end{aligned}$$

for each $i \in \{1,\ldots,n\}$, where we used $\mathsf{u}+\mathsf{v}=\mathsf{e}$ in the second equality. We conclude that

$$\mathsf{b}^k = -\mathsf{c}^k, \quad \text{for all } k \in K\,. \tag{4.13}$$

In a similar manner, by using $\operatorname{diag}(\mathsf{Z}^{\mathsf{xx}})+\operatorname{diag}(\mathsf{Z}^{\mathsf{yy}})+\operatorname{diag}(\mathsf{Z}^{\mathsf{uu}})+2\left[\operatorname{diag}(\mathsf{Z}^{\mathsf{xy}}) - \operatorname{diag}(\mathsf{Z}^{\mathsf{xu}}) - \operatorname{diag}(\mathsf{Z}^{\mathsf{uy}})\right] = \mathsf{o}$, we arrive at

$$(x_i + y_i - u_i)^2 + \sum_{k\in K}\left(a_i^k + d_i^k - b_i^k\right)^2 = 0$$

for each $i \in \{1,\ldots,n\}$, which, together with $\mathsf{x}+\mathsf{y}=\mathsf{u}$, implies that

$$\mathsf{a}^k + \mathsf{d}^k = \mathsf{b}^k, \quad \text{for all } k \in K\,. \tag{4.14}$$

Therefore,

$$\mathsf{x}\mathsf{e}^\top - \mathsf{Z}^{\mathsf{xu}} = \mathsf{x}\mathsf{e}^\top - \mathsf{x}\mathsf{u}^\top - \sum_{k\in K}\mathsf{a}^k\left(\mathsf{b}^k\right)^\top = \mathsf{x}\mathsf{v}^\top + \sum_{k\in K}\mathsf{a}^k\left(\mathsf{c}^k\right)^\top = \mathsf{Z}^{\mathsf{xv}} \geq \mathsf{O},$$

where we used $\mathsf{u}+\mathsf{v}=\mathsf{e}$, (4.13), and $\mathsf{Z} \in \mathcal{DNN}^{4n+1}$. This establishes (4.3). Arguing similarly,

$$-\mathsf{Z}^{\mathsf{xx}} + \mathsf{Z}^{\mathsf{xu}} = -\mathsf{x}\mathsf{x}^\top - \sum_{k\in K}\mathsf{a}^k\left(\mathsf{a}^k\right)^\top + \mathsf{x}\mathsf{u}^\top + \sum_{k\in K}\mathsf{a}^k\left(\mathsf{b}^k\right)^\top = \mathsf{x}\mathsf{y}^\top + \sum_{k\in K}\mathsf{a}^k\left(\mathsf{d}^k\right)^\top = \mathsf{Z}^{\mathsf{xy}} \geq \mathsf{O},$$

where we used $\mathsf{x}+\mathsf{y}=\mathsf{u}$, (4.14), and $\mathsf{Z} \in \mathcal{DNN}^{4n+1}$. This establishes (4.4). Next,

$$\begin{aligned} \mathsf{Z}^{\mathsf{xx}} - \mathsf{Z}^{\mathsf{xu}} - (\mathsf{Z}^{\mathsf{xu}})^\top + \mathsf{Z}^{\mathsf{uu}} &= \mathsf{x}\mathsf{x}^\top + \sum_{k\in K}\mathsf{a}^k\left(\mathsf{a}^k\right)^\top - \mathsf{x}\mathsf{u}^\top - \sum_{k\in K}\mathsf{a}^k\left(\mathsf{b}^k\right)^\top \\ &\qquad - \mathsf{u}\mathsf{x}^\top - \sum_{k\in K}\mathsf{b}^k\left(\mathsf{a}^k\right)^\top + \mathsf{u}\mathsf{u}^\top + \sum_{k\in K}\mathsf{b}^k\left(\mathsf{b}^k\right)^\top \\ &= \mathsf{y}\mathsf{y}^T + \sum_{k\in K}\mathsf{d}^k\left(\mathsf{d}^k\right)^\top \\ &= \mathsf{Z}^{\mathsf{yy}} \geq \mathsf{O}, \end{aligned}$$

where we used $\mathsf{x}+\mathsf{y}=\mathsf{u}$, (4.14), and $\mathsf{Z} \in \mathcal{DNN}^{4n+1}$. Similarly,

$$\mathsf{e}\mathsf{e}^\top - \mathsf{e}\mathsf{u}^\top - \mathsf{u}\mathsf{e}^\top + \mathsf{Z}^{\mathsf{uu}} = \mathsf{e}\mathsf{e}^\top - \mathsf{e}\mathsf{u}^\top - \mathsf{u}\mathsf{e}^\top + \mathsf{u}\mathsf{u}^\top + \sum_{k\in K}\mathsf{b}^k\left(\mathsf{b}^k\right)^\top = \mathsf{v}\mathsf{v}^\top + \sum_{k\in K}\mathsf{c}^k\left(\mathsf{c}^k\right)^\top = \mathsf{Z}^{\mathsf{vv}} \geq \mathsf{O},$$

where we used $\mathsf{u}+\mathsf{v}=\mathsf{e}$, (4.13), and $\mathsf{Z} \in \mathcal{DNN}^{4n+1}$, establishing (4.6). Next,

$$\begin{aligned} -\mathsf{e}\mathsf{x}^\top + (\mathsf{Z}^{\mathsf{xu}})^\top + \mathsf{e}\mathsf{u}^\top - \mathsf{Z}^{\mathsf{uu}} &= -\mathsf{e}\mathsf{x}^\top + \mathsf{u}\mathsf{x}^\top + \sum_{k\in K}\mathsf{b}^k\left(\mathsf{a}^k\right)^\top + \mathsf{e}\mathsf{u}^\top - \mathsf{u}\mathsf{u}^\top - \sum_{k\in K}\mathsf{b}^k\left(\mathsf{b}^k\right)^\top \\ &= -\mathsf{v}\mathsf{x}^\top + \mathsf{v}\mathsf{u}^\top + \sum_{k\in K}\mathsf{c}^k\left(\mathsf{d}^k\right)^\top \\ &= \mathsf{Z}^{\mathsf{vy}} \geq \mathsf{O}, \end{aligned}$$

where we used $\mathsf{u}+\mathsf{v}=\mathsf{e}$, $\mathsf{x}+\mathsf{y}=\mathsf{u}$, (4.13), (4.14), and $\mathsf{Z}\in\mathcal{DNN}^{4n+1}$, establishing (4.7).

Using diag $(\mathsf{Z}^{\mathsf{uu}})=\mathsf{u}$, (4.3), and (4.7),

$$\begin{aligned}\mathsf{o}\le\operatorname{diag}\left(\mathsf{x}\mathsf{e}^\top-\mathsf{Z}^{\mathsf{xu}}\right) &= \mathsf{x}-\operatorname{diag}(\mathsf{Z}^{\mathsf{xu}}) \quad\text{and}\\ \mathsf{o}\le\operatorname{diag}\left(-\mathsf{e}\mathsf{x}^\top+(\mathsf{Z}^{\mathsf{xu}})^\top+\mathsf{e}\mathsf{u}^\top-\mathsf{Z}^{\mathsf{uu}}\right) &= -\mathsf{x}+\operatorname{diag}(\mathsf{Z}^{\mathsf{xu}})+\mathsf{u}-\operatorname{diag}(\mathsf{Z}^{\mathsf{uu}})=-\mathsf{x}+\operatorname{diag}(\mathsf{Z}^{\mathsf{xu}}),\end{aligned}$$

which together yield (4.8). By (4.8) and diag $(\mathsf{Z}^{\mathsf{uu}})=\mathsf{u}$, it is easy to see that (4.9) and (4.10) follow from (4.3) and (4.7), respectively. Using (4.6), diag $(\mathsf{Z}^{\mathsf{uu}})=\mathsf{u}$ and $\mathsf{u}+\mathsf{v}=\mathsf{e}$, we obtain

$$\operatorname{diag}(\mathsf{e}\mathsf{e}^\top-\mathsf{e}\mathsf{u}^\top-\mathsf{u}\mathsf{e}^\top+\mathsf{Z}^{\mathsf{uu}})=\mathsf{e}-2\,\mathsf{u}+\mathsf{u}=\mathsf{v}=\operatorname{diag}(\mathsf{Z}^{\mathsf{vv}}),$$

which establishes (4.11). Finally, (4.12) follows from (4.11), diag $(\mathsf{Z}^{\mathsf{uu}})=\mathsf{u}$, diag $(\mathsf{Z}^{\mathsf{uu}})+2\operatorname{diag}(\mathsf{Z}^{\mathsf{uv}})+\operatorname{diag}(\mathsf{Z}^{\mathsf{vv}})=\mathsf{e}$, and $\mathsf{u}+\mathsf{v}=\mathsf{e}$. This completes the proof. □

In view of Lemma 26, let us now consider the following optimization problem:

$$\begin{aligned}\nu(\mathrm{D1B}(\rho)):=\min_{\mathsf{W}\in\mathcal{S}_+^{2n+1}}\quad & \langle\mathsf{Q},\mathsf{W}^{\mathsf{xx}}\rangle\\ \text{s.t.}\quad & \mathsf{e}^\top\mathsf{x}=1\\ & \mathsf{e}^\top\mathsf{u}=\rho\\ & \langle\mathsf{E},\mathsf{W}^{\mathsf{xx}}\rangle=1\\ & \langle\mathsf{E},\mathsf{W}^{\mathsf{uu}}\rangle=\rho^2\\ & \operatorname{diag}(\mathsf{W}^{\mathsf{uu}})=\mathsf{u}\\ & \mathsf{x}\mathsf{e}^\top-\mathsf{W}^{\mathsf{xu}}\ge\mathsf{O}\\ & -\mathsf{W}^{\mathsf{xx}}+\mathsf{W}^{\mathsf{xu}}\ge\mathsf{O}\\ & \mathsf{W}^{\mathsf{xx}}-\mathsf{W}^{\mathsf{xu}}-(\mathsf{W}^{\mathsf{xu}})^\top+\mathsf{W}^{\mathsf{uu}}\ge\mathsf{O}\\ & \mathsf{e}\mathsf{e}^\top-\mathsf{e}\mathsf{u}^\top-\mathsf{u}\mathsf{e}^\top+\mathsf{W}^{\mathsf{uu}}\ge\mathsf{O}\\ & -\mathsf{e}\mathsf{x}^\top+(\mathsf{W}^{\mathsf{xu}})^\top+\mathsf{e}\mathsf{u}^\top-\mathsf{W}^{\mathsf{uu}}\ge\mathsf{O}\\ & \mathsf{W}^{\mathsf{xx}}\ge\mathsf{O}\\ & \mathsf{W}:=\begin{pmatrix}1 & \mathsf{x}^\top & \mathsf{u}^\top\\ \mathsf{x} & \mathsf{W}^{\mathsf{xx}} & \mathsf{W}^{\mathsf{xu}}\\ \mathsf{u} & (\mathsf{W}^{\mathsf{xu}})^\top & \mathsf{W}^{\mathsf{uu}}\end{pmatrix}\succeq\mathsf{O}.\end{aligned}\tag{D1B(ρ)}$$

It is easy to see that (D1B(ρ)) is a convex relaxation of (P1(ρ)) since, for any feasible solution $(\mathsf{x},\mathsf{u})\in\mathbb{R}^n\times\mathbb{R}^n$ of (P1(ρ)),

$$\mathsf{W}=\begin{pmatrix}1\\ \mathsf{x}\\ \mathsf{u}\end{pmatrix}\begin{pmatrix}1\\ \mathsf{x}\\ \mathsf{u}\end{pmatrix}^\top\in\mathcal{S}_+^{2n+1}$$

is a feasible solution of (D1B(ρ)) with the same objective function value. Note that (D1B(ρ)) has $n+4$ linear equality constraints, $(9/2)n^2+(3/2)n$ linear inequality constraints, and a PSD constraint in $\mathcal{S}^{2n+1}$. It is worth noticing that the inequality constraints of (D1B(ρ)) are precisely given by all the RLT (reformulation-linearization technique) constraints obtained from the inequalities $\mathsf{x}\ge\mathsf{o}$, $\mathsf{x}\le\mathsf{u}$, and $\mathsf{u}\le\mathsf{e}$ (see, e.g., [SA99]).

First, we present some properties of feasible solutions of (D1B(ρ)).

Lemma 27. *Let* $\mathsf{W} \in \mathcal{S}_+^{2n+1}$ *be* (D1B(ρ))*-feasible. Then,*

$$\begin{aligned} \mathsf{W} &\in \mathcal{DNN}^{2n+1}, && (4.15)\\ -(\mathsf{W}^{\mathsf{xu}})^\top + \mathsf{W}^{\mathsf{uu}} &\geq \mathsf{O}, && (4.16)\\ \mathsf{u}\mathsf{e}^\top - \mathsf{W}^{\mathsf{uu}} &\geq \mathsf{O}, && (4.17)\\ \mathsf{u} &\leq \mathsf{e}, && (4.18)\\ \mathsf{x} &\leq \mathsf{u}. && (4.19) \end{aligned}$$

Proof. Let $\mathsf{W} \in \mathcal{S}^{2n+1}$ be (D1B(ρ))-feasible. Since $\mathsf{W}^{\mathsf{xx}} \geq \mathsf{O}$ and $-\mathsf{W}^{\mathsf{xx}} + \mathsf{W}^{\mathsf{xu}} \geq \mathsf{O}$, we obtain $\mathsf{W}^{\mathsf{xu}} \geq \mathsf{O}$. By $\mathsf{x}\mathsf{e}^\top - \mathsf{W}^{\mathsf{xu}} \geq \mathsf{O}$, we arrive at $\mathsf{x} \geq \mathsf{o}$. Since diag $(\mathsf{W}^{\mathsf{uu}}) = \mathsf{u}$ and $\mathsf{W} \succeq \mathsf{O}$, we have $\mathsf{u} \geq \mathsf{o}$. Then, adding the inequalities $-\mathsf{W}^{\mathsf{xx}} + \mathsf{W}^{\mathsf{xu}} \geq \mathsf{O}$ and $\mathsf{W}^{\mathsf{xx}} - \mathsf{W}^{\mathsf{xu}} - (\mathsf{W}^{\mathsf{xu}})^\top + \mathsf{W}^{\mathsf{uu}} \geq \mathsf{O}$ yields (4.16), which, in turn, establishes (4.15) since $\mathsf{W}^{\mathsf{uu}} \geq (\mathsf{W}^{\mathsf{xu}})^\top \geq \mathsf{O}$. Similarly, adding the inequalities $\left(\mathsf{x}\mathsf{e}^\top - \mathsf{W}^{\mathsf{xu}}\right)^\top = \mathsf{e}\mathsf{x}^\top - (\mathsf{W}^{\mathsf{xu}})^\top \geq \mathsf{O}$ and $-\mathsf{e}\mathsf{x}^\top + (\mathsf{W}^{\mathsf{xu}})^\top + \mathsf{e}\mathsf{u}^\top - \mathsf{W}^{\mathsf{uu}} \geq \mathsf{O}$, we arrive at (4.17). Adding (4.17) to $\mathsf{e}\mathsf{e}^\top - \mathsf{e}\mathsf{u}^\top - \mathsf{u}\mathsf{e}^\top + \mathsf{W}^{\mathsf{uu}} \geq \mathsf{O}$, we get $\mathsf{e}(\mathsf{e}-\mathsf{u})^\top \geq \mathsf{O}$, which yields (4.18). Similarly, using diag $(\mathsf{W}^{\mathsf{uu}}) = \mathsf{u}$,

$$\begin{aligned} \mathsf{o} \leq \text{diag}\,\left(\mathsf{x}\mathsf{e}^\top - \mathsf{W}^{\mathsf{xu}}\right) &= \mathsf{x} - \text{diag}\,(\mathsf{W}^{\mathsf{xu}}) \quad \text{and}\\ \mathsf{o} \leq \text{diag}\,\left(-\mathsf{e}\mathsf{x}^\top + (\mathsf{W}^{\mathsf{xu}})^\top + \mathsf{e}\mathsf{u}^\top - \mathsf{W}^{\mathsf{uu}}\right) &= -\mathsf{x} + \text{diag}\,(\mathsf{W}^{\mathsf{xu}}) + \mathsf{u} - \text{diag}\,(\mathsf{W}^{\mathsf{uu}}) = -\mathsf{x} + \text{diag}\,(\mathsf{W}^{\mathsf{xu}}), \end{aligned}$$

which together yield diag $(\mathsf{W}^{\mathsf{xu}}) = \mathsf{x}$. Combining this with (4.16), we get

$$\mathsf{o} \leq \text{diag}\,\left(-(\mathsf{W}^{\mathsf{xu}})^\top + \mathsf{W}^{\mathsf{uu}}\right) = -\text{diag}\,(\mathsf{W}^{\mathsf{xu}}) + \text{diag}\,(\mathsf{W}^{\mathsf{uu}}) = -\mathsf{x} + \mathsf{u},$$

where we used diag $(\mathsf{W}^{\mathsf{uu}}) = \mathsf{u}$. This establishes (4.19) and completes the proof. □ □

Our next result shows that (D1B(ρ)) is equivalent to the DNN relaxation (D1A(ρ)) of (CP1A(ρ)).

Proposition 28. (D1A(ρ)) *and* (D1B(ρ)) *are equivalent to each other. Therefore,* $\nu(D1A(\rho)) = \nu(D1B(\rho))$*.*

Proof. Let $\mathsf{W} \in \mathcal{S}^{2n+1}$ be (D1B(ρ))-feasible. Let us define $\mathsf{Z} \in \mathcal{S}^{4n+1}$ as follows:

$$\mathsf{Z} = \begin{pmatrix} 1 & \mathsf{o}^\top & \mathsf{o}^\top \\ \mathsf{o} & \mathsf{I} & \mathsf{o} \\ \mathsf{o} & \mathsf{O} & \mathsf{I} \\ \mathsf{e} & \mathsf{O} & -\mathsf{I} \\ \mathsf{o} & -\mathsf{I} & \mathsf{I} \end{pmatrix} \begin{pmatrix} 1 & \mathsf{x}^\top & \mathsf{u}^\top \\ \mathsf{x} & \mathsf{W}^{\mathsf{xx}} & \mathsf{W}^{\mathsf{xu}} \\ \mathsf{u} & (\mathsf{W}^{\mathsf{xu}})^\top & \mathsf{W}^{\mathsf{uu}} \end{pmatrix} \begin{pmatrix} 1 & \mathsf{o}^\top & \mathsf{o}^\top \\ \mathsf{o} & \mathsf{I} & \mathsf{o} \\ \mathsf{o} & \mathsf{O} & \mathsf{I} \\ \mathsf{e} & \mathsf{O} & -\mathsf{I} \\ \mathsf{O} & -\mathsf{I} & \mathsf{I} \end{pmatrix}^\top. \tag{4.20}$$

Therefore,

$$\mathsf{Z} = \begin{pmatrix} 1 & \mathsf{x}^\top & \mathsf{u}^\top & \mathsf{e}^\top - \mathsf{u}^\top & -\mathsf{x}^\top + \mathsf{u}^\top \\ \mathsf{x} & \mathsf{W}^{\mathsf{xx}} & \mathsf{W}^{\mathsf{xu}} & \mathsf{x}\mathsf{e}^\top - \mathsf{W}^{\mathsf{xu}} & -\mathsf{W}^{\mathsf{xx}} + \mathsf{W}^{\mathsf{xu}} \\ \mathsf{u} & (\mathsf{W}^{\mathsf{xu}})^\top & \mathsf{W}^{\mathsf{uu}} & \mathsf{u}\mathsf{e}^\top - \mathsf{W}^{\mathsf{uu}} & -(\mathsf{W}^{\mathsf{xu}})^\top + \mathsf{W}^{\mathsf{uu}} \\ \mathsf{e}-\mathsf{u} & \mathsf{e}\mathsf{x}^\top - (\mathsf{W}^{\mathsf{xu}})^\top & \mathsf{e}\mathsf{u}^\top - \mathsf{W}^{\mathsf{uu}} & \mathsf{e}\mathsf{e}^\top - \mathsf{e}\mathsf{u}^\top - \mathsf{u}\mathsf{e}^\top + \mathsf{W}^{\mathsf{uu}} & -\mathsf{e}\mathsf{x}^\top + (\mathsf{W}^{\mathsf{xu}})^\top + \mathsf{e}\mathsf{u}^\top - \mathsf{W}^{\mathsf{uu}} \\ -\mathsf{x}+\mathsf{u} & -\mathsf{W}^{\mathsf{xx}} + (\mathsf{W}^{\mathsf{xu}})^\top & -\mathsf{W}^{\mathsf{xu}} + \mathsf{W}^{\mathsf{uu}} & -\mathsf{x}\mathsf{e}^\top + \mathsf{W}^{\mathsf{xu}} + \mathsf{u}\mathsf{e}^\top - \mathsf{W}^{\mathsf{uu}} & \mathsf{W}^{\mathsf{xx}} - \mathsf{W}^{\mathsf{xu}} - (\mathsf{W}^{\mathsf{xu}})^\top + \mathsf{W}^{\mathsf{uu}} \end{pmatrix}. \tag{4.21}$$

We claim that $\mathsf{Z} \in \mathcal{S}^{4n+1}$ is (D1A(ρ))-feasible. By Lemma 27, (4.20), and (4.21), we conclude that $\mathsf{Z} \in \mathcal{DNN}^{4n+1}$. Since $\mathsf{v} = \mathsf{e} - \mathsf{u}$, $\mathsf{y} = \mathsf{u} - \mathsf{x}$, $\mathsf{Z}^{\mathsf{xx}} = \mathsf{W}^{\mathsf{xx}}$, and $\mathsf{Z}^{\mathsf{uu}} = \mathsf{W}^{\mathsf{uu}}$, the first seven sets of

equality constraints of (D1A(ρ)) are satisfied. Considering the eighth set of equality constraints of (D1A(ρ)) and substituting the corresponding submatrices in (4.21), we obtain

$$\begin{aligned}
\operatorname{diag}(\mathsf{Z}^{\mathsf{xx}}) &= \operatorname{diag}(\mathsf{W}^{\mathsf{xx}}) \\
\operatorname{diag}(\mathsf{Z}^{\mathsf{yy}}) &= \operatorname{diag}(\mathsf{W}^{\mathsf{xx}}) - 2\operatorname{diag}(\mathsf{W}^{\mathsf{xu}}) + \operatorname{diag}(\mathsf{W}^{\mathsf{uu}}) \\
\operatorname{diag}(\mathsf{Z}^{\mathsf{uu}}) &= \operatorname{diag}(\mathsf{W}^{\mathsf{uu}}) \\
\operatorname{diag}(\mathsf{Z}^{\mathsf{xy}}) &= -\operatorname{diag}(\mathsf{W}^{\mathsf{xx}}) + \operatorname{diag}(\mathsf{W}^{\mathsf{xu}}) \\
\operatorname{diag}(\mathsf{Z}^{\mathsf{xu}}) &= \operatorname{diag}(\mathsf{W}^{\mathsf{xu}}) \\
\operatorname{diag}(\mathsf{Z}^{\mathsf{uy}}) &= -\operatorname{diag}(\mathsf{W}^{\mathsf{xu}}) + \operatorname{diag}(\mathsf{W}^{\mathsf{uu}}),
\end{aligned}$$

which implies that

$$\operatorname{diag}(\mathsf{Z}^{\mathsf{xx}}) + \operatorname{diag}(\mathsf{Z}^{\mathsf{yy}}) + \operatorname{diag}(\mathsf{Z}^{\mathsf{uu}}) + 2\,[\operatorname{diag}(\mathsf{Z}^{\mathsf{xy}}) - \operatorname{diag}(\mathsf{Z}^{\mathsf{xu}}) - \operatorname{diag}(\mathsf{Z}^{\mathsf{uy}})] = \mathsf{o}\,.$$

Finally, consider the last set of equality constraints of (D1A(ρ)):

$$\operatorname{diag}(\mathsf{Z}^{\mathsf{uu}}) + 2\operatorname{diag}(\mathsf{Z}^{\mathsf{uv}}) + \operatorname{diag}(\mathsf{Z}^{\mathsf{vv}}) = \operatorname{diag}(\mathsf{W}^{\mathsf{uu}}) + 2\,(\mathsf{u} - \operatorname{diag}(\mathsf{W}^{\mathsf{uu}})) + \mathsf{e} - 2\,\mathsf{u} + \operatorname{diag}(\mathsf{W}^{\mathsf{uu}}) = \mathsf{e}\,,$$

which implies that $\mathsf{Z} \in \mathcal{S}^{4n+1}$ is (D1A(ρ))-feasible and achieves the same objective function value as W in (D1B(ρ)).

Conversely, let $\mathsf{Z} \in \mathcal{S}^{4n+1}$ be (D1A(ρ))-feasible. Let $\mathsf{W} \in \mathcal{S}^{2n+1}$ be given by the top left 3×3-block of Z, i.e.,

$$\mathsf{W} = \begin{pmatrix} 1 & \mathsf{x}^\top & \mathsf{u}^\top \\ \mathsf{x} & \mathsf{Z}^{\mathsf{xx}} & \mathsf{Z}^{\mathsf{xu}} \\ \mathsf{u} & (\mathsf{Z}^{\mathsf{xu}})^\top & \mathsf{Z}^{\mathsf{uu}} \end{pmatrix}. \tag{4.22}$$

Clearly, $\mathsf{W} \in \mathcal{DNN}^{2n+1}$ and satisfies all linear equality constraints and $\mathsf{W}^{\mathsf{xx}} = \mathsf{Z}^{\mathsf{xx}} \geq \mathsf{O}$ in (D1B(ρ)). By Lemma 26, we conclude that W satisfies each of the remaining linear inequality constraints and achieves the same objective function value as Z in (D1A(ρ)). It follows that ν(D1A(ρ)) = ν(D1B(ρ)). □ □

Recall that (D1A(ρ)) consists of $5n + 4$ linear equality constraints and a doubly nonnegative constraint in $\mathcal{S}^{4n+1}$. In contrast, (D1B(ρ)) has $n + 4$ linear equality constraints, $(9/2)n^2 + (3/2)n$ linear inequality constraints, and a PSD constraint in $\mathcal{S}^{2n+1}$. By Proposition 28, we conclude that the lower bound arising from the DNN relaxation of (CP1A(ρ)) can be computed by solving a conic optimization problem in a much smaller dimension.

