Title
Strong formulations for quadratic optimization with M-matrices and indicator variables

Journal
Mathematical Programming, 170(1)

ISSN
0025-5610

Authors
Atamtürk, Alper
Gómez, Andrés

Publication Date
2018-07-01

DOI
10.1007/s10107-018-1301-5

Peer reviewed

Math. Program., Ser. B (2018) 170:141–176
https://doi.org/10.1007/s10107-018-1301-5

FULL LENGTH PAPER

Strong formulations for quadratic optimization with M-matrices and indicator variables

Alper Atamtürk[1] · **Andrés Gómez**[2]

Received: 1 February 2018 / Accepted: 16 May 2018 / Published online: 25 May 2018

Abstract We study quadratic optimization with indicator variables and an M-matrix, i.e., a PSD matrix with non-positive off-diagonal entries, which arises directly in image segmentation and portfolio optimization with transaction costs, as well as a substructure of general quadratic optimization problems. We prove, under mild assumptions, that the minimization problem is solvable in polynomial time by showing its equivalence to a submodular minimization problem. To strengthen the formulation, we decompose the quadratic function into a sum of simple quadratic functions with at most two indicator variables each, and provide the convex-hull descriptions of these sets. We also describe strong conic quadratic valid inequalities. Preliminary computational experiments indicate that the proposed inequalities can substantially improve the strength of the continuous relaxations with respect to the standard perspective reformulation.

Keywords Quadratic optimization · Submodularity · Perspective formulation · Conic quadratic cuts · Convex piecewise nonlinear inequalities

Mathematics Subject Classification 90C11 · 90C20 · 90C57

A. Atamtürk was supported, in part, by Grant FA9550-10-1-0168 from the Office of the Assistant Secretary of Defense for Research and Engineering.

✉ Alper Atamtürk
atamturk@berkeley.edu

Andrés Gómez
agomez@pitt.edu

[1] Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720, USA

[2] Department of Industrial Engineering , Swanson School of Engineering, University of Pittsburgh, Pittsburgh, PA 15261, USA

1 Introduction

Consider the quadratic optimization problem with indicator variables

$$\text{(QOI)} \quad \min \left\{ a'x + b'y + y'Ay \ : \ (x, y) \in C, \ 0 \le y \le x, \ x \in \{0, 1\}^N \right\},$$

where $N = \{1, \ldots, n\}$, a and b are n-vectors, A is an $n \times n$ symmetric matrix and $C \subseteq \mathbb{R}^{N \times N}$. Binary variables x indicate a selected subset of N and are often used to model non-convexities such as cardinality constraints and fixed charges. (QOI) arises in linear regression with best subset selection [10], control [23], filter design [47] problems, and portfolio optimization [11], among others. In this paper, we give strong convex relaxations for the related mixed-integer set

$$S = \left\{ (x, y, t) \in \{0, 1\}^N \times \mathbb{R}^N \times \mathbb{R} : y'Qy \le t, \ 0 \le y_i \le x_i \text{ for all } i \in N \right\},$$

where Q is an M-matrix [43], i.e., $Q \succeq 0$ and $Q_{ij} \le 0$ if $i \neq j$. M-matrices arise in the analysis of Markov chains [30]. Convex quadratic programming with an M-matrix is also studied on its own right [37]. Quadratic minimization with an M-matrix arises directly in a variety of applications including portfolio optimization with transaction costs [33] and image segmentation [27].

There are numerous approaches in the literature for deriving strong formulations for (QOI) and S. Dong and Linderoth [19] describe lifted inequalities for (QOI) from its continuous quadratic optimization counterpart over bounded variables. Bienstock and Michalka [12] give a characterization linear inequalities obtained by strengthening gradient inequalities of a convex objective function over a non-convex set. Convex relaxations of S can also be constructed from the mixed-integer epigraph of the bilinear function $\sum_{i \neq j} Q_{ij} y_i y_j$. There is an increasing amount of recent work focusing on bilinear functions [e.g., [13,14,35]]. However, the convex hull of such functions is not fully understood even in the continuous case. More importantly, considering the bilinear functions independent from the quadratic function $\sum_{i \in N} Q_{ii} y_i^2$ may result in weaker formulations for S. Another approach, applicable to general mixed-integer optimization, is to derive a strong formulation based on disjunctive programming [8,17,45]. Specifically, if a set is defined as the disjunction of convex sets, then its convex hull can be represented in an extended formulation using perspective functions. Such extended formulations, however, require creating a copy of each variable for each disjunction, and lead to prohibitively large formulations even for small-scale instances. There is also a increasing body of work on characterizing the convex hulls in the original space of variables, but such descriptions may be highly complex even for a single disjunction, e.g., see [7,9,31,39].

The convex hull of S is well-known for a couple of special cases. When the matrix Q is diagonal, the quadratic function $y'Qy$ is separable and the convex hull of S can be described using the *perspective reformulation* [21]. This perspective formulation has a compact conic quadratic representation [2,24] and is by now a standard model strengthening technique for mixed-integer nonlinear optimization [15,25,38,48]. In particular, a convex quadratic function $y'Ay$ is decomposed as $y'Dy + y'Ry$, where

$A = D + R$, $D, R \succeq 0$ and D is diagonal and then each diagonal term $D_{ii} y_i^2 \le t_i$, $i \in N$, is reformulated as $y_i^2 \le t_i x_i$. Such decomposition and strengthening of the diagonal terms are also standard for the binary restriction, where $y_i = x_i$, $i \in N$, in which case $x'Ax \Leftrightarrow \sum_{i \in N} D_{ii} x_i + x'Rx$ [e.g.[3,44]]. The binary restriction of S, where $y_i = x_i$ and $Q_{ij} \le 0$, $i \ne j$, is also well-understood, since in that case the quadratic function $x'Qx$ is submodular [40] and min $\{a'x + x'Qx : x \in \{0,1\}^n\}$ is a minimum cut problem [28,42] and, therefore, is solvable in poynomial time.

Whereas the set S with an M-matrix is interesting on its own, the convexification results on S can also be used to strengthen a general quadratic $y'Ay$ by decomposing A as $A = Q + R$, where Q is an M-matrix, and then applying the convexification results in this paper only on the $y'Qy$ term with negative off-diagonal coefficients, generalizing the perspective reformulation approach above. We demonstrate this approach for portfolio optimization problems with negative as well as positive correlations through computations that indicate significant additional strengthening over the perspective formulation through exploiting the negative correlations.

The key idea for deriving strong formulations for S is decompose the quadratic function in the definition of S as the sum of quadratic functions involving one or two variables:

$$y'Qy = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} Q_{ij} \right) y_i^2 - \sum_{i=1}^{n} \sum_{j=i+1}^{n} Q_{ij} (y_i - y_j)^2. \tag{1}$$

Since a univariate quadratic function with an indicator is well-understood, we turn our attention to studying the mixed-integer set with two continuous and two indicator variables:

$$X = \left\{ (x, y, t) \in \{0,1\}^2 \times \mathbb{R}^2 \times \mathbb{R} : (y_1 - y_2)^2 \le t,\ 0 \le y_i \le x_i,\ i = 1, 2 \right\}.$$

Frangioni et al [22] also construct strong formulations for (QOI) based on 2×2 decompositions. In particular, they characterize quadratic functions that can be decomposed as the sum of convex quadratic functions with at most two variables. They utilize the disjunctive convex extended formulation for the mixed-integer quadratic set

$$\hat{X} = \left\{ (x, y, t) \in \{0,1\}^2 \times \mathbb{R}^2 \times \mathbb{R} : q(y) \le t,\ 0 \le y_i \le x_i,\ i = 1, 2 \right\},$$

where $q(y)$ is a general convex quadratic function. The authors report that the formulations are weaker when the matrix A is an M-matrix, and remark on the high computational burden of solving the convex relaxations due the large number of additional variables. Additionally, Jeon et al. [29] give conic quadratic valid inequalities for $\hat{X}$, which can be easily projected into the original space of variables, and demonstrate their effectiveness via computations. However, a convex hull description of $\hat{X}$ in the original space of variable is unknown.

In this paper, we improve upon previous results for the sets S and X. In particular, our main contributions are (i) showing, under mild assumptions, that the minimization of a quadratic function with an M-matrix and indicator variables is equivalent to a submodular minimization problem and, hence, solvable in polynomial time; (ii)

giving the convex hull description of X in the original space of variables—the resulting formulations for S are at least as strong as the ones used by Frangioni et al. and require substantially fewer variables; (iii) proposing conic quadratic inequalities amenable to use with conic quadratic MIP solvers—the proposed inequalities dominate the ones given by Jeon et al.; (iv) demonstrating the strength and performance of the resulting formulations for (QOI).

Outline The rest of the paper is organized as follows. In Sect. 2 we review the previous results for S and X. In Sect. 3 we study the relaxations of S and X, where the constraints $0 \le y_i \le x_i$ are relaxed to $y_i(1 - x_i) = 0$, and the related optimization problem. In Sect. 4 we give the convex hull description of X. The convex hulls obtained in Sects. 3 and 4 cannot be immediately implemented with off-the-shelf solvers in the original space of variables. Thus, in Sect. 5 we propose valid conic quadratic inequalities and discuss their strength. In Sect. 6 we give extensions to quadratic functions with positive off-diagonal entries and continuous variables unrestricted in sign. In Sect. 7 we provide a summary computational experiments and in Sect. 8 we conclude the paper.

Notation Throughout the paper, we use the following convention for division by 0: ${}^0/_0 = 0$ and ${}^a/_0 = \infty$ if $a > 0$. In particular, the function $p : [0, 1] \times \mathbb{R}_+ \to \mathbb{R}_+$ given by $p(x, y) = {}^{y^2}/_x$ is the closure of the perspective function of the quadratic function $q(y) = y^2$, and is convex [e.g. [26], p.160] . For a set $X \subseteq \mathbb{R}^N$, $\text{conv}(X)$ denotes the convex hull of X. Throughout, Q denotes an $n \times n$ M-matrix, i.e., $Q \succeq 0$ and $Q_{ij} \le 0$ for $i \ne j$.

2 Preliminaries

In this section we briefly review the relevant results on the binary restriction of S and the previous results on set X.

2.1 The binary restriction of S

Let S_B be the binary restriction of S, i.e. $y = x \in \{0, 1\}^n$. In this case, the decomposition

$$x'Qx = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} Q_{ij} \right) x_i^2 - \sum_{i=1}^{n} \sum_{j=i+1}^{n} Q_{ij}(x_i - x_j)^2 \le t \tag{2}$$

leads to $\text{conv}(S_B)$, by simply taking the convex hull of each term. Indeed, the quadratic problem $\min \left\{ x'Qx : x \in \{0, 1\}^n \right\}$ is equivalent to an undirected min-cut problem [e.g. [42]] and can be formulated as

$$\min \sum_{i=1}^{n} \left(\sum_{j=1}^{n} Q_{ij} \right) x_i - \sum_{i=1}^{n} \sum_{j=i+1}^{n} Q_{ij} t_{ij} : x_i - x_j \le t_{ij},\ x_j - x_i \le t_{ij},\ 0 \le x \le 1.$$

Decomposition (2) leading to a simple convex hull description of S_B in the binary case is our main motivation for studying decomposition (1) with the indicator variables.

2.2 Previous results for set X

Here we review the valid inequalities of Jeon et al. [29] for X. Although their construction is not directly applicable as they assume a strictly convex function, one can utilize it to obtain limiting inequalities. For $q(y) = y'Ay$ the inequalities of Jeon et al. are described via the inverse of the Cholesky factor of A. However, for X, we have $q(y) = (y_1 - y_2)^2$ or $q(y) = y'Ay$, where $A = \left[\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right]$ is a singular matrix and the Cholesky factor is not invertible.

However, if the matrix is given by $A = \left[\begin{smallmatrix} d_1 & -1 \\ -1 & d_2 \end{smallmatrix}\right]$ with $d_1, d_2 > 1$, then their approach yields three valid inequalities:

$$
\begin{aligned}
& d_2 \frac{y_2^2}{x_2} - \frac{1}{d_1} x_1 + \left(\frac{d_1 d_2 - 1}{d_1}\right) \frac{y_2^2}{x_2} \le t \\
& (d_2 - 1) \frac{y_2^2}{x_2} + d_1 \frac{y_1^2}{x_1} + \frac{x_2}{d_1} - 2x_2 \le t \\
& \left(\frac{d_1 d_2 - 1}{d_1}\right) \frac{y_2^2}{x_2} + \frac{\left(\sqrt{d_1} y_1 - \sqrt{\frac{1}{d_1}} y_2\right)^2}{x_1 + x_2} \le t.
\end{aligned}
$$

As $d_1, d_2 \to 1$, we arrive at three limiting valid inequalities for X.

Proposition 1 *The following convex inequalities are valid for X:*

$$
\frac{y_2^2}{x_2} - x_1 \le t, \tag{3}
$$

$$
\frac{y_1^2}{x_1} - x_2 \le t, \tag{4}
$$

$$
\frac{(y_1 - y_2)^2}{x_1 + x_2} \le t. \tag{5}
$$

For completeness, we verify here the validity of the limiting inequalities directly. The validity of inequality (3) is easy to see: observe that $y_2^2/x_2 \le 1$ for $(x, y) \in X$; then, for $x_1 = 0$, (3) reduces to the perspective formulation for the quadratic constraint $y_2^2 \le t$, and for $x_1 = 1$ we have $y_2^2/x_2 - x_1 \le 0 \le t$. The validity of inequality (4) is proven identically. Finally, inequality (5) is valid since it forces $y_1 = y_2$ when $x_1 = x_2 = 0$, and is dominated by the original inequality $(y_1 - y_2)^2 \le t$ for other integer values of x.