4.3.2. DNN relaxations of the complementarity constraint formulation

In this section, we focus on the DNN relaxation of the copositive optimization problem (CP2A(ρ)). By replacing the intractable conic constraint $\mathsf{Z} \in \mathcal{CP}^{3n+1}$ in (CP2A(ρ)) by $\mathsf{Z} \in \mathcal{DNN}^{3n+1}$, we obtain the following convex optimization problem:

$$
\begin{aligned}
\nu(\mathrm{D2A}(\rho)) := \min_{\mathsf{Y}\in\mathcal{S}^{3n+1}_{+}} \quad & \langle \mathsf{Q}, \mathsf{Y}^{\mathsf{xx}}\rangle \\
\text{s.t.} \quad & \mathsf{e}^{\top}\mathsf{x} = 1 \\
& \mathsf{e}^{\top}\mathsf{u} = \rho \\
& \mathsf{u} + \mathsf{v} = \mathsf{e} \\
& \langle \mathsf{E}, \mathsf{Y}^{\mathsf{xx}}\rangle = 1 \\
& \langle \mathsf{E}, \mathsf{Y}^{\mathsf{uu}}\rangle = \rho^2 \\
& \mathrm{diag}\,(\mathsf{Y}^{\mathsf{uu}}) = \mathsf{u} \\
& \mathsf{e}^{\top}\mathrm{diag}\,(\mathsf{Y}^{\mathsf{xv}}) = 0 \\
& \mathrm{diag}\,(\mathsf{Y}^{\mathsf{uu}}) + 2\,\mathrm{diag}\,(\mathsf{Y}^{\mathsf{uv}}) + \mathrm{diag}\,(\mathsf{Y}^{\mathsf{vv}}) = \mathsf{e} \\
& \mathsf{Y} := \begin{pmatrix} 1 & \mathsf{x}^{\top} & \mathsf{u}^{\top} & \mathsf{v}^{\top} \\ \mathsf{x} & \mathsf{Y}^{\mathsf{xx}} & \mathsf{Y}^{\mathsf{xu}} & \mathsf{Y}^{\mathsf{xv}} \\ \mathsf{u} & (\mathsf{Y}^{\mathsf{xu}})^{\top} & \mathsf{Y}^{\mathsf{uu}} & \mathsf{Y}^{\mathsf{uv}} \\ \mathsf{v} & (\mathsf{Y}^{\mathsf{xv}})^{\top} & (\mathsf{Y}^{\mathsf{uv}})^{\top} & \mathsf{Y}^{\mathsf{vv}} \end{pmatrix} \in \mathcal{DNN}^{3n+1}.
\end{aligned}
\tag{D2A(ρ)}
$$

Once again, it is easy to verify that (D2A(ρ)) is a tractable convex relaxation of (P2A(ρ)). Note that (D2A(ρ)) consists of $3n + 5$ linear equality constraints and a doubly nonnegative constraint in $\mathcal{S}^{3n+1}$.

First, we establish several properties of feasible solutions of (D2A(ρ)).

Lemma 29. *Let $\mathsf{Y} \in \mathcal{S}^{3n+1}$ be (D2A(ρ))-feasible. Then, the following relations hold:*

$$
\begin{aligned}
\mathsf{x}\mathsf{e}^{\top} - \mathsf{Y}^{\mathsf{xu}} &= \mathsf{Y}^{\mathsf{xv}} \geq \mathsf{O}, && (4.23)\\
\mathsf{e}\mathsf{e}^{\top} - \mathsf{e}\mathsf{u}^{\top} - \mathsf{u}\mathsf{e}^{\top} + \mathsf{Y}^{\mathsf{uu}} &= \mathsf{Y}^{\mathsf{vv}} \geq \mathsf{O}, && (4.24)\\
\mathsf{u}\mathsf{e}^{\top} - \mathsf{Y}^{\mathsf{uu}} &= \mathsf{Y}^{\mathsf{uv}} \geq \mathsf{O}, && (4.25)\\
\mathit{diag}\,(\mathsf{Y}^{\mathsf{xv}}) &= \mathsf{o}, && (4.26)\\
\mathit{diag}\,(\mathsf{Y}^{\mathsf{xu}}) &= \mathsf{x}, && (4.27)\\
\mathit{diag}\,(\mathsf{Y}^{\mathsf{vv}}) &= \mathsf{v}, && (4.28)\\
\mathit{diag}\,(\mathsf{Y}^{\mathsf{uv}}) &= \mathsf{o}. && (4.29)
\end{aligned}
$$

Proof. Let $\mathsf{Y} \in \mathcal{S}^{3n+1}$ be (D2A(ρ))-feasible. By using the Schur complementation, we obtain

$$
\begin{pmatrix} \mathsf{Y}^{\mathsf{xx}} & \mathsf{Y}^{\mathsf{xu}} & \mathsf{Y}^{\mathsf{xv}} \\ (\mathsf{Y}^{\mathsf{xu}})^{\top} & \mathsf{Y}^{\mathsf{uu}} & \mathsf{Y}^{\mathsf{uv}} \\ (\mathsf{Y}^{\mathsf{xv}})^{\top} & (\mathsf{Y}^{\mathsf{uv}})^{\top} & \mathsf{Y}^{\mathsf{vv}} \end{pmatrix} = \begin{pmatrix} \mathsf{x} \\ \mathsf{u} \\ \mathsf{v} \end{pmatrix} \begin{pmatrix} \mathsf{x} \\ \mathsf{u} \\ \mathsf{v} \end{pmatrix}^{\top} + \sum_{k\in K} \begin{pmatrix} \mathsf{a}^k \\ \mathsf{b}^k \\ \mathsf{c}^k \end{pmatrix} \begin{pmatrix} \mathsf{a}^k \\ \mathsf{b}^k \\ \mathsf{c}^k \end{pmatrix}^{\top},
$$

where K is a finite set and $\{\mathsf{a}^k, \mathsf{b}^k, \mathsf{c}^k\} \subset \mathbb{R}^n$ for each $k \in K$. Arguing similarly to the proof of Lemma 26, we obtain $\mathsf{b}^k = -\mathsf{c}^k$ for each $k \in K$. The relations (4.23), (4.24), and (4.25) can be established in a very similar manner. Since $\mathsf{e}^{\top}\mathrm{diag}\,(\mathsf{Y}^{\mathsf{xv}}) = 0$ and $\mathsf{Y} \in \mathcal{DNN}^{3n+1}$, we obtain (4.26). By using $\mathsf{u} + \mathsf{v} = \mathsf{e}$, $\mathsf{b}^k = -\mathsf{c}^k$ for each $k \in K$, and the decomposition above, it is easy to verify that $\mathrm{diag}\,(\mathsf{Y}^{\mathsf{xu}}) + \mathrm{diag}\,(\mathsf{Y}^{\mathsf{xv}}) = \mathsf{x}$, which establishes (4.27) due to (4.26). Similarly, we obtain $\mathrm{diag}\,(\mathsf{Y}^{\mathsf{vv}}) = \mathsf{e} - 2\,\mathsf{u} + \mathrm{diag}\,(\mathsf{Y}^{\mathsf{uu}}) = \mathsf{e} - \mathsf{u} = \mathsf{v}$ since $\mathrm{diag}\,(\mathsf{Y}^{\mathsf{uu}}) = \mathsf{u}$, which yields (4.28). Finally, we obtain (4.29) from $\mathrm{diag}\,(\mathsf{Y}^{\mathsf{uu}}) = \mathsf{u}$, $\mathrm{diag}\,(\mathsf{Y}^{\mathsf{uu}}) + 2\,\mathrm{diag}\,(\mathsf{Y}^{\mathsf{uv}}) + \mathrm{diag}\,(\mathsf{Y}^{\mathsf{vv}}) = \mathsf{e}$, $\mathsf{u} + \mathsf{v} = \mathsf{e}$, and (4.28). This completes the proof. □

Once again, we can utilize Lemma 29 to obtain the following optimization problem:

$$\begin{aligned}
\nu(\text{D2B}(\rho)) := \min_{\mathsf{S}\in\mathcal{S}^{2n+1}} \quad & \langle \mathsf{Q}, \mathsf{S}^{\mathsf{xx}} \rangle \\
\text{s.t.} \quad & \mathsf{e}^\top \mathsf{x} = 1 \\
& \mathsf{e}^\top \mathsf{u} = \rho \\
& \langle \mathsf{E}, \mathsf{S}^{\mathsf{xx}} \rangle = 1 \\
& \langle \mathsf{E}, \mathsf{S}^{\mathsf{uu}} \rangle = \rho^2 \\
& \text{diag}\,(\mathsf{S}^{\mathsf{uu}}) = \mathsf{u} \\
& \text{diag}\,(\mathsf{S}^{\mathsf{xu}}) = \mathsf{x} \\
& \mathsf{x}\mathsf{e}^\top - \mathsf{S}^{\mathsf{xu}} \geq \mathsf{O} \\
& \mathsf{e}\mathsf{e}^\top - \mathsf{e}\mathsf{u}^\top - \mathsf{u}\mathsf{e}^\top + \mathsf{S}^{\mathsf{uu}} \geq \mathsf{O} \\
& \mathsf{u}\mathsf{e}^\top - \mathsf{S}^{\mathsf{uu}} \geq \mathsf{O} \\
& \mathsf{S} := \begin{pmatrix} 1 & \mathsf{x}^\top & \mathsf{u}^\top \\ \mathsf{x} & \mathsf{S}^{\mathsf{xx}} & \mathsf{S}^{\mathsf{xu}} \\ \mathsf{u} & (\mathsf{S}^{\mathsf{xu}})^\top & \mathsf{S}^{\mathsf{uu}} \end{pmatrix} \in \mathcal{DNN}^+_{2n+1}.
\end{aligned} \tag{D2B(ρ))}$$

Note that (D2B(ρ)) has $2n+4$ linear equality constraints, $(5/2)n^2 + (1/2)n$ linear inequality constraints, and a doubly nonnegative constraint in $\mathcal{S}^{2n+1}$. Once again, it is worth noticing that (D2B(ρ)) contains all the RLT constraints obtained from the inequalities $\mathsf{x} \geq \mathsf{o}$, $\mathsf{u} \geq \mathsf{o}$, and $\mathsf{u} \leq \mathsf{e}$.

Our next result establishes the equivalence between (D2A(ρ)) and (D2B(ρ)).

Proposition 30. (D2A(ρ)) *and* (D2B(ρ)) *are equivalent to each other. Therefore,* ν*(D2A(*ρ*)) =* ν*(D2B(*ρ*)).*

Proof. The proof is very similar to that of Proposition 28. Let $\mathsf{S} \in \mathcal{S}^{2n+1}$ be (D2B(ρ))-feasible. Let us define $\mathsf{Y} \in \mathcal{S}^{3n+1}$ as follows:

$$\mathsf{Y} = \begin{pmatrix} 1 & \mathsf{o}^\top & \mathsf{o}^\top \\ \mathsf{o} & \mathsf{I} & \mathsf{o} \\ \mathsf{o} & \mathsf{O} & \mathsf{I} \\ \mathsf{e} & \mathsf{O} & -\mathsf{I} \end{pmatrix} \begin{pmatrix} 1 & \mathsf{x}^\top & \mathsf{u}^\top \\ \mathsf{x} & \mathsf{S}^{\mathsf{xx}} & \mathsf{S}^{\mathsf{xu}} \\ \mathsf{u} & (\mathsf{S}^{\mathsf{xu}})^\top & \mathsf{S}^{\mathsf{uu}} \end{pmatrix} \begin{pmatrix} 1 & \mathsf{o}^\top & \mathsf{o}^\top \\ \mathsf{o} & \mathsf{I} & \mathsf{o} \\ \mathsf{o} & \mathsf{O} & \mathsf{I} \\ \mathsf{e} & \mathsf{O} & -\mathsf{I} \end{pmatrix}^\top. \tag{4.30}$$

Therefore,

$$\mathsf{Y} = \begin{pmatrix} 1 & \mathsf{x}^\top & \mathsf{u}^\top & \mathsf{e}^\top - \mathsf{u}^\top \\ \mathsf{x} & \mathsf{S}^{\mathsf{xx}} & \mathsf{S}^{\mathsf{xu}} & \mathsf{x}\mathsf{e}^\top - \mathsf{S}^{\mathsf{xu}} \\ \mathsf{u} & (\mathsf{S}^{\mathsf{xu}})^\top & \mathsf{S}^{\mathsf{uu}} & \mathsf{u}\mathsf{e}^\top - \mathsf{S}^{\mathsf{uu}} \\ \mathsf{e} - \mathsf{u} & \mathsf{e}\mathsf{x}^\top - (\mathsf{S}^{\mathsf{xu}})^\top & \mathsf{e}\mathsf{u}^\top - \mathsf{S}^{\mathsf{uu}} & \mathsf{e}\mathsf{e}^\top - \mathsf{e}\mathsf{u}^\top - \mathsf{u}\mathsf{e}^\top + \mathsf{S}^{\mathsf{uu}} \end{pmatrix}. \tag{4.31}$$

We claim that $\mathsf{Y} \in \mathcal{S}^{3n+1}$ is (D2A(ρ))-feasible. Since $\text{diag}\,(\mathsf{S}^{\mathsf{uu}}) = \mathsf{u}$, and $\mathsf{S}^{\mathsf{uu}} - \mathsf{u}\mathsf{u}^\top \succeq 0$, we obtain $\mathsf{u} \leq \mathsf{e}$. Therefore, by (4.30) and (4.31), we conclude that $\mathsf{Y} \in \mathcal{DNN}^{4n+1}$. Since $\mathsf{v} = \mathsf{e} - \mathsf{u}$, $\mathsf{Y}^{\mathsf{xx}} = \mathsf{S}^{\mathsf{xx}}$, $\mathsf{Y}^{\mathsf{xu}} = \mathsf{S}^{\mathsf{xu}}$, and $\mathsf{Y}^{\mathsf{uu}} = \mathsf{S}^{\mathsf{uu}}$, the first six sets of equality constraints of (D2A(ρ)) are satisfied. Consider the seventh set of equality constraints of (D2A(ρ)):

$$\mathsf{e}^\top \text{diag}\,(\mathsf{Y}^{\mathsf{xv}}) = \mathsf{e}^\top \text{diag}\,(\mathsf{x}\mathsf{e}^\top - \mathsf{S}^{\mathsf{xu}}) = \mathsf{e}^\top [\mathsf{x} - \text{diag}\,(\mathsf{S}^{\mathsf{xu}})] = 0,$$

where we used $\text{diag}\,(\mathsf{S}^{\mathsf{xu}}) = \mathsf{x}$. Finally, the last set of equality constraints of (D2A(ρ)) can be directly verified using (4.31). Therefore, $\mathsf{Y} \in \mathcal{S}^{3n+1}$ is (D2A(ρ))-feasible and achieves the same objective function value as S in (D2B(ρ)).

Conversely, let $\mathsf{Y} \in \mathcal{S}^{3n+1}$ be (D2A(ρ))-feasible. Let $\mathsf{S} \in \mathcal{S}^{2n+1}$ be given by the top left 3×3-block of Z, i.e.,

$$\mathsf{S} = \begin{pmatrix} 1 & \mathsf{x}^\top & \mathsf{u}^\top \\ \mathsf{x} & \mathsf{Y}^{\mathsf{xx}} & \mathsf{Y}^{\mathsf{xu}} \\ \mathsf{u} & (\mathsf{Y}^{\mathsf{xu}})^\top & \mathsf{Y}^{\mathsf{uu}} \end{pmatrix}. \tag{4.32}$$

Clearly, $\mathsf{S} \in \mathcal{DNN}^{2n+1}$ and satisfies all linear equality and inequality constraints in (D2B(ρ)) by Lemma 29. Furthermore, S has the same objective function value in (D2B(ρ)) as that of Y in (D2A(ρ)). It follows that ν(D2A(ρ)) = ν(D2B(ρ)). □ □

A comparison of (D2A(ρ)) and (D2B(ρ)) reveals that the former consists of $3n+5$ linear equality constraints and a doubly nonnegative constraint in $\mathcal{S}^{3n+1}$ whereas the latter has $2n + 4$ linear equality constraints, $(5/2)n^2 + (1/2)n$ linear inequality constraints, and a doubly nonnegative constraint in $\mathcal{S}^{2n+1}$. Proposition 30 implies that the lower bound arising from the DNN relaxation of (CP2A(ρ)) can be computed by solving a conic optimization problem in a smaller dimension.

4.3.3. Comparing the two DNN relaxations

Now we compare the lower bounds arising from the two DNN relaxations of (CP1A(ρ)) and (CP2A(ρ)).

Proposition 31. (D1A(ρ)) *(or equivalently,* (D1B(ρ))*) is at least as tight as* (D2A(ρ)) *(or equivalently,* (D2B(ρ))*), i.e.,*

$$\nu(D2A(\rho)) = \nu(D2B(\rho)) \leq \nu(D1A(\rho)) = \nu(D1B(\rho)) \leq \ell_\rho(\mathsf{Q})\,.$$

Proof. Let $\mathsf{Z} \in \mathcal{S}^{4n+1}$ be (D1A(ρ))-feasible. Let $\mathsf{Y} \in \mathcal{S}^{3n+1}$ be given by the top left 4×4-block of Z, i.e.,

$$\mathsf{Y} = \begin{pmatrix} 1 & \mathsf{x}^\top & \mathsf{v}^\top & \mathsf{u}^\top \\ \mathsf{x} & \mathsf{Z}^{\mathsf{xx}} & \mathsf{Z}^{\mathsf{xv}} & \mathsf{Z}^{\mathsf{xu}} \\ \mathsf{u} & (\mathsf{Z}^{\mathsf{xu}})^\top & \mathsf{Z}^{\mathsf{uu}} & \mathsf{Z}^{\mathsf{uv}} \\ \mathsf{v} & (\mathsf{Z}^{\mathsf{xv}})^\top & (\mathsf{Z}^{\mathsf{uv}})^\top & \mathsf{Z}^{\mathsf{vv}} \end{pmatrix}.$$

We claim that $\mathsf{Y} \in \mathcal{S}^{3n+1}$ is (D2A(ρ))-feasible. Clearly, $\mathsf{Y} \in \mathcal{DNN}^{3n+1}$ and all constraints of (D2A(ρ)) are satisfied by Lemma 26. Therefore, for every (D1A(ρ))-feasible solution, there exists a corresponding (D2A(ρ))-feasible solution with the same objective function value. By Propositions 28 and 30, we conclude that ν(D2A(ρ)) = ν(D2B(ρ)) $\leq$ ν(D1A(ρ)) = ν(D1B(ρ)) $\leq$ $\ell_\rho(\mathsf{Q})$. □ □

Our next example illustrates that we can have ν(D2A(ρ)) < ν(D1A(ρ)). i.e., (D2A(ρ)) can be strictly weaker than (D1A(ρ)).

Example 6. *Consider the following instance, where $n = 6$ and $\rho = 3$:*

$$\mathsf{Q} = \begin{pmatrix} 2.6947 & -0.2028 & -1.1144 & -2.4230 & -2.1633 & 0.7710 \\ -0.2028 & 5.1998 & 0.5005 & -2.0941 & 0.7828 & -3.3611 \\ -1.1144 & 0.5005 & 5.2918 & 1.4119 & -1.1526 & -1.9723 \\ -2.4230 & -2.0941 & 1.4119 & 4.1140 & -0.2025 & 0.3132 \\ -2.1633 & 0.7828 & -1.1526 & -0.2025 & 7.6645 & -0.0170 \\ 0.7710 & -3.3611 & -1.9723 & 0.3132 & -0.0170 & 3.3040 \end{pmatrix}.$$

For this instance, we have $\nu(D2A(\rho)) = \nu(D2B(\rho)) \approx 0.1320 < \nu(D1A(\rho)) = \nu(D1B(\rho)) \approx 0.1333$. In addition, we observe that the computed optimal solution S^* of (D2B(ρ)) is not feasible for (D1B(ρ)). For instance, $(S^{\mathsf{xu}}_{31})^* - (S^{\mathsf{xx}}_{31})^* \approx -0.00137 < 0$.

4.4. Generating nontrivial instances of sparse StQPs

In this section, we discuss procedures for generating nontrivial instances of the sparse StQP problem (StQP(ρ)). To this end, we say that an instance of (StQP(ρ)) is *nontrivial* if the sparsity constraint is violated by each optimal solution of the corresponding instance of StQP given by (StQP) (or, equivalently, (StQP(ρ)) with $\rho = n$). For a nontrivial instance of the sparse StQP, it follows that the sparsity constraint cuts off all optimal solutions of the corresponding StQP instance without the sparsity constraints.

We aim to construct a nontrivial instance of (StQP(ρ)) in two steps. First, we discuss how to construct an instance of (StQP) with a prespecified *unique* optimal solution. We then choose the sparsity parameter ρ to be strictly smaller than the sparsity of the designated unique optimal solution of (StQP). It follows that the resulting instance of (StQP(ρ)) is nontrivial since the unique optimal solution of (StQP) is cut off by the sparsity constraint.

In Section 4.4.1, we consider generating an instance of (StQP) with a unique optimal solution. We then present simple algorithms for generating nontrivial instances of (StQP(ρ)) in Section 4.4.2.

4.4.1. Generating StQP instances with a unique optimal solution

Let us next focus on generating an instance of (StQP) with a unique optimal solution. To that end, we first recall that an StQP can be reformulated as the following copositive optimization problem [BDDK$^+$00]:

$$\ell_n(\mathsf{Q}) = \min_{\mathsf{x}\in\mathbb{R}^n} \left\{ \mathsf{x}^\top \mathsf{Q}\mathsf{x} : \mathsf{x} \in F_n \right\} = \min_{\mathsf{X}\in\mathcal{S}^n} \left\{ \langle \mathsf{Q}, \mathsf{X}\rangle : \langle \mathsf{E}, \mathsf{X}\rangle \ = 1, \quad \mathsf{X} \in \mathcal{CP}_n \right\} ,$$

where F_n is given by (4.1) and $\mathsf{E} = \mathsf{e}\mathsf{e}^\top \in \mathcal{S}^n$.

The doubly nonnegative (DNN) relaxation of an StQP is therefore given by

$$\mu(\mathsf{Q}) = \min_{\mathsf{X}\in\mathcal{S}^n} \left\{ \langle \mathsf{Q}, \mathsf{X}\rangle : \langle \mathsf{E}, \mathsf{X}\rangle \ = 1, \quad \mathsf{X} \in \mathcal{DNN}^n \right\} . \tag{4.33}$$

Note that $\mu(\mathsf{Q}) \le \ell_n(\mathsf{Q})$. For a given instance of StQP, we say that its DNN relaxation is *exact* if $\mu(\mathsf{Q}) = \ell_n(\mathsf{Q})$ and *inexact* otherwise.

We will review the following useful results, which will play a fundamental role in our instance construction procedures.

Theorem 32. *Let $\mathsf{x} \in F_n$, where F_n is given by* (4.1).

(i) x is a globally optimal solution of (StQP) *if and only if there exist $\mathsf{M} \in \mathbf{bd}\ \mathcal{COP}^n$ and $\lambda \in \mathbb{R}$ such that $\mathsf{x} \in \mathbf{V}^{\mathsf{M}}$ and $\mathsf{Q} = \mathsf{M} + \lambda\,\mathsf{E}$, where $\mathbf{bd}\ \mathcal{COP}^n$ and $\mathbf{V}^{\mathsf{M}}$ are given by*

$$\mathbf{bd}\ \mathcal{COP}^n = \left\{ \mathsf{M} \in \mathcal{COP}^n : \exists\ \mathsf{u} \in F_n \ \textit{s.t.}\ \mathsf{u}^\top \mathsf{M}\mathsf{u} = 0 \right\}, \tag{4.34}$$

and

$$\mathbf{V}^{\mathsf{M}} = \left\{ \mathsf{u} \in F_n : \mathsf{u}^\top \mathsf{M}\mathsf{u} = 0 \right\}. \tag{4.35}$$

respectively. Furthermore, in this case, λ is the optimal value of (StQP)*, and the set of optimal solutions of* (StQP) *is given by $\mathbf{V}^{\mathsf{M}}$.*

(ii) x is a globally optimal solution of (StQP) *and its doubly nonnegative relaxation is exact if and only if there exists $\mathsf{R} \succeq \mathsf{O}$, $\mathsf{N} \ge \mathsf{O}$, and $\lambda \in \mathbb{R}$ such that $\mathsf{x}^\top \mathsf{N}\mathsf{x} = 0$ and $\mathsf{Q} = (\mathsf{I} - \mathsf{e}\mathsf{x}^\top)\mathsf{R}(\mathsf{I} - \mathsf{x}\mathsf{e}^\top) + \mathsf{N} + \lambda\,\mathsf{E}$.*

(iii) x *is a globally optimal solution of* (StQP) *and its doubly nonnegative relaxation is inexact if and only if there exist* $\mathsf{M} \in \mathbf{bd}\, \mathcal{COP}^n \backslash \mathcal{SPN}^n$ *and* $\lambda \in \mathbb{R}$ *such that* $\mathsf{x} \in \mathbf{V}^{\mathsf{M}}$ *and* $\mathsf{Q} = \mathsf{M} + \lambda\, \mathsf{E}$*, where* $\mathbf{bd}\, \mathcal{COP}^n$ *and* $\mathbf{V}^{\mathsf{M}}$ *are given by* (4.34) *and* (4.35)*, respectively.*

Proof. The reader is referred to [Bom97] for (i), and to [GY22] for (ii) and (iii). □ □

In view of Theorem 32, we first give a simple recipe to construct an instance of (StQP) such that it has a unique optimal solution and its doubly nonnegative relaxation is exact.

Proposition 33. *Let* $\mathsf{x} \in F_n$*, where* F_n *is given by* (4.1)*. Let* $\mathbb{A} = \{j \in \{1, \ldots, n\} : \mathsf{x}_j > 0\}$ *and* $\mathbb{B} = \{j \in \{1, \ldots, n\} : \mathsf{x}_j = 0\}$*. Then, for any* $\mathsf{R} \succ \mathsf{O}$*, any* $\mathsf{N} \in \mathcal{S}^n$ *is such that* $\mathsf{N}_{\mathbb{AA}} = \mathsf{O}$*,* $\mathsf{N}_{\mathbb{AB}} \geq \mathsf{O}$*, and* $\mathsf{N}_{\mathbb{BB}} \geq \mathsf{O}$*, and any* $\lambda \in \mathbb{R}$*, if* $\mathsf{Q} = (\mathsf{I} - \mathsf{e}\mathsf{x}^\top)\mathsf{R}(\mathsf{I} - \mathsf{x}\mathsf{e}^\top) + \mathsf{N} + \lambda\, \mathsf{E}$*, then* x *is the unique globally optimal solution of* (StQP) *and its DNN relaxation is exact.*

Proof. It follows from the hypothesis and Theorem 32(ii) that x is a globally optimal solution of (StQP) and its DNN relaxation is exact. We next show the uniqueness. For any $\mathsf{y} \in F_n$, where F_n is given by (4.1), we have

$$\mathsf{y}^\top \mathsf{Q}\mathsf{y} = (\mathsf{y} - \mathsf{x})^\top \mathsf{R}(\mathsf{y} - \mathsf{x}) + \mathsf{y}^\top \mathsf{N}\mathsf{y} + \lambda \geq \lambda = \mathsf{x}^\top \mathsf{Q}\mathsf{x},$$

where we used $\mathsf{x} \in F_n$, $\mathsf{y} \in F_n$, $\mathsf{R} \succ \mathsf{O}$, $\mathsf{y} \geq \mathsf{o}$, and $\mathsf{N} \geq \mathsf{O}$. Since $\mathsf{R} \succ \mathsf{O}$, the inequality above holds with equality if and only if $\mathsf{y} = \mathsf{x}$. We conclude that x is the unique globally optimal solution of (StQP). □ □

In Proposition 33, if one chooses $\mathsf{N} = \mathsf{O}$ and $\lambda \geq \mathsf{o}$, we obtain $\mathsf{Q} \succeq \mathsf{O}$, which implies that the resulting instance of (StQP) is a convex optimization problem. On the other hand, by choosing N and λ in such a way that $\mathsf{Q} \not\succeq \mathsf{O}$, Proposition 33 allows us to construct a nonconvex instance of StQP that admits an exact DNN relaxation. We will utilize both of these observations in our computational experiments.