Inequalities (3)–(5) are not sufficient to describe conv(X) though. In the next two sections we describe conv(X) and give new conic quadratic valid inequalities dominating (3)–(5) for X.

3 The unbounded relaxation

In this section we study the unbounded relaxations of S and X obtained by dropping the upper bound on the continuous variables:

$$S_U = \left\{(x, y, t) \in \{0, 1\}^N \times \mathbb{R}^N_+ \times \mathbb{R} : y'Qy \leq t,\ \ y_i(1 - x_i) = 0 \text{ for all } i \in N\right\},$$
$$X_U = \left\{(x, y, t) \in \{0, 1\}^2 \times \mathbb{R}^2_+ \times \mathbb{R} : (y_1 - y_2)^2 \leq t : y_i(1 - x_i) = 0,\ \ i = 1, 2\right\}.$$

In Sect. 3.1 we show that the minimization of a linear function over S_U is equivalent to a submodular minimization problem and, consequently, solvable in polynomial time. In Sect. 3.2, we describe $\text{conv}(X_U)$ and in Sect. 3.3 we use the results in Sect. 3.2 to derive valid inequalities for S_U.

3.1 Optimization over S_U

We now show that the optimization of a linear function over S_U can be solved in polynomial time under a mild assumption on the objective function. Consider the problem

$$\text{(P)} \quad \min\left\{a'x + b'y + t : (x, y, t) \in S_U\right\},$$

where Q is a positive definite M-matrix and $b \leq 0$. We show that (P) is a submodular minimization problem. The positive definiteness assumption on Q ensures that an optimal solution exists. Otherwise, if there is $y \geq 0$ with $y'Qy = 0$, the problem may be unbounded. The assumption $b \leq 0$ is satisfied in most applications (e.g., see Sects. 7.1 and 7.3). If $b > 0$, then $y = 0$ in any optimal solution.

Proposition 2 (Characterization 15 [43]) *A positive definite M-matrix Q is* inverse-positive*, i.e., its inverse satisfies $Q^{-1}_{ij} \geq 0$ for all i, j.*

Proposition 3 *Problem (P) is equivalent to a submodular minimization problem and it is, therefore, solvable in polynomial time.*

Proof We assume that $a \geq 0$ (otherwise $x = 1$ in any optimal solution) and that an optimal solution exists. Given an optimal solution (x^*, y^*) to (P), let $T = \left\{i \in N : y^*_i > 0\right\}$, b_T the subvector of b induced by T, and by Q_T the submatrix of Q induced by T. Then, from KKT conditions, we find $b_T + 2Q_T y_T = 0 \Leftrightarrow y_T = -Q^{-1}_T b_T/2\cdot$ Thus, an optimal solution satisfies $b'y^* + y^{*\prime}Qy^* = -\frac{b'_T Q^{-1}_T b_T}{4}\cdot$

Consequently, defining $\theta_{ij} : 2^N \to \mathbb{R}$ for $i, j \in N$ as $\theta_{ij}(T) = (Q^{-1}_T)_{ij}$ if $i, j \in T$ and 0 o.w., observe that (P) is equivalent to the binary minimization problem

$$\min_{T \subseteq N} \ a(T) - \frac{1}{4}\sum_{i \in N}\sum_{j \in N} b_i b_j \theta_{ij}(T)\cdot$$

Note that since Q_T is a positive definite M-matrix for any $T \subseteq N$, $Q_T = \mu I_T - P_T$, where P_T is a nonnegative matrix and the largest eigenvalue of P_T is less than μ. By scaling, we may assume that $\mu = 1$. Moreover, $Q_T^{-1} = (I - P_T)^{-1} = \sum_{\ell=0}^{\infty} P_T^{\ell}$ [e.g. [49]]. For $\ell \in \mathbb{Z}_+$ and all $i, j \in N$ let $\bar{\theta}_{ij}^{\ell}(T) = (P_T^{\ell})_{ij}$ if $i, j \in T$, and 0 o.w. Note that $\theta_{ij}(T) = \sum_{\ell=0}^{\infty} \bar{\theta}_{ij}^{\ell}(T)$. Finally, define for $k \in N$ and $T \subseteq N \backslash \{k\}$ the increment function $\rho_{ij}^{\ell}(k, T) = \bar{\theta}_{ij}^{\ell}(T \cup \{k\}) - \bar{\theta}_{ij}^{\ell}(T)$.

Claim For all $i, j \in N$ and $\ell \in \mathbb{Z}_+$, $\bar{\theta}_{ij}^{\ell}$ is a monotone supermodular function.

Proof The claim is proved by induction on ℓ.

- Base case, $\ell = 0$: Let $k \in N$ and $T \subseteq N \backslash \{k\}$. Note that $P_T^0 = I_T$. Thus $\rho_{kk}^0(k, T) = 1$, and $\rho_{ij}^0(k, T) = 0$ for all cases except $i = j = k$. Thus, the marginal contributions are constant and $\bar{\theta}_{ij}^0$ is supermodular. Monotonicity can be checked easily.
- Induction step: Suppose $\bar{\theta}_{ij}^{\ell}$ is supermodular and monotone for all $i, j \in N$. Observe that $\bar{\theta}_{ij}^{\ell+1}(T) = \sum_{t \in N} \bar{\theta}_{it}^{\ell}(T) P_{tj}$ if $i, j \in T$ and $\bar{\theta}_{ij}^{\ell+1}(T) = 0$ otherwise. Monotonocity of $\bar{\theta}_{ij}^{\ell+1}$ follows immediately from the monotonicity of the functions $\bar{\theta}_{it}^{\ell}$. Now let $k \in N$ and $T_1 \subseteq T_2 \subseteq N \backslash \{k\}$. To prove supermodularity, we check that $\rho_{ij}^{\ell+1}(k, T_2) - \rho_{ij}^{\ell+1}(k, T_1) \geq 0$ by considering all cases:

  $k \notin \{i, j\}$: If $\{i, j\} \subseteq T_1$ then $\rho_{ij}^{\ell+1}(k, T_2) - \rho_{ij}^{\ell+1}(k, T_1) = \sum_{t \in N} (\rho_{it}^{\ell}(k, T_2) - \rho_{it}^{\ell}(k, T_1)) P_{tj} \geq 0$ by supermodularity of functions $\bar{\theta}_{it}^{\ell}$; if $\{i, j\} \nsubseteq T_1$ and $\{i, j\} \subseteq T_2$ then $\rho_{ij}^{\ell+1}(k, T_2) - \rho_{ij}^{\ell+1}(k, T_1) = \rho_{ij}^{\ell+1}(k, T_2) \geq 0$ by monotonicity; finally, if $\{i, j\} \nsubseteq T_2$ then $\rho_{ij}^{\ell+1}(k, T_2) - \rho_{ij}^{\ell+1}(k, T_1) = 0$.

  $k = i$: If $j \in T_1$ then $\rho_{kj}^{\ell+1}(k, T_2) - \rho_{kj}^{\ell+1}(k, T_1) = \sum_{t \in N} (\rho_{kt}^{\ell}(k, T_2) - \rho_{kt}^{\ell}(k, T_1)) P_{tj} \geq 0$ by supermodularity of functions $\bar{\theta}_{kt}^{\ell}$; if $j \notin T_1$ and $j \in T_2$ then $\rho_{kj}^{\ell+1}(k, T_2) - \rho_{kj}^{\ell+1}(k, T_1) = \bar{\theta}_{kj}^{\ell+1}(T_2 \cup \{k\}) \geq 0$; finally, if $j \notin T_2$ then $\rho_{kj}^{\ell+1}(k, T_2) - \rho_{kj}^{\ell+1}(k, T_1) = 0$. The case $k = j$ is identical. □

As $\theta_{ij}(T) = \sum_{\ell=0}^{\infty} \bar{\theta}_{ij}^{\ell}(T)$ is a sum of supermodular functions, it is supermodular. Consequently, $1/4 \sum_{i \in N} \sum_{j \in N} b_i b_j \theta_{ij}(T)$ is a supermodular function and (P) is a submodular minimization problem, solvable with a strongly polynomial number of calls to a value oracle [e.g.[41]]. Evaluating the submodular function for a given set T, i.e., computing $a(T) - b_T' Q_T^{-1} b_T / 4$, requires only matrix multiplication and inversion, and can be done in strongly polynomial time. Therefore (P) is solvable in strongly polynomial time. □

3.2 Convex hull of X_U

Consider the function $f : [0, 1]^2 \times \mathbb{R}_+^2 \to \mathbb{R}_+$ defined as

$$f(x, y) = \begin{cases} \frac{(y_1 - y_2)^2}{x_1} & \text{if } y_1 \geq y_2 \\ \frac{(y_2 - y_1)^2}{x_2} & \text{if } y_1 \leq y_2 \end{cases} \tag{6}$$

and the corresponding nonlinear inequality

$$f(x, y) \leq t. \tag{7}$$

Remark 1 Observe that that inequality (7) dominates inequality (5) since

$$\frac{(y_1 - y_2)^2}{x_1 + x_2} \leq \frac{(y_1 - y_2)^2}{\max\{x_1, x_2\}} \leq f(x, y).$$

Inequalities (3)–(4) are not valid for the unbounded relaxation as the conditions $y_i^2/x_i \leq 1$ are not satisfied by all feasible points in X_U. For example, feasible points with $x_1 = x_2 = 1$, $y_1 = y_2 > 1$ and $t = 0$ are cut off by (3)–(4).

Proposition 4 *Inequality* (7) *is valid for* X_U.

Proof There are four cases to consider. If $x_1 = x_2 = 1$, then $f(x, y)$ reduces to the original quadratic inequality $(y_1 - y_2)^2$, thus the inequality is valid. If $x_1 = x_2 = 0$, then the points in X_U satisfy $y_1 = y_2 = 0$ and $t \geq 0$; since $f(0, 0) = 0$, none of these points are cut off by (7). If $x_1 = 1$ and $x_2 = 0$, then $y_2 = 0$ in any point in X_U and, in particular, $y_1 \geq y_2$; thus $f(x, y)$ reduces to the original inequality. The case where $x_1 = 0$ and $x_2 = 1$ is similar. □

Observe that function f is a piecewise nonlinear function, where each piece is conic quadratic representable. However, the pieces are not valid outside of the region where they are defined, e.g., $(y_1 - y_2)^2 \leq tx_1$ is invalid when $y_2 > y_1$ as it cuts off feasible points with $x_1 = y_1 = 0$ and $y_2 > 0$. Thus, inequality (7) is not equivalent to the system given by $(y_1 - y_2)^2 \leq tx_i$, $i = 1, 2$. Nevertheless, as shown in Proposition 5 below, (7) is a convex inequality.

Proposition 5 *The function* f *is convex on its domain.*

Proof Let $(\bar{x}, \bar{y}), (\hat{x}, \hat{y}) \in [0, 1]^2 \times \mathbb{R}_+^2$ and let $(x^*, y^*) = (1 - \lambda)(\bar{x}, \bar{y}) + \lambda(\hat{x}, \hat{y})$ for $0 \leq \lambda \leq 1$ be a convex combination of $(\bar{x}, \bar{y})$ and $(\hat{x}, \hat{y})$. We need to prove that

$$f(x^*, y^*) \leq (1 - \lambda) f(\bar{x}, \bar{y}) + \lambda f(\hat{x}, \hat{y}). \tag{8}$$

If $\bar{y}_1 \geq \bar{y}_2$ and $\hat{y}_1 \geq \hat{y}_2$, or $\bar{y}_1 \leq \bar{y}_2$ and $\hat{y}_1 \leq \hat{y}_2$, inequality (8) holds by convexity of the individual functions in the definition of f. Otherwise, assume, without loss of generality, that $\bar{y}_1 \geq \bar{y}_2$, $\hat{y}_1 \leq \hat{y}_2$, and $y_1^* \leq y_2^*$. Letting $\gamma = \lambda - (1 - \lambda)\frac{\bar{y}_1 - \bar{y}_2}{\hat{y}_2 - \hat{y}_1}$, observe that

- $\gamma \leq \lambda \leq 1$.
- $\gamma \geq 0$, which is equivalent to $y_2^* - y_1^* \geq 0$.
- $y_2^* - y_1^* = \gamma(\hat{y}_2 - \hat{y}_1)$.
- $\gamma \hat{x}_2 \leq \lambda \hat{x}_2 \leq x_2^*$.