Next, we consider constructing an instance of StQP with a unique optimal solution but an inexact DNN relaxation. In contrast with the previous case, such a construction is more involved and requires additional care. Our next result presents such a procedure.

Proposition 34. *Let* $\mathsf{x} \in F_n$*, where* F_n *is given by* (4.1)*. Let* $\mathbb{A} = \{j \in \{1, \ldots, n\} : \mathsf{x}_j > 0\}$ *and* $\mathbb{B} = \{j \in \{1, \ldots, n\} : \mathsf{x}_j = 0\}$*. Let* $\mathsf{Q} = (\mathsf{I} - \mathsf{e}\mathsf{x}^\top)\mathsf{R}(\mathsf{I} - \mathsf{x}\mathsf{e}^\top) + \mathsf{N} + \lambda\, \mathsf{E}$*, where* $\lambda \in \mathbb{R}$*,* $\mathsf{R} \in \mathcal{S}^n$ *is such that* $\mathsf{R}_{\mathbb{AA}} \succeq \mathsf{O}$*,* $\mathsf{R}_{\mathbb{BB}} \in \mathcal{COP}^{|\mathbb{B}|}$*,* $\mathsf{R}_{\mathbb{AB}} = \mathsf{O}$*, and* $\mathsf{N} \in \mathcal{S}^n$ *is such that* $\mathsf{N}_{\mathbb{AA}} = \mathsf{O}$*,* $\mathsf{N}_{\mathbb{AB}} \geq \mathsf{O}$*, and* $\mathsf{N}_{\mathbb{BB}} \geq \mathsf{O}$*.*

(i) x *is a globally optimal solution of* (StQP)*. Furthermore, if* $\mathsf{R}_{\mathbb{AA}} \succ \mathsf{O}$*, then* x *is the unique globally optimal solution.*

(ii) If $\mathsf{R}_{\mathbb{BB}} \in \mathcal{COP}^{|\mathbb{B}|} \backslash \mathcal{SPN}^{|\mathbb{B}|}$ *and* $\mathsf{N} = \mathsf{O}$*, there exists an* $\epsilon > 0$ *such that, for all* $\mathsf{R}_{\mathbb{AA}} \succ \mathsf{O}$ *with* $\|\mathsf{R}_{\mathbb{AA}}\| < \epsilon$*, where* $\|\cdot\|$ *denotes the operator norm,* x *is the unique globally optimal solution of* (StQP) *and its doubly nonnegative relaxation is inexact.*

Proof. (i) Let $\mathsf{M} = (\mathsf{I} - \mathsf{e}\mathsf{x}^\top)\mathsf{R}(\mathsf{I} - \mathsf{x}\mathsf{e}^\top) + \mathsf{N}$. By Theorem 32(i), it suffices to show that $\mathsf{M} \in \mathbf{bd}\, \mathcal{COP}^n$ and $\mathsf{x} \in \mathbf{V}^{\mathsf{M}}$, where $\mathbf{V}^{\mathsf{M}}$ is given by (4.35). For any $\mathsf{y} \in F_n$,

$$\mathsf{y}^\top \mathsf{M}\mathsf{y} = (\mathsf{y} - \mathsf{x})^\top \mathsf{R}(\mathsf{y} - \mathsf{x}) + \mathsf{y}^\top \mathsf{N}\mathsf{y} = (\mathsf{y}_{\mathbb{A}} - \mathsf{x}_{\mathbb{A}})^\top \mathsf{R}_{\mathbb{AA}}(\mathsf{y}_{\mathbb{A}} - \mathsf{x}_{\mathbb{A}}) + \mathsf{y}_{\mathbb{B}}^\top \mathsf{R}_{\mathbb{BB}} \mathsf{y}_{\mathbb{B}} + \mathsf{y}^\top \mathsf{N}\mathsf{y} \geq 0 = \mathsf{x}^\top \mathsf{M}\mathsf{x}, \tag{4.36}$$

where we used $\mathsf{y} \in F_n$ in the first equality, $\mathsf{R}_{\mathbb{AB}} = \mathsf{O}$ and $\mathsf{x}_{\mathbb{B}} = \mathsf{o}$ in the second equality, $\mathsf{R}_{\mathbb{AA}} \succeq \mathsf{O}$, $\mathsf{R}_{\mathbb{BB}} \in \mathcal{COP}^{|\mathbb{B}|}$, $\mathsf{N} \geq \mathsf{O}$, and $\mathsf{y} \geq \mathsf{o}$ to derive the inequality, and $\mathsf{x} \in F_n$ in the

last equality. We conclude that $\mathsf{M} \in \mathbf{bd}\, \mathcal{COP}^n$ and $\mathsf{x} \in \mathbf{V}^{\mathsf{M}}$. By Theorem 32(i), x is a globally optimal solution of (StQP). If, in addition, $\mathsf{R}_{\mathbb{AA}} \succ \mathsf{O}$, then it follows from (4.36) that $\mathsf{y}^\top \mathsf{M}\mathsf{y} = 0$ if and only if $(\mathsf{y}_{\mathbb{A}} - \mathsf{x}_{\mathbb{A}})^\top \mathsf{R}_{\mathbb{AA}}(\mathsf{y}_{\mathbb{A}} - \mathsf{x}_{\mathbb{A}}) = \mathsf{y}_{\mathbb{B}}^\top \mathsf{R}_{\mathbb{BB}}\mathsf{y}_{\mathbb{B}} = \mathsf{y}^\top \mathsf{N}\mathsf{y} = 0$. Since $\mathsf{R}_{\mathbb{AA}} \succ \mathsf{O}$, we conclude that $\mathsf{y}_{\mathbb{A}} = \mathsf{x}_{\mathbb{A}}$. Since $\mathsf{x} \in F_n$, $\mathsf{x}_{\mathbb{B}} = \mathsf{o}$, and $\mathsf{y}_{\mathbb{A}} = \mathsf{x}_{\mathbb{A}}$, we obtain $1 = \mathsf{e}^\top \mathsf{x} = \mathsf{e}_{\mathbb{A}}^\top \mathsf{x}_{\mathbb{A}} + \mathsf{e}_{\mathbb{B}}^\top \mathsf{x}_{\mathbb{B}} = \mathsf{e}_{\mathbb{A}}^\top \mathsf{x}_{\mathbb{A}} = \mathsf{e}_{\mathbb{A}}^\top \mathsf{y}_{\mathbb{A}}$, which implies that $\mathsf{y}_{\mathbb{B}} = \mathsf{o}$ since $\mathsf{y} \in F_n$. It follows that $\mathsf{y} = \mathsf{x}$. Therefore, $\mathbf{V}^{\mathsf{M}} = \{\mathsf{x}\}$. By Theorem 32(i), x is the unique globally optimal solution of (StQP).

(ii) Let $\mathsf{M} = (\mathsf{I} - \mathsf{e}\mathsf{x}^\top)\mathsf{R}(\mathsf{I} - \mathsf{x}\mathsf{e}^\top)$. By Theorem 32(ii), it suffices to show that $\mathsf{M} \in \mathbf{bd}\, \mathcal{COP}^n \backslash \mathcal{SPN}^n$ and $\mathsf{x} \in \mathbf{V}^{\mathsf{M}}$. By (i), $\mathsf{M} \in \mathbf{bd}\, \mathcal{COP}_n$ and $\mathsf{x} \in \mathbf{V}^{\mathsf{M}}$. Since $\mathsf{R}_{\mathbb{BB}} \in \mathcal{COP}^{|\mathbb{B}|} \backslash \mathcal{SPN}^{|\mathbb{B}|}$, $\mathcal{CP}_{|\mathbb{B}|}$ and $\mathcal{DNN}^{|\mathbb{B}|}$ are the dual cones of $\mathcal{COP}^{|\mathbb{B}|}$ and $\mathcal{SPN}^{|\mathbb{B}|}$, respectively, it follows from (4.36) that there exists $\mathsf{T} \in \mathcal{DNN}^{|\mathbb{B}|} \backslash \mathcal{CP}^{|\mathbb{B}|}$ such that $\langle \mathsf{R}_{\mathbb{BB}}, \mathsf{T} \rangle = -\delta < 0$. Let $\mathsf{U} \in \mathcal{S}^n$ be such that $\mathsf{U}_{\mathbb{AA}} = \mathsf{O}$, $\mathsf{U}_{\mathbb{AB}} = \mathsf{O}$, and $\mathsf{U}_{\mathbb{BB}} = \mathsf{T}$. Clearly, $\mathsf{U} \in \mathcal{DNN}^n$. Furthermore,

$$\begin{aligned} \langle \mathsf{M}, \mathsf{U} \rangle &= \langle \mathsf{M}_{\mathbb{BB}}, \mathsf{T} \rangle \\ &= \left\langle \left(\mathsf{x}_{\mathbb{A}}^\top \mathsf{R}_{\mathbb{AA}} \mathsf{x}_{\mathbb{A}} \right) \mathsf{e}_{\mathbb{B}} \mathsf{e}_{\mathbb{B}}^\top + \mathsf{R}_{\mathbb{BB}}, \mathsf{T} \right\rangle \\ &= \left(\mathsf{x}_{\mathbb{A}}^\top \mathsf{R}_{\mathbb{AA}} \mathsf{x}_{\mathbb{A}} \right) \left(\mathsf{e}_{\mathbb{B}}^\top \mathsf{T} \mathsf{e}_{\mathbb{B}} \right) - \delta \\ &\leq \|\mathsf{R}_{\mathbb{AA}}\| \mu - \delta, \end{aligned}$$

where $\mu = \mathsf{e}_{\mathbb{B}}^\top \mathsf{T} \mathsf{e}_{\mathbb{B}} > 0$ and we used $\|\mathsf{x}_{\mathbb{A}}\| \leq 1$ to derive the last inequality. We conclude that $\langle \mathsf{M}, \mathsf{U} \rangle < 0$ provided that $\|\mathsf{R}_{\mathbb{AA}}\| < \epsilon$, where $\epsilon = \delta / \mu > 0$. Therefore, for any $\mathsf{R}_{\mathbb{AA}} \succeq \mathsf{O}$ such that $\|\mathsf{R}_{\mathbb{AA}}\| < \epsilon$, we obtain $\mathsf{M} \in \mathbf{bd}\, \mathcal{COP}^n \backslash \mathcal{SPN}^n$ since $\mathsf{U} \in \mathcal{DNN}^n$. By Theorem 32(ii), it follows that the DNN relaxation of (StQP) is inexact. If, in addition, $\mathsf{R}_{\mathbb{AA}} \succ \mathsf{O}$, then x is the unique globally optimal solution of (StQP) by part (i). This concludes the proof.

□

4.4.2. Two algorithms for generating nontrivial sparse StQP instances

In this section, we present two algorithms for generating nontrivial instances of (StQP(ρ)). Given $\mathsf{x} \in F_n$ with $\|\mathsf{x}\|_0 \geq 2$, both algorithms initially generate an instance of (StQP) such that x is the unique optimal solution. Then, a nontrivial instance of (StQP(ρ)) is constructed by choosing $\rho \in \{1, \ldots, \|\mathsf{x}\|_0 - 1\}$.

The details of our algorithms are presented in Algorithm 2 and Algorithm 3. The two algorithms differ only in terms of whether the StQP instance generated in the first stage admits an exact or inexact DNN relaxation.

Data: $n \geq 2$, $\mathsf{x} \in F_n$ such that $\|\mathsf{x}\|_0 \geq 2$, $\lambda \in \mathbb{R}$
Result: A nontrivial instance of (StQP(ρ))
$\mathbb{A} \leftarrow \{j \in \{1, \ldots, n\} : \mathsf{x}_j > 0\}$, $\mathbb{B} \leftarrow \{j \in \{1, \ldots, n\} : \mathsf{x}_j = 0\}$
Choose an arbitrary $\mathsf{R} \in \mathcal{S}^n$ such that $\mathsf{R} \succ \mathsf{O}$
Choose an arbitrary $\mathsf{N} \in \mathcal{S}^n$ such that $\mathsf{N}_{\mathbb{AA}} = \mathsf{O}$, $\mathsf{N}_{\mathbb{AB}} \geq \mathsf{O}$, and $\mathsf{N}_{\mathbb{BB}} \geq \mathsf{O}$
$\mathsf{Q} \leftarrow (\mathsf{I} - \mathsf{e}\mathsf{x}^\top)\mathsf{R}(\mathsf{I} - \mathsf{x}\mathsf{e}^\top) + \mathsf{N} + \lambda\, \mathsf{E}$
Choose an arbitrary $\rho \in \{1, \ldots, \|\mathsf{x}\|_0 - 1\}$.

Algorithm 2: Nontrivial instance of (StQP(ρ)) with (StQP) admitting an exact DNN relaxation

---

Our first algorithm, given by Algorithm 2, requires as input $n \geq 2$, $\mathsf{x} \in F_n$ such that $\|\mathsf{x}\|_0 \geq 2$, and $\lambda \in \mathbb{R}$, and generates an instance of (StQP(ρ)). By Proposition 33, the instance generated

by Algorithm 2 is a nontrivial instance of (StQP(ρ)) such that x is the unique optimal solution of (StQP), which admits an exact DNN relaxation. We remark that each step of Algorithm 2 can be implemented in a fairly straightforward way.

Data: $n \geq 7$, $\mathsf{x} \in F_n$ such that $2 \leq \|\mathsf{x}\|_0 \leq n-5$, $\lambda \in \mathbb{R}$
Result: A nontrivial instance of (StQP(ρ))
$\mathbb{A} \leftarrow \{j \in \{1,\ldots,n\} : \mathsf{x}_j > 0\}$, $\mathbb{B} \leftarrow \{j \in \{1,\ldots,n\} : \mathsf{x}_j = 0\}$
Construct $\mathsf{R} \in \mathcal{S}^n$ using the following steps.
$\mathsf{R}_{\mathbb{AB}} \leftarrow \mathsf{O}$
Choose an arbitrary $\mathsf{R}_{\mathbb{BB}} \in \mathcal{COP}^{|\mathbb{B}|} \backslash \mathcal{SPN}^{|\mathbb{B}|}$.
Choose an arbitrary $\mathsf{T} \in \mathcal{DNN}^{|\mathbb{B}|} \backslash \mathcal{CP}^{|\mathbb{B}|}$ such that $\langle \mathsf{R}_{\mathbb{BB}}, \mathsf{T} \rangle < 0$.
$\delta \leftarrow -\langle \mathsf{R}_{\mathbb{BB}}, \mathsf{T} \rangle$, $\mu \leftarrow \mathsf{e}_{\mathbb{B}}^\top \mathsf{T} \mathsf{e}_{\mathbb{B}}$, $\epsilon \leftarrow \frac{\delta}{\mu}$
Choose any $\mathsf{R}_{\mathbb{AA}} \succ \mathsf{O}$ such that $\|\mathsf{R}_{\mathbb{AA}}\| < \epsilon$.
$\mathsf{Q} \leftarrow (\mathsf{I} - \mathsf{e}\mathsf{x}^\top)\mathsf{R}(\mathsf{I} - \mathsf{x}\mathsf{e}^\top) + \lambda \mathsf{E}$
Choose an arbitrary $\rho \in \{1,\ldots,\|\mathsf{x}\|_0 - 1\}$.

Algorithm 3: Nontrivial instance of (StQP(ρ)) with (StQP) admitting an inexact DNN relaxation

---

Our second algorithm, given by Algorithm 3, takes as input $n \geq 7$, $\mathsf{x} \in F_n$ such that $2 \leq \|\mathsf{x}\|_0 \leq n-5$, and $\lambda \in \mathbb{R}$, and outputs an instance of (StQP(ρ)). Similarly, it follows from Proposition 34 that the instance generated by Algorithm 3 is a nontrivial instance of (StQP(ρ)) such that x is the unique optimal solution of (StQP), which admits an inexact DNN relaxation.

In contrast with Algorithm 2, the practical implementation of Algorithm 3 merits further discussion. Step 4 of Algorithm 3 requires $\mathsf{R}_{\mathbb{BB}} \in \mathcal{COP}^{|\mathbb{B}|} \backslash \mathcal{SPN}^{|\mathbb{B}|}$. Such matrices are called *exceptional*. Recall that $\mathcal{SPN}^n \subseteq \mathcal{COP}^n$, and $\mathcal{SPN}^n = \mathcal{COP}^n$ if and only if $n \leq 4$ (see [Dia62]). Therefore, such an exceptional matrix exists if and only if $|\mathbb{B}| \geq 5$, which is ensured by the choices of the input parameters of Algorithm 3. A well-known exceptional 5×5 matrix is the Horn matrix given by

$$\mathsf{H} = \begin{pmatrix} 1 & -1 & 1 & 1 & -1 \\ -1 & 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 \\ -1 & 1 & 1 & -1 & 1 \end{pmatrix} \in \mathcal{COP}_5 \backslash \mathcal{SPN}_5. \tag{4.37}$$

Furthermore, H is an extreme ray of $\mathcal{COP}^5$ [HN63].

We next present a practical procedure for constructing an exceptional matrix in higher dimensions. Our procedure is inspired by a similar procedure outlined in [GY22, Section 6.2].

Let $n \geq 7$ and let $\mathsf{x} \in F_n$ satisfy $\|\mathsf{x}\|_0 \leq n-5$, which ensures that $|\mathbb{B}| \geq 5$. We now construct $\mathsf{R}_{\mathbb{BB}} \in \mathcal{COP}^{|\mathbb{B}|} \backslash \mathcal{SPN}^{|\mathbb{B}|}$ as follows. Choose an arbitrary $\mathsf{B} \in \mathcal{COP}^{|\mathbb{B}|-5}$, which can, for instance, be ensured by choosing $\mathsf{B} \in \mathcal{SPN}^{|\mathbb{B}|-5}$ (or even $\mathsf{B} \geq \mathsf{O}$ or $\mathsf{B} \succeq \mathsf{O}$), and an arbitrary $\mathsf{C} \in \mathbb{R}^{(|\mathbb{B}|-5)\times|\mathbb{B}|}$ such that $\mathsf{C} \geq \mathsf{O}$. Let us define

$$\mathsf{R}_{\mathbb{BB}} = \begin{bmatrix} \mathsf{B} & \mathsf{C} \\ \mathsf{C}^\top & \mathsf{H} \end{bmatrix} \in \mathcal{S}^{|\mathbb{B}|}. \tag{4.38}$$

It is easy to see that $\mathsf{R}_{\mathbb{BB}} \in \mathcal{COP}^{|\mathbb{B}|}$ (see also [SM16, Lemma 3.4 (a)]). Since $\mathsf{H} \in \mathcal{COP}^5 \backslash \mathcal{SPN}^5$, there exists $\mathsf{F} \in \mathcal{DNN}^5 \backslash \mathcal{CP}^5$ such that $\langle \mathsf{H}, \mathsf{F} \rangle = -\delta < 0$. Let

$$\mathsf{T} = \begin{pmatrix} \mathsf{O} & \mathsf{O} \\ \mathsf{O} & \mathsf{F} \end{pmatrix}. \tag{4.39}$$

Note that $\mathsf{T} \in \mathcal{DNN}^{|\mathbb{B}|}$ and $\langle \mathsf{R}_{\mathbb{BB}}, \mathsf{T} \rangle = \langle \mathsf{H}, \mathsf{F} \rangle = -\delta < 0$, which implies that $\mathsf{R}_{\mathbb{BB}} \notin \mathcal{SPN}^{|\mathbb{B}|}$. We conclude that $\mathsf{R}_{\mathbb{BB}} \in \mathcal{COP}^{|\mathbb{B}|} \backslash \mathcal{SPN}^{|\mathbb{B}|}$. We then define $\mu = \mathsf{e}_{\mathbb{B}}^{\top} \mathsf{T} \mathsf{e}_{\mathbb{B}} > 0$ and $\epsilon = \delta / \mu$.

While any choice of $\mathsf{F} \in \mathcal{DNN}^5 \backslash \mathcal{CP}^5$ such that $\langle \mathsf{H}, \mathsf{F} \rangle = -\delta < 0$ works in the above construction, it may be preferable in practice to choose F such that the corresponding value of ϵ is as large as possible. This can be easily achieved in practice by solving the following small semidefinite programming (SDP) problem:

$$\epsilon := \max_{\mathsf{F} \in \mathcal{DNN}^5 \backslash \{\mathsf{O}\}} \frac{-\langle \mathsf{H}, \mathsf{F} \rangle}{\mathsf{e}_{\mathbb{B}}^{\top} \mathsf{F} \mathsf{e}_{\mathbb{B}}} = \max \{ -\langle \mathsf{H}, \mathsf{F} \rangle : \mathsf{e}_{\mathbb{B}}^{\top} \mathsf{F} \mathsf{e}_{\mathbb{B}} = 1 \, , \, \mathsf{F} \in \mathcal{DNN}_5 \} \, , \tag{4.40}$$

where the second equality follows from the homogeneity of the objective function. An approximate optimal solution of (4.40) is given by

$$\mathsf{F} = \frac{1}{78.2} \begin{pmatrix} 7 & 4.32 & 0 & 0 & 4.32 \\ 4.32 & 7 & 4.32 & 0 & 0 \\ 0 & 4.32 & 7 & 4.32 & 0 \\ 0 & 0 & 4.32 & 7 & 4.32 \\ 4.32 & 0 & 0 & 4.32 & 7 \end{pmatrix} , \tag{4.41}$$

which yields $\epsilon \approx 0.1049$.

Remark 1. *The above construction works for any exceptional* $\mathsf{R}_{\mathbb{BB}}$ *once we can control* T*, the normal separating* $\mathsf{R}_{\mathbb{BB}}$ *(corresponding to* H *above) from* $\mathcal{SPN}^{|\mathbb{B}|}$*, in the sense to solve the (effectively small) SDP problem*

$$\epsilon := \max_{\mathsf{T} \in \mathcal{DNN}^{|\mathbb{B}|} \backslash \{\mathsf{O}\}} \frac{-\langle \mathsf{R}_{\mathbb{BB}}, \mathsf{T} \rangle}{\mathsf{e}_{\mathbb{B}}^{\top} \mathsf{T} \mathsf{e}_{\mathbb{B}}} = \max \{ -\langle \mathsf{R}_{\mathbb{BB}}, \mathsf{T} \rangle : \mathsf{e}_{\mathbb{B}}^{\top} \mathsf{T} \mathsf{e}_{\mathbb{B}} = 1 \, , \, \mathsf{T} \in \mathcal{DNN}^{|\mathbb{B}|} \} \, . \tag{4.42}$$

Observe that any extremal ray of $\mathcal{COP}^{|\mathbb{B}|}$ *different from a rank-one matrix or a symmetric componentwise nonnegative matrix with exactly two positive entries – which constitute the extremal rays of the cone of PSD matrices and the cone of componentwise nonnegative matrices – is necessarily exceptional, by extremality. While for larger* $|\mathbb{B}|$*, these extremal rays are (yet) unknown, we can imitate the above extension strategy and build upon the known ones for* $|\mathbb{B}| \in \{5, 6\}$*, e.g., using Hildebrand matrices [AHD21, Hil12]. All these used in* $\mathsf{R}_{\mathbb{BB}}$ *would render* (4.42) *feasible (recommended only for extremal rays of moderate order to keep this effort low), and proceeding as above with* H *and* F*, this approach provides more nontrivial examples. Finally, note that extremality is used in this context only to ensure exceptionality; any non-extremal but exceptional matrix would work as well.*

4.4.3. Set of instances

To assess the impact of the choice of the instance set on the solution time of each exact model and each convex relaxation as well as on the quality of the lower bound, we conduct extensive experiments on a variety of carefully constructed instances of (StQP(ρ)). Let us now describe the set of instances in detail.

Using Algorithm 2 and Algorithm 3, we generated three sets of nontrivial instances of (StQP(ρ)) with different characteristics.

(i) *PSD Instances:* This set of instances of (StQP(ρ)) is constructed using Algorithm 2 in such a way that $\mathsf{Q} \succeq \mathsf{O}$, i.e., the corresponding instance of (StQP) is a convex optimization

problem. For instance, such instances may arise in portfolio optimization problems. We remark that the resulting instances of (StQP(ρ)) are, in general, still NP-hard. In Step 2 of Algorithm 2, we construct $\mathsf{R} \succ \mathsf{O}$ so that all of its eigenvalues are uniformly distributed in $(0, 3)$. By choosing $\mathsf{N} = \mathsf{O}$ and $\lambda = 0$, we ensure that $\mathsf{Q} = (\mathsf{I} - \mathsf{e}\mathsf{x}^\top)\mathsf{R}(\mathsf{I} - \mathsf{x}\mathsf{e}^\top) \succeq \mathsf{O}$. Since $\mathsf{x}^\top \mathsf{Q}\mathsf{x} = 0$, Q is PSD but not positive definite. By Proposition 33, each PSD instance satisfies

$$\ell_\rho(\mathsf{Q}) > \ell_n(\mathsf{Q}) = \mu(\mathsf{Q}) = 0, \tag{4.43}$$

where $\mu(\mathsf{Q})$ is defined as in (4.33).

(ii) *SPN Instances:* This set of instances of (StQP(ρ)) is generated using Algorithm 2 in a similar manner to PSD instances, i.e., we set $\lambda = 0$ and use the same procedure to construct R. However, instead of choosing $\mathsf{N} = \mathsf{O}$ as in PSD instances, each entry of $\mathsf{N}_{\mathbb{A}\mathbb{B}} \in \mathbb{R}^{|\mathbb{A}|\times|\mathbb{B}|}$ and $\mathsf{N}_{\mathbb{B}\mathbb{B}} \in \mathcal{S}^{|\mathbb{B}|}$ is uniformly generated in $(0, 3)$. In contrast with PSD instances, $\mathsf{Q} = (\mathsf{I} - \mathsf{e}\mathsf{x}^\top)\mathsf{R}(\mathsf{I} - \mathsf{x}\mathsf{e}^\top) + \mathsf{N}$ is, in general, not PSD. In fact, we numerically verified that Q was an indefinite matrix for each SPN instance. By Proposition 33, each SPN instance satisfies

$$\ell_\rho(\mathsf{Q}) > \ell_n(\mathsf{Q}) = \mu(\mathsf{Q}) = 0. \tag{4.44}$$

(iii) *COP Instances:* This set of instances of (StQP(ρ)) is generated by Algorithm 3. In Step 2, we construct $\mathsf{R} \in \mathcal{S}^n$ using the procedure outlined in Section 4.4.2. In particular, $\mathsf{R}_{\mathbb{B}\mathbb{B}}$ is constructed as in (4.38), where $\mathsf{B} \succ \mathsf{O}$ is generated with eigenvalues uniformly distributed in $(0, 3)$, and each entry of $\mathsf{C} \geq \mathsf{O}$ is uniformly distributed in $(0, 1)$. Our choices are, in part, motivated by the observation that the eigenvalues of the Horn matrix H given by (4.37) lie in $[-1.236, 3.236]$ so that the eigenvalues of $\mathsf{B} \succ \mathsf{O}$ have similar magnitude. The separating matrix T is constructed as in (4.39), where F is given by (4.41), which yields $\epsilon \approx 0.1049$. Finally, $\mathsf{R}_{\mathbb{A}\mathbb{A}} \succ \mathsf{O}$ is constructed with eigenvalues uniformly distributed in $(0, 0.99\epsilon)$, ensuring that $\|\mathsf{R}_{\mathbb{A}\mathbb{A}}\| < \epsilon$. We then choose $\lambda = 0$ and set $\mathsf{Q} = (\mathsf{I} - \mathsf{e}\mathsf{x}^\top)\mathsf{R}(\mathsf{I} - \mathsf{x}\mathsf{e}^\top)$. By Proposition 34, this procedure ensures that $\mathsf{Q} \not\succeq \mathsf{O}$ and each COP instance satisfies

$$\ell_\rho(\mathsf{Q}) > \ell_n(\mathsf{Q}) = 0 > \mu(\mathsf{Q}). \tag{4.45}$$

In summary, each of the three sets of instances of (StQP(ρ)) is comprised of nontrivial instances that exhibit different characteristics. While each PSD instance and each SPN instance has a corresponding StQP instance that admits an exact doubly nonnegative (DNN) relaxation, they differ in terms of the convexity of the objective function. On the other hand, each COP instance has a nonconvex objective function with the additional property that the DNN relaxation of the corresponding StQP instance is inexact.