Then, we find

$$f(x^*, y^*) = \frac{(y_2^* - y_1^*)^2}{x_2^*} \le \frac{(y_2^* - y_1^*)^2}{\gamma \hat{x}_2}$$
$$= \gamma \frac{(\hat{y}_2 - \hat{y}_1)^2}{\hat{x}_2} \le \lambda f(\hat{x}, \hat{y}) + (1-\lambda) f(\bar{x}, \bar{y}).$$

□

A consequence of Proposition 5 is that the convex inequality (7) can be implemented (with off-the-shelf solvers) using subgradient inequalities as for a subgradient $\xi \in \partial f(\bar{x}, \bar{y})$ at a given point $(\bar{x}, \bar{y})$, we have $f(\bar{x}, \bar{y}) + \xi'(x - \bar{x}, y - \bar{y}) \le f(x, y)$, for all points (x, y) in the domain of the convex function f. In particular, the linear cuts

$$f(\bar{x}, \bar{y}) + \xi'(x - \bar{x}, y - \bar{y}) \le t \text{ for } \xi \in \partial f(\bar{x}, \bar{y}) \tag{9}$$

provide an outer-approximation of $f(x, y) \le t$ at $(\bar{x}, \bar{y})$ and are valid everywhere on the domain. A subgradient ξ can be found simply by taking the gradient of the relevant piece of the function at $(\bar{x}, \bar{y})$. In particular, for $\bar{y}_1 \ge \bar{y}_2$ and $\bar{x}_1 > 0$, a subgradient inequality is

$$-\left(\frac{\bar{y}_1 - \bar{y}_2}{\bar{x}_1}\right)^2 x_1 + 2\left(\frac{\bar{y}_1 - \bar{y}_2}{\bar{x}_1}\right)(y_1 - y_2) \le t. \tag{10}$$

The process outlined here to find subgradient cuts (9) for f can be utilized for any convex piecewise nonlinear function, and will be used for other functions in the rest of the paper. Convex piecewise nonlinear functions also arise in strong formulations for mixed-integer conic quadratic optimization [5], and subgradient linear cuts for such functions were recently used in the context of the pooling problem [36].

As Theorem 1 below states, inequality (7) and bound constraints for the binary variables describe the convex hull of X_U.

Theorem 1 (Convex hull of X_U)

$$conv(X_U) = \left\{(x, y, t) \in [0, 1]^2 \times \mathbb{R}_+^2 \times \mathbb{R} : f(x, y) \le t\right\}.$$

Proof Consider the optimization problems

$$(P_0) \qquad \min_{(x,y,t)\in X_U} a'x + b'y + ct;$$
$$(P_1) \qquad \min_{(x,y,t)\in[0,1]^2\times\mathbb{R}_+^2\times\mathbb{R}} a'x + b'y + ct \text{ s.t. } f(x, y) \le t.$$

To prove the result we show that for any value of a, b, c, either (P_0) and (P_1) are both unbounded, or there exists a solution integral in x that is optimal for both problems. If $c < 0$, then (P_0) and (P_1) are both unbounded, and if $c = 0$ then (P_1) corresponds to an optimization problem over an integral polyhedron and it is easily checked that (P_0) and (P_1) are equivalent. Thus, the interesting case is $c > 0$ or, by scaling, $c = 1$. Note that $t = (y_1 - y_2)^2$ in any optimal solution of (P_0), and $t = f(x, y)$ in any optimal solution of (P_1). If $b_1, b_2 \geq 0$, then $y_1 = y_2 = 0$ is optimal with corresponding integer x optimal for both (P_0) and (P_1).

Moreover, if $b_1 + b_2 < 0$, then both problems are unbounded: $x_1 = x_2 = 1$, $y_1 = y_2 = \lambda$ is feasible for any $\lambda > 0$ for both problems. Thus, one needs to consider only the case where $b_1 + b_2 \geq 0$ and $b_1 < 0$ or $b_2 < 0$. Without loss of generality, let $b_1 < 0$ and $b_2 > 0$.

Optimal solutions of (P_0). There exists an optimal solution with $y_2 = 0$ (if $0 < y_2 \leq y_1$, subtracting $\epsilon > 0$ from both y_1 and y_2 does not increase the objective – and if $y_2 > y_1$, then swapping the values of y_1 and y_2 reduces the objective). Thus, $y_2 = 0$, $x_2 = 0$ if $a_2 \geq 0$ and $x_2 = 1$ otherwise, and either $x_1 = y_1 = 0$ or $x_1 = 1$ and $y_1 = -\frac{b_1}{2}$, which is the stationary point of $b_1 y_1 + y_1^2$.

Optimal solutions of (P_1). Note that there exists an optimal solution of (P_1) where at least one of the continuous variables is 0 (if $0 < y_1, y_2$, subtracting $\epsilon > 0$ from both variables does not increase the objective value — this operation does not change the relative order of y_1 and y_2). Then, we conclude that $y_2 = 0$ in an optimal solution (if $y_1 = 0$ and $y_2 > 0$, then setting $y_2 = 0$ reduces the objective value). Moreover, when $y_2 = 0$, then $f(x, y) = y_1^2/x_1$. Thus, in the optimal solution $y_1 = -b_1 x_1/2$. Substituting in the objective, we see that (P_1) simplifies to $\min_{0 \leq x_1, x_2 \leq 1} a_2 x_2 + \big(a_1 - b_1^2/4\big)x_1$. For an optimal solution, $x_2 = 0$ if $a_2 \geq 0$ and $x_2 = 1$ otherwise, and $x_1 = 0$ if $a_1 - b_1^2/4 \geq 0$ and $x_1 = 1$ otherwise. And, if $x_1 = 1$, then $y_1 = -b_1/2$. Hence, the optimal solutions coincide. □

3.3 Valid inequalities for S_U

Inequalities in an extended formulation Let $\bar{Q}_i = \sum_{j=1}^n Q_{ij}$ and $P = \{i \in N : \bar{Q}_i > 0\}$ and $\bar{P} = N \backslash P$. Using decomposition (1) and introducing t_{ij}, $1 \leq i \leq j \leq n$, one can write a convex relaxation of S_U as

$$\begin{aligned}
&\sum_{i \in \bar{P}} \bar{Q}_i y_i + \sum_{i \in P} \bar{Q}_i y_i^2/x_i - \sum_{i=1}^n \sum_{j=i+1}^n Q_{ij} t_{ij} \leq t \\
&\qquad\qquad\qquad\qquad\qquad f(x_i, x_j, y_i, y_j) \leq t_{ij}, \quad 1 \leq i \leq j \leq n.
\end{aligned}$$

Inequalities in the original space of variables By projecting out the auxiliary variables t_{ij} one obtains valid inequalities in the original space of variables. By re-indexing variables if necessary, assume that $y_1 \geq y_2 \geq \ldots \geq y_n$ to obtain the convex inequality

$$\sum_{i \in \bar{P}} \bar{Q}_i y_i + \sum_{i \in P} \bar{Q}_i y_i^2/x_i - \sum_{i=1}^n \sum_{j=i+1}^n Q_{ij} (y_i - y_j)^2/x_i \leq t. \tag{11}$$

Observe that the nonlinear inequality (11) is valid only if $y_1 \geq \ldots \geq y_n$ holds. However, we can obtain linear inequalities that are valid for S_U by underestimating the convex function $\sum_{i \in \bar{P}} \bar{Q}_i y_i + \sum_{i \in P} \bar{Q}_i y_i^2/x_i - \sum_{i=1}^n \sum_{j=i+1}^n Q_{ij} f(x_i, x_j, y_i, y_j)$

by its subgradients. Let $(\bar{x}, \bar{y}) \in [0,1]^N \times \mathbb{R}^N_+$ be such that $\bar{y}_1 \geq \ldots \geq \bar{y}_n$ and $\bar{x} > 0$. Then, the subgradient inequality

$$
\begin{aligned}
&-\sum_{i \in P} \bar{Q}_i \left(\frac{\bar{y}_i}{\bar{x}_i}\right)^2 x_i + \sum_{i=1}^{n} \left(\sum_{j=i+1}^{n} \frac{Q_{ij}(\bar{y}_i - \bar{y}_j)^2}{\bar{x}_i^2} \right) x_i \\
&+ 2\sum_{i \in P} \bar{Q}_i \frac{\bar{y}_i}{\bar{x}_i} y_i + \sum_{i \in \bar{P}} \bar{Q}_i y_i \\
&+ 2\sum_{i=1}^{n} \left(\sum_{j=1}^{i-1} \frac{Q_{ij}(\bar{y}_j - \bar{y}_i)}{\bar{x}_j} - \sum_{j=i+1}^{n} \frac{Q_{ij}(\bar{y}_i - \bar{y}_j)}{\bar{x}_i} \right) y_i \leq t,
\end{aligned}
$$

corresponding to a first order approximation of (11) around $(\bar{x}, \bar{y})$, is valid for S_U (regardless of the ordering of the variables).

4 The bounded set X

Let $g : [0,1]^2 \times \mathbb{R}^2_+ \to \mathbb{R}_+$ be defined as

$$
g(x, y) = \begin{cases} \frac{(y_1 - x_2)^2}{x_1 - x_2} + \frac{(x_2 - y_2)^2}{x_2} & \text{if } y_2 \leq x_2 \leq y_1 \text{ and } x_2(x_1 - y_1) \leq y_2(x_1 - x_2) \\ \frac{(y_2 - x_1)^2}{x_2 - x_1} + \frac{(x_1 - y_1)^2}{x_1} & \text{if } y_1 \leq x_1 \leq y_2 \text{ and } x_1(x_2 - y_2) \leq y_1(x_2 - x_1) \\ f(x, y) & \text{otherwise,} \end{cases} \tag{12}
$$

where f is the function defined in (6). This section is devoted to proving the main result:

Theorem 2 (Convex hull of X)

$$
conv(X) = \left\{ (x, y, t) \in [0,1]^2 \times \mathbb{R}^3_+ : g(x, y) \leq t,\ y_i \leq x_i,\ i = 1, 2 \right\}.
$$

Remark 2 Observe that for the binary restriction X_B with $y_i = x_i$, $i = 1, 2$, $g(x, y) \leq t$ reduces to $|x_1 - x_2| \leq t$, which together with the bound constraints describe $\text{conv}(X_B)$.

The rest of this section is organized as follows. In Sect. 4.1 we give the convex hull description of the intermediate set with two continuous variables and one indicator variable:

$$
X_1 = \left\{ (x, y, t) \in \{0,1\} \times \mathbb{R}^2_+ \times \mathbb{R} : (y_1 - y_2)^2 \leq t,\ y_1 \leq x,\ y_2 \leq 1 \right\}.
$$

In Sect. 4.2 we use this results to prove Theorem 2. Finally, in Sect. 4.3 we give valid inequalities for S. Unlike in Sect. 3, the convex hull proofs in this section are constructive, i.e., we show how g is constructed from the mixed-binary description of X, instead of just verifying that g does indeed result in $\text{conv}(X)$.

4.1 Convex hull description of X_1

Let $g_1 : [0, 1] \times \mathbb{R}^2_+ \to \mathbb{R}_+$ be given by

$$g_1(x, y_1, y_2) = \begin{cases} \frac{(y_2-x)^2}{1-x} + \frac{(x-y_1)^2}{x} & \text{if } x - y_1 \le x(y_2 - y_1) \\ \frac{(y_1-y_2)^2}{x} & \text{if } y_2 \le y_1 \\ (y_2 - y_1)^2 & \text{otherwise.} \end{cases}$$

Proposition 6 *conv*$(X_1) = \{(x, y, t) \in [0, 1] \times \mathbb{R}^2_+ \times \mathbb{R} : g_1(x, y_1, y_2) \le t, y_1 \le x, y_2 \le 1\}$.

Proof Note that a point (x, y, t) belongs to conv(X_1) if and only if there exists $(\bar{x}, \bar{y}, \bar{t})$, $(\hat{x}, \hat{y}, \hat{t})$ and $0 \le \lambda \le 1$ such that

$$t = (1-\lambda)\bar{t} + \lambda\hat{t} \tag{13}$$
$$x = (1-\lambda)\bar{x} + \lambda\hat{x} \tag{14}$$
$$y_1 = (1-\lambda)\bar{y}_1 + \lambda\hat{y}_1 \tag{15}$$
$$y_2 = (1-\lambda)\bar{y}_2 + \lambda\hat{y}_2 \tag{16}$$
$$\bar{x} = 0,\ \hat{x} = 1 \tag{17}$$
$$\bar{y}_1 = 0,\ 0 \le \hat{y}_1 \le 1 \tag{18}$$
$$0 \le \bar{y}_2,\ \hat{y}_2 \le 1 \tag{19}$$
$$\bar{t} \ge \bar{y}_2^2 \tag{20}$$
$$\hat{t} \ge (\hat{y}_1 - \hat{y}_2)^2. \tag{21}$$

The non-convex system (13)–(21) follows directly from the definition of the convex hull. Note that a convex extended formulation of conv(X_1) could also be obtained using the approach proposed by [17]. See also Vielma [46] for a recent approach to eliminate the auxiliary variables using Cayley embedding. We now show how to project out the additional variables $(\bar{x}, \bar{y}, \bar{t})$, $(\hat{x}, \hat{y}, \hat{t})$ to find conv(X_1) in the original space of variables, which can be done directly from the non-convex formulation above.