4.5. Computational results

In this section, we report and discuss the results of our computational experiments. After describing the experimental setup in Section 4.5.1, we report our results in detail in Sections 4.5.2 and 4.5.3.

4.5.1. Experimental setup

The set of instances was already described in Section 4.4.3. Here, we outline the details of our experimental setup.

In our experiments, we chose $n \in \{25, 50\}$. For each choice of n, we identified three sparsity levels for the designated optimal solution $\mathsf{x} \in F_n$ of (StQP), denoted by $\rho_0 = \|\mathsf{x}\|_0$. In particular, we considered $\rho_0 \in \{\lfloor 0.25n \rceil, \lfloor 0.5n \rceil, \lfloor 0.75n \rceil\}$, where $\lfloor \cdot \rceil$ denotes the nearest integer function. Finally, for each choice of ρ_0, we chose $\rho \in \{\lfloor 0.25\rho_0 \rceil, \lfloor 0.5\rho_0 \rceil, \lfloor 0.75\rho_0 \rceil\}$. The parameters are summarized in Table 4.1.

n	ρ_0	ρ
25	6	{2,3,4}
	12	{3,6,9}
	19	{5,10,14}
50	12	{3,6,9}
	25	{6,12,19}
	38	{10,19,28}

Table 4.1.: Choices of the parameters n, ρ_0, and ρ

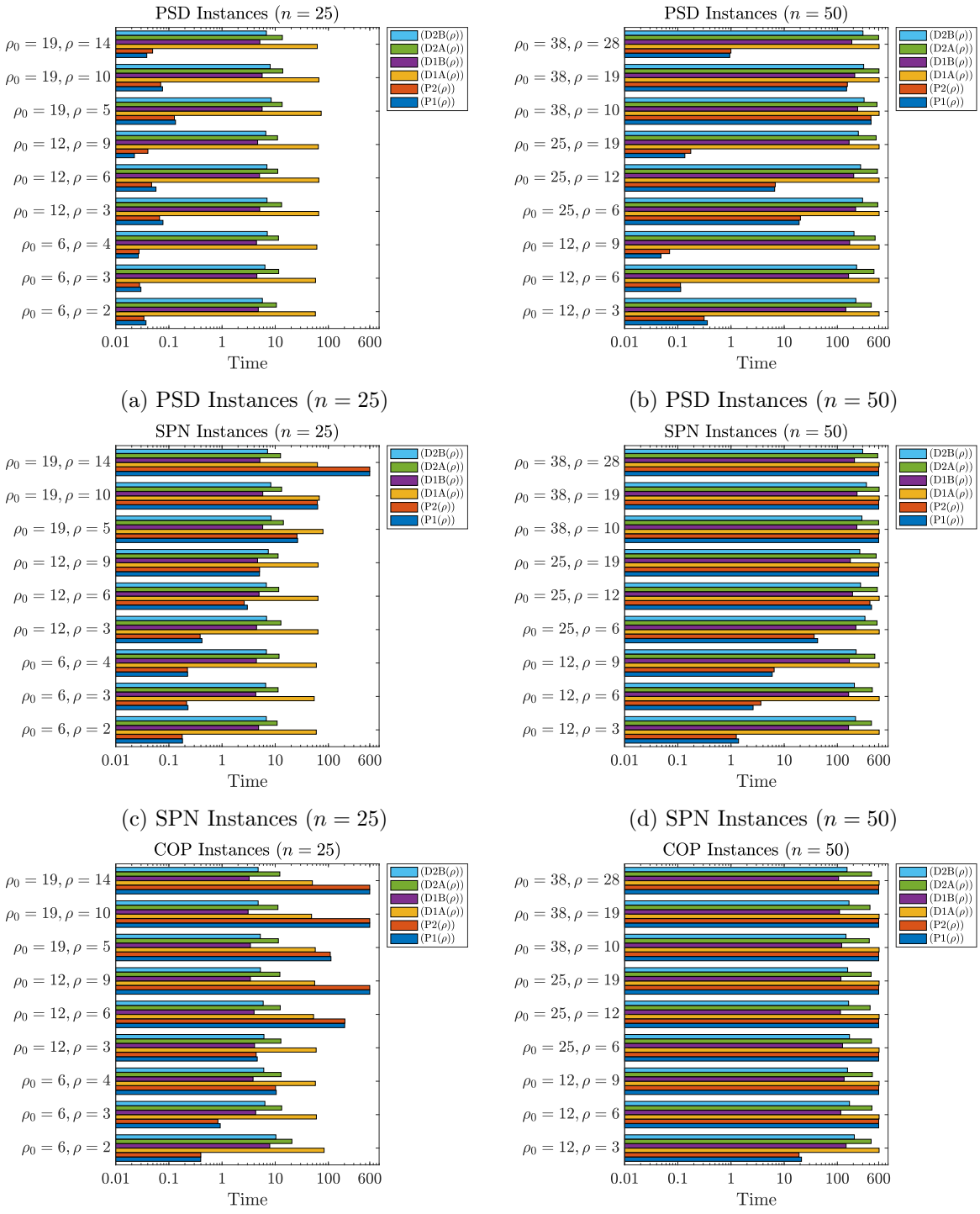
For each choice of n, ρ_0, and ρ presented in Table 4.1, we generated 25 instances from each of the three sets of instances described in Section 4.4.3, which gave rise to a total of 1,350 nontrivial instances of (StQP(ρ)) comprising of 450 PSD instances, 450 SPN instances, and 450 COP instances. For each instance of (StQP(ρ)), we solved (P1(ρ)), (P2(ρ)), (D1A(ρ)), (D1B(ρ)), (D2A(ρ)), and (D2B(ρ)).

Algorithm 2 and Algorithm 3 were implemented in Julia version 1.8.5 [BEKS17]. Each instance of (P1(ρ)) and (P2(ρ)) was solved by `Gurobi` version 11.0.0 [Gur23] via the modeling language `JuMP` v1.19.0 [LDD+23]. We solved (D1A(ρ)), (D1B(ρ)), (D2A(ρ)), and (D2B(ρ)) by `MOSEK` version 10.1.24 [ApS24] via `JuMP`. Our computational experiments were carried out on a 64-bit HP workstation with 24 threads (2 sockets, 6 cores per socket, 2 threads per core) running `Ubuntu Linux` with 96 GB of RAM and Intel Xeon CPU E5-2667 processors with a clock speed of 2.90 GHz. We imposed a time limit of 600 seconds on each solve. The default settings were used for all of the other parameters of `Gurobi` and `MOSEK`.

4.5.2. Solution time

Now we focus on the solution time of each of the six models (P1(ρ)), (P2(ρ)), (D1A(ρ)), (D1B(ρ)), (D2A(ρ)), and (D2B(ρ)).

Figure 4.1 depicts the average solution times of each of the six models on each of the three instance sets. The results for PSD instances, SPN instances, and COP instances are presented in the first, second, and third rows of Figure 4.1, respectively. In each row, the results for $n = 25$ and $n = 50$ are given in the first and second column, respectively. In each graph, the horizontal axis denotes the average solution time (in seconds) presented in logarithmic scale. We used identical axis limits in each graph to facilitate a better comparison across different graphs. The vertical axis is comprised of nine sets of bar charts, each of which represents a particular choice of the tuple (ρ_0, ρ) corresponding to the choice of n as outlined in Table 4.1. Finally, for each choice of (ρ_0, ρ), the corresponding bar chart represents the average solution times of the six models (P1(ρ)), (P2(ρ)), (D1A(ρ)), (D1B(ρ)), (D2A(ρ)), and (D2B(ρ)) over 25 instances generated using the procedure in Section 4.4.3. Note that the solution times of instances that were terminated due to the time limit of 600 seconds were also included in the average solution times. Therefore, each bar chart represents the average computational requirement of the corresponding model. We



(a) PSD Instances ($n = 25$)

(b) PSD Instances ($n = 50$)

(c) SPN Instances ($n = 25$)

(d) SPN Instances ($n = 50$)

(e) COP Instances ($n = 25$)

(f) COP Instances ($n = 50$)

Figure 4.1.: Average solution times (in seconds) of (P1(ρ)), (P2(ρ)), (D1A(ρ)), (D1B(ρ)), (D2A(ρ)), and (D2B(ρ)) for PSD, SPN, and COP instances

recall that (P1(ρ)) and (P2(ρ)) were solved by `Gurobi` whereas (D1A(ρ)), (D1B(ρ)), (D2A(ρ)), and (D2B(ρ)) by `MOSEK`.

Figure 4.1 reveals several interesting relations about average solution times. We first outline our observations concerning the exact MIQP models (P1(ρ)) and (P2(ρ)):

(i) Average solution times of the two exact models exhibit very similar behavior for each fixed choice of the instance set and the triple (n, ρ_0, ρ). However, there seems to be a very strong correlation between the average solution time and the choices of (n, ρ_0, ρ) as well as the instance set.

(ii) Considering each row of Figure 4.1 separately, the average solution time of each of the two exact models across all choices of the tuple (ρ_0, ρ) increases as n increases. However, the rate of increase seems to be highly dependent on the choice of the instance set.

We next present the corresponding observations for the average solution times of the convex relaxations (D1A(ρ)), (D1B(ρ)), (D2A(ρ)), and (D2B(ρ)):

(i) In contrast with the exact models, for each choice of n, Figure 4.1 reveals that the average solution times of the convex relaxations exhibit very similar behavior across all three sets of instances and all choices of (ρ_0, ρ).

(ii) The average solution time of our reduced formulation (D1B(ρ)) consistently achieves the lowest average solution times, followed by the reduced formulation (D2B(ρ)), which, in turn, is followed by (D2A(ρ)) and (D1A(ρ)), respectively.

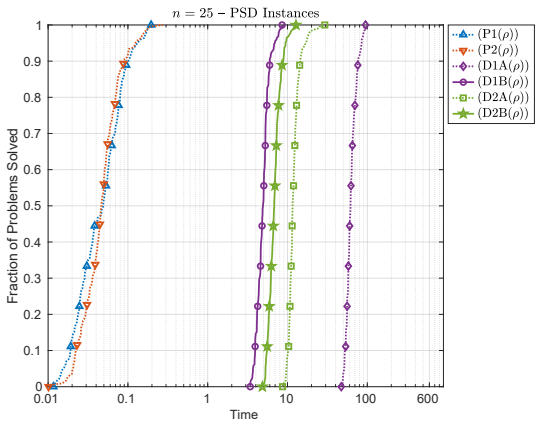
(iii) The average solution time of our reduced formulation (D1B(ρ)) exhibits at least an order of magnitude improvement over that of (D1A(ρ)). On the other hand, the corresponding improvement of (D2B(ρ)) over (D2A(ρ)) seems to be less pronounced, especially in smaller dimensions.

(iv) As expected, the average solution time of each relaxation increases with n. While the rate of increase does not seem to be influenced by the specific instance set, we observe that it does depend on the specific relaxation.

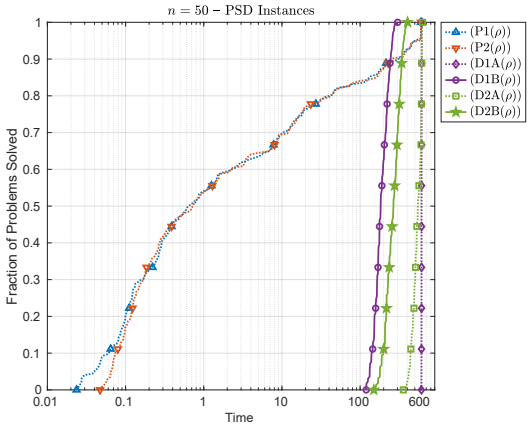
Since Figure 4.1 is based only on average solution times, it cannot capture the variability in the data. In an attempt to shed more light on the distribution of the solution times across all instance sets, we present the empirical cumulative distribution functions of all six models in Figure 4.2, which is organized similarly to Figure 4.1. The three rows display the results for PSD, SPN, and COP instances, respectively, whereas the two columns are devoted to the results for $n = 25$ and $n = 50$, respectively. Each of the six graphs presents six plots corresponding to the empirical cumulative distribution functions of solution times of each of the six models on all 225 instances for the corresponding choice of the instance set and n (see Table 4.1). The horizontal axis denotes the solution time (in seconds) on a logarithmic scale, and the vertical axis represents the fraction of the instances solved. The markers represent the data points for every 25^{th} instance. Once again, we employed identical axis limits in each of the six graphs.

Figure 4.2 clearly reveals the difference between the distributions of solution times of the exact models (P1(ρ)) and (P2(ρ)) and those of the convex relaxations (D1A(ρ)), (D1B(ρ)), (D2A(ρ)), and (D2B(ρ)). We outline our observations below:

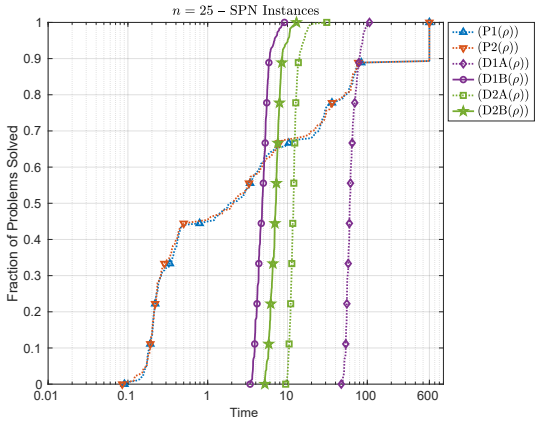
(i) The solution times of the two exact models, in general, exhibit a very similar distribution but a significant variability across different instance sets.



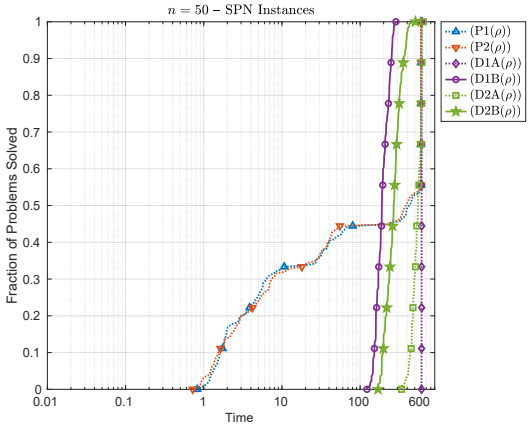
(a) PSD Instances ($n = 25$)



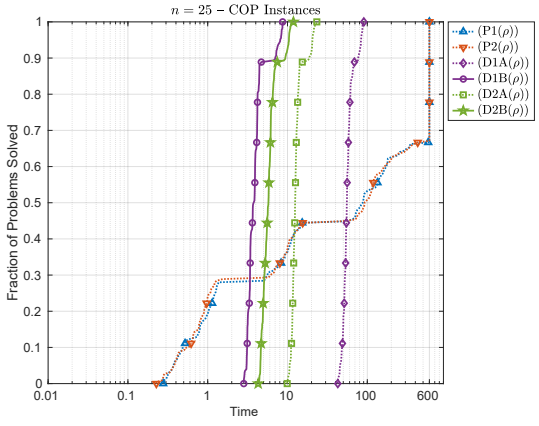
(b) PSD Instances ($n = 50$)



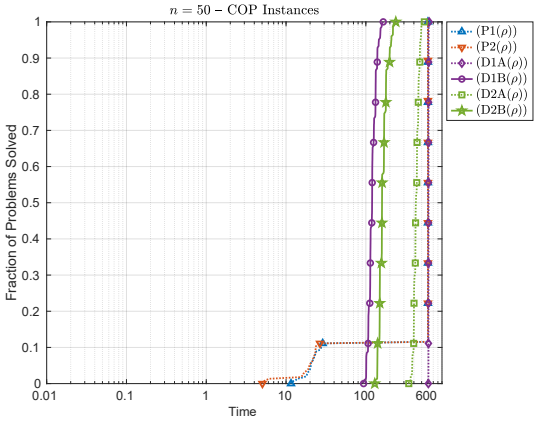
(c) SPN Instances ($n = 25$)



(d) SPN Instances ($n = 50$)



(e) COP Instances ($n = 25$)



(f) COP Instances ($n = 50$)

Figure 4.2.: Empirical cumulative distribution functions of solution times of (P1(ρ)), (P2(ρ)), (D1A(ρ)), (D1B(ρ)), (D2A(ρ)), and (D2B(ρ))

(ii) In contrast, we observe that the solution times of the convex relaxations tend to have a much smaller variance and exhibit very similar distributions across different instance sets.

Therefore, in contrast with the exact models (P1(ρ)) and (P2(ρ)), we conclude that the solution time of each convex relaxation seems to be very robust with respect to the choice of the instance set and the choice of (ρ_0, ρ) in our setting.

In the sequel, we present a more detailed discussion of the behavior of the solution times of the exact models and convex relaxations.

Exact MIQP models

As illustrated by Figures 4.1 and 4.2, the solution time of the exact models is highly dependent on the choices of the instance set and on the triple (n, ρ_0, ρ). Here, we aim to shed more light on this dependence.

In Figures 4.3 and 4.4, we present the empirical cumulative distribution functions of solution times of (P1(ρ)) and (P2(ρ)) for all instances with $n = 25$ and $n = 50$, respectively. Each graph in each figure consists of six plots, each of which corresponds to the solution times of (P1(ρ)) and (P2(ρ)) on each of PSD, SPN, and COP instances for a fixed choice of (n, ρ_0, ρ). In each column of each figure, we fix the tuple (n, ρ_0) and present the empirical cumulative distributions of the solution times of 25 instances corresponding to the three different choices of ρ as presented in Table 4.1. On the other hand, each row corresponds to a different ratio of $\rho/\rho_0 \in \{0.25, 0.5, 0.75\}$. Again, we use a logarithmic scale for the solution time and ensure that all axis limits are identical for a meaningful comparison across different graphs. Finally, we note that the markers represent the data points at every 5^{th} instance.

A close examination of Figures 4.3 and 4.4 reveal the following observations about the two exact models (P1(ρ)) and (P2(ρ)):

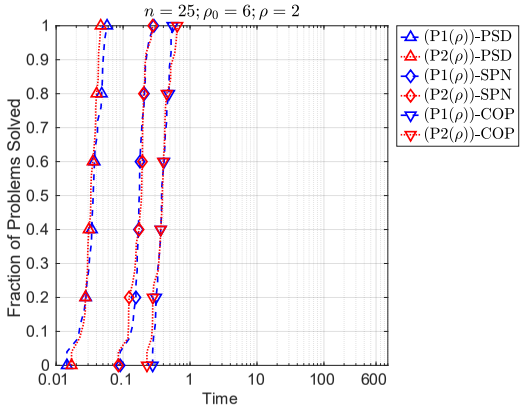
(i) The two models, in general, exhibit very similar distributions across all parameter choices, except for some small differences for a few choices of (n, ρ_0, ρ) on PSD instances, where (P1(ρ)) seems to be solved slightly faster than (P2(ρ)).

(ii) We clearly observe an empirical first-order stochastic dominance among PSD, SPN, and COP instances. For each choice of (n, ρ_0, ρ), PSD instances tend to achieve the lowest solution times, followed by SPN instances, whereas COP instances exhibit the highest solution times.

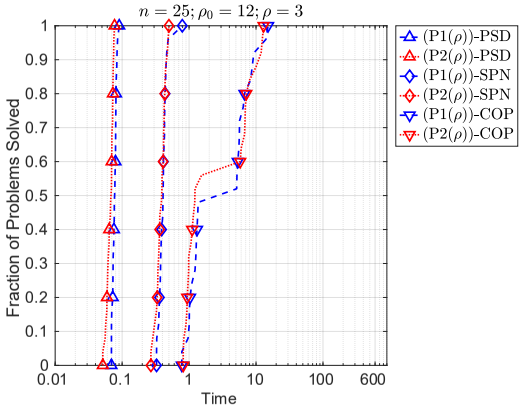
(iii) For each fixed (n, ρ_0), we observe that the solution time tends to decrease as ρ increases for PSD instances, especially for $n = 50$, whereas it increases for each of SPN and COP instances.

(iv) If we fix n and the ratio $\rho/\rho_0 \in \{0.25, 0.5, 0.75\}$, the solution time increases as ρ_0 increases across all three sets of instances. For fixed n, we therefore conclude that the solution time of the exact models seems to be also highly influenced by the choice of the ratio ρ/ρ_0.

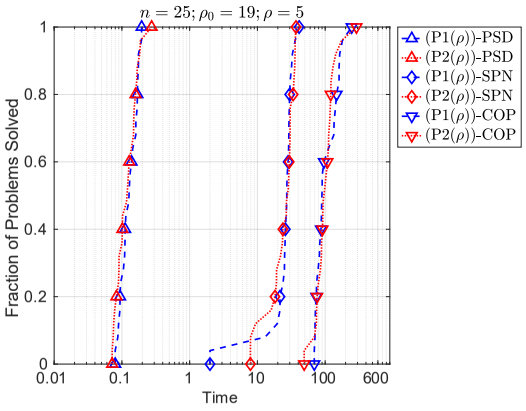
(v) For $n = 25$, both exact models can be solved very quickly for all PSD instances. In contrast, for SPN and COP instances, each of (P1(ρ)) and (P2(ρ)) is terminated due to the time limit of 600 seconds on certain subsets of the instances. Therefore, we observe that SPN and COP instances can be particularly challenging for (P1(ρ)) and (P2(ρ)), even for $n = 25$ with particular choices of (ρ_0, ρ). For $n = 50$, we observe that even some PSD instances were terminated due to the time limit. Recall that each instance in this set has a convex objective function. For each choice of (n, ρ_0, ρ), a comparison of PSD, SPN, and COP instances reveals that the number of instances terminated due to the time limit exhibits a monotonically increasing behavior.



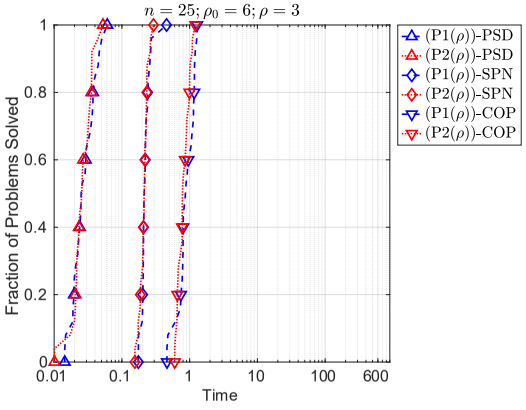
(a) $n = 25, \rho_0 = 6, \rho = 2$



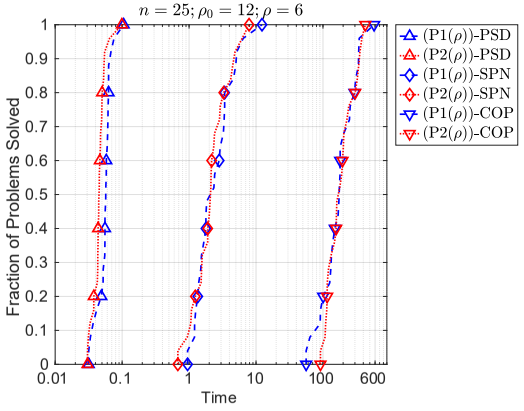
(b) $n = 25, \rho_0 = 12, \rho = 3$



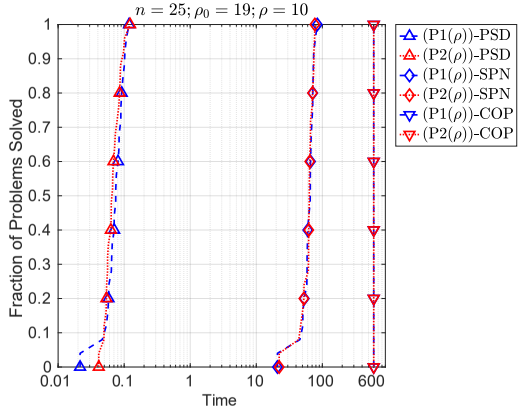
(c) $n = 25, \rho_0 = 19, \rho = 5$



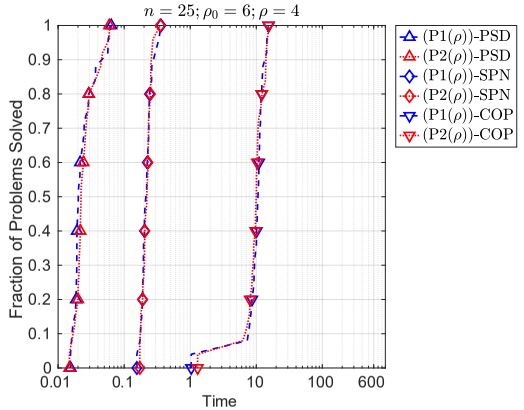
(d) $n = 25, \rho_0 = 6, \rho = 3$



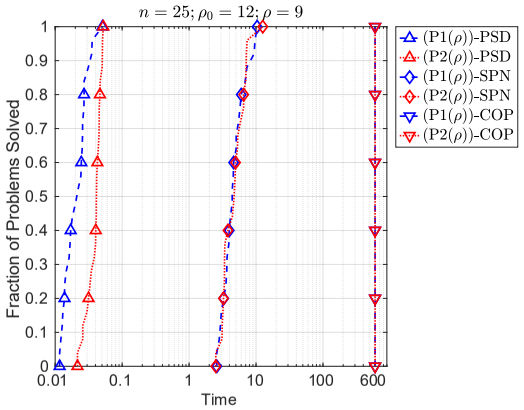
(e) $n = 25, \rho_0 = 12, \rho = 6$



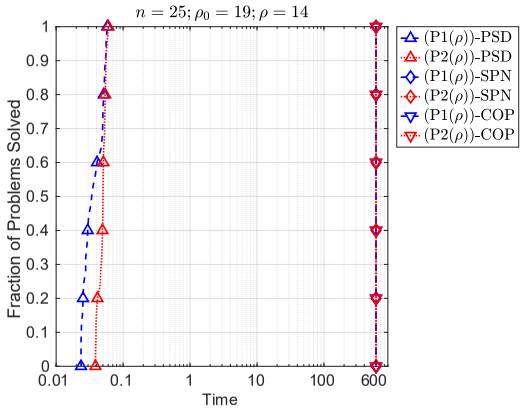
(f) $n = 25, \rho_0 = 19, \rho = 10$



(g) $n = 25, \rho_0 = 6, \rho = 4$

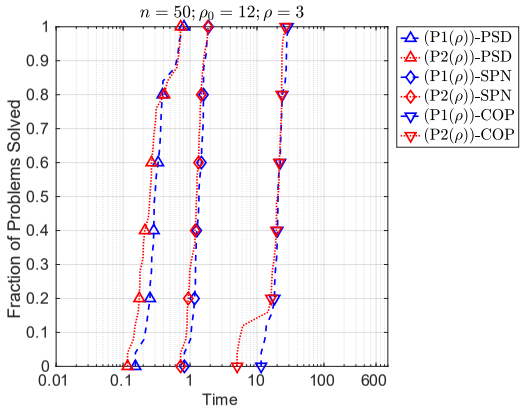


(h) $n = 25, \rho_0 = 12, \rho = 9$

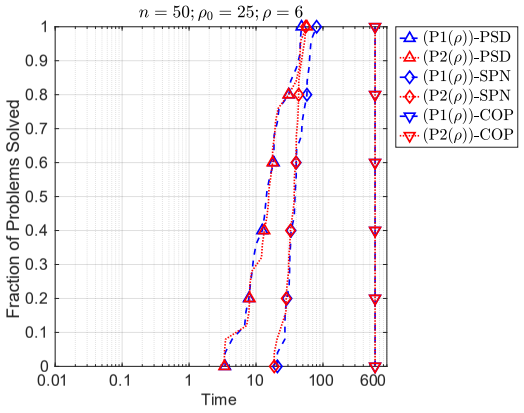


(i) $n = 25, \rho_0 = 19, \rho = 14$

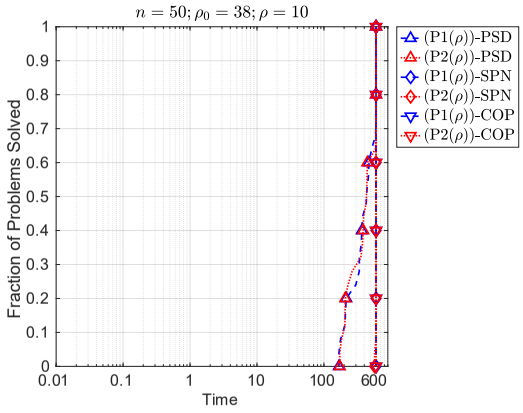
Figure 4.3.: Empirical cumulative distribution functions of solution times of (P1(ρ)) and (P2(ρ)) for all instances with $n = 25$



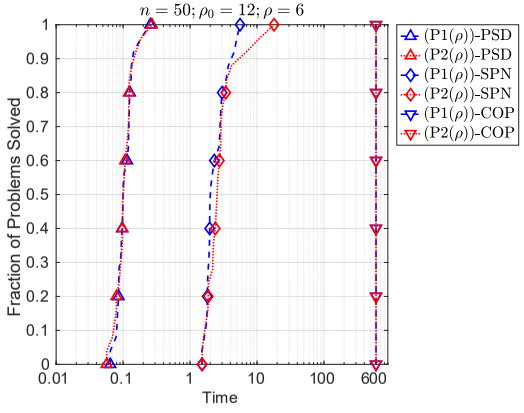
(a) $n = 50, \rho_0 = 12, \rho = 3$



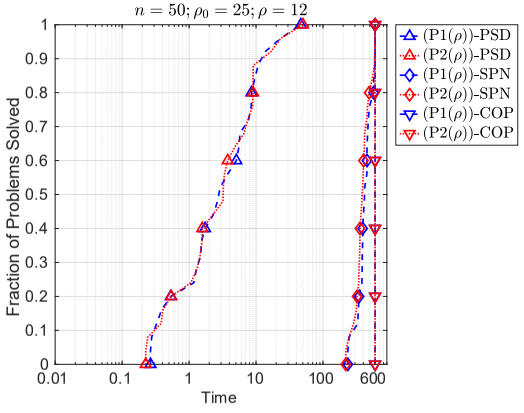
(b) $n = 50, \rho_0 = 25, \rho = 6$



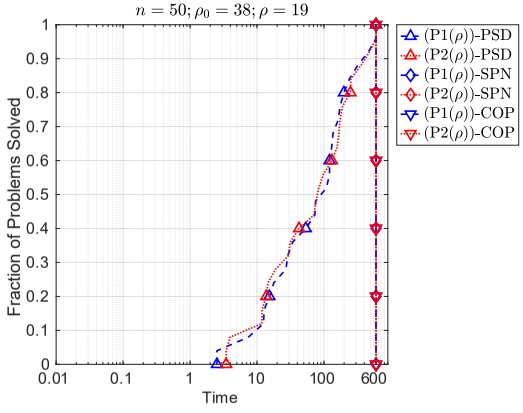
(c) $n = 50, \rho_0 = 38, \rho = 10$



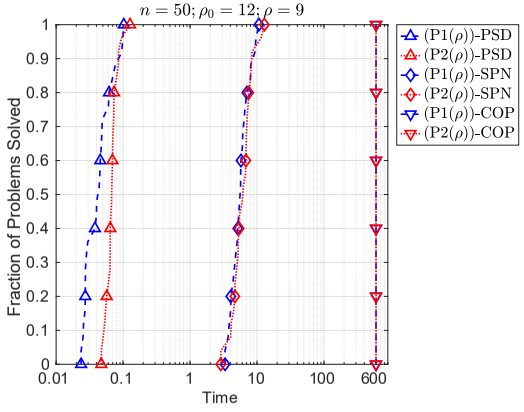
(d) $n = 50, \rho_0 = 12, \rho = 6$



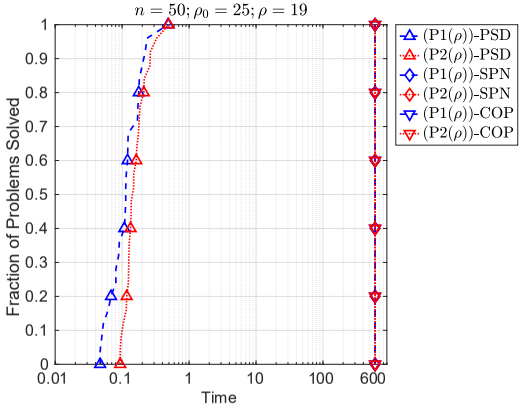
(e) $n = 50, \rho_0 = 25, \rho = 12$



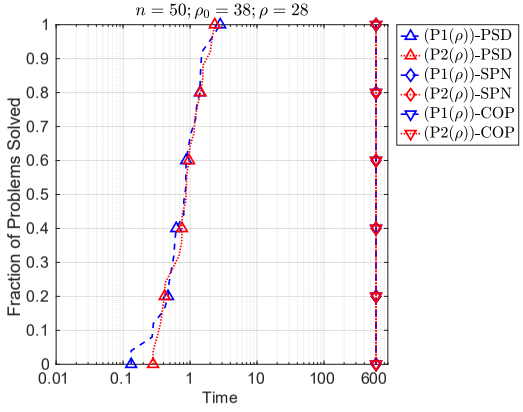
(f) $n = 50, \rho_0 = 38, \rho = 19$



(g) $n = 50, \rho_0 = 12, \rho = 9$



(h) $n = 50, \rho_0 = 25, \rho = 19$



(i) $n = 50, \rho_0 = 38, \rho = 28$

Figure 4.4.: Empirical cumulative distribution functions of solution times of (P1(ρ)) and (P2(ρ)) for all instances with $n = 50$

In summary, the solution time of exact models (P1(ρ)) and (P2(ρ)) seem to be highly dependent on each of the choices of the instance set, (n, ρ_0, ρ), as well as on the ratio ρ/ρ_0. COP instances tend to be the most challenging for the exact models, followed by SPN instances, which, in turn, are followed by PSD instances. These results illustrate that Algorithms 2 and 3 are capable of generating instances of (StQP(ρ)) that are particularly challenging for `Gurobi`, even in relatively small dimensions.

Convex DNN relaxations

Now let us focus on solution times of the four convex relaxations (D1A(ρ)), (D1B(ρ)), (D2A(ρ)), and (D2B(ρ)).

In contrast with the exact models, Figure 4.2 illustrates that the solution times of convex relaxations do not exhibit a strong correlation with the choices of the instance set and the tuple (ρ_0, ρ). As such, we do not present the counterparts of Figures 4.3 and 4.4 for the convex relaxations due to space considerations.

Based on Figure 4.2, we make the following observations about the convex relaxations (D1A(ρ)), (D1B(ρ)), (D2A(ρ)), and (D2B(ρ)):

(i) The distributions of the solution times clearly illustrate the computational advantages of the reduced formulations (D2A(ρ)) and (D2B(ρ)) over their counterparts.

(ii) We observe an empirical first-order stochastic dominance among the distributions of the solution times of (D1A(ρ)), (D1B(ρ)), (D2A(ρ)), and (D2B(ρ)) on each of the six graphs. In particular, we observe that the reduced formulation (D1B(ρ)) consistently achieves the lowest solution times, followed by the reduced formulation (D2B(ρ)), which, in turn, is followed by (D2A(ρ)) and (D1A(ρ)), respectively. In addition to being the theoretically tightest relaxation, it is worth noticing that (D1B(ρ)) also outperforms (D2B(ρ)) in terms of the solution time.

(iii) For $n = 25$, each of the four relaxations can be solved to optimality within the time limit for each instance set and each choice of ρ_0 and ρ. For $n = 50$, we observe that (D1A(ρ)) was terminated due to the time limit on every instance whereas (D2A(ρ)) was terminated due to the time limit on some subsets of the PSD and SPN instances. In contrast, each of the two reduced formulations (D1B(ρ)) and (D2B(ρ)) was solved to optimality within the time limit on all instances. These observations clearly demonstrate the computational advantages of the reduced formulations (D2A(ρ)) and (D2B(ρ)) over their counterparts, especially in higher dimensions.

(iv) For PSD instances with $n = 25$, the distributions of the solution times of each of the two exact models exhibit a first-order stochastic dominance on the corresponding distribution of each of the four convex relaxations. In contrast, we observe that the solution times of the convex relaxations achieve better performance on a larger subset of instances for each instance set with a larger n. For fixed n, a comparison of PSD, SPN, and COP instances reveals an increasingly better performance of the convex relaxations in comparison with the exact models.

Table 4.2 reports the number of instances on which (D2A(ρ)) is terminated due to the time limit. Note that we omit (D1A(ρ)) since all instances with $n = 50$ were terminated due to the time limit. It is worth noticing that (D2A(ρ)) was solved to optimality within the time limit on all COP instances with $n = 50$. In contrast with $n = 25$, we therefore conclude that the solution

time of (D2A(ρ)) for $n = 50$ seems to be somewhat sensitive to the set of instances and the choices of the parameters (ρ_0, ρ).

Instance Set	n	ρ_0	ρ	(D2A(ρ))
PSD	50	12	9	1
		25	6	14
		25	12	10
		25	19	3
		38	10	8
		38	19	22
		38	28	18
SPN	50	12	3	1
		25	6	12
		25	12	8
		25	19	2
		38	10	20
		38	19	24
		38	28	10
COP	50	–	–	–

Table 4.2.: Number of instances (out of 25) terminated due to the time limit (excluding (D1A(ρ)) that was terminated due to the time limit on *all* instances with $n = 50$)

Finally, we recall that optimality gaps are not reported on instances terminated due to the time limit since they are not particularly meaningful in our setting.

In conclusion, the solution time of each relaxation seems to be very robust with respect to the choice of the instance set and the choice of (ρ_0, ρ) in our setting. The reduced formulations (D2A(ρ)) and (D2B(ρ)) yield significant computational advantages over their counterparts, especially in higher dimensions.

4.5.3. Quality of lower bounds

In this section, we discuss the quality of the lower bounds arising from the convex relaxations. To that end, we compare the absolute gaps of the exact models with those of the convex relaxations. For an instance, we define the MIQP gap as the difference between the best upper bound and the best lower bound obtained from (P1(ρ)) and (P2(ρ)). Similarly, we define the DNN gap to be the difference between the best upper bound obtained from the two exact models, (P1(ρ)) and (P2(ρ)), and the best lower bound obtained from the reduced formulations (D2A(ρ)) and (D2B(ρ)). Note that we use the lower bounds from the reduced formulations since all of them can be solved to optimality on all instances.

If at least one of (P1(ρ)) and (P2(ρ)) can solve an instance to optimality, then the MIQP gap will be very close to zero. On such an instance, we deem that our relaxation is *exact* if the DNN gap is of similar magnitude to that of the MIQP gap. If each of (P1(ρ)) and (P2(ρ)) is terminated due to the limit on an instance, then the MIQP gap will be sufficiently away from zero. In this case, if the DNN gap is smaller than or equal to the MIQP gap, we say that our relaxations are better than the exact modes. Otherwise, our relaxations are *worse* than the exact models.

In Figure 4.5, which is organized similarly to Figure 4.2, we plot the DNN and MIQP gaps. To ease reading, we have accumulated all instances of the same dimensions from each instance set across all different parameter constellations (ρ_0, ρ), yielding 225 such instances in total for every graph (a)-(f), depicting the situation in the different hardness classes of the generated instances. In each graph, the horizontal axis represents the instances ordered in nondecreasing MIQP gaps, which are represented by the blue curve, and the vertical axis denotes the absolute gap. Finally, the markers indicate the data points at every 25^{th} instance.

We outline our observations:

(i) For PSD instances, our relaxations are exact on the vast majority of all instances solved to optimality by an exact model and are better than the exact models on all instances with a positive MIQP gap.

(ii) For SPN instances, our relaxations are exact on a relatively smaller proportion of all instances solved to optimality by an exact model and are better than the exact models on all instances with a positive MIQP gap. Furthermore, it is worth noticing that our relaxations continue to be exact on several instances with a positive MIQP gap.

(iii) On COP instances, our relaxations are worse than the exact models on the vast majority of instances. However, we remark that the DNN gap is smaller on a small subset of instances with $n = 50$ that admit a positive MIQP gap.

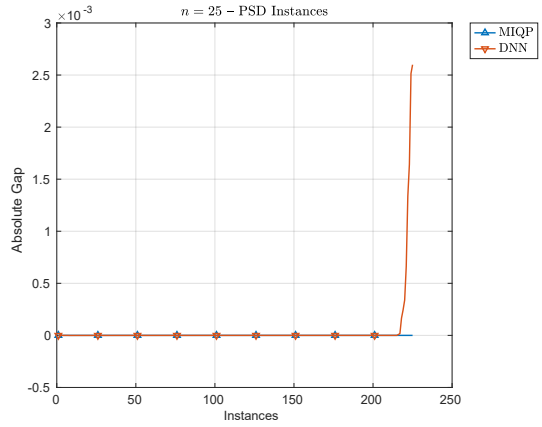
The results show clearly that the DNN relaxations provide tight lower bounds on a large number of PSD and SPN instances and that there is a significant number of instances where the MIQP gap is larger than the gaps achieved by the DNN relaxations, sometimes drastically so (one or even two orders of magnitude). The dominance of MIQP models on the extremely difficult COP instances in graphs (e) and (f) may be explained by the fact that additional cuts may be needed to tighten the DNN gaps without resorting to higher-order relaxations of the (intractable) exact conic reformulations.

We close this part by reporting that we have not observed a significant difference between the quality of the lower bounds arising from the provably tighter relaxation (D1B(ρ)) and the weaker (D2B(ρ)). However, we still recommend using the tighter (D1B(ρ)) since it also outperforms (D2B(ρ)) in terms of the solution time on our instance sets.

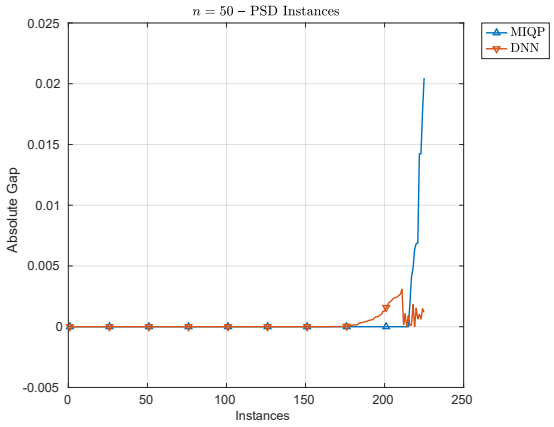
In this section, we studied the sparse StQP. By utilizing exact copositive reformulations of two mixed integer quadratic optimization models, we proposed two tractable convex relaxations. For both relaxations, we established equivalent tractable reformulations in significantly smaller dimensions. We also presented a theoretical comparison of the two relaxations. Finally, we proposed an instance generation scheme that is guaranteed to construct nontrivial instances. Our computational results clearly illustrate the computational advantages of our reduced relaxations as well as the quality of the relaxation bounds.

4.6. Conclusion

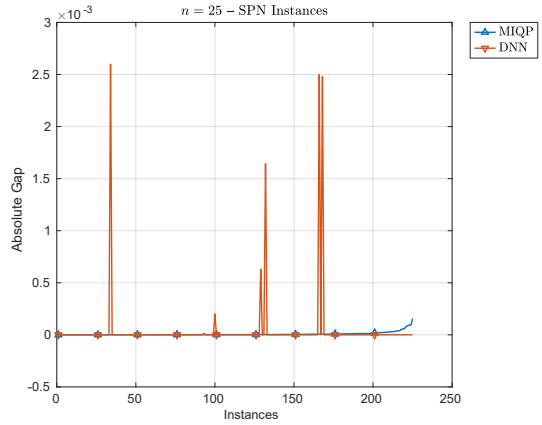
In this chapter, we studied the sparse StQP. By utilizing exact copositive reformulations of two mixed integer quadratic optimization models, we proposed two tractable convex relaxations. For both relaxations, we established equivalent tractable reformulations in significantly smaller dimensions. We also presented a theoretical comparison of the two relaxations. Finally, we proposed an instance generation scheme that is guaranteed to construct nontrivial instances. Our computational results clearly illustrate the computational advantages of our reduced relaxations as well as the quality of the relaxation bounds.



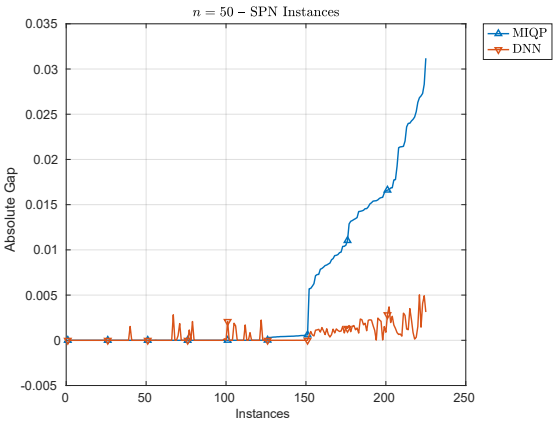
(a) PSD Instances ($n = 25$)



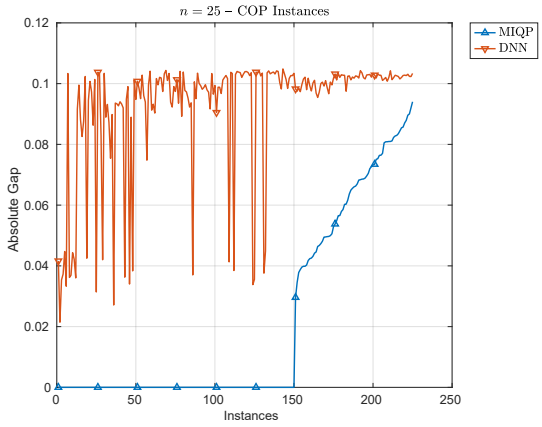
(b) PSD Instances ($n = 50$)



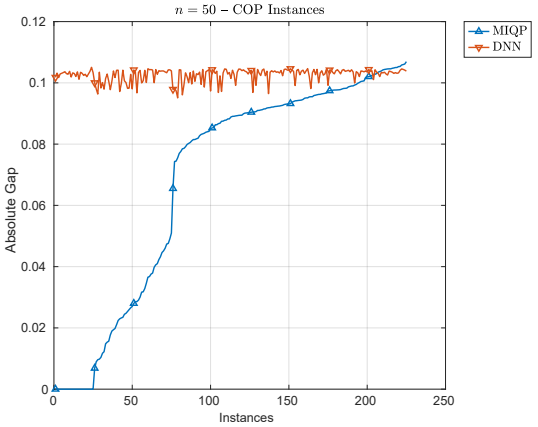
(c) SPN Instances ($n = 25$)



(d) SPN Instances ($n = 50$)



(e) COP Instances ($n = 25$)



(f) COP Instances ($n = 50$)

Figure 4.5.: Absolute gaps of the best of (P1(ρ)) and (P2(ρ)) versus the best of (D1B(ρ)) and (D2B(ρ))

5. A scalable conic approach for feature section in linear SVMs

In this chapter, we study the embedded feature selection problem in linear Support Vector Machines (SVMs), in which a cardinality constraint is employed, leading to an explainable selection model. To handle the hard problem, we first introduce two mixed-integer formulations for which novel semidefinite relaxations are proposed. In addition, we propose both heuristics and exact algorithms using the information of its optimal solution. The content of this chapter is based on the submitted paper [BDPP24] for which revisions were requested by the editor after the first round of evaluations.

5.1. The problem

Given a dataset with m samples, and n features, $\{(\mathsf{x}^i, y^i) \in \mathbb{R}^n \times \{-1,1\} : i \in [1\!:\!m]\}$, the linear SVM classification problem defines a classifier as the function $f : \mathbb{R}^n \to \{-1,1\}$,

$$f(\mathsf{x}) = \text{sign}\ ({\mathsf{w}^*}^\top \mathsf{x} + b^*)\,.$$

In traditional SVMs [CV95], coefficients $(\mathsf{w}^*, b^*) \in \mathbb{R}^n \times \mathbb{R}$ identify a separating hyperplane (if it exists) that maximizes its margin, i.e. the distance between the hyperplane and the closest data points of each class. Accounting for classes not linearly separable, the tuple (w^*, b^*) is found by solving the following convex quadratic problem:

$$\begin{array}{lll} \text{(SVM)} & \min\limits_{\mathsf{w},b,\boldsymbol{\xi}} & \frac{1}{2}\|\mathsf{w}\|_2^2 + C\sum_{i=1}^m \xi_i \\ & \text{s.t.} & y^i(\mathsf{w}^\top \mathsf{x}^i + b) \geq 1 - \xi_i\,, \quad i \in [1\!:\!m]\,, \\ & & \boldsymbol{\xi} \geq \mathsf{o}\,. \end{array} \tag{5.1}$$

The non-negative variables ξ allow points to be on the wrong side of their "soft margin", as well as being "too close" to the decision boundary. The sum $\sum_{i=1}^m \xi_i$ in the objective function represents an upper bound on the total misclassification error, which is weighted by the hyperparameter $C > 0$ so to balance the maximization of the margin, which is inversely proportional to $\|\mathsf{w}\|_2$.

The optimal solution w^* of (5.1) is usually dense, in the sense that usually all components of w^* are nonzero and thus every feature contributes to the definition of the classifier. Indeed, through duality theory, w^* can be expressed as the linear combination of a subset of samples, called support vectors. More precisely, $\mathsf{w}^* = \sum_{i=1}^m \alpha_i^* y^i \mathsf{x}^i$, with $\alpha_i^* \geq 0$ being the dual optimal solution of problem (5.1) with nonzero value only if sample i represents a support vector. Thus, it is unlikely to obtain $w_j^* = 0$ for some feature j, and sparsity in the vector w^* must be forced. In this work, we aim to strictly control sparsity by imposing hard cardinality constraints on the number of nonzero components in w^* and this leads to the following problem:

$$\begin{array}{lll} \text{(FS-SVM)} & \min\limits_{\mathsf{w},b,\boldsymbol{\xi}} & \frac{1}{2}\|\mathsf{w}\|_2^2 + C\sum_{i=1}^m \xi_i \\ & \text{s.t.} & y^i(\mathsf{w}^\top \mathsf{x}^i + b) \geq 1 - \xi_i\,, \quad i \in [1\!:\!m]\,, \\ & & \|\mathsf{w}\|_0 \leq B \\ & & \boldsymbol{\xi} \geq \mathsf{o}\,, \end{array} \tag{5.2}$$

where the user-defined parameter $B \in \mathbb{Z}_+$ constitutes a budget on the number of features used, indicating that the number of features involved in the classifier is at most B. Even though objective function and all other constraints are convex quadratic or linear, the cardinality constraint $\|\mathsf{w}\|_0 \leq B$ is of combinatorial type, which renders (5.2) NP-hard.

The integration of feature selection in linear SVM has been intensively studied over the past few decades. There are essentially two types of models for feature selection. The first approach involves using surrogates to pursue sparse solutions, such as well-known ℓ_1 regularization and its variants, such as [FM04, HRTZ04, NDlT10, YCHL10, WZZ06, ZH05]. The second approach directly addresses the cardinality constraint. Several studies [MPWL14, LMMRC19] focus on the representing the constraint through mixed-integer modelling, incorporating a large constant (referred to as big-M), and solving the problem using LP-based branch and bound methods. However, the numerical performance of this approach is highly dependent on the choice of the big-M constant. Estimating this constant is often as challenging as solving the original problem itself, leading to inefficiencies, particularly when limited information is available regarding the optimal value of big-M.

In this chapter, we propose several novel semidefinite relaxations based on the complementarity reformulation of the problem (5.2). Exploiting the sparsity pattern of the relaxations, we decompose the problems and obtain equivalent relaxations in a much smaller cone, making the conic approaches scalable. To make the best usage of the decomposed relaxations, we propose heuristics using the information of its optimal solution. Moreover, an exact procedure is proposed by solving a sequence of mixed-integer decomposed semidefinite optimization problems. Numerical results on classical benchmarking datasets are reported, showing the efficiency and effectiveness of our approach. Before proceeding to the details, we introduce several sets that will be frequently used throughout the remainder of the text.

The margin set is denoted by

$$\Xi = \{(\mathsf{w}, b, \boldsymbol{\xi}) \in \mathbb{R}^{n+1+m} : y^i(\mathsf{w}^\top \mathsf{x}^i + b) \geq 1 - \xi_i\,,\ i \in [1\!:\!m]\,, \boldsymbol{\xi} \geq \mathsf{o}\}\,, \tag{5.3}$$

so that problem (5.2) reads as:

$$\begin{array}{ll} \min\limits_{\mathsf{w},b,\boldsymbol{\xi}} & \frac{1}{2}\|\mathsf{w}\|_2^2 + C\sum_{i=1}^m \xi_i \\ \text{s.t.} & (\mathsf{w}, b, \boldsymbol{\xi}) \in \Xi \\ & \|\mathsf{w}\|_0 \leq B. \end{array}$$

In the next sections, we will also make use of big-M reformulations, so it is convenient to introduce the two sets

$$\Omega_M = \{(\mathsf{w}, \mathsf{v}) \in \mathbb{R}^{2n} : -M\mathsf{v} \leq \mathsf{w} \leq M\mathsf{v}\,,\ \mathsf{e}^\top \mathsf{v} = B\} \tag{5.4}$$

and

$$\mathcal{B}_M = \{(\mathsf{w}, \mathsf{u}) \in \mathbb{R}^{2n} : -M(1-\mathsf{u}) \leq \mathsf{w} \leq M(1-\mathsf{u})\,,\ \mathsf{e}^\top \mathsf{u} = n - B\}\,, \tag{5.5}$$

where the big-M parameter $M > 0$ is large enough (see Remark 4 below) and the variables $\mathsf{u} = 1 - \mathsf{v}$ are the negated v, sometimes more convenient in notation.