From constraints (14) and (17) we see $\lambda = x$, from constraint (15) $\hat{y}_1 = \frac{y_1}{x}$, from (18) $y_1 \le x$, from (16) we find $\bar{y}_2 = \frac{y_2 - x\hat{y}_2}{1-x}$, and from (19) we get $0 \le \hat{y}_2 \le 1$ and $0 \le \frac{y_2 - x\hat{y}_2}{1-x} \le 1$. Thus, (13)–(21) is feasible if and only if $0 \le y_1 \le x$, $0 \le y_2 \le 1$ and there exists $\hat{y}_2$ such that

$$t \ge \frac{\left(y_2 - x\hat{y}_2\right)^2}{1-x} + \frac{\left(x\hat{y}_2 - y_1\right)^2}{x},\quad 0 \le \hat{y}_2 \le 1,\quad \frac{y_2}{x} - \frac{1-x}{x} \le \hat{y}_2 \le \frac{y_2}{x}.$$

The existence of such $\hat{y}_2$ can be checked by solving the convex optimization problem

$$\text{(M1)} \qquad \min\ \varphi(\hat{y}_2) := \frac{\left(y_2 - x\hat{y}_2\right)^2}{1-x} + \frac{\left(x\hat{y}_2 - y_1\right)^2}{x}$$

$$\text{s.t. } \max\left\{0, \frac{y_2}{x} - \frac{1-x}{x}\right\} \leq \hat{y}_2 \leq \min\left\{1, \frac{y_2}{x}\right\}.$$

The equation $\varphi'(\hat{y}_2) = 0$ yields

$$-\frac{(y_2 - x\hat{y}_2)}{1-x} + \frac{(x\hat{y}_2 - y_1)}{x} = 0$$
$$\Leftrightarrow \hat{y}_2 = y_2 + y_1\frac{1-x}{x} := \eta(x, y).$$

Let $\hat{y}_2^*$ be an optimal solution to (M1). Note that $\hat{y}_2^* > 0$ whenever $\eta(x, y) > 0$. Moreover, $\eta(x, y) \leq \frac{y_2}{x} - \frac{1-x}{x} \implies y_1 + 1 \leq y_2$, which can only happen if $y_1 = 0$ and $y_2 = 1$, in which case $\frac{y_2}{x} - \frac{1-x}{x} = 1$. Thus, we may assume that $\hat{y}_2^*$ is not equal to one of its lower bounds.

Now observe that $\frac{y_2}{x} \leq \eta(x, y) \Leftrightarrow y_2 \leq y_1$, in which case $\eta(x, y) \leq \frac{y_1}{x} \leq 1$. Additionally, if $1 \leq \eta(x, y)$, then $x \leq y_2$ and in particular $y_1 \leq y_2$. Therefore, the cases $\eta(x, y) \leq \min\{1, \frac{y_2}{x}\}$, $\eta(x, y) \geq 1$, and $\eta(x, y) \geq \frac{y_2}{x}$ are mutually exclusive if $\frac{y_2}{x} \neq x$, and the optimal solution of (M1) corresponds to setting $\hat{y}_2^* = \eta(x, y)$, $\hat{y}_2^* = 1$, or $\hat{y}_2^* = \frac{y_2}{x}$, respectively. By calculating the objective function of (M1) with the appropriate value of $\hat{y}_2^*$, we find $\varphi(\hat{y}_2^*) = g_1(x, y_1, y_2)$. Hence, $(x, y, t) \in \text{conv}(X_1)$ if and only if $t \geq g_1(x, y_1, y_2)$ and $0 \leq y_1 \leq x \leq 1$, $0 \leq y_2 \leq 1$. □

4.2 Convex hull description of X

We use a similar argument as in the proof of Proposition 6 to prove Theorem 2. Let (x, y, t) be a point such that $0 \leq y_i \leq x_i \leq 1$ and *we additionally assume that* $y_1 \geq y_2$. A point (x, y, t) belongs to $\text{conv}(X)$ if and only if there exists $(\bar{x}, \bar{y}, \bar{t})$, $(\hat{x}, \hat{y}, \hat{t})$, and $0 \leq \lambda \leq 1$ such that

$$t = (1-\lambda)\bar{t} + \lambda\hat{t} \tag{22}$$
$$x_1 = (1-\lambda)\bar{x}_1 + \lambda\hat{x}_1 \tag{23}$$
$$x_2 = (1-\lambda)\bar{x}_2 + \lambda\hat{x}_2 \tag{24}$$
$$y_1 = (1-\lambda)\bar{y}_1 + \lambda\hat{y}_1 \tag{25}$$
$$y_2 = (1-\lambda)\bar{y}_2 + \lambda\hat{y}_2 \tag{26}$$
$$\bar{x}_2 = 0, \ \hat{x}_2 = 1 \tag{27}$$
$$\bar{y}_2 = 0, \ 0 \leq \hat{y}_2 \leq 1 \tag{28}$$
$$0 \leq \bar{y}_1 \leq \bar{x}_1 \leq 1, \ 0 \leq \hat{y}_1 \leq \hat{x}_1 \leq 1 \tag{29}$$
$$\bar{t} \geq \bar{y}_1^2/\bar{x}_1 \tag{30}$$
$$\hat{t} \geq g_1(\hat{x}_1, \hat{y}_1, \hat{y}_2). \tag{31}$$

The system (22)–(31) corresponds to $\text{conv}(K_0 \cup K_1)$, where $K_0 = \{(x, y, t) \in [0, 1]^2 \times \mathbb{R}_+^2 \times \mathbb{R} : y_1^2/x_1 \leq t, \ y_2 = x_2 = 0\}$ and $K_1 = \{(x, y, t) \in [0, 1]^2 \times \mathbb{R}_+^2 \times \mathbb{R} :$

$g_1(x_1, y_1, y_2) \le t,\ x_2 = 1\}$. Observe that K_0 and K_1 are the convex hulls of the restrictions of X, where $x_2 = 0$ and $x_2 = 1$, respectively.

Using a similar reasoning as in the proof of Proposition 6, we find $\lambda = x_2$, $\hat{y}_2 = \frac{y_2}{x_2}$, $\bar{x}_1 = \frac{x_1 - x_2\hat{x}_1}{1-x_2}$, $\bar{y}_1 = \frac{y_1 - x_2\hat{y}_1}{1-x_2}$, and

$$
\begin{aligned}
\text{(M2)} \qquad t \ge \min_{\hat{x}_1, \hat{y}_1} \ & \psi(\hat{x}_1, \hat{y}_1) \\
\text{s.t. } & 0 \le \hat{y}_1 \le \hat{x}_1 \le 1 && (32) \\
& \hat{y}_1 \le \frac{y_1}{x_2},\ \hat{x}_1 - \hat{y}_1 \le \frac{x_1 - y_1}{x_2},\ \frac{x_1}{x_2} - \frac{1 - x_2}{x_2} \le \hat{x}_1, && (33)
\end{aligned}
$$

where

$$\psi(\hat{x}_1, \hat{y}_1) := \frac{\left(y_1 - x_2\hat{y}_1\right)^2}{x_1 - x_2\hat{x}_1} + x_2 g_1(\hat{x}_1, \hat{y}_1, y_2/x_2)\cdot$$

Thus, to find the convex hull of X, we need to compute in closed form the solutions of the optimization problem (M2).

Lemma 1 *There exists an optimal solution $(\hat{x}_1^*, \hat{y}_1^*)$ to (M2) such that $\hat{y}_1^* \ge \frac{y_2}{x_2}$.*

Proof Note that if $\hat{y}_1 < \frac{y_2}{x_2}$, the function ψ is non-increasing in $\hat{y}_1$ for any value of $\hat{x}_1$. Thus there exists an optimal solution where $\hat{y}_1$ is set to one of its upper bounds, i.e., either $\hat{y}_1^* = y_1/x_2$ or $\hat{y}_1^* = \hat{x}_1^*$. Since we assume $y_1 \ge y_2$ and $\hat{y}_1 < y_2/x_2$, the case $\hat{y}_1^* = y_1/x_2$ is not possible.

Now suppose that $\hat{y}_1 = \hat{x}_1$. Then observe that $1 \le \frac{y_2}{x_2} + \hat{y}_1 \frac{1-\hat{x}_1}{\hat{x}_1} \Leftrightarrow \hat{x}_1 \le \frac{y_2}{x_2}$. Thus

$$\psi(\hat{x}_1) = \frac{\left(y_1 - x_2\hat{x}_1\right)^2}{x_1 - x_2\hat{x}_1} + \frac{\left(y_2 - x_2\hat{x}_1\right)^2}{x_2 - x_2\hat{x}_1}$$

in this case (substituting $\hat{y}_1 = \hat{x}_1$). Taking the derivative, we find

$$
\begin{aligned}
\psi'(\hat{x}_1) = {} & x_2 \frac{y_1 - x_2\hat{x}_1}{(x_1 - x_2\hat{x}_1)^2}\left(-2x_1 + x_2\hat{x}_1 + y_1\right) \\
& + x_2 \frac{(y_2 - x_2\hat{x}_1)}{(x_2 - x_2\hat{x}_1)^2}\left(-2x_2 + x_2\hat{x}_1 + y_2\right)\cdot
\end{aligned}
$$

Note that $y_1 - x_2\hat{x}_1 \ge 0$ since $\hat{x}_1 = \hat{y}_1 \le y_1/x_2$ in any feasible solution, and $y_2 - x_2\hat{x}_1 \ge 0$, by assumption. Additionally

- since $y_1 \le x_1$ and $\hat{x}_1 = \hat{y}_1 \le y_1/x_2 \le x_1/x_2$, we find that $-2x_1 + x_2\hat{x}_1 + y_1 \le 0$,
- since $y_2 \le x_2$ and $\hat{x}_1 \le 1$, we find that $-2x_2 + x_2\hat{x}_1 + y_2 \le 0$.

Therefore, $\psi'(x_1)$ is non-positive, i.e., ψ is non-increasing. Then, increasing $\hat{y}_1 = \hat{x}_1$ another optimal solution can be found. In particular, an optimal solution with $\hat{y}_1^* \ge y_2/x_2$ exits. □

From Lemma 1 we can assume, without loss of generality, that

$$\psi(\hat{x}_1, \hat{y}_1) = \frac{(y_1 - x_2\hat{y}_1)^2}{x_1 - x_2\hat{x}_1} + \frac{(x_2\hat{y}_1 - y_2)^2}{x_2\hat{x}_1}. \tag{34}$$

Taking partial derivatives, we find that

$$\begin{aligned}
\frac{\partial \psi}{\partial \hat{y}_1}(\hat{x}_1, \hat{y}_1) &= 2x_2\left(-\frac{y_1 - x_2\hat{y}_1}{x_1 - x_2\hat{x}_1} + \frac{x_2\hat{y}_1 - y_2}{x_2\hat{x}_1}\right),\\
\frac{\partial \psi}{\partial \hat{x}_1}(\hat{x}_1, \hat{y}_1) &= x_2\left(\frac{y_1 - x_2\hat{y}_1}{x_1 - x_2\hat{x}_1}\right)^2 - x_2\left(\frac{x_2\hat{y}_1 - y_2}{x_2\hat{x}_1}\right)^2.
\end{aligned}$$

Lemmas 2–4 characterize the optimal solutions of (M2), depending on the values of (x, y). Note that if

$$\hat{y}_1 = \frac{y_2}{x_2} + \frac{\hat{x}_1}{x_1}(y_1 - y_2), \tag{35}$$

then $\frac{\partial \psi}{\partial \hat{y}_1}(\hat{x}_1, \hat{y}_1) = \frac{\partial \psi}{\partial \hat{x}_1}(\hat{x}_1, \hat{y}_1) = 0$, independently of the values of $\hat{x}_1$ and $\hat{y}_1$. Thus, any feasible point that satisfies (35) is an optimal solution of (M2), as is the case for Lemmas 2 and 3. In contrast, under the conditions of Lemma 4, no feasible point satisfies (35) as it would violate upper bound constraints.

Lemma 2 *If* $x_1 \le x_2$ *then* $\hat{x}_1^* = \frac{x_1 - \epsilon}{x_2}$, *where* $\epsilon > 0$ *is a sufficiently small number, and* $\hat{y}_1^* = \frac{y_2}{x_2} + \frac{\hat{x}_1^*}{x_1}(y_1 - y_2)$ *is an optimal solution to* $(M2)$ *with objective* $\psi(\hat{x}_2^*, \hat{y}_2^*) = \frac{(y_1 - y_2)^2}{x_1}$.

Proof We have $\frac{\partial \psi}{\partial \hat{y}_1}(\hat{x}_1^*, \hat{y}_1^*) = \frac{\partial \psi}{\partial \hat{x}_1}(\hat{x}_1^*, \hat{y}_1^*) = 0$ and (x_1^*, y_1^*) satisfies all constraints (32)–(33). Thus, (x_1^*, y_1^*) is a KKT point and, by convexity, is an optimal solution. Substituting in (34), we get the result. □

Lemma 3 *If* $x_1 > x_2$ *and* $y_2(x_1 - x_2) + y_1 x_2 \le x_2 x_1$, *then* $\hat{x}_1^* = 1$ *and* $\hat{y}_1^* = \frac{y_2}{x_2} + \frac{\hat{x}_1^*}{x_1}(y_1 - y_2)$ *is an optimal solution to* $(M2)$ *with objective* $\psi(\hat{x}_2^*, \hat{y}_2^*) = \frac{(y_1 - y_2)^2}{x_1}$.