5.2. Exact Mixed-Integer approaches

In this section, we propose two exact mixed-integer formulations of the original problem (5.2), using two different mixed-integer systems to present the cardinality constraint with auxillary

binary variables introduced as an indicator, namely a big-M strategy and a complementarity constraint strategy.

Differently from [LMMRC19], we study the following MIQP model incorporating the standard ℓ_2-norm on the hyperplane weights as the regularization term:

$$\begin{aligned} \text{(BigMP)} \quad \min_{\mathsf{w},b,\boldsymbol{\xi},\mathsf{v}} \quad & \tfrac{1}{2}\|\mathsf{w}\|_2^2 + C\sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & (\mathsf{w}, b, \boldsymbol{\xi}) \in \Xi \\ & (\mathsf{w}, \mathsf{v}) \in \Omega_M \\ & \mathsf{v} \in \{0,1\}^n\,, \end{aligned} \tag{5.6}$$

where $\boldsymbol{\Xi}$ and Ω_M are defined in (5.3) and (5.4).

To show the equivalence between (5.2) and (5.6), we consider a B-sparse w, e.g., $\|\mathsf{w}\|_0 \le B$. If $\|\mathsf{w}\|_0 = B$, there exists a $\mathsf{v} \in \{0,1\}^n$ such that $(\mathsf{w}, \mathsf{v}) \in \Omega_M$, where v is given by the rule that $v_j = 0$ if $w_j = 0$ and $v_j = 1$ if $w_j \neq 0$; If $\|\mathsf{w}\|_0 < B$, there also exists a $\mathsf{v} \in \{0,1\}^n$ such that $(\mathsf{w}, \mathsf{v}) \in \Omega_M$, where v is given by the rule that $v_j = 1$ for $j \in I$ and $v_j = 0$ otherwise, where I is a index set covers all of features with $w_j \neq 0$, satisfying $|I| = B$.

Remark 2. *It is worth noticing that the optimal value of model* (5.6) *remains the same if we substitute the equality constraint* $\mathsf{e}^\top \mathsf{v} = B$ *by the budget constraint* $\mathsf{e}^\top \mathsf{v} \le B$*, which is derived directly from the targeted problem* (5.2).

Alternatively, instead of expressing the cardinality constraint by the big-M formulation, we employ a set of binary variables $\mathsf{u} \in \{0,1\}^n$ and a complementarity constraint to count the nonzero elements of w and end up with another mixed-integer reformulation of (5.2):

$$\begin{aligned} \text{(CoP)} \quad \min_{\mathsf{w},b,\boldsymbol{\xi},\mathsf{u}} \quad & \tfrac{1}{2}\|\mathsf{w}\|_2^2 + C\sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & (\mathsf{w}, b, \boldsymbol{\xi}) \in \Xi \\ & u_j w_j = 0\,, \qquad\qquad j \in [1\!:\!n] \\ & \mathsf{e}^\top \mathsf{u} = n - B \\ & \mathsf{u} \in \{0,1\}^n\,. \end{aligned} \tag{5.7}$$

Unlike (5.6), the binary variables u_j count the number of zero elements of w. For a (5.7)-feasible point $(\mathsf{w}, b, \boldsymbol{\xi}, \mathsf{u})$, we let $\mathsf{v} = 1 - \mathsf{u}$ and the point $(\mathsf{w}, b, \boldsymbol{\xi}, \mathsf{v})$ is (5.6)-feasible. To show this, we only need to show $(\mathsf{w}, \mathsf{v}) \in \Omega_M$, namely, that (w, v) satisfies the big-M constraint in (5.4). It is clear that the constraint holds for the coordinates of w with $w_j = 0$. On the other hand, the complementarity constraint $(1 - v_j)w_j = 0$ implies $v_j = 1$ if $w_j \neq 0$, and therefore the point $(\mathsf{w}, \mathsf{v}) \in \Omega_M$. Hence, the problem (5.7) is equivalent to the problem (5.6).

Remark 3. *In the above problem* (5.7)*, the constraint* $\mathsf{e}^\top \mathsf{u} = n - B$ *can be equivalently replaced by the constraint* $\mathsf{e}^\top \mathsf{u} \ge n - B$ *for the same reason as Remark 2, and the binary constraint* $u_j \in \{0,1\}$ *can be equivalently reduced to the box constraint* $0 \le u_j \le 1$ *due to the existence of the complementarity constraint. To show this, we consider the model with the box constraint and assume that* w *is* ρ*-sparse with* $\rho > B$*. Due to the constraint* $u_j w_j = 0$*,* $u_j = 0$ *if* w_j *is nonzero and* $0 \le u_j \le 1$ *if* w_j *is zero. Hence,* $\mathsf{e}^\top \mathsf{u} \le n - \rho < n - B$*, which contradicts the budget constraint. Therefore all* w *vectors with cardinality strictly larger than* B *are rendered infeasible to the model* (5.7) *and its relaxation* (5.8) *below.*

In the first formulation, we need a big-M parameter M for the reformulation. It is known that the choice of M has a profound influence on the computational performance if we solve

the problem with off-the-shell global solvers. In a branch-and-bound framework, we solve LP relaxation at each node, and in many cases, the LP relaxation can be as weak as the SVM problem without feature selection, if too large M is chosen, as shown later in Theorem 35. In contrast, in the second formulation, the difficulty of the problem mainly goes to the complementarity constraint, and no big-M needed.

5.3. Tractable Relaxations

In this section, we propose several tractable relaxations, e.g., LP relaxations, and SDP relaxations, of the FS-SVM model, and a comparison is made among some of them. Moreover, a decomposition strategy is proposed by exploring the sparse pattern involved in the both objective and feasible region of the original problem, leading to a much smaller and scalable SDP-based relaxation.

5.3.1. Box relaxation

In this subsection, we study the box relaxation (also known as a type of LP relaxation) of two exact mixed-integer formulations presented in the last chapter and point out the crucial role of M that it plays in the first formulation.

After relaxing the binarity constraints to the box constraints in the problem (5.6), we arrive at:

$$\begin{aligned} \text{(BoxMP)} \quad \min_{\mathsf{w},b,\boldsymbol{\xi},\mathsf{v}} \quad & \tfrac{1}{2}\|\mathsf{w}\|_2^2 + C\sum_{i=1}^m \xi_i \\ & (\mathsf{w}, b, \boldsymbol{\xi}) \in \Xi \\ & (\mathsf{w}, \mathsf{v}) \in \Omega_M \\ & \mathsf{o} \leq \mathsf{v} \leq \mathsf{e}\,. \end{aligned} \tag{5.8}$$

The relaxation is a convex quadratic optimization problem solvable in polynomial time, therefore it has been widely used in many previous works, see references in [LMMRC19]. However, the tightness of the relaxation is highly dependent on the choice of M and there is a possibility that it can be as weak as the original SVM model without feature selection (5.1). To be more precise, if M is larger than a data-based threshold, the relaxation (5.8) is equivalent to the problem (5.1), as presented in the following theorem.

Theorem 35. *If the parameter M in the problem* (5.8) *satisfies*

$$M \geq \frac{\|\mathsf{w}^*\|_1}{B}\,,$$

where $(\mathsf{w}^*, b^*, \boldsymbol{\xi}^*)$ is an (5.1)*-optimal solution, then* (5.8) *is equivalent to* (5.1)*.*

Proof. It is clear to see that (5.1) is a relaxation of (5.8), as the former includes the constraint $(\mathsf{w}, b, \boldsymbol{\xi}) \in \Xi$. Hence the optimal value of (5.1) is not larger than that of (5.8). To show the converse, we assume that $(\mathsf{w}^*, b^*, \boldsymbol{\xi}^*)$ is optimal to the problem (5.1). Let $v_j = \frac{|w_j^*|}{M}$ for all $j \in [1\!:\!n]$. Then the point $(\mathsf{w}^*, b^*, \boldsymbol{\xi}^*, \mathsf{v})$ is (5.8)-feasible. Indeed, due to the construction of v_j, the constraint $-Mv_j \leq w_j \leq Mv_j$ is satisfied automatically for all $j \in [1\!:\!n]$. Since $\mathsf{e}^\top \mathsf{v} = \frac{\|\mathsf{w}^*\|_1}{M}$, the constraint $\mathsf{e}^\top \mathsf{v} \leq B$ is implied by the assumption we made. Since the objective functions in both problems coincide, the claim follows. □

On the other hand, M can not be chosen smaller than:

$$M \geq \|\overline{\mathsf{w}}^*\|_\infty \,,$$

where $\overline{\mathsf{w}}^*$ is part of an optimal solution to (5.2). Otherwise, the optimal solution will be cut off in the sense that the problem (5.6) is not a reformulation for an insufficient choice of M. Hence, the proper range of the parameter M has to be:

$$M \in \left[\|\overline{\mathsf{w}}^*\|_\infty, \frac{\|\mathsf{w}^*\|_1}{B} \right). \tag{5.9}$$

However, this set can be empty depending on the problem data itself, which means no matter what M we chose, the box relaxation (5.8) ends up with one of the following situations: (i) no improvement compared to the original SVM problem (5.1); (ii) the optimal solution of the targeted problem (5.2) is cut off, because M is too small to render (5.8) a reformulation of (5.2). This is quite problematic in practice and it encourages us to explore the other formulation (5.7).

Remark 4. *Since estimating the range of $\overline{\mathsf{w}}^*$ is as difficult as solving the problem itself, we usually choose M under a safe threshold, which is much larger than the tightest upper bound. In most practical cases, the condition $M \geq \frac{\|\mathsf{w}^*\|_1}{B}$ holds then as well, which means the LP relaxation* (5.8) *does not provide any improvement compared to the plain problem* (5.1)*.*

Due to the presence of complementarity constraints, the box relaxation of (5.7) is equivalent to (5.7) itself, as mentioned in Remark 3. However, the combinatorial nature of a complementarity constraint is the main difficulty in both the unrelaxed and the relaxed version, so this continuous reformulation does not solve the problem immediately without any further processing.

5.3.2. The Shor relaxation

In this subsection, we study the Shor relaxation, a type of SDP relaxation, of the two proposed mixed-integer reformulations.

Letting $\mathsf{W} = \mathsf{w}\mathsf{w}^\top$, $\mathsf{V} = \mathsf{v}\mathsf{v}^\top$ and dropping the rank one constraint, the Shor relaxation of (5.6) reads as:

$$\begin{aligned} \text{(SMP)} \quad \min_{\mathsf{w},\mathsf{W},b,\boldsymbol{\xi},\mathsf{v},\mathsf{V}} \quad & \tfrac{1}{2}\text{trace}\,(\mathsf{W}) + C\sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & (\mathsf{w}, b, \boldsymbol{\xi}) \in \Xi \\ & (\mathsf{w}, \mathsf{v}) \in \Omega_M \\ & \text{diag}\,\mathsf{V} = \mathsf{v} \\ & \begin{bmatrix} 1 & \mathsf{w}^\top \\ \mathsf{w} & \mathsf{W} \end{bmatrix} \in \mathcal{S}_+^{n+1}, \; \begin{bmatrix} 1 & \mathsf{v}^\top \\ \mathsf{v} & \mathsf{V} \end{bmatrix} \in \mathcal{S}_+^{n+1}. \end{aligned} \tag{5.10}$$

Theorem 36. *The Shor relaxation* (5.10) *is equivalent to the box relaxation* (5.8)*.*

Proof. For a (5.10)-feasible point $(\mathsf{w}, \mathsf{W}, b, \boldsymbol{\xi}, \mathsf{v}, \mathsf{V})$, the projected point $(\mathsf{w}, b, \boldsymbol{\xi}, \mathsf{v})$ satisfies all of the constraints of (5.8). Moreover, we have $\text{trace}\,(\mathsf{W}) \geq \|\mathsf{w}\|_2^2$, hence the relaxation (5.8) is weaker than the relaxation (5.10). To show the other direction, we suppose that the point $(\mathsf{w}, b, \boldsymbol{\xi}, \mathsf{v})$ is feasible to the problem (5.8), and let $\mathsf{W} = \mathsf{w}\mathsf{w}^\top$ and $\mathsf{V} = \mathsf{v}\mathsf{v}^\top + \mathsf{Z}$, where $\mathsf{Z} \in \mathcal{S}_+^n$ is a diagonal matrix with $Z_{jj} = v_j - v_j^2 \geq 0$ for all $j \in [1\!:\!n]$. The constructed point $(\mathsf{w}, \mathsf{W}, b, \boldsymbol{\xi}, \mathsf{v}, \mathsf{V})$ then satisfies all constraints of (5.10). Indeed, the last positive-semidefiniteness constraint follows by the construction of V due to the Schur complement, which therefore completes the proof. □

Letting $\mathsf{W} = \mathsf{w}\mathsf{w}^\top$, $\mathsf{R} = \mathsf{w}\mathsf{u}^\top$, $\mathsf{U} = \mathsf{u}\mathsf{u}^\top$ and dropping the rank-one constraint, we arrive at the Shor relaxation of (5.7):

$$\begin{aligned} (\text{SCoP}) \qquad \min_{\substack{\mathsf{w},\mathsf{W},b,\boldsymbol{\xi},\\ \mathsf{R},\mathsf{u},\mathsf{U}}} \quad & \tfrac{1}{2}\text{trace }(\mathsf{W}) + C\textstyle\sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & (\mathsf{w}, b, \boldsymbol{\xi}) \in \Xi \\ & \mathsf{e}^\top \mathsf{u} = n - B \\ & \text{diag } \mathsf{R} = \mathsf{o}, \quad \text{diag } \mathsf{U} = \mathsf{u} \\ & \begin{bmatrix} 1 & \mathsf{w}^\top & \mathsf{u}^\top \\ \mathsf{w} & \mathsf{W} & \mathsf{R} \\ \mathsf{u} & \mathsf{R}^\top & \mathsf{U} \end{bmatrix} \in \mathcal{S}_+^{2n+1} \end{aligned} \tag{5.11}$$

Theorem 37. *If there exists part* w *of an optimal solution to* (5.1) *with* $\|\mathsf{w}\|_0 > B$*, then the optimal value of* (5.11) *is strictly larger than the optimal value of* (5.1)*.*

Proof. For a given (5.1)-feasible point $(\mathsf{w}, b, \boldsymbol{\xi})$ with $\|\mathsf{w}\|_0 > B$, we restrict the corresponding variables of (5.11) to the same value and estimate the optimal value of the restricted version of (5.11). The matrix

$$\begin{bmatrix} 1 & \mathsf{w}^\top & \mathsf{u}^\top \\ \mathsf{w} & \mathsf{W} & \mathsf{R} \\ \mathsf{u} & \mathsf{R}^\top & \mathsf{U} \end{bmatrix} \in \mathcal{S}_+^{2n+1}, \quad \text{with} \quad \text{diag } \mathsf{R} = \mathsf{o} \text{ and } \text{diag } \mathsf{U} = \mathsf{u}\,,$$

has a principal submatrix

$$\begin{bmatrix} 1 & w_j & u_j \\ w_j & W_{jj} & 0 \\ u_j & 0 & u_j \end{bmatrix} \in \mathcal{S}_+^3, \quad \text{for all } j \in [1\!:\!n]\,,$$

which implies that

$$\begin{cases} 0 \le u_j \le 1\,, \quad W_{jj} \ge w_j^2, & j \in [1\!:\!n] \\ (W_{jj} - w_j^2)(u_j - u_j^2) - w_j^2 u_j^2 \ge 0\,, & j \in [1\!:\!n]\,. \end{cases}$$

It follows that

$$\begin{cases} 0 \le u_j \le 1\,, & \text{if} \quad j \in S_0 := \{j : w_j = 0\}\,, \\ 0 \le u_j \le \dfrac{W_{jj} - w_j^2}{W_{jj}} < 1 & \text{if} \quad j \in S_+ := [1\!:\!n] \setminus S_0 \end{cases}$$

(observe that $w_j > 0$ implies $W_{jj} > 0$). Recalling the assumption $\|\mathsf{w}\|_0 > B$, it follows that

$$\sum_{j \in S_0} u_j < n - B\,.$$

Together with the constraint

$$\mathsf{e}^\top \mathsf{u} = n - B\,,$$

we have $u_j > 0$ for some $j \in S_+$, which implies that $W_j > w_j^2$ for some $j \in S_+$. Hence, the optimal value of (5.11) is strictly larger than the optimal value of (5.1). □

There is no direct comparison between the relaxation (5.10) and the relaxation (5.11). But given the case we have a valid estimate of M at hand satisfying (5.9), the relaxation (5.11) can be tightened with a small computational cost by adding the big-M constraint via $\mathcal{B}_M$, which reads as:

$$\begin{aligned}
\text{(SCoMP)} \quad \min_{\substack{\mathsf{w},\mathsf{W},b,\boldsymbol{\xi},\\ \mathsf{R},\mathsf{u},\mathsf{U}}} \quad & \tfrac{1}{2}\text{trace }(\mathsf{W}) + C\textstyle\sum_{i=1}^m \xi_i \\
\text{s.t.} \quad & (\mathsf{w}, b, \boldsymbol{\xi}) \in \Xi \\
& (\mathsf{w}, \mathsf{u}) \in \mathcal{B}_M \\
& \text{diag } \mathsf{R} = \mathsf{o}\,, \quad \text{diag } \mathsf{U} = \mathsf{u}\,, \\
& \begin{bmatrix} 1 & \mathsf{w}^\top & \mathsf{u}^\top \\ \mathsf{w} & \mathsf{W} & \mathsf{R} \\ \mathsf{u} & \mathsf{R}^\top & \mathsf{U} \end{bmatrix} \in \mathcal{S}_+^{2n+1}\,.
\end{aligned} \tag{5.12}$$

Theorem 38. *The relaxation* (5.12) *is tighter than the relaxations* (5.10) *and* (5.11).

Proof. It is clear that the relaxation (5.12) is tighter than the relaxation (5.11) due to the additional big-M constraint. To show the other claim, we construct a (5.10)-feasible point, given a (5.12)-feasible point. Suppose that the point $(\mathsf{w}, \mathsf{W}, b, \boldsymbol{\xi}, \mathsf{R}, \mathsf{u}, \mathsf{U})$ is (5.12)-feasible, let $\mathsf{v} = \mathsf{e} - \mathsf{u}$ and $\mathsf{V} = \mathsf{v}\mathsf{v}^\top + \mathsf{Z}$, where $\mathsf{Z} \in \mathcal{S}_+^n$ is a diagonal matrix with $Z_{jj} = v_j - v_j^2 \geq 0$ for all $j \in [1\!:\!n]$. Then $(\mathsf{w}, \mathsf{W}, b, \boldsymbol{\xi}, \mathsf{v}, \mathsf{V})$ is (5.10)-feasible with the same objective value. □

5.3.3. Decomposed SDP-based relaxations

In this subsection, we propose a procedure for obtaining equivalent and much smaller SDP-based relaxations by exploiting the sparsity patterns of the Shor relaxations.

We notice that, for example: in problem (5.11), some sparsity patterns appear both in the objective and the constraints. e.g., only diagonal elements of matrices W, R, and U are involved. Based on this observation, we write down the following decomposed version of (5.11):

$$\begin{aligned}
\text{(DSCoP)} \quad \min_{\mathsf{w},\text{diag }\mathsf{W},b,\boldsymbol{\xi},\mathsf{u}} \quad & \tfrac{1}{2}\text{trace }(\mathsf{W}) + C\textstyle\sum_{i=1}^m \xi_i \\
\text{s.t.} \quad & (\mathsf{w}, b, \boldsymbol{\xi}) \in \Xi \\
& \mathsf{e}^\top \mathsf{u} = n - B \\
& \begin{bmatrix} 1 & w_j & u_j \\ w_j & W_{jj} & 0 \\ u_j & 0 & u_j \end{bmatrix} \in \mathcal{S}_+^3\,, \quad j \in [1\!:\!n]
\end{aligned} \tag{5.13}$$

Theorem 39. *The relaxation* (5.11) *is equivalent to the relaxation* (5.13).

Proof. Suppose that $(\mathsf{w}, \mathsf{W}, b, \boldsymbol{\xi}, \mathsf{u})$ is (5.11)-feasible. It is clear that $(\mathsf{w}, \text{diag }\mathsf{W}, b, \boldsymbol{\xi}, \mathsf{u})$ is (5.13)-feasible. In the opposite direction, we assume that the point $(\mathsf{w}, \text{diag }\mathsf{W}, b, \boldsymbol{\xi}, \mathsf{u})$ is (5.13)-feasible. It is clear that all linear constraints in (5.11) are satisfied if we let diag $\mathsf{R} = \mathsf{o}$ and diag $\mathsf{U} = \mathsf{u}$. The only thing that needs to be checked is if the following partial positive-semidefinite matrix is

positive-semidefinite completable:

$$\begin{bmatrix} 1 & w_1 & \dots & w_n & u_1 & \dots & u_n \\ w_1 & W_{11} & * & * & 0 & * & * \\ \vdots & * & \ddots & * & * & \ddots & * \\ w_n & * & * & W_{nn} & * & * & 0 \\ u_1 & 0 & * & * & u_1 & * & * \\ \vdots & * & \ddots & * & * & \ddots & * \\ u_n & * & * & 0 & * & * & u_n \end{bmatrix}.$$

By $\mathcal{G}$ we denote the specification graph of the above partial positive-semidefinite matrix. Since the vertex w_i is not connected with w_j and u_j for all $i \neq j$, and the vertex u_i is not connected with u_j and w_j for all $i \neq j$, the only cycle we can find in the graph is the triangle $1 \to w_j \to u_j \to 1$ for all $[1\,{:}\,n]$. Therefore $\mathcal{G}$ is chordal. Due to [BSM03, Theorem 1.39], if the specification graph of a partial positive-semidefinite matrix is chordal, it is always completable to a positive-semidefinite matrix. □

Following a similar fashion, we propose the decomposed version of the relaxation (5.12):

$$\begin{aligned} \text{(DSCoMP)} \qquad \min_{\mathsf{w},\operatorname{diag}\mathsf{W},b,\boldsymbol{\xi},\mathsf{u}} \quad & \tfrac{1}{2}\operatorname{trace}(\mathsf{W}) + C\textstyle\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & (\mathsf{w}, b, \boldsymbol{\xi}) \in \Xi \\ & (\mathsf{w}, \mathsf{u}) \in \mathcal{B}_M \\ & \begin{bmatrix} 1 & w_j & u_j \\ w_j & W_{jj} & 0 \\ u_j & 0 & u_j \end{bmatrix} \in \mathcal{S}^3_+, \quad j \in [1\,{:}\,n]. \end{aligned} \tag{5.14}$$

Corollary 40. *The relaxation* (5.14) *is equivalent to the relaxation* (5.12).

Proof. See the proof of Theorem 39. □

For the formulation (5.6), the decomposed version of its Shor relaxation (5.15) reads as:

$$\begin{aligned} \text{(DSMP)} \qquad \min_{\mathsf{w},\operatorname{diag}\mathsf{W},b,\boldsymbol{\xi},\mathsf{v}} \quad & \tfrac{1}{2}\operatorname{trace}(\mathsf{W}) + C\textstyle\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & (\mathsf{w}, b, \boldsymbol{\xi}) \in \Xi \\ & (\mathsf{w}, \mathsf{v}) \in \Omega_M \\ & \begin{bmatrix} 1 & w_j \\ w_j & W_{jj} \end{bmatrix} \in \mathcal{S}^2_+, \begin{bmatrix} 1 & v_j \\ v_j & v_j \end{bmatrix} \in \mathcal{S}^2_+, \quad j \in [1\,{:}\,n]. \end{aligned} \tag{5.15}$$

Corollary 41. *The relaxation* (5.15) *is equivalent to the relaxation* (5.10).

Proof. See the proof of Theorem 39. □

5.4. The algorithms

In this section, we introduce both a heuristic and an exact algorithm for solving the feature selection problem we are interested in. In particular, we will first present two heuristic strategies

to find good upper bound values for our problem and a heuristic strategy to estimate a tight value for the big-M parameter M for the problem. All strategies are based on the resolution of the decomposed SDP-based relaxations presented in the previous section. These strategies will then be embedded in a heuristic algorithm which can ultimately improve the heuristic upper bound. Finally, we present an exact algorithm that implements one of the two upper bound procedures as a subroutine and consists of the resolution of a sequence of Mixed-Integer Second-Order Cone Optimization problems (MISOCPs).

5.4.1. Two upper bounding strategies

In this subsection, we propose two strategies for obtaining an upper bound, namely a feasible objective value of a feasible solution $(\widehat{\mathsf{w}}, \widehat{b}, \widehat{\boldsymbol{\xi}})$ to (5.2), based on the optimal solution of relaxations.

Local Search

The first strategy, denoted as Local Search, consists of three main steps and employs a user-specified excess parameter k which will temporarily allow to work on just $k + B$ instead of all features. We first solve the relaxation by off-the-shelf software such as `Mosek`, an interior point solver. With the optimal solution u^* in hand, we want to search for a feasible point nearby. We next sort features according to increasing u^*. Namely, if u_j^* is close to 1, then the feature j is unlikely to be selected because of the complementarity constraint $w_j u_j = 0$, while if u_j^* is closer to 0, then the feature j is more likely to be selected. Based on it, we select the first $k + B$ features with small corresponding u, denoted by a set $K \subseteq [1\!:\!n]$ with $|K| = k + B$. In the end, we solve the restricted mixed-integer problem:

$$
\begin{array}{lll}
(\mathrm{CoP}(K)) \quad \min\limits_{\mathsf{w},b,\boldsymbol{\xi},\mathsf{u}} & \frac{1}{2}\|\mathsf{w}\|_2^2 + C\sum_{i=1}^m \xi_i & \\
\text{s.t.} & (\mathsf{w}, b, \boldsymbol{\xi}) \in \Xi & \\
& \sum\limits_{j\in K} u_j = |K| - B & \\
& u_j w_j = 0\,, & j \in K\,, \\
& u_j \in \{0,1\}\,, & j \in K\,.
\end{array}
\tag{5.16}
$$

which is much smaller than the original problem (5.7), as it only involves $|K|$ binary variables u_j. The problem is solved by the global solver `Gurobi`. This first upper bounding strategy is described in Algorithm 4.

Data: $(\mathsf{x}^i, y^i) \in \mathbb{R}^n \times \{1, -1\}$, B, C, parameter $k \in [0\!:\!n - B]$
Result: The set of selected features and an upper bound UB.
Solve the relaxation (5.13) and return the optimal solution u^*;
Generate a subset $K \subseteq [1\!:\!n]$ with $|K| = k + B$ such that $u_i^* \leq u_j^*$ for all $i \in K$ and for all $j \in [1\!:\!n] \setminus K$
Solve the mixed-integer problem (5.16)

Algorithm 4: Local Search

Kernel Search

Slightly enlarging the search set at each iteration, we propose the following search strategy for a feasible solution of (5.2). Our Kernel Search strategy consists in solving a sequence of

small MIQPs considering an initial ranking on the features. This strategy takes inspiration from a heuristic proposed by [AMS10] for the general multi-dimensional knapsack problem which has been applied to different kinds of problems such as location problems [GS12] and portfolio optimization [AMS12]. Recently, [LMMRC19] also applied it to solve their embedded feature selection for ℓ_1-SVMs. In this strategy, based on some ranking, a restricted set of promising features is kept and updated throughout each iteration. Usually, for ranking variables in general MILPs, LP theory is exploited using both relaxed solution values and reduced costs, a strategy followed by [LMMRC19] for ranking features.

By contrast, in our case we decided to use as a ranking criterion the coordinates of an optimal solution u^* to the decomposed SDP relaxation (5.13). Features are sorted with respect to this vector u^*: the smaller the value of u_j^*, the higher the relevance of feature j. Since in practice (5.13) leads to much tighter lower bound solutions in contrast to the plain relaxation (5.8) similar to the model used by [LMMRC19], there is a justified hope that the ranking provided by u^* is closer to the relevance of features.

Let K be the ordered set of features, ρ a user-defined parameter and $N := \lceil \frac{n-\rho}{\rho} \rceil$. Then K is divided into N subsets denoted as K_i for $i \in [1\!:\!N]$. In particular, each subset K_i, $i \in [1\!:\!N-1]$, will be composed of ρ features, and K_N will contain the remaining features.

An initial value of the UB is set to ∞. Also define a feature subset $\overline{K} \subseteq [1\!:\!n]$ as the *kernel set*, containing features that are kept at the next iteration. To initialize, set $\overline{K} = \emptyset$. At each iteration k, the heuristic considers set $\mathcal{K}_k = \overline{K} \cup K_k$ i.e. the union of the features in the kernel set and the features in the set K_k. To update the UB, at each iteration we solve problem CoP($\mathcal{K}_k$) plus the following two constraints:

$$\frac{1}{2}\|\mathsf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \leq \text{UB}, \tag{5.17}$$

$$\sum_{j \in K_k} u_j \leq |K_k| - 1. \tag{5.18}$$

Constraint (5.17) restricts the objective function to take a value smaller than or equal to the current upper bound, while constraint (5.18) makes sure that at least one feature from the new set K_k is selected. Note that, due to the addition of these constraints, the optimization problems may potentially be infeasible. If the optimization problems are feasible, the new features selected are then added to the kernel set $\overline{K}$ used in the next iteration since adding these features obtains an identical or better upper bound. Conversely, the set of features of $\overline{K}$ that have not been chosen in the optimal solution in the previous iterations is removed from the kernel set $\overline{K}$. The removal of some of the features from the kernel set is decisive in that it does not excessively increase the number of binary variables considered in each iteration. We decided to remove the features that were not selected in the previous two iterations. The set of added features is denoted as K^+ and the set of removed features as K^-. The resulting kernel set for the next iteration is $\overline{K} \cup K^+ \backslash K^-$. Conversely, if the problem is infeasible, the kernel set is not modified and the procedure skips to the next iteration. The Kernel Search strategy is described in Algorithm 5.

Data: $(\mathsf{x}^i, y^i) \in \mathbb{R}^n \times \{1, -1\}$, B, C, parameter $\rho \in [1\!:\!n]$.
Result: The set of selected features and an upper bound UB.
Solve the relaxation (5.13) and return the optimal solution u^*;
Sort features according to the numerical size of corresponding u^* increasingly, denoted by a vector $K \in \mathbb{Z}_+^n$;
Separate K into $N := \lfloor \frac{n}{\rho} \rfloor$ groups evenly such that $K = [K_1, \cdots, K_N]$ and set $\overline{K} = \emptyset$, $k = 1$, UB $= \infty$;
while $k \leq N$ **do**
  $\mathcal{K}_k \leftarrow K_k \cup \overline{K}$;
  Solve the problem CoP($\mathcal{K}_k$) + (5.17) + (5.18);
  if *problem is feasible* **then**
    Return solution u^* and optimal value UB*;
    UB $\leftarrow$ UB*;
    Build $K^+ := \{j \in K_k : j$ selected at current iteration$\}$;
    Build $K^- := \{j \in \overline{K} : j$ not selected in the solution of the last two iterations$\}$;
    Update $\overline{K} \leftarrow \overline{K} \cup K^+ \backslash K^-$;
  end
  $k \leftarrow k + 1$;
end

Algorithm 5: Kernel Search

5.4.2. A strategy to tighten the big-M parameter

Given an upper bound UB of problem (5.2), one can estimate the upper bound of w_j by solving the following problem:

$$\begin{aligned} \overline{M}_j := \max_{\mathsf{w},\mathrm{diag}\,\mathsf{W},b,\boldsymbol{\xi},\mathsf{u}} \quad & w_j \\ \text{s.t.} \quad & (\mathsf{w}, b, \boldsymbol{\xi}) \in \Xi \\ & \tfrac{1}{2}\|\mathsf{w}\|_2^2 + C\sum_{i=1}^m \xi_i \leq \mathrm{UB} \\ & \mathsf{e}^\top \mathsf{u} = n - B \\ & \begin{bmatrix} 1 & w_j & u_j \\ w_j & W_{jj} & 0 \\ u_j & 0 & u_j \end{bmatrix} \in \mathcal{S}_+^3, \qquad j \in [1\!:\!n]\,. \end{aligned} \tag{5.19}$$

Similarly, the lower bound of w_j can be obtained by solving the same optimization problem with the max replaced by min and we denote the optimal value by $\underline{M}_j$. The proper choice of M is $\max\limits_{j\in[1:n]}\{\max\{|\overline{M}_j|, |\underline{M}_j|\}\}$.

Proposition 42. *Let* $[\widehat{\mathsf{w}}, \widehat{b}, \widehat{\boldsymbol{\xi}}]$ *be a feasible solution of* (5.2) *and* $UB := \frac{1}{2}\|\widehat{\mathsf{w}}\|_2^2 + C\sum_{i=1}^m \widehat{\xi}_i$ *be the corresponding objective value. Then*

$$M = \max_{j\in[1:n]} \{\max\{|\overline{M}_j|, |\underline{M}_j|\}\}$$

is a valid bound, where $\overline{M}_j$ *and* $\underline{M}_j$ *are obtained by* (5.19).

Proof. obvious. □

Remark 5. *The proposed bound M is tighter than most of the currently used big-Ms. Nevertheless, empirically, the tightening of any such bound is of little relevance unfortunately, because they still fall outside of the range specified in* (5.9).

5.4.3. The heuristic algorithm

In this subsection, we propose a procedure for generating a good feasible solution of (5.2) taking one of the two upper bound strategies and the big-M estimation strategy into account. At first, a feasible solution is computed with the chosen upper bound strategy. Then, according to Proposition 42, we solve a series of $2n$ many SDP problems to find, for each feature j, values $\overline{M}_j$,$\underline{M}_j$ and then obtain the big-M parameter M. If this value satisfies (5.9), we know that the SDP formulation (5.14) is better than (5.13), thus we compute a new relaxed solution u^* of (5.14). We thus use vector u^* as a ranking criterion for the chosen upper bound strategy, we run the strategy again and we update the heuristic solution if a better one is found.

Data: $(\mathsf{x}^i,y^i)\in \mathbb{R}^n \times \{1,-1\}$, B, C, the parameter $k\in[0\,{:}\,n-B]$ or the parameter $\rho\in[1\,{:}\,n]$.
Result: $(\mathsf{w}^*, b^*, \boldsymbol{\xi}^*)$ solution of (5.2).
Run Algorithm 4 or Algorithm 5 to obtain a feasible solution $(\widehat{\mathsf{w}}, \widehat{b}, \widehat{\boldsymbol{\xi}})$ to (5.2) with upper bound value $\widehat{\mathrm{UB}}$;
Set $(\mathsf{w}^*, b^*, \boldsymbol{\xi}^*) \leftarrow (\widehat{\mathsf{w}}, \widehat{b}, \widehat{\boldsymbol{\xi}})$;
Obtain M according to Proposition 42;
Solve problem (5.1) and obtain w^*;
if $M < \frac{\|\mathsf{w}^*\|_1}{B}$ **then**
  Solve problem (5.14) and obtain u^*;
  Run Algorithm 4 or Algorithm 5 starting with u^*, get $(\overline{\mathsf{w}}, \overline{b}, \overline{\boldsymbol{\xi}})$ feasible to (5.2) and upper bound $\overline{\mathrm{UB}}$;
  if $\overline{UB} \leq \widehat{UB}$ **then**
    $(\mathsf{w}^*, b^*, \boldsymbol{\xi}^*) \leftarrow (\overline{\mathsf{w}}, \overline{b}, \overline{\boldsymbol{\xi}})$
  end
end

Algorithm 6: Heuristic procedure

5.4.4. The exact algorithm

In this subsection, we propose an exact procedure for (5.2) by solving a sequence of MISOCPs. With a slight abuse of notation, let us define, given a subset of features $K \subseteq [1\,{:}\,n]$, the set

$$\mathcal{B}_M(K) = \{(\mathsf{w},\mathsf{u}) \in \mathbb{R}^{2n} : -M(1-u_j) \leq w_j \leq M(1-u_j),\, j \in K\,,\ \mathsf{e}^\top \mathsf{u} = n - B\}\,.$$

The subproblem considered in the exact procedure is

$$
\begin{aligned}
(\text{SR-DSCoP}(K)) \qquad \min_{\mathsf{w},\mathrm{diag}\,\mathsf{W},b,\boldsymbol{\xi},\mathsf{u}} \quad & \tfrac{1}{2}\mathrm{trace}\,(\mathsf{W}) + C\sum_{i=1}^{m}\xi_i \\
\text{s.t.} \quad & (\mathsf{w}, b, \boldsymbol{\xi}) \in \Xi \\
& (\mathsf{w}, \mathsf{u}) \in \mathcal{B}_M(K) \\
& \begin{bmatrix} 1 & w_j & u_j \\ w_j & W_{jj} & 0 \\ u_j & 0 & u_j \end{bmatrix} \in \mathcal{S}^3_+\,, \quad j\in[1\!:\!n]\setminus K\,, \\
& u_j \in \{0,1\}\,, \qquad\qquad\qquad\quad j \in K\,, \\
& \begin{bmatrix} 1 & w_j \\ w_j & W_{jj} \end{bmatrix} \in \mathcal{S}^2_+\,, \qquad\quad j \in K\,,
\end{aligned}
\tag{5.20}
$$

where SR is the abbreviation for semi-relaxation. In the problem (5.20), we essentially relax the complementarity constraints associated with features in the set $[1\!:\!n]\setminus K$ via decomposed SDP relaxation, and express the other by the big-M mixed-integer formulation. We note that the problem (5.20) is a mixed-integer positive semi-definite optimization problem (MISDP) and there is no off-the-shelf solver dealing with it. In the following part, we show a way to express the problem in the form of MISOCP, which can be handled by the off-the-shelf solvers like `copt`.

Before specifying the MISOCP reformulation, we present a useful result: some special slices $\mathcal{F}$ of the positive-semidefinite cone are second-order cone representable, due to Lemma 43 below.

Lemma 43.

$$
\mathcal{F} := \left\{ (w,u,W) \in \mathbb{R}^3 : \begin{bmatrix} 1 & w & u \\ w & W & 0 \\ u & 0 & u \end{bmatrix} \in \mathcal{S}^3_+ \right\} = \left\{ (w,u,W) \in \mathbb{R}^3 : \begin{array}{r} \sqrt{(W-s)^2+4w^2} \le W+s \\ u-t\le 0\,, \quad u\ge 0 \\ s+t=1\,, \quad \textit{some } (s,t)\in\mathbb{R}^2\,. \end{array} \right\}
$$

Proof. Noticing that the matrix in the left set has a chordal sparsity pattern, due to [AHMR88], we have the decomposition:

$$
\begin{bmatrix} 1 & w & u \\ w & W & 0 \\ u & 0 & u \end{bmatrix} \in \mathcal{S}^3_+ \quad\Longleftrightarrow\quad \exists (s,t)\in\mathbb{R}^2 \quad \text{s.t.} \quad \begin{bmatrix} s & w \\ w & W \end{bmatrix} \in \mathcal{S}^2_+, \begin{bmatrix} t & u \\ u & u \end{bmatrix} \in \mathcal{S}^2_+, \; s+t=1.
$$

It follows that

$$
\begin{bmatrix} s & w \\ w & W \end{bmatrix} \in \mathcal{S}^2_+ \quad\Longleftrightarrow\quad s+W\ge 0 \;\text{ and }\; sW\ge w^2 \quad\Longleftrightarrow\quad \sqrt{(W-s)^2+4w^2} \le W+s\,.
$$

Finally, noticing the fact that

$$
\begin{bmatrix} t & u \\ u & u \end{bmatrix} \in \mathcal{S}^2_+ \quad\Longleftrightarrow\quad u-t\le 0, \quad u\ge 0\,,
$$

we complete the proof. □

Applying Lemma 43, the problem (5.20) can be rewritten of the form

$$
\begin{array}{lll}
\text{(SR-DLMP}(K)) & \min\limits_{\mathsf{w},\mathrm{diag}\ \mathsf{W},b,\boldsymbol{\xi},\mathsf{u},\mathsf{s},\mathsf{t}} & \frac{1}{2}\mathrm{trace}\ (\mathsf{W}) + C\sum_{i=1}^{m}\xi_i \\
& \text{s.t.} & (\mathsf{w}, b, \boldsymbol{\xi}) \in \Xi \\
& & (\mathsf{w}, \mathsf{u}) \in \mathcal{B}_M(K) \\
& & \mathsf{d}_j := \begin{bmatrix} W_{jj} - s_j \\ 2w_j \end{bmatrix} \in \mathbb{R}^2\,, \quad j \in [1\!:\!n] \setminus K\,, \\
& & \|\mathsf{d}_j\|_2 \le s_j + W_{jj}\,, \qquad j \in [1\!:\!n] \setminus K\,, \\
& & u_j - t_j \le 0\,, \qquad j \in [1\!:\!n] \setminus K\,, \\
& & s_j + t_j = 1\,, \qquad j \in [1\!:\!n] \setminus K\,, \\
& & u_j \in \{0,1\}\,, \qquad j \in K\,, \\
& & w_j^2 \le W_{jj}\,, \qquad j \in K\,,
\end{array}
\tag{5.21}
$$

where $2n$ (actually, $2(n - |K|)$ more variables $\mathsf{s} \in \mathbb{R}^n$ and $\mathsf{t} \in \mathbb{R}^n$ are introduced.

Now we can introduce the exact algorithm reported as Algorithm 7. In this exact procedure, we solve a sequence of semi-relaxed problems SR-DLMP(K) associated with a subset $K \subseteq [1\!:\!n]$ of features. In such problems, only variables u_j with $j \in K$ will be considered as binary and the remaining ones are relaxed. To obtain initial bounds on the objective value, the exact procedure exploits the upper bound strategies detailed above. The main step of the exact procedure consists of solving the sequence of semi-relaxed problems (5.21) to improve the lower bound of the objective value.

To start with, we must select a subset of features $K \subseteq [1\!:\!n]$ whose associated u variables will be considered as binary variables in the first semi-relaxed problem. The upper-bound strategy run at the beginning provides a subset of features that allows us to obtain a good bound on the optimal objective value. Therefore, the exact procedure will consider the set provided by the heuristic as the initial K and it will obtain an initial LB solving SR-DLMP(K). Then the set K is updated by adding and removing some of the features, in order to improve the lower bound of the objective value. To this end, two sets (denoted by K^+ and K^-) are built in each iteration. Set K^+ consists of some of the features in $[1\!:\!n] \setminus K$ whose associated u variables will be considered as binary in the next iteration, i.e. features of K^+ will be added to K. Similarly, K^- consists of features in K that will not be considered as binary in the next iteration.

In addition and if possible, we will update the UB in the main step running the upper bound strategies and using vector $\widetilde{\mathsf{u}}$ as the initial ranking. There could be many different ways of defining set K^+ and updating K. In the strategy we adopted, we set K^+ as the set of s features j such that the relaxed solution $\widetilde{\mathsf{u}}$ of SR-DLMP(K) was "less binary". This is why we first create a ranking on the features based on terms $\widetilde{u}_j - \widetilde{u}_j^2$ in descending order, and then select the first s, which is a user-defined parameter.

Data: $(\mathsf{x}^i, y^i) \in \mathbb{R}^n \times \{1, -1\}$, B, C, $s \in [1\!:\!n]$, LB $= -\infty$, UB $= \infty$, $K = \emptyset$.
Result: $(\mathsf{w}^*, b^*, \boldsymbol{\xi}^*)$ solution of (5.2).
Run Algorithm 4 or Algorithm 5, obtain set K as the set of the selected features and update the UB value;
while $\frac{UB-LB}{UB} \cdot 100 \geq 0.01$ **do**
  Solve problem SR-DLMP(K);
  if *a time limit of* 1800 *seconds is reached* **then**
    STOP
  end
  Obtain optimal value $\widehat{\text{LB}}$ and optimal solution $\widetilde{\mathsf{u}}$;
  Run Algorithm 4 or Algorithm 5 with the first step skipped vector $\widetilde{\mathsf{u}}$ used for initial ranking;
  Obtain the solution $(\widehat{\mathsf{w}}, \widehat{b}, \widehat{\boldsymbol{\xi}}, \widehat{\mathsf{u}})$ and upper bound value $\widehat{\text{UB}}$;
  Sort features using terms $\widetilde{u}_j - \widetilde{u}_j^2$ in descending order;
  Build set K^+ with the first s features;
  Build $K^- := \{j \in K : j$ not selected in the solution of the last two iterations$\}$;
  $K \leftarrow K \cup K^+ \setminus K^-$;
  if $\widehat{UB} < UB$ **then**
    UB $\leftarrow \widehat{\text{UB}}$;
    $(\mathsf{w}^*, b^*, \boldsymbol{\xi}^*, \mathsf{u}^*) \leftarrow (\widehat{\mathsf{w}}, \widehat{b}, \widehat{\boldsymbol{\xi}}, \widehat{\mathsf{u}})$;
  end
  if $\widehat{LB} > LB$ **then**
    LB $\leftarrow \widehat{\text{LB}}$;
  end
end

Algorithm 7: Exact procedure

5.5. Numerical experiments

The various computational experiments were performed on an Apple M1 CPU 16 GB RAM computer. All models and algorithms were implemented in `Python`. Results were performed using `Gurobi 11.0` for the resolution of the MIQPs, `Mosek 10.1` for the resolution of the SDP relaxations and `Copt` for the resolution of the MISOCP models solved in the Exact algorithm.

The experiments were carried out on fifteen different datasets. Nine of them can be found in the UCI repository [DG17], (see Table 5.1), where m is the number of data points, n is the number of features and the last column shows the percentage of samples in each class. As can be observed, they contain a small number of features. The other six datasets used in the experiments have a larger number of features, as shown in Table 5.2. The Arrythmia, Madelon, and MFeat datasets are available in the UCI repository as well. The remaining three datasets, Colorectal, DLBCL, and Lymphoma, are microarray datasets containing thousands of features but smaller sample sizes. Further descriptions of these last datasets can be found in [ABN$^+$99, GWBV02, SRW$^+$02].

Dataset	m	n	Class (%)
Breast Cancer D.	569	30	63/37
Breast Cancer W.	683	9	65/35
Breast Cancer P.	198	33	65/35
Cleveland	297	13	54/46
Diabetes	768	8	65/35
German	1000	24	70/30
Ionosphere	351	33	36/64
Parkinsons	195	22	25/75
Wholesale	440	7	68/32

Table 5.1.: First datasets

Dataset	m	n	Class (%)
Arrhythmia	452	274	54/46
Colorectal	62	2000	65/35
DLBCL	77	7129	25/75
Lymphoma	96	4026	35/65
Madelon	2000	500	50/50
Mfeat	2000	649	10/90

Table 5.2.: Second datasets

5.5.1. Computational analysis on the MIQP models

As starting results, we solved the two proposed MIQP models BigMP and CoP with `Gurobi`. Initially, we created different instances of the problems for each dataset in Table 5.1 ranging B values among all the possible values from 1 to n and ranging C values in the set $\{10^r : r \in [-3:3]\}$. For this first set of results, `Gurobi` was able to reach global optimum in every case but formulation CoP was the one which was easier to solve. Figure 5.1 shows average computation times for solving the two MIQPs, CoP and BigMP. On average, among datasets in Table 5.1, solving model CoP with `Gurobi` was 1.4 times faster.

		BigMP		CoP	
Dataset	B	ObjFun	MipGap	ObjFun	MipGap
Arrhythmia	10	2398.06	**0.64**	**2151.56**	0.82
	20	2047.67	**0.67**	**1712.64**	0.81
	30	1733.28	**0.67**	**1441.63**	0.78
Colorectal	10	2.99	**0.64**	**2.12**	0.98
	20	**0.64**	**0.13**	0.68	0.93
	30	**0.41**	**0.04**	0.43	0.89
DLBCL	10	0.63	**1.00**	**0.56**	**1.00**
	20	0.18	**0.99**	**0.15**	**0.99**
	30	0.16	0.99	**0.09**	**0.98**
Lymphoma	10	18.07	1.00	**1.16**	**0.99**
	20	0.71	0.99	**0.29**	**0.98**
	30	0.34	0.98	**0.16**	**0.96**
Madelon	10	**16680.48**	**0.32**	16800.07	0.33
	20	**16419.87**	**0.32**	18558.98	0.40
	30	**16075.70**	**0.30**	18425.44	0.39
Mfeat	10	0.92	**0.59**	**0.68**	0.93
	20	**0.26**	**0.21**	**0.26**	0.81
	30	**0.18**	**0.15**	**0.18**	0.71

Table 5.3.: Comparison of BigMP and CoP on the large datasets

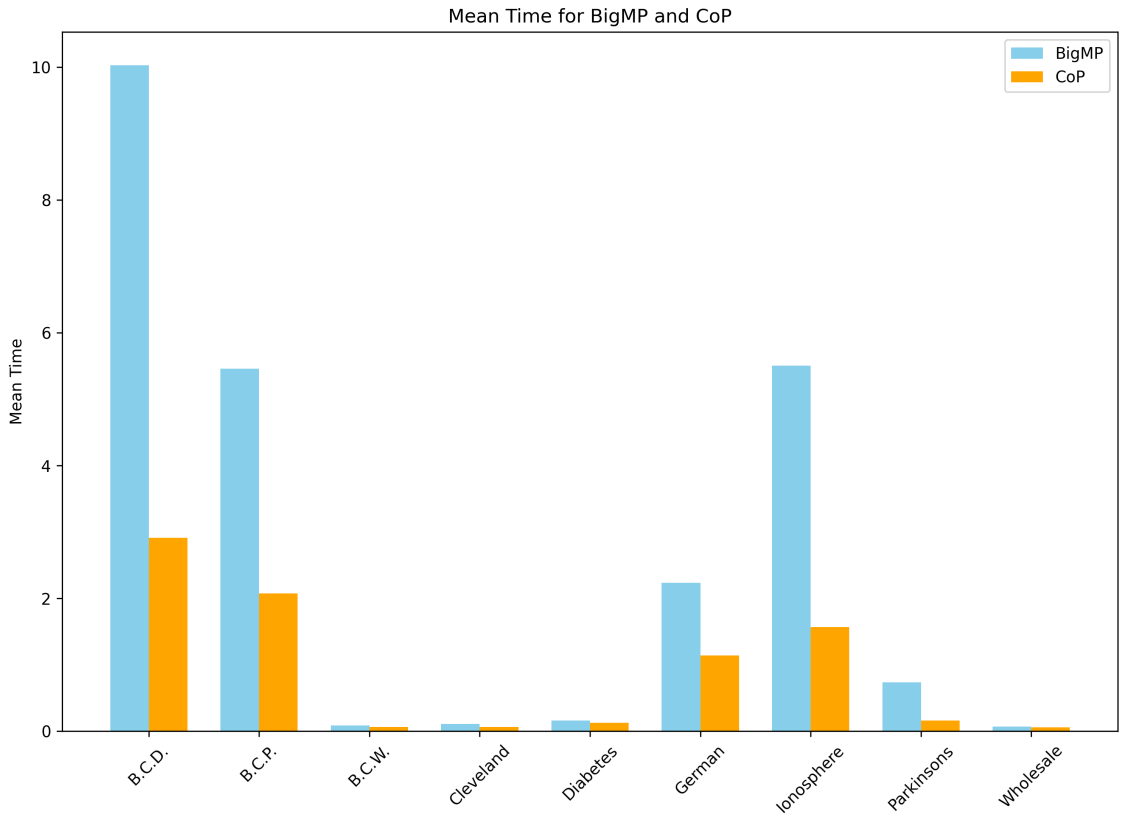


Figure 5.1.: Comparison of solving BigMP and CoP with `Gurobi` for the first set of datasets: mean computational times (in seconds)

The reason behind this first comparisons is that `Gurobi` is used in the two upper bound strategies we proposed. The two strategies, the Local Search and the Kernel Search, require the resolution of the FS-SVM model on instances created on a smaller subset of features, and this can be either done by using the BigMP or the CoP formulation. From these initial tests, we understand that CoP is the model that is faster to solve with `Gurobi` when dealing with datasets with a small number of features, thus, we will use the resolution of model CoP as the subroutine used by the heuristic algorithms to find ad update the upper bounds values.

In Table 5.3 we show the computational performances of `Gurobi` when tested on the larger instances related to the datasets with the larger number of features. For these tests, we set the hyperparameter $C = 10$, and, similarly to [LMMRC19], we ranged B values in $\{10, 20, 30\}$. We can notice how the resulting optimization problems are much harder to solve in contrast to the problem related to the smaller datasets. A time limit of 1 hour was set, and the final MipGap values for the majority of these dataset are very high. This is certainly due to the fact that the lower bound values computed in the Branch & Bound tree of `Gurobi` are not very effective for these type of formulations. This is particularly the case for model CoP whose final MipGap values tend to be worse than the ones of model BigMP. In contrast, `Gurobi` manages to find much better incumbent solutions when solving model CoP for all datasets, apart from Colorectal and Madelon.

5.5.2. Results on the heuristics

In this section we analyse the performances of the heuristic algorithm we proposed and we compare with `Gurobi`. We have two versions of the heuristic algorithm depending on whether the Local Search or the Kernel Search upper bound strategies are used. We denote by `H-LS` the heuristic algorithm which implements the Local Search strategy, and by `H-KS` the heuristic implementing Kernel Search.

Table 5.4 shows the optimization performance of `H-LS` and `H-KS`. Both heuristics where given a time limit of 600 seconds. In particular for `H-LS` we set the parameter $k = 10$, which means that after getting the relaxed solution we solved model CoP on the $B + k$ most promising features. For `H-KS` we also set parameter $s = 10$, which means that the overall set of features was divided in subsets of ten features each. A smaller time limit of 60 seconds was given for the resolution of each CoP subproblem solved in the Kernel Search strategy.

As expected, `H-KS` manages to find better upper bounds compared to `H-LS`, at the cost of generally longer computational times. In some instances like in Lymphoma, Madelon and Mfeat, both heuristics found the same solutions. The time limit of 600 seconds was reached by `H-KS` for the instances related to datasets Madelon and Mfeat. For the rest of the datasets, the average computational times do not exceed 300 seconds for both heuristics. In general, compared to the objective function values of the solutions found by `Gurobi` in 1 hour (Table 5.3), the solutions found by the heuristics are much better.