Proof Observe that $(\hat{x}_1^*, \hat{y}_1^*)$ is feasible as $\hat{y}_1^* = \frac{y_2}{x_2} + \frac{y_1 - y_2}{x_1} \le \frac{y_2}{x_2} + \frac{y_1 - y_2}{x_2} = \frac{y_1}{x_2}$; $\hat{y}_1^* = \frac{y_2}{x_2} + \frac{y_1 - y_2}{x_1} = \frac{y_2 x_1 + y_1 x_2 - y_2 x_2}{x_1 x_2} \le 1 = \hat{x}_1^*$; $\hat{x}_1^* - \hat{y}_1^* = 1 - \frac{y_2}{x_2} - \frac{y_1 - y_2}{x_1} \le 1 - \frac{y_2}{x_1} - \frac{y_1 - y_2}{x_1} = \frac{x_1 - y_1}{x_1} \le \frac{x_1 - y_1}{x_2}$; $\frac{x_1}{x_2} - \frac{1 - x_2}{x_2} = \frac{x_1 - 1}{x_2} + 1 \le 1 = \hat{x}_1^*$. Additionally, note that $\frac{\partial \psi}{\partial \hat{y}_1}(\hat{x}_1^*, \hat{y}_1^*) = \frac{\partial \psi}{\partial \hat{x}_1}(\hat{x}_1^*, \hat{y}_1^*) = 0$. Thus, (x_1^*, y_1^*) is a KKT point and, by convexity, is an optimal solution. Substituting in (34), we find the result. □

Lemma 4 *If* $x_1 > x_2$ *and* $y_2(x_1 - x_2) + y_1 x_2 \ge x_2 x_1$, *then* $\hat{x}_1^* = 1$ *and* $\hat{y}_1^* = 1$ *is an optimal solution to* $(M2)$ *with objective* $\psi(\hat{x}_2^*, \hat{y}_2^*) = \frac{(y_1 - x_2)^2}{x_1 - x_2} + \frac{(x_2 - y_2)^2}{x_2}$.

Proof Note that since $x_2 \ge y_2$ and $y_2(x_1 - x_2) + y_1 x_2 \ge x_2 x_1$, we have $x_2(x_1 - x_2) + y_1 x_2 \ge x_2 x_1 \Leftrightarrow y_1 \ge x_2$ and, in particular, $\hat{y}_1^* \le \frac{y_1}{x_2}$. Additionally, it is easily checked that all other constraints (32)–(33) are satisfied. From $y_2(x_1 - x_2) + y_1 x_2 \geq x_2 x_1$ we find that $\frac{x_2-y_2}{x_2} \leq \frac{y_1-x_2}{x_1-x_2}$. Now let μ_1 and μ_2 be the dual variables associated with constraints $\hat{y}_1 \leq \hat{x}_1$ and $\hat{x}_1 \leq 1$, respectively. Since both constraints are satisfied at equality at $(\hat{x}_1^*, \hat{y}_1^*)$, then we see that the dual variables μ_1 and μ_2 may take positive values without violating complementary slackness. In particular, let $\mu_1^* = 2x_2\left(\frac{y_1-x_2}{x_1-x_2} - \frac{x_2-y_2}{x_2}\right) \geq 0$ and $\mu_2^* = x_2\left(\frac{y_1-x_2}{x_1-x_2} - \frac{x_2-y_2}{x_2}\right)\left(\frac{x_1-y_1}{x_1-x_2} + \frac{y_2}{x_2}\right) \geq 0$. Then, $\frac{\partial\psi}{\partial\hat{y}_1}(\hat{x}_1^*, \hat{y}_1^*) = \mu_1^*$ and $\frac{\partial\psi}{\partial\hat{x}_1}(\hat{x}_1^*, \hat{y}_1^*) = -\mu_1^* + \mu_2^*$. Thus $(\hat{x}_1^*, \hat{y}_1^*)$ corresponds to a KKT point and, by convexity, is optimal. Substituting in (34) gives the result. □

Note that Lemmas 2, 3 and 4 cover all cases with $y_1 \geq y_2$. We can now prove the main result.

Proof (Theorem 2) If $y_1 \geq y_2$, the description of the convex hull follows directly from Lemmas 2, 3 and 4. If $y_1 \leq y_2$, the result follows from symmetry. □

4.3 Valid inequalities for S

Similar to the discussion in Sect. 3.3, the description of $\text{conv}(X)$ can be used to derive strong extended convex relaxations for S. In order to obtain (nonlinear) inequalities in the original space of variables, we project out the auxiliary variables for a given ordering $y_1 \geq \cdots \geq y_n$ of the continuous variables with additional restrictions corresponding to conditions $x_j(x_i - y_i) \leq y_j(x_i - x_j)$ in (12). Finally, to obtain linear inequalities valid independent of the conditions, we derive the first order approximations.

Suppose $y_1 \geq \cdots \geq y_n$, and $x_j(x_i - y_i) \leq y_j(x_i - x_j)$ for $j > i$, which holds, in particular, if $x = y$. By eliminating the auxiliary variables under these conditions we obtain the inequality

$$\phi(x, y) = \sum_{i\in\bar{P}} \bar{Q}_i y_i + \sum_{i\in P} \bar{Q}_i y_i^2/x_i - \sum_{i=1}^{n}\sum_{j=i+1}^{n} Q_{ij}\left(\frac{(y_1 - x_2)^2}{x_1 - x_2} + \frac{(x_2 - y_2)^2}{x_2}\right) \leq t. \tag{36}$$

Inequality (36) is only valid for the particular permutation of the continuous variables and when conditions $x_j(x_i - y_i) \leq y_j(x_i - x_j)$ for $j > i$ hold. Since $\sum_{i\in\bar{P}} \bar{Q}_i\bar{y}_i + \sum_{i\in P} \bar{Q}_i\bar{y}_i^2/\bar{x}_i - \sum_{i=1}^{n}\sum_{j=i+1}^{n} Q_{ij} g(\bar{x}_i, \bar{x}_j, \bar{y}_i, \bar{y}_j) = \phi(\bar{x}, \bar{y})$, we can find valid subgradient inequalities by taking gradients of the left-hand-side of (36). Let $\pi_i = Q_{ii} + 2\sum_{j=i}^{i-1} Q_{ij}$ and $\alpha_i = 2\sum_{j=1}^{i} Q_{ij}$, and recall $\bar{Q}_i = \sum_{j=1}^{n} Q_{ij}$. The partial derivatives of ϕ evaluated at a point $(\bar{x}, \bar{y})$ where $\bar{x} = \bar{y}$ are as follows:

$$\frac{\partial\phi}{\partial x_i}(\bar{x}, \bar{y}) = \sum_{j=i+1}^{n} Q_{ij} + \sum_{j=i+1}^{i-1} Q_{ij} - \bar{Q}_i = -Q_{ii} = \pi - \alpha_i, \qquad i \in P$$

$$\frac{\partial\phi}{\partial x_i}(\bar{x}, \bar{y}) = \sum_{j=i+1}^{n} Q_{ij} + \sum_{j=i+1}^{i-1} Q_{ij} = \pi - \alpha_i + \bar{Q}_i, \qquad i \in \bar{P}$$

$$\frac{\partial \phi}{\partial y_i}(\bar{x},\bar{y}) = -2\sum_{j=i+1}^{n} Q_{ij} + 2\bar{Q}_i = -\alpha_i, \qquad i \in P$$

$$\frac{\partial \phi}{\partial y_i}(\bar{x},\bar{y}) = -2\sum_{j=i+1}^{n} Q_{ij} + \bar{Q}_i = -\alpha_i - \bar{Q}_i, \qquad i \in \bar{P}.$$

Thus, since $\phi(\bar{x},\bar{y}) + \nabla\phi(\bar{x},\bar{y})(x-\bar{x}, y-\bar{y}) \le g(x,y) \le t$, we obtain the linear inequality

$$\sum_{i=1}^{n} \pi_i x_i \le t + \sum_{i=1}^{n} \alpha_i (x_i - y_i) - \sum_{i \in \bar{P}} \bar{Q}_i (x_i - y_i). \tag{37}$$

Observe that inequality (37) depends only on the ordering of $\bar{x}$, but not on the actual values.

Remark 3 Consider the submodular function given by $q(x) = x'Qx$. The extreme points of the extended polymatroid [20] associated with q, Π, correspond to the vectors π in inequality (37); thus, the convex lower envelope of q is described by the function $\bar{q}(x) = \max_{\pi\in\Pi} \pi' x$ [34]. Atamtürk and Bhardwaj [4] employ these polymatroid inequalities for the binary case. For the mixed-integer case, the inequality (37) is tight for the binary restriction $x = y$, and the right hand side is relaxed as the distance between x and y increases.

Remark 4 The values α_i in inequality (37) corresponds to the value of derivative of $q(x)$ with respect to x_i when $x_j = 1$ for all $j \le i$ and $x_j = 0$ for $j > i$. Atamtürk and Jeon [6] use lifting to derive similar inequalities for another class of nonlinear functions with indicator variables and submodular binary restriction.

5 Valid conic quadratic inequalities for X

The inequalities $f(x,y) \le t$ and $g(x,y) \le t$ derived in Sects. 3 and 4 for X_U and X, respectively, cannot be directly used within off-the-shelf solvers in the original space of variables as they are piecewise functions. However, since they are convex, they can be implemented using gradient outer-approximations at differentiable points (as discussed in Sects. 3.3 and 4.3): given a fractional point $(\bar{x},\bar{y})$ with $\bar{x} > 0$ and a subgradient $\xi \in \partial g(\bar{x},\bar{y})$, the inequality

$$g(\bar{x},\bar{y}) + \xi'(x-\bar{x}, y-\bar{y}) \le t \tag{38}$$

can be used as a cutting plane to improve the continuous relaxation. However, such an approach may require adding too many inequalities (38) to the formulation, possibly resulting in poor performance (see also Sects. 7.1 and 7.3 for additional discussion on computations). Alternatively, an extended formulation could be used [e.g., [17,22]]; however, such formulations may require a prohibitively large number of variables, resulting in hard-to-solve convex formulations and poor performance in branch-and-bound algorithms. Therefore, in this section we give valid conic quadratic inequalities that provide a strong approximation of $\text{conv}(X)$ and can be readily used within conic quadratic solvers.

5.1 Derivation of the inequalities

Let $L_2 = \{(x, y, t) \in X : x_2 = 0\}$ and observe that

$$\text{conv}(L_2) = \left\{(x, y, t) \in [0, 1]^2 \times \mathbb{R}^2_+ \times \mathbb{R} : \frac{y_1^2}{x_1} \le t,\ \ y_1 \le x_1,\ \ x_2 = y_2 = 0\right\}.$$

We now consider inequalities obtained by lifting the valid inequality $\frac{y_1^2}{x_1} \le t$ for $\text{conv}(L_2)$, i.e., inequalities of the form

$$\frac{y_1^2}{x_1} + h(x_2, y_2) \le t \tag{39}$$

for X, where $h : [0, 1] \times \mathbb{R}_+ \to \mathbb{R}$. We additionally require the left hand side of (39) to be convex, which is the case if and only if h is convex.

Proposition 7 *Inequality*

$$\frac{y_1^2}{x_1} + \frac{y_2^2}{x_2} - 2y_2 \le t \tag{40}$$

is valid for X and is the strongest convex inequality of the form (39).

Proof Any valid inequality of the form (39) needs to satisfy

$$h(x_2, y_2) \le \alpha = \min \left\{(y_1 - y_2)^2 - \frac{y_1^2}{x_1}\ :\ 0 \le y_1 \le x_1,\ \ x_1 \in \{0, 1\}\right\}.$$

If $x_1 = 0$, then $\alpha = y_2^2$; else, $\alpha = -2y_1y_2 + y_2^2$. Thus, $y_1 = x_1 = 1$ is a minimizer. We also find that $h(x_2, y_2) \le y_2^2 - 2y_2$ for $x_2 \in \{0, 1\}$. To find the strongest convex inequality, we compute $\text{conv}(W)$, where $W = \left\{(x_2, y_2, t_2) \in \{0, 1\} \times \mathbb{R}^2_+ : y_2^2 - 2y_2 \le t_2,\ \ y_2 \le x_2\right\}$. Using the perspective reformulation, one sees that

$$\text{conv}(W) = \left\{(x_2, y_2, t_2) \in [0, 1] \times \mathbb{R}^2_+ : \frac{y_2^2}{x_2} - 2y_2 \le t_2,\ \ y_2 \le x_2\right\},$$

and we get inequality (40). □

By changing the lifting order, we also get that valid inequality $\frac{y_1^2}{x_1} + \frac{y_2^2}{x_2} - 2y_1 \le t$, or, writing the inequalities more compactly, we arrive at the convex valid inequality

$$\frac{y_1^2}{x_1} + \frac{y_2^2}{x_2} - 2\min\{y_1, y_2\} \le t. \tag{41}$$

Remark 5 Observe that inequality (41) dominates inequality (4) since

$$\frac{y_1^2}{x_1} - x_2 = \frac{y_1^2}{x_1} - y_2 - (x_2 - y_2) \le \frac{y_1^2}{x_1} - y_2 - (x_2 - y_2)\frac{y_2}{x_2} = \frac{y_1^2}{x_1} + \frac{y_2^2}{x_2} - 2y_2.$$

Similarly, we find that (41) dominates inequality (3).

Remark 6 For the binary case, $y_i = x_i$, $i = 1, 2$, (41) reduces to $|x_1 - x_2| \le t$.