		H-LS		H-KS	
Dataset	B	ObjFun	Time	ObjFun	Time
Arrhythmia	10	2491.04	**5.37**	**2142.67**	41.24
	20	1982.08	**9.55**	**1652.83**	168.83
	30	1729.55	**7.59**	**1380.19**	311.62
Colorectal	10	2.13	**15.38**	**2.04**	15.88
	20	0.67	**20.12**	**0.65**	57.51
	30	0.411	**37.26**	**0.406**	69.46
DLBCL	10	0.25	51.01	**0.23**	**26.56**
	20	0.104	**46.97**	**0.101**	67.74
	30	0.065	134.95	**0.063**	**117.29**
Lymphoma	10	**0.69**	**28.73**	**0.69**	98.64
	20	0.24	**40.87**	**0.23**	54.42
	30	0.14	**98.25**	**0.13**	109.33
Madelon	10	16660.85	**25.07**	**16615.43**	607.32
	20	**16236.87**	48.57	**16236.87**	612.62
	30	**15917.52**	176.92	**15917.99**	605.93
Mfeat	10	0.70	**21.27**	**0.66**	341.87
	20	**0.25**	**76.68**	**0.25**	621.16
	30	**0.17**	**276.90**	**0.17**	606.96

Table 5.4.: Results on large datasets of the Heuristic algorithm implementing either the Local search (H-LS) or the Kernel Search (H-KS) as the upper bound strategies

Another set of results we carried out was the one of using `H-KS` as a warm start heuristic in order to evaluate if `Gurobi` was able to find better solution. In Table 5.5 we compare the optimal value of the solution of `Gurobi` after 1 hour for model CoP, the value of the `H-KS` heuristic,

and the value of the solution found by `Gurobi` using as a warm start solution the one of `H-KS` (denoted by CoP*).

Regarding Table 5.5, we can see that `Gurobi` was unable to find a better solution than the heuristic after 1 hour for all datasets, apart from Arrhythmia, and this is probably due to the fact that the heuristic solutions are already optimal or nearly-optimal. In this table, column d_f indicate the percentage of improvement of the objective function found by `Gurobi` using `H-KS` starting solution, against the objective function of the solution found by `Gurobi` alone.

		ObjFun			
Dataset	B	CoP	`H-KS`	CoP*	d_f
Arrhythmia	10	2151.56	2142.67	**2098.55**	2%
	20	1712.64	**1652.83**	**1652.83**	3%
	30	1441.63	1380.19	**1366.94**	5%
Colorectal	10	2.12	**2.04**	**2.04**	4%
	20	0.68	**0.65**	**0.65**	5%
	30	0.43	**0.41**	**0.41**	6%
DLBCL	10	0.56	**0.23**	**0.23**	60%
	20	0.15	**0.10**	**0.10**	35%
	30	0.09	**0.06**	**0.06**	29%
Lymphoma	10	1.16	**0.69**	**0.69**	41%
	20	0.29	**0.22**	**0.22**	21%
	30	0.16	**0.13**	**0.13**	16%
Madelon	10	16800.07	**16615.43**	**16615.43**	1%
	20	18558.98	**16236.86**	**16236.86**	13%
	30	18425.44	**15917.51**	**15917.51**	14%
Mfeat	10	0.68	**0.66**	**0.66**	3%
	20	0.26	**0.25**	**0.25**	4%
	30	0.17	**0.17**	**0.17**	3%

Table 5.5.: Comparison of the objective functions of CoP returned by `Gurobi` with and without using the Heuristic Kernel Search algorithm as a warm start for large datasets

5.5.3. Results on the exact approach

As shown in Table 5.7, the proposed exact algorithm performs much better than `Gurobi` with a warm start heuristic, both on optimal value and computation time. In particular, the exact algorithm is able to globally solve four problems under a time limit of 3600 seconds. For the unsolved problems, the exact algorithm provides a better objective value and a smaller MipGap.

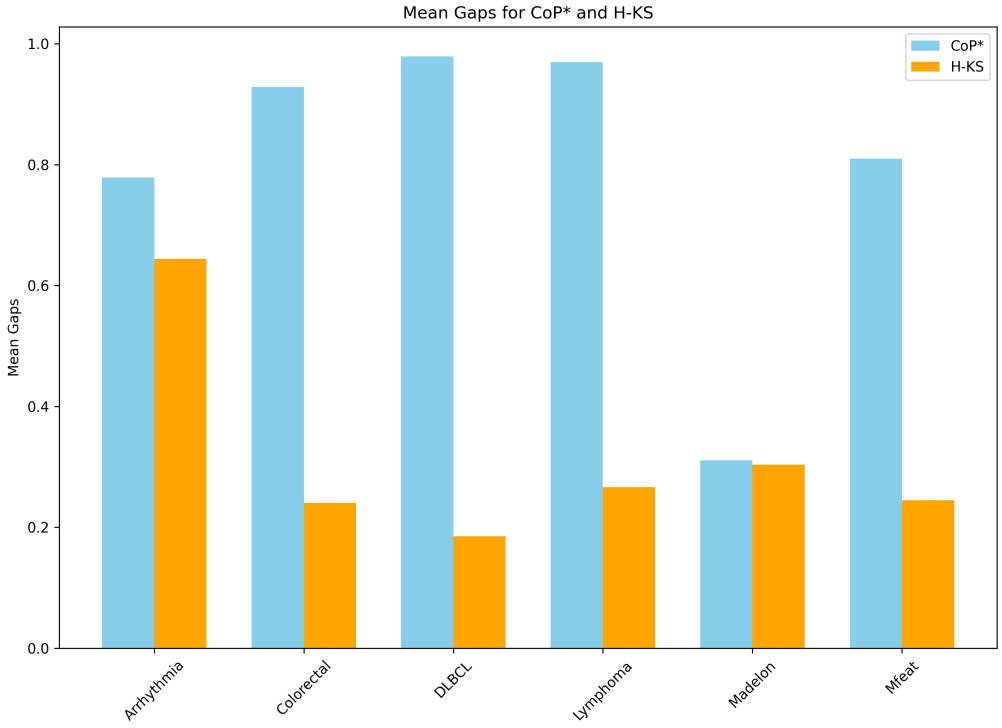


Figure 5.2.: Comparison of the mean MipGap values of the solution found by `Gurobi` solving CoP after 1 hour, and the Relaxed Gap of the Heuristic Kernel Search solution using the solution of DSCoP as a lower bound

5.6. Conclusion

In this chapter, we addressed the NP-hard feature selection problem in linear SVMs under a cardinality constraint, which ensures that the model remains interpretable by selecting a limited number of features. Our approach involves formulating the problem as an MIQP problem and introducing novel SDP relaxations to handle its complexity and enhance scalability.

We developed two MIQP formulations: one employing a big-M method and another one using a complementarity constraint. These formulations accommodate the cardinality constraint by integrating binary variables to select features. To solve these challenging formulations, we proposed several tractable SDP relaxations and a decomposed SDP approach, exploiting the sparsity pattern inherent to the problems. This decomposition significantly reduces computational complexity, making the conic relaxations scalable even for datasets with a large number of features.

Numerical experiments on various benchmark datasets demonstrate the effectiveness of our approach. The proposed methods not only outperform `Gurobi` but also provide competitive classification accuracy. By allowing a flexible adjustment of the number of features, our approach enhances interpretability without significantly compromising predictive performance.

		Initial Bounds		
Dataset	B	DSCoP	BigMP	CoP
Arrhythmia	10	**754.46**	451.10	274.79
	20	**622.44**	395.92	274.79
	30	**538.86**	303.66	274.79
Colorectal	10	**1.07**	0.05	0.05
	20	**0.54**	0.05	0.05
	30	**0.38**	0.05	0.05
DLBCL	10	**0.16**	0.00	0.00
	20	**0.08**	0.00	0.00
	30	**0.06**	0.00	0.00
Lymphoma	10	**0.35**	0.00	0.01
	20	**0.18**	0.00	0.01
	30	**0.12**	0.00	0.01
Madelon	10	**11243.14**	0.00	0.00
	20	**11209.44**	0.00	0.00
	30	**11197.73**	0.00	0.00
Mfeat	10	**0.36**	0.05	0.05
	20	**0.20**	0.05	0.05
	30	**0.15**	0.05	0.05

Table 5.6.: Lower bounds comparison between the SDP one and the starting bound found by `Gurobi` for the BigMP and CoP models for different datasets

		CoP*			Exact KS		
Dataset	B	ObjFun	Time	MipGap	ObjFun	Time	MipGap
Arrhythmia	10	2098.55	3600	0.82	**2003.98**	3600	**0.22**
	20	1652.83	3600	0.80	**1632.32**	3600	**0.27**
	30	1366.94	3600	0.77	**1362.75**	3600	**0.32**
Colorectal	10	2.04	3600	0.98	**2.04**	**407.2**	**0**
	20	0.65	3600	0.93	**0.65**	**431.5**	**0**
	30	0.41	3600	0.88	**0.41**	**407.6**	**0**
DLBCL	10	**0.23**	3600	0.99	**0.23**	**451.3**	**0**
	20	**0.10**	3600	0.98	**0.10**	**542.7**	**0**
	30	**0.06**	3600	0.97	**0.06**	**746.2**	**0**
Lymphoma	10	**0.69**	3600	0.99	**0.69**	**314.6**	**0**
	20	**0.22**	3600	0.97	**0.22**	**532.3**	**0**
	30	**0.13**	3600	0.95	**0.13**	**823.4**	**0**
Madelon	10	16615.43	3600	0.33	**16499.38**	3600	**0.17**
	20	16236.86	3600	0.31	**16142.54**	3600	**0.18**
	30	15917.51	3600	0.30	**15765.83**	3600	**0.21**
Mfeat	10	0.66	3600	0.93	**0.64**	**397.5**	**0**
	20	0.25	3600	0.80	**0.24**	**573.5**	**0**
	30	**0.17**	3600	0.71	**0.17**	**1045.5**	**0**

Table 5.7.: Comparison of the `Gurobi` solution to the CoP model using the Heuristic Kernel Search solution as a warm start heuristic, and the Exact algorithm using the Kernel Search procedure as the upper bound strategy.

Bibliography

[ABN+99] Uri Alon, Naama Barkai, Daniel A Notterman, Kurt Gish, Suzanne Ybarra, Daniel Mack, and Arnold J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, 1999.

[AHD21] Andrey Afonin, Roland Hildebrand, and Peter J.C. Dickinson. The extreme rays of the 6×6 copositive cone. *Journal of Global Optimization*, 79(1):153–190, 2021.

[AHMR88] Jim Agler, William Helton, Scott McCullough, and Leiba Rodman. Positive semidefinite matrices with a given sparsity pattern. *Linear Algebra and its Applications*, 107:101–149, 1988.

[AMS10] Enrico Angelelli, Renata Mansini, and M. Grazia Speranza. Kernel search: A general heuristic for the multi-dimensional knapsack problem. *Computers & Operations Research*, 37(11):2017–2026, 2010.

[AMS12] Enrico Angelelli, Renata Mansini, and M. Grazia Speranza. Kernel search: A new heuristic framework for portfolio selection. *Computational Optimization and Applications*, 51:345–361, 2012.

[Ans21] Kurt M. Anstreicher. Testing copositivity via mixed–integer linear programming. *Linear Algebra and Its Applications*, 609:218–230, 2021.

[AOHE19] Tiago Andrade, Fabricio Oliveira, Silvio Hamacher, and Andrew Eberhard. Enhancing the normalized multiparametric disaggregation technique for mixed-integer quadratic programming. *Journal of Global Optimization*, 73:701–722, 2019.

[ApS24] MOSEK ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 10.1.*, 2024.

[BDDK+00] Immanuel M. Bomze, Mirjam Dür, Etienne De Klerk, Cornelis Roos, Arie J. Quist, and Tamás Terlaky. On copositive programming and standard quadratic optimization problems. *Journal of Global Optimization*, 18(4):301–320, 2000.

[BDDM+09] Joshua Brodie, Ingrid Daubechies, Christine De Mol, Domenico Giannone, and Ignace Loris. Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272, 2009.

[BDPP24] Immanuel M. Bomze, Federico D'Onofrio, Laura Palagi, and Bo Peng. Feature selection in linear SVMs via hard cardinality constraint: a scalable SDP decomposition approach. *arXiv preprint arXiv:2404.10099*, 2024.

[BEKS17] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

[Bie96] Daniel Bienstock. Computational study of a family of mixed-integer quadratic programming problems. *Mathematical Programming*, 74:121–140, 1996.

[BKL21] Immanuel M. Bomze, Michael Kahr, and Markus Leitner. Trust your data or not – StQP remains StQP: Community detection via robust Standard Quadratic Optimization. *Mathematics of Operations Research*, 46:301–316, 2021.

[BKS16] Oleg P. Burdakov, Christian Kanzow, and Alexandra Schwartz. Mathematical programs with cardinality constraints: reformulation by complementarity-type conditions and a regularization method. *SIAM Journal on Optimization*, 26(1):397–425, 2016.

[BLT08] Immanuel M. Bomze, Marco Locatelli, and Fabio Tardella. New and old bounds for standard quadratic optimization: dominance, equivalence and incomparability. *Mathematical Programming*, 115(1):31–64, 2008.

[BM05] Stephen Braun and John E. Mitchell. A semidefinite programming heuristic for quadratic programming problems with complementarity constraints. *Computational Optimization and Applications*, 31(1):5–29, 2005.

[BMP16] Lijie Bai, John E. Mitchell, and Jong-Shi Pang. On conic QPCCs, conic QCQPs and completely positive programs. *Mathematical Programming*, 159:109–136, 2016.

[Bom97] Immanuel M. Bomze. Evolution towards the maximum clique. *Journal of Global Optimization*, 10(2):143–164, 1997.

[Bom02] Immanuel M. Bomze. Regularity versus degeneracy in dynamics, games, and optimization: A unified approach to different aspects. *SIAM Review*, 44(3):394–414, 2002.

[Bom12] Immanuel M. Bomze. Copositive optimization–recent developments and applications. *European Journal of Operational Research*, 216(3):509–520, 2012.

[BP23] Immanuel M. Bomze and Bo Peng. Conic formulation of QPCCs applied to truly sparse QPs. *Computational Optimization and Applications*, 84:703–735, 2023.

[BPQY24] Immanuel M. Bomze, Bo Peng, Yuzhou Qiu, and E. Alper Yıldırım. Tighter yet more tractable relaxations and nontrivial instance generation for sparse standard quadratic optimization. *arXiv preprint arXiv:2406.01239*, 2024.

[BPQY25] Immanuel M. Bomze, Bo Peng, Yuzhou Qiu, and E. Alper Yıldırım. On tractable convex relaxations of standard quadratic optimization problems under sparsity constraints. *Journal of Optimization Theory and Applications*, to appear 2025.

[BS09] Dimitris Bertsimas and Romy Shioda. Algorithm for cardinality-constrained quadratic optimization. *Computational Optimization and Applications*, 43(1):1–22, 2009.

[BSM03] Abraham Berman and Naomi Shaked-Monderer. *Completely Positive Matrices*. World Scientific, 2003.

[BSU12] Immanuel M. Bomze, Werner Schachinger, and Gabriele Uchida. Think co(mpletely)positive! Matrix properties, examples and a clustered bibliography on copositive optimization. *Journal of Global Optimization*, 52(3):423–445, 2012.

[BSU15] Immanuel M. Bomze, Werner Schachinger, and Reinhard Ullrich. New lower bounds and asymptotics for the cp-rank. *SIAM Journal on Matrix Analysis and Applications*, 36(1):20–37, 2015.

[Bur09] Samuel Burer. On the copositive representation of binary and continuous nonconvex quadratic programs. *Mathematical Programming*, 120(2):479–495, 2009.

[Bur15] Samuel Burer. A gentle, geometric introduction to copositive optimization. *Mathematical Programming*, 151:89–116, 2015.

[CPS92] Richard W. Cottle, Jong-Shi Pang, and Richard E. Stone. *The Linear Complementarity Problem*, volume 60. SIAM, 1992.

[CPZ13] Xin Chen, Jiming Peng, and Shuzhong Zhang. Sparse solutions to random standard quadratic optimization problems. *Mathematical Programming*, 141:273–293, 2013.

[CRT06] Emmanuel J. Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[CT06] Emmanuel J. Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on Information Theory*, 52(12):5406–5425, 2006.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273—297, 1995.

[CW08] Emmanuel J. Candès and Michael B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[DDM17] Alberto Del Pia, Santanu S. Dey, and Marco Molinaro. Mixed-integer quadratic programming is in NP. *Mathematical Programming*, 162:225–240, 2017.

[DG17] Dheeru Dua and Casey Graff. UCI Machine Learning Repository. `http://archive.ics.uci.edu/ml`, 2017.

[DGW23] Jin-Hong Du, Yifeng Guo, and Xueqin Wang. High-dimensional portfolio selection with cardinality constraints. *Journal of the American Statistical Association*, 118(542):779–791, 2023.

[Dia62] Palahenedi H. Diananda. On non-negative forms in real variables some or all of which are non-negative. *Mathematical Proceedings of the Cambridge Philosophical Society*, 58(1):17–25, 1962.

[Dic19] Peter J.C. Dickinson. A new certificate for copositivity. *Linear Algebra and its Applications*, 569:15–37, 2019.

[DLLR+12] David Di Lorenzo, Giampaolo Liuzzi, Francesco Rinaldi, Fabio Schoen, and Marco Sciandrone. A concave optimization-based approach for sparse portfolio selection. *Optimization Methods and Software*, 27(6):983–1000, 2012.

[DR21] Mirjam Dür and Franz Rendl. Conic optimization: a survey with special focus on copositive optimization and binary quadratic problems. *EURO Journal on Computational Optimization*, 9:100021, 2021.

[FG07] Antonio Frangioni and Claudio Gentile. SDP diagonalizations and perspective cuts for a class of nonseparable MIQP. *Operations Research Letters*, 35(2):181–185, 2007.

[FM04] Glenn Fung and Olvi L. Mangasarian. A feature selection Newton method for support vector machine classification. *Computational Optimization and Applications*, 28:185–202, 2004.

[FMP+18] Mingbin Feng, John J. Mitchell, Jong Shi Pang, Xin Shen, and Andreas Waechter. Complementarity formulations of l0-norm optimization. *Pacific Journal of Optimization*, 14(2):273–305, 2018.

[FW56] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

[GS12] Gianfranco Guastaroba and M. Grazia Speranza. Kernel search for the capacitated facility location problem. *Journal of Heuristics*, 18(6):877–917, 2012.

[Gur23] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023.

[GWBV02] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[GY21] Jacek Gondzio and E. Alper Yıldırım. Global solutions of nonconvex standard quadratic programs via mixed integer linear programming reformulations. *Journal of Global Optimization*, 81(2):293–321, 2021.

[GY22] Y. Görkem Gökmen and E. Alper Yıldırım. On standard quadratic programs with exact and inexact doubly nonnegative relaxations. *Mathematical Programming*, 193(Ser. A):365–403, 2022.

[Hil12] Roland Hildebrand. The extreme rays of the 5×5 copositive cone. *Linear Algebra and its Applications*, 437(7):1538–1547, 2012.

[HN63] Marshall Hall and Morris Newman. Copositive and completely positive quadratic forms. *Mathematical Proceedings of the Cambridge Philosophical Society*, 59(2):329–339, 1963.

[HP73] Alan J. Hoffman and Francisco Pereira. On copositive matrices with $-1, 0, 1$ entries. *Journal of Combinatorial Theory, Series A*, 14(3):302–309, 1973.

[HRTZ04] Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5(Oct):1391–1415, 2004.

[HV19] Nikolaus Hautsch and Stefan Voigt. Large-scale portfolio allocation under transaction costs and model uncertainty. *Journal of Econometrics*, 212(1):221–240, 2019.

[Jol02] Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[LDD+23] Miles Lubin, Oscar Dowson, Joaquim Dias Garcia, Joey Huchette, Benoît Legat, and Juan Pablo Vielma. JuMP 1.0: Recent improvements to a modeling language for mathematical optimization. *Mathematical Programming Computation*, 2023.

[LMMRC19] Martine Labbé, Luisa I. Martínez-Merino, and Antonio M. Rodríguez-Chía. Mixed integer linear programming for feature selection in support vector machine. *Discrete Applied Mathematics*, 261:276–304, 2019.

[LPR96] Zhi-Quan Luo, Jong-Shi Pang, and Daniel Ralph. *Mathematical programs with equilibrium constraints*. Cambridge University Press, 1996.

[Mar52] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[MD19] Luca Mencarelli and Claudia D'Ambrosio. Complex portfolio selection via convex mixed-integer quadratic programming: A survey. *International Transactions in Operational Research*, 26(2):389–414, 2019.

[Mil84] Alan J. Miller. Selection of subsets of regression variables. *Journal of the Royal Statistical Society Series A*, 147(3):389–410, 1984.

[MK87] Katta G. Murty and Santosh N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39(2):117–129, 1987.

[Mot52] Theodore S. Motzkin. Copositive quadratic forms. *National Bureau of Standards Report 1818*, pages 11–12, 1952.

[MPWL14] Sebastián Maldonado, Juan Pérez, Richard Weber, and Martine Labbé. Feature selection for support vector machines via mixed integer linear programming. *Information Sciences*, 279:163–175, 2014.

[MS65] Theodore S. Motzkin and Ernst G. Straus. Maxima for graphs and a new proof of a theorem of Turán. *Canadian Journal of Mathematics*, 17:533–540, 1965.

[MS12] Walter Murray and Howard Shek. A local relaxation method for the cardinality constrained portfolio optimization problem. *Computational Optimization and Applications*, 53:681–709, 2012.

[NDlT10] Minh Hoai Nguyen and Fernando De la Torre. Optimal feature selection for support vector machines. *Pattern Recognition*, 43(3):584–591, 2010.

[Pen22] Bo Peng. Performance comparison of two recently proposed copositivity tests. *EURO Journal on Computational Optimization*, 10:100037, 2022.

[Per84] Andre F. Perold. Large-scale portfolio optimization. *Management Science*, 30(10):1143–1160, 1984.

[PVZ15] Javier Pena, Juan C. Vera, and Luis F. Zuluaga. Completely positive reformulations for polynomial optimization. *Mathematical Programming*, 151:405–431, 2015.

[QdKRT98] Arie J. Quist, Etienne de Klerk, Cornelis Roos, and Tamas Terlaky. Copositive relaxation for general quadratic programming. *Optimization Methods and Software*, 9(1-3):185–208, 1998.

[RTGMS10] Rubén Ruiz-Torrubiano, Sergio García-Moratilla, and Alberto Suárez. Optimization problems with cardinality constraints. In *Computational Intelligence in Optimization: Applications and Implementations*, pages 105–130. Springer, 2010.

[SA99] Hanif D. Sherali and Warren P. Adams. *A Reformulation-Linearization Technique for Solving Discrete and Continuous Nonconvex Problems*. Springer US, Boston, MA, 1999.

[Sha48] Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[SM16] Naomi Shaked-Monderer. SPN graphs: when copositive = SPN. *Linear Algebra and its Applications*, 509:82–113, 2016.

[SMB21] Naomi Shaked-Monderer and Abraham Berman. *Copositive and completely positive matrices*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2021.

[SRW+02] Margaret A. Shipp, Pablo Ross, Ken N .and Tamayo, Andrew P. Weng, Jeffery L. Kutok, Ricardo C.T. Aguiar, Michelle Gaasenbeek, Michael Angelo, Michael Reich, Geraldine S. Pinkus, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74, 2002.

[SZL13] Xiaoling Sun, Xiaojin Zheng, and Duan Li. Recent advances in mathematical programming with semi-continuous variables and cardinality constraint. *Journal of the Operations Research Society of China*, 1(1):55–77, 2013.

[The24] The MathWorks Inc. `MATLAB` Optimization Toolbox version 24.1 (R2024a), 2024.

[WZ21] Angelika Wiegele and Shudian Zhao. Tight SDP relaxations for cardinality-constrained problems. In Norbert Trautmann and Mario Gnägi, editors, *Operations Research Proceedings 2021*, Lecture Notes in Operations Research, pages 167–172. Springer, 2021.

[WZZ06] Li Wang, Ji Zhu, and Hui Zou. The doubly regularized support vector machine. *Statistica Sinica*, 16(2):589–615, 2006.

[XTYP24] Wei Xu, Jie Tang, Ka Fai Cedric Yiu, and Jian Wen Peng. An efficient global optimal method for cardinality constrained portfolio optimization. *INFORMS Journal on Computing*, 36(2):690–704, 2024.

[YCHL10] Guo-Xun Yuan, Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. A comparison of optimization methods and software for large-scale l1-regularized linear classification. *Journal of Machine Learning Research*, 11:3183–3234, 2010.

[ZH05] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67(2):301–320, 2005.

[ZSL14] Xiaojin Zheng, Xiaoling Sun, and Duan Li. Improving the performance of miqp solvers for quadratic programs with cardinality and minimum threshold constraints: A semidefinite program approach. *INFORMS Journal on Computing*, 26(4):690–703, 2014.

[ZSLS14] Xiaojin Zheng, Xiaoling Sun, Duan Li, and Jie Sun. Successive convex approximations to cardinality-constrained convex programs: a piecewise-linear DC approach. *Computational Optimization and Applications*, 59(1):379–397, 2014.

A. Appendix

A.1. Notation

Throughout the thesis, by serif font lower case, e.g. x, and serif font upper case letter, e.g. X, we denote a vector and a matrix respectively. For instance, o and O stand for zero vectors and matrices of suitable size, respectively. For $\mathsf{x} \in \mathbb{R}^n$, the n-dimensional Euclidean space, and any nonempty $I \subset [1\!:\!n]$, we denote by $\mathsf{x}_I = [x_i]_{i\in I}$ the restricted subvector. We also define sets by:

$$
\begin{aligned}
\Delta^n &= \{\mathsf{x} \in \mathbb{R}^n_+ : \|\mathsf{x}\|_1 = 1\}\,,\\
\mathcal{S}^n &= \{\mathsf{X} \in \mathbb{R}^{n\times n} : \mathsf{X} = \mathsf{X}^\top\}\,,\\
\mathcal{N}^n &= \mathcal{S}^n \cap \mathbb{R}^{n\times n}_+\,,\\
\mathcal{S}^n_+ &= \{\mathsf{X} \in \mathcal{S}^n : \mathsf{v}^\top \mathsf{X}\mathsf{v} \geq 0 \text{ for all } \mathsf{v} \in \mathbb{R}^n\},\\
\mathcal{COP}^n &= \{\mathsf{X} \in \mathcal{S}^n : \mathsf{v}^\top \mathsf{X}\mathsf{v} \geq 0 \text{ for all } \mathsf{v} \in \mathbb{R}^n_+\}\,,\\
\mathcal{CP}^n &= \{\sum_{i\in I} \mathsf{x}_i\mathsf{x}_i^\top \in \mathcal{S}^n : \mathsf{x}_i \in \mathbb{R}^n_+\,,\ \text{for all } i \in I\}\,,\\
\mathcal{DNN}^n &= \mathcal{S}^n_+ \cap \mathcal{N}^n\,,\\
\mathcal{SPN}^n &= \mathcal{S}^n_+ + \mathcal{N}^n\,.
\end{aligned}
$$