5.2 Strength of the inequalities

In order to assess the strength of inequality (41), we consider the optimization problem

$$\begin{aligned}
&\min\ a_1x_1 + a_2x_2 + b_1y_1 + b_2y_2 + t\\
&\text{s.t. } (y_1 - y_2)^2 \le t\\
\text{(SR)}\qquad &\frac{y_1^2}{x_1} + \frac{y_2^2}{x_2} - 2\min\{y_1, y_2\} \le t\\
&0 \le y_1 \le x_1 \le 1\\
&0 \le y_2 \le x_2 \le 1.
\end{aligned}$$

Inequalities (41) are not sufficient to guarantee the integrality of x in the optimal solutions of (SR) for all values of a and b, since they do not describe $\text{conv}(X)$ (given in Sect. 4). However, we now show that optimal solutions of (SR) are indeed integral under mild assumptions on the coefficients a and b. First, we prove an auxiliary lemma.

Lemma 5 *If there exists an optimal solution to (SR) with $y_i \in \{0, 1\}$ for some $i \in \{1, 2\}$, then there exists an optimal solution that is integral in x.*

Proof If $y_1 = 0$, then clearly there is an optimal solution with $x_1 \in \{0, 1\}$, depending on the sign of a_1. Moreover, (SR) reduces to $\min_{0\le y_2\le x_2\le 1}\left\{a_2x_2 + b_2y_2 + y_2^2/x_2\right\}$, which has an optimal integral solution in x_2. On the other hand, if $y_1 = x_1 = 1$, then (SR) reduces to $\min_{0\le y_2\le x_2\le 1}\left\{a_2x_2 + (b_2 - 2)y_2 + y_2^2/x_2\right\}$, which, again, has an optimal integral solution in x_2. The case with $y_2 \in \{0, 1\}$ is symmetric. □

Proposition 8 *If a_1, a_2 have the same sign and b_1, b_2 have the same sign, then (SR) has an optimal solution that is integral in x.*

Proof Note that if $a_1, a_2 \le 0$, then $x_1 = x_2 = 1$ for an optimal solution of (SR). Also, if $b_1, b_2 \ge 0$, then $y_1 = y_2 = 0$ in an optimal solution of (SR), in which case x is integral in extreme point solutions. It remains to show that if $a_1, a_2 \ge 0$ and $b_1, b_2 \le 0$, then there exists an optimal solution of (SR) that is integral in x.

Suppose that $y_1 = y_2 = y$ in an optimal solution. Then $(y_1 - y_2)^2 = 0$ and $\frac{y^2}{x_1} + \frac{y^2}{x_2} - 2y \le 0$. Thus, $t = 0$ and (SR) reduces to $\min\{a_1x_1 + a_2x_2 + (b_1 + b_2)y : 0 \le y \le \min\{x_1, x_2\} \le 1\}$, which has an optimal solution integral in x.

Now suppose, without loss of generality, there is an optimal solution with $1 > y_1 > y_2 > 0$ (if $y_1 = 1$ or $y_2 = 0$ then by Lemma 5 the solution is integral in x). Then observe that, in this case, the functions $(y_1 - y_2)^2$ and $y_2^2/x_2 - 2y_2$ are non-increasing in y_2. Since $b_2 \le 0$, there exists a solution where y_2 is at its upper bound, i.e., $y_2 = x_2$. Thus problem (SR) reduces to

$$(\mathrm{SR}') \quad \min\left\{a_1x_1 + b_1y_1 + (a_2+b_2)y_2 + t : (y_1-y_2)^2 \right.$$
$$\left. \le t, \frac{y_1^2}{x_1} - y_2 \le t, y_1 \le x_1 \le 1\right\}.$$

Let $(\lambda, \mu, \alpha, \beta)$ be the dual variables associated with the $\le$ constraints displayed in the order above and consider the dual feasibility conditions of problem (SR$'$)

$$\begin{aligned}
&- a_1 = -\mu_1 \frac{y_1^2}{x_1^2} - \alpha + \beta \\
&- b_1 = 2\lambda(y_1 - y_2) + 2\mu \frac{y_1}{x_1} + \alpha \\
&- (a_2 + b_2) = -2\lambda(y_1 - y_2) - \mu \\
&1 = \lambda + \mu \\
&0 \le \lambda, \mu, \alpha, \beta.
\end{aligned}$$

Let $(\bar{x_1}, \bar{y_1}, \bar{y_2}, \bar{t})$ be a KKT point with multipliers $(\bar{\lambda}, \bar{\mu}, \bar{\alpha}, \bar{\beta})$ and suppose that $\bar{x}_1 < 1$. Then observe that for small $\epsilon > 0$, $(\frac{\bar{y}_1+\epsilon}{\bar{y}_1}\bar{x}_1, \bar{y_1} + \epsilon, \bar{y_2} + \epsilon, \bar{t})$ is also a KKT point with the same multipliers. In particular, by choosing ϵ so that $1 = \frac{\bar{y}+\epsilon}{\bar{y}}\bar{x}$, we see that there is an optimal solution with $x_1 = 1$. Then, problem (SR$'$) further simplifies to

$$(\mathrm{SR}'') \quad \min\{b_1y_1 + (a_2 + b_2)y_2 + t : (y_1 - y_2)^2 \le t, y_1^2 - y_2 \le t\}\cdot$$

It remains to show that $y_2 = x_2$ is integral. Note that

$$y_1^2 - 2y_1y_2 + y_2^2 = y_1^2 - y_2(2y_1 - 1) \ge y_1^2 - y_2,$$

and, therefore, constraint $y_1^2 - y_2 \le t$ is not binding when $y_1 < 1$. So, (SR$''$) is equivalent to $\min b_1y_1 + (a_2+b_2)y_2 + (y_1-y_2)^2$. However, by increasing or decreasing y_1 and y_2 by the same amount it is easy to check that there exists an optimal solution where either $y_1 = 1$ or $y_2 = 0$, and from Lemma 5 there exists an optimal integral solution. $\square$

Proposition 8 provides some insight on the problems for which inequalities (41) may be particularly effective: if the coefficients of the binary variables and the continuous variables have the same sign, then the relaxation induced by (41) may be close to ideal; otherwise, using subgradient inequalities may be required to find strong formulations. In our computations, this simple rule of thumb indeed results in the best performance.

6 Extensions to other quadratic functions with two indicator variables

In this paper we focus on the set X, i.e., a mixed-integer set with non-negative continuous variables and non-positive off-diagonal entries in the quadratic matrix. Although an in-depth study of more general quadratic functions is outside the scope of this paper, the approach used in Sect. 5 can be naturally extended to other quadratic functions. We briefly discuss two such extensions.

6.1 General quadratic functions

Observe that a general quadratic function $y'Ay$ can be decomposed as

$$y'Ay = \sum_{i=1}^{n} \left(\left(A_{ii} - \sum_{j \neq i} |A_{ij}| \right) y_i^2 - \sum_{j>i:A_{ij}<0} A_{ij}(y_i - y_j)^2 \right.$$
$$\left. + \sum_{j>i:A_{ij}>0} A_{ij}(y_i + y_j)^2 \right).$$

Thus, stronger formulations for general quadratic functions may be obtained by studying the set with two continuous and two indicator variables and positive off-diagonal term

$$X_+ = \left\{ (x, y, t) \in \{0, 1\}^2 \times \mathbb{R}_+^2 \times \mathbb{R} : (y_1 + y_2)^2 \leq t,\ \ y_i \leq x_i,\ \ i = 1, 2 \right\}.$$

Proposition 9 *Inequality*

$$\frac{y_1^2}{x_1} + \frac{y_2^2}{x_2} \leq t \tag{42}$$

is valid for X_+ and is the strongest among inequalities of the form (39).

The proof is analogous the the proof of Proposition 7 as is omitted for brevity. Although inequality (42) is similar in spirit to (40), and that it is the strongest among inequalities of the form (39), it is not as strong as (40) for X. In particular, an integrality result similar to Proposition 8 does not hold for (42).

6.2 Quadratic functions with continuous variables unrestricted in sign

Consider the set

$$X_\pm = \left\{ (x, y, t) \in \{0, 1\}^2 \times \mathbb{R}^2 \times \mathbb{R} : (y_1 \pm y_2)^2 \leq t,\ \ -x_i \leq y_i \leq x_i \text{ for } i = 1, 2 \right\}.$$

Observe that, since the continuous variables can be positive or negative, the sign inside the quadratic expression does not matter (e.g., it can be flipped via the transformation $\bar{y}_2 = -y_2$). Thus we assume, without loss of generality, that it is a minus sign.

Proposition 10 *Inequality* (4)*, originally proposed by Jeon et al.* [29]*, is valid for* $X_\pm$ *and is the strongest among inequalities of the form* (39).

Proof Any valid inequality for $X_\pm$ of the form (39) needs to satisfy

$$h(x_2, y_2) \leq \alpha = \min \left\{ (y_1 - y_2)^2 - \frac{y_1^2}{x_1} \;:\; -x_1 \leq y_1 \leq x_1,\; x_1 \in \{0, 1\} \right\}.$$

If $x_1 = 0$, then $\alpha = y_2^2$. Else, $\alpha = -2y_1y_2 + y_2^2$; in this case, the minimum is attained at $y_1^* = 1$ if $y_2 \geq 0$ and at $y_1^* = -1$ otherwise. Thus, we find that $h(x_2, y_2) \leq y_2^2 - 2|y_2|$ for $x_2 \in \{0, 1\}$. To find the strongest convex inequality, we compute conv$(W_\pm)$, where $W_\pm = \left\{(y_2, x_2, t_2) \in \{0, 1\} \times \mathbb{R} \times \mathbb{R} : y_2^2 - 2|y_2| \leq t_2,\; -x_2 \leq y_2 \leq x_2\right\}$. The convex lower envelope corresponding to the one-dimensional non-convex function $h_1(y_2) = y_2^2 - 2|y_2|$ for $y_2 \in [-1, 1]$ is the constant function equal to -1. Moreover, it can be shown that

$$\text{conv}(W_\pm) = \left\{(y_2, x_2, t_2) \in [0, 1] \times \mathbb{R}_+^2 : -x_2 \leq t_2,\; -x_2 \leq y_2 \leq x_2\right\}$$

and we get the convex valid inequality $\frac{y_1^2}{x_1} - x_2 \leq t$ for $X_\pm$. □

In light of Proposition 10, inequalities (40)-(41) can be interpreted as inequalities that additionally account for the non-negativity of the continuous variables, with respect to the valid inequalities proposed by Jeon et al. [29]. Moreover, although not explicitly considered by Jeon et al., their inequalities may be particularly effective for quadratic optimization problems with indicator variables and continuous variables unrestricted in sign. Observe that inequalities (3)–(5) are indeed valid even if the variables are not required to be non-negative – in contrast with the inequalities $f(x, y) \leq t$, $g(x, y) \leq t$ and (41), which account for the non-negativity of the variables and are only valid in that case.

7 Computations

In this section we report a summary of computational experiments performed to test the effectiveness of the proposed inequalities in a branch-and-bound algorithm. All experiments are conducted using Gurobi 7.5 solver on a workstation with a 3.60GHz Intel® Xeon® E5-1650 CPU and 32 GB main memory with a single thread. The time limit is set to one hour and Gurobi's default settings are used (except for the parameter "PreCrush", which is set to 1 in order to use cuts). Cuts (if used) are added only at the root node using the callback features of Gurobi, and the reported times include the time used to add cuts.

7.1 Image segmentation with ℓ-0 penalty

Given a finite set N, functions $d_i : \mathbb{R} \to \mathbb{R}_+$ for $i \in N$ and $s_{ij} : \mathbb{R} \to \mathbb{R}_+$ for $i \neq j$, consider

$$(D) \qquad \min_{y \in Y} \sum_{i \in N} d_i(y_i) + \sum_{i \neq j} s_{ij}(y_i - y_j),$$

where $Y \subseteq \mathbb{R}_+^N$. Problem (D) arises as the Markov Random Fields (MRF) problem for image segmentation, see [16,32]. In the MRF context, d_i are the *deviation* penalty functions, used to model the cost of changing the value of a pixel from the observed value p_i to y_i, e.g., $d_i(y_i) = c_i(p_i - y_i)^2$ with $c_i \in \mathbb{R}_+$; functions s_{ij} are the *separation* penalty functions, used to model the cost of having adjacent pixels with different values, e.g., $s_{ij}(y_i - y_j) = c_{ij}(y_i - y_j)^2$ with $c_{ij} > 0$ if pixels i and j are adjacent, and $s_{ij}(y_i - y_j) = 0$ otherwise. Often, $Y = [0, 1]^N$ or is given by a suitable discretization, i.e., y is a vector of integer multiples of a parameter ε. We consider in our computations the case $Y = [0, 1]^N$, but the proposed approach can be used with any Y.

Problem (D) can be cast as the nonlinear dual of the undirected minimum cost network flow problem [1] and efficient algorithms exist when all functions are convex [27]. In contrast, we consider here the case where the deviation functions involve a non-convex ℓ-0 penalty, which is often used to induce sparsity, e.g., restricting the number of pixels that can have a color different from the background color. In particular, $d_i(y_i) = a_i \|y_i\|_0 + \bar{d}_i(y_i)$ with $\bar{d}_i = c_i(p_i - y_i)^2$. Thus, the problem can be formulated as

$$\min \sum_{i \in N} a_i x_i + \sum_{i \in N} c_i(p_i - y_i)^2 + \sum_{i \neq j} c_{ij} t_{ij} \text{ s. t. } (x_i, x_j, y_i, y_j, t_{ij}) \in X, \ \forall i \neq j. \tag{43}$$

Instances The instances are constructed as follows. The elements of N correspond to points in a $k \times k$ grid, thus $n = k^2$, and separation functions s_{ij} are non-zero whenever the corresponding points are adjacent in the grid. The parameters p_i for $i \in N$, and c_{ij} for each pair of adjacent points $i, j \in N$ are drawn uniformly between 0 and 1. We set $a_i = c_i$, where c_i is generated as follows: first we draw $\tilde{c}_i$ uniformly between 0 and 1 for all $i \in N$, let $C_1 = \sum_{i \in N} \tilde{c}_i$ and $C_2 = \sum_{i: p_i \geq 0.5}(2p_i - 1)$; then we set $c_i = \tilde{c}_i \frac{C_1}{C_2}$. Instances generated with these parameters are observed to have large integrality gaps.

Formulations We test the following formulations for solving problem (43):

`Basic` The natural formulation

$$\min \sum_{i \in N} a_i x_i + \sum_{i \in N} c_i(p_i - y_i)^2 + \sum_{i \neq j} c_{ij}(y_i - y_j)^2 \text{ s.t. } 0 \leq y \leq x, \ x \in \{0, 1\}^N.$$

`Perspective` The perspective reformulation implemented with rotated cone constraints

$$\sum_{i\in N} c_i p_i^2 + \min \sum_{i\in N} a_i x_i + \sum_{i\in N} c_i\,(-2p_i y_i + z_i) + \sum_{i\neq j} c_{ij}(y_i - y_j)^2$$
$$\text{s.t. } y_i^2 \le z_i x_i,\ \ \forall i\in N$$
$$0 \le y \le x,\ z \ge 0,\ x\in\{0,1\}^N.$$

`Conic` The formulation with the conic quadratic inequalities (41)

$$\sum_{i\in N} c_i p_i^2 + \min \sum_{i\in N} a_i x_i + \sum_{i\in N} c_i\,(-2p_i y_i + z_i) + \sum_{i\neq j} c_{ij} t_{ij}$$
$$\text{s.t. } y_i^2 \le z_i x_i,\ \forall i\in N$$
$$(y_i - y_j)^2 \le t_{ij},\ z_i + z_j - 2y_i \le t_{ij},\ z_i + z_j - 2y_j \le t_{ij},\ \forall i\neq j$$
$$0 \le y \le x,\ z \ge 0,\ x\in\{0,1\}^N.$$

Furthermore, we also test models `Perspective+cuts` and `Conic+cuts`, where the subgradient inequalities (38) are used as cutting planes to strengthen the `Perspective` and `Conic` formulations, respectively. If $\bar{x}_i = 0$ for some $i \in N$ then we use the first-order expansion around $\bar{x}_i = 10^{-5}$ instead.

Results Table 1 shows a comparison of the performance of the algorithm for each formulation for varying grid sizes. Each row in the table represents the average for five instances for a grid size. Table 1 displays the initial gap (`igap`), the root gap improvement (`rimp`), the number of branch and bound nodes (`nodes`), the elapsed time in seconds (`time`), and the end gap at termination (`egap`) (in brackets, we report the number of instances solved to optimality within the time limit). The initial gap is computed as $\texttt{igap} = \frac{\texttt{obj}_{\texttt{best}} - \texttt{obj}_{\texttt{cont}}}{|\texttt{obj}_{\texttt{best}}|} \times 100$, where $\texttt{obj}_{\texttt{best}}$ is the objective value of the best feasible solution found and $\texttt{obj}_{\texttt{cont}}$ is the objective of the continuous relaxation of `Basic`. The root improvement is computed as $\texttt{rimp} = \frac{\texttt{obj}_{\texttt{relax}} - \texttt{obj}_{\texttt{cont}}}{\texttt{obj}_{\texttt{best}} - \texttt{obj}_{\texttt{cont}}} \times 100$, where $\texttt{obj}_{\texttt{relax}}$ is the objective value of the relaxation obtained after processing the first node of the branch-and-bound tree for a given formulation, obtained by querying Gurobi's attribute "ObjBound" at the root node using a callback.

We observe that the `Basic` formulation requires a substantial amount of branching before proving optimality, resulting in long solution times. The `Perspective` formulation results in a root gap improvement close to 50% and better times and end gaps than the `Basic` formulation. However, even with the `Perspective` formulation, instances with $k \times k = 400$ and larger cannot be solved to optimality leaving end gaps 15.3% or more. In contrast, formulation `Conic` results in root gap improvements close to 100%, and the performance of the branch-and-bound algorithm is orders-of-magnitude better than with the `Basic` and `Perspective` formulations: instances with $k \times k = 400$ that are not close to being solved after one hour of computation with `Basic` and `Perspective` are solved to optimality in one second; while formulation `Basic` is able to solve in five minutes instances with 100 variables, formulation `Conic` is able to solve in the same amount of time formulations with 2, 500 variables, i.e., instances 250 times larger.

Formulation `Conic+cuts` results in very modest improvement in the strength of the continuous relaxation when compared with `Conic` (less than 0.3% additional

Table 1 Experiments with image segmentation with ℓ-0 penalty

$k \times k$	igap	Basic			Perspective				Perspective + cuts				Conic				Conic+cuts			
		Nodes	Time	Egap	Rimp	Nodes	Time	Egap	Rimp	Nodes	Time	Egap	Rimp	Nodes	Time	Egap	Rimp	Nodes	Time	Egap
100	51.0	2,065,285	301	0.0[5]	47.9	70,898	17	0.0[5]	99.6	27,006	601	0.0[5]	99.4	7	0	0.0[5]	99.7	7	0	0.0[5]
400	47.7	9,520,774	3,600	34.0[0]	48.6	5,277,876	3600	15.3[0]	93.2	305	2	0.0[5]	99.5	59	1	0.0[5]	99.5	58	1	0.0[5]
2,500	47.9	1,091,872	3,600	46.3[0]	45.6	682,406	3600	25.6[0]	47.2	38,989	2235	9.9[2]	99.3	17,561	393	0.0[5]	99.6	9220	210	0.0[5]
10,000	47.4	167,529	3,600	47.2[0]	45.9	131,986	3600	25.9[0]	32.4	25,992	3600	0.2[0]	99.5	25,842	3600	0.1[0]	99.6	26,695	3600	0.1[0]

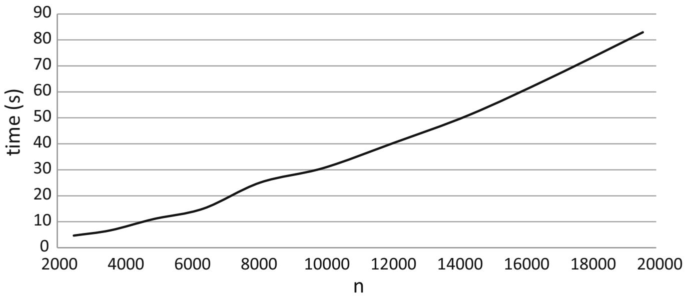

Fig. 1 Time to prove an optimality gap of 1% with `Conic` as a function of the dimension $n = k \times k$

root gap improvement) and almost no difference in terms of nodes, times or end gaps. Observe that in (43) the coefficients of the linear objective terms corresponding to the discrete and continuous variables have the same sign, and the experimental results are consistent with Proposition 8—`Conic` indeed is a very close approximation of inequalities (38) in this case.

Note that if cuts are added without the approximation given by inequalities (41) (formulation `Perspective+cuts`), the root improvement is substantial for small instances but it degrades as the size increases. We conjecture that the required number of cuts to obtain an adequate relaxation increases with the size of the instances. Thus, for larger instances, Gurobi may stop adding cuts before obtaining a strong relaxation. Additionally, to solve second-order conic subproblems in branch-and-bound, solvers like Gurobi construct a linear outer approximation of the convex sets; adding a large number of cuts may interfere with the construction of the outer approximation, leading to weak relaxations of the convex set, which is observed for instances with $k \times k = 10,000$. Using the approximation of the convex hull derived in Sect. 5 as a starting point appears to circumvent such numerical difficulties.

Finally, we remark that for the larger instances that are not solved to optimality by `Conic`, high quality solutions and tight lower bounds are found within a few seconds, but branching is ineffective to close the remaining gap. To illustrate, Figure 1 presents the time to prove an optimality gap of at most 1%, as a function of the dimension n of the problem. We see that the proposed approach scales very well (almost linearly) up to $n = 20{,}000$. In particular, the lower bound found corresponds to the one obtained at the root node, and the feasible solutions are found within a small number (50–60) of branch-and-bound nodes. Memory limit is reached for instances with $n > 20{,}000$.

7.2 Portfolio optimization with transaction costs

Consider a simple portfolio optimization problem with transaction costs similar to the one discussed in [18, p.146]. However, in our case, transactions have a fixed cost and there is a restricted number of transactions. For simplicity, we first consider assets with uncorrelated returns. In this context, an M-matrix arises directly due to the buying and selling decisions. In Sect. 7.3 we present computations with a general covariance matrix, from which an M-matrix corresponding to the negatively correlated assets can be extracted to apply the reformulations.

Let N be the set of assets, $\mu, \sigma \in \mathbb{R}^N_+$ be the vectors of expected returns and standard deviations of returns. Let $w \in \mathbb{R}^N_+$ denote the current holdings in each asset, let $a^+, a^- \in \mathbb{R}^N_+$ be the fixed transaction costs associated with buying and selling any quantity, $c^+, c^- \in \mathbb{R}^N$ be the variable transaction costs and profits of buying and selling each asset, let $u^+, u^- \in \mathbb{R}^N_+$ be the upper bounds on the transactions, and let k be the maximum number of transactions. Then the problem of finding a minimum risk portfolio that satisfies a given expected return $b \in \mathbb{R}$ with at most k transactions can be formulated as the mixed-integer quadratic problem:

$$\begin{aligned}
\min\ & v(y) = \sum_{i \in N} \sigma_i^2 (w_i + y_i^+ - y_i^-)^2 \\
\text{s.t. } & \sum_{i \in N} \left(\mu_i w_i + y_i^+ (\mu_i - c_i^+) - y_i^- (\mu_i - c_i^-) - a_i^+ x_i^+ - a_i^- x_i^- \right) \geq b \\
& \sum_{i \in N} (x_i^+ + x_i^-) \leq k \\
& 0 \leq y_i^+ \leq u_i^+ x_i^+, \ 0 \leq y_i^- \leq u_i^- x_i^-, \ x_i^+ + x_i^- \leq 1, \quad \forall i \in N \\
& (x^+, x^-, y^+, y^-) \in \{0,1\}^N \times \{0,1\}^N \times \mathbb{R}^N_+ \times \mathbb{R}^N_+,
\end{aligned}$$

where $v(y)$ is the variance of the new portfolio, the decision variables y_i^+ (y_i^-) indicate the amount bought (sold) in asset i and the variables x_i^+ (x_i^-) indicate whether asset i is bought (sold). Note that the quadratic objective function is nonseparable and the corresponding quadratic matrix is positive semi-definite but not positive definite; therefore, the classical perspective reformulation cannot be used. Additionally, observe that the portfolio optimization problem can be reformulated by adding continuous variables $t \in \mathbb{R}^N_+$, constraints $(x_i^+, x_i^-, y_i^+, y_i^-, t_i) \in X$ for all $i \in N$ to minimize the linear objective

$$\sum_{i \in N} \sigma_i^2 (2 w_i (y_i^+ - y_i^-) + t_i) \cdot \tag{44}$$

Note that since each continuous variable is involved in exactly one term in the objective, the extended formulation given by (44) and constraints $(x_i^+, x_i^-, y_i^+, y_i^-, t_i) \in \text{conv}(X)$ results in the convex envelope of $v(y)$.

Instances The instances are constructed as follows. We set $w_i = u_i^+ = u_i^- = 1$ for all $i \in N$. Coefficients σ_i are drawn uniformly between 0 and 1, μ_i are drawn uniformly between 0 and $2\sigma_i$, the transactions costs and profits c_i^+ and c_i^- are drawn uniformly between 0 and μ_i, the fixed costs a_i^+ and a_i^- are drawn uniformly between 0 and $(\mu_i - c_i^+)$ and $(\mu_i - c_i^-)$, respectively. The target return is set to $\beta \sum_{i \in N} \mu_i$ where $\beta > 0$ is a parameter; k is set to $n/10$.

Formulations We test the formulations `Basic`, `Basic+cuts`, `Conic`, and `Conic+cuts`, as defined in Sect. 7.1. As mentioned above, the perspective reformulation cannot be used for these instances.

Results Table 2 shows the results for varying number of assets n and values of the expected return β. Observe that instances with lower values of β are more difficult to solve for the `Basic` formulation: low β results in more feasible solutions, and more branch-and-bound nodes need to be explored before proving optimality. We also see that the `Basic` formulation is not effective for instances with 250 or more assets, where most instances (27 out of 30) are not solved to optimality within the time limit and leaving large end gaps at termination. On the other hand, the other three formulations achieve root improvements of over 90% in most cases, and lead to much lower solution times and end gaps.

Observe that for the portfolio problem, the coefficients of y_i^+ and y_i^- in the objective and return constraints have opposite signs. Thus, we expect the approximation given by `Conic` not to be as effective as in Sect. 7.1 and, therefore, the cuts to have a larger impact in closing the root gaps. Indeed, we see in these experiments that adding cuts leads to an additional 2% to 4% root improvement (compared to the 0.3% improvement observed in Sect. 7.1)[1]. In particular, formulation `Basic+cuts` is able to solve all instances in seconds, even instances with low values of β where all other formulations struggle.

7.3 General convex quadratic functions

The quadratic matrices used in the previous computations had specific structures, given by the applications considered. Although our results are for M-matrices, in this section, we test the strength of the formulations for more general problems, with dense matrices having positive and negative off-diagonal entries. To employ the results developed for M-matrices, we simply apply the strengthening on the pairs of variables with a negative off-diagonal entry. Toward this end, we consider the mean-variance portfolio optimization

$$
\begin{aligned}
(MV) \qquad \min\ & y'Ay \\
\text{s.t. } & b'y \geq r \\
& 1'x \leq k \\
& 0 \leq y \leq x \\
& x \in \{0,1\}^n.
\end{aligned}
$$

where the objective is to minimize the portfolio variance $y'Ay$, where A is a covariance matrix, subject to meeting a target return and satisfying sparsity constraints.

Instances In order test the effect of positive off-diagonal elements and diagonal dominance, the matrix A is constructed as follows: Let $\rho \geq 0$ be a parameter that controls the magnitude of the positive off-diagonal entries of A, and $\delta \geq 0$ be a parameter that controls the diagonal dominance of A. First, we construct a factor covariance matrix $F = GG'$, where each entry in $G_{20\times 20}$ in drawn uniformly from $[-1, 1]$, and factor

[1] The root gap improvements of 95% achieved by `Conic` indicate that the approximation given in Sect. 5 is strong and considerably better than the natural continuous relaxation.

Table 2 Experiments with portfolio optimization with fixed transaction costs

n	β	igap	Basic				Basic + cuts				Conic				Conic + cuts			
			Rimp	Nodes	Time	Egap	Rimp	Nodes	Time	Egap	Rimp	Nodes	Time	Egap	Rimp	Nodes	Time	Egap
100	0.95	30.8	0.0	39,963	11	0.0[5]	98.9	57	0	0.0[5]	86.9	822	4	0.0[5]	92.8	1,069	4	0.0[5]
	0.98	27.9	0.0	6,926	2	0.0[5]	93.2	130	1	0.0[5]	98.4	35	0	0.0[5]	94.4	167	1	0.0[5]
	1.00	32.7	0.0	3,229	1	0.0[5]	97.9	32	0	0.0[5]	96.9	49	0	0.0[5]	97.2	37	0	0.0[5]
Average			0.0	16,706	5	0.0[15]	96.7	76	0	0.0[15]	94.1	302	2	0.0[15]	94.8	425	2	0.0[15]
250	0.95	32.4	0.0	5,344,016	3600	15.0[0]	98.8	176	0	0.0[5]	94.0	175,859	2,880	1.7[1]	96.0	233,024	2880	1.1[1]
	0.98	26.0	0.0	4,831,484	3,227	6.2[1]	97.8	210	1	0.0[5]	99.1	27	0	0.0[5]	98.4	50,689	720	0.3[4]
	1.00	29.4	0.0	4,518,960	2970	4.0[1]	97.3	2061	49	0.0[5]	97.4	3597	38	0.0[5]	97.0	3,858	130	0.0[5]
Average			0.0	4,898,153	3,265	8.4[2]	98.0	816	17	0.0[15]	96.8	59,827	973	0.6[11]	97.2	95,857	1,243	0.5[10]
500	0.95	32.3	0.0	2,906,338	3,600	24.5[0]	97.6	387	2	0.0[5]	95.2	26,640	1441	0.6[3]	97.2	139,686	3600	0.9[0]
	0.98	26.1	0.0	3,096,026	3,600	16.4[0]	98.0	343	3	0.0[5]	96.4	295	2	0.0[5]	99.1	182	1	0.0[5]
	1.00	32.8	0.0	3,076,324	3600	18.8[0]	97.5	328	2	0.0[5]	93.4	330	2	0.0[5]	97.0	254	1	0.0[5]
Average			0.0	3,026,229	3600	19.9[0]	97.7	353	2	0.0[15]	95.0	9,088	481	0.2[13]	97.7	46,707	1201	0.3[10]

exposure matrix $X_{n\times 20}$ such that $X_{ij}=0$ with probability 0.8, and X_{ij} is drawn uniformly from [0, 1], otherwise. Then we construct an auxiliary matrix $\bar{A}=XFX'$. Then, for $i\neq j$, we set $A_{ij}=\bar{A}_{ij}$ if $\bar{A}_{ij}\leq 0$, and we set $A_{ij}=\rho\bar{A}_{ij}$ otherwise[2]. Finally, υ_i is drawn uniformly from $[0,\delta\bar{\sigma}]$, where $\bar{\sigma}=\frac{1}{n}\sum_{i\neq j}|A_{ij}|$, and $A_{ii}=\sum_{j\in N}|A_{ij}|+\upsilon_i$. Observe that the auxiliary matrix $\bar{A}$ represents a low-rank matrix obtained from a 20-factor model, and $\text{diag}(\upsilon)$ is a diagonal matrix representing the residual variances not explained by the factor model. The matrix A is obtained by scaling the positive off-diagonals of $\bar{A}$ by ρ, and updating the diagonal entries to ensure positive definiteness by imposing diagonal dominance. Additionally, b_i is drawn uniformly between $0.5U_{ii}$ and $1.5U_{ii}$. Finally, we let $r=0.25\times\sum_{i\in N}b_i$ and $k=n/5$ for "small" instances, and $r=0.125\times\sum_{i\in N}b_i$ and $k=n/10$ for "large" instances.

Formulations We test the same formulations as in Sect. 7.1. In this case, the diagonal matrix $\text{diag}(\upsilon)$ is used for the `Perspective` formulation. In particular, formulations `Perspective+cuts`, `Conic` and `Conic+cuts` are based on the decomposition of the objective function given by

$$\begin{aligned}
&\min \sum_{i\in N}\upsilon_i z_i+\sum_{A_{ij}<0}|A_{ij}|t_{ij}+y'(A-Q-\text{diag}(\upsilon))y\\
&\text{s.t. } y_i^2\leq z_i x_i,\ \forall i\in N,\quad (x_i,x_j,y_i,y_j,t_{ij})\in X,\ \forall i\neq j: A_{ij}<0,
\end{aligned}$$

where $Q_{ij}=\min\{0,A_{ij}\}$ for $i\neq j$ and $Q_{ii}=-\sum_{j\neq i}Q_{ij}$. By construction, $A-Q-diag(\upsilon)$ is positive semi-definite.

Results Table 3 presents the results for matrices with non-positive off diagonal entries (i.e., $\rho=0$) and varying diagonal dominance δ. Table 4 presents the results for matrices with fixed diagonal dominance and varying magnitudes for positive off-diagonal entries ρ. We see that, in all cases formulation `Conic` results in better root gap improvements than `Perspective` and `Basic`. The gap improvements depend on the parameters δ and ρ. In Table 3 we see that `Conic` formulation closes an additional 30% to 40% gap with respect to `Perspective` (independent of the diagonal dominance δ). In Table 4 we observe that, as expected, `Conic` formulation is more effective at closing root gaps when the magnitude ρ for the positive off-diagonal entries is small. Nevertheless, for all instances formulations `Conic` and `Conic+cuts` result in significantly stronger root improvements than `Perspective` (at least 15%, and often much more) and the number of nodes required to solve the instances is decreased by at least an order of magnitude.

Observe that the stronger formulations of `Conic` and `Conic+cuts` do not necessarily lead to better solution times for small instances. Nevertheless, for the larger instances ($n=100$), using the `Conic` formulation leads to faster solution times, lower end gaps and more instances solved to optimality for all values of δ and ρ. As in

[2] The matrices generated this way have only 20.1% of the off-diagonal entries negative on average – the rest are positive if $\rho>0$ and 0 if $\rho=0$. The ratio of the magnitude of the negative entries vs. the total, i.e., $\frac{\sum_{i\neq j:A_{ij}<0}|A_{ij}|}{\sum_{i\neq j}|A_{ij}|}$, is on average 0.72 if $\rho=0.1$, 0.57 if $\rho=0.2$ and 0.34 if $\rho=0.5$.

Table 3 Experiments with non-positive off diagonal entries and varying diagonal dominance, $k = n/5$

n	δ	igap	Basic			Perspective				Perspective+cuts				Conic				Conic + cuts			
			Nodes	Time	Egap	Rimp	Nodes	Time	Egap	Rimp	Nodes	Time	Egap	Rimp	Nodes	Time	Egap	Rimp	Nodes	Time	Egap
60	0.1	88.2	$4 \cdot 10^5$	86	0.0[5]	7.2	$4 \cdot 10^5$	99	0.0[5]	19.2	15,230	544	0.0[5]	43.6	3,704	107	0.0[5]	43.9	4,653	154	0.0[5]
	0.5	80.2	$5 \cdot 10^5$	103	0.0[5]	28.0	$2 \cdot 10^5$	47	0.0[5]	38.9	3,243	92	0.0[5]	66.1	1,783	44	0.0[5]	66.6	1,567	49	0.0[5]
	1.0	74.0	$6 \cdot 10^5$	121	0.0[5]	44.4	$6 \cdot 10^4$	18	0.0[5]	52.8	1,335	35	0.0[5]	81.5	863	14	0.0[5]	82.3	709	19	0.0[5]
Average			$5 \cdot 10^5$	103	0.0[15]	26.5	$2 \cdot 10^5$	55	0.0[15]	37.0	6,603	224	0.0[15]	63.7	2,117	55	0.0[15]	64.3	2,310	74	0.0[15]
80	0.1	90.3	$1 \cdot 10^7$	3600	9.7[0]	7.2	$9 \cdot 10^6$	3600	10.1[0]	4.0	31,194	3600	16.1[0]	37.0	26,657	2,758	5.7[2]	37.3	36,998	2,776	4.6[2]
	0.5	82.8	$1 \cdot 10^7$	3600	10.5[0]	28.2	$6 \cdot 10^6$	2902	2.8[3]	16.8	29,220	3,017	4.0[2]	60.2	11,367	1,108	0.0[5]	60.4	13,898	1,208	0.0[5]
	1.0	77.0	$1 \cdot 10^7$	3600	9.5[0]	44.1	$2 \cdot 10^6$	988	0.0[5]	27.2	4,889	566	0.0[5]	78.4	2,689	183	0.0[5]	79.0	3,395	233	0.0[5]
Average			$1 \cdot 10^7$	3600	9.9[0]	26.5	$5 \cdot 10^6$	2,496	4.3[8]	16.0	21,768	2,394	6.7[7]	58.5	13,571	1,350	1.9[12]	58.9	18,097	1,406	1.5[12]
100	0.1	90.2	$1 \cdot 10^7$	3600	30.0[0]	6.4	$6 \cdot 10^6$	3600	29.3[0]	2.8	14,855	3600	35.8[0]	37.1	19,660	3600	19.6[0]	37.0	17,047	3600	21.6[2]
	0.5	83.0	$1 \cdot 10^7$	3600	27.5[0]	25.2	$5 \cdot 10^6$	3600	18.7[0]	12.8	11,912	3600	16.4[0]	58.6	16,398	3,432	7.7[1]	58.7	18,645	3600	7.9[0]
	1.0	77.3	$1 \cdot 10^7$	3600	25.0[0]	39.9	$6 \cdot 10^6$	3600	10.0[0]	19.7	16,144	3,236	4.8[1]	75.0	11,376	1,824	2.1[3]	75.4	10,588	1,822	2.5[3]
Average			$1 \cdot 10^7$	3600	27.5[0]	23.8	$6 \cdot 10^6$	3600	19.3[0]	11.8	14,304	3,479	19.0[1]	56.9	15,811	2,952	9.8[4]	57.1	15,426	3,007	10.7[3]

Table 4 Experiments with constant diagonal dominance and varying positive off-diagonal entries, $k = n/5$

n	ρ	igap	Basic			Perspective				Perspective + cuts				Conic				Conic + cuts			
			Nodes	Time	Egap	Rimp	Nodes	Time	Egap	Rimp	Nodes	Time	Egap	Rimp	Nodes	Time	Egap	Rimp	Nodes	Time	Egap
60	0.1	62.4	$7 \cdot 10^5$	153	0.0[5]	46.0	$7 \cdot 10^4$	22	0.0[5]	56.1	10,165	62	0.0[5]	77.6	2,141	19	0.0[5]	78.1	2,065	23	0.0[5]
	0.2	57.3	$7 \cdot 10^5$	144	0.0[5]	46.8	$7 \cdot 10^4$	22	0.0[5]	56.4	16,642	89	0.0[5]	73.5	3,314	20	0.0[5]	73.9	3,261	24	0.0[5]
	0.5	51.2	$6 \cdot 10^5$	128	0.0[5]	48.0	$6 \cdot 10^4$	19	0.0[5]	53.6	22,526	137	0.0[5]	65.1	8,635	36	0.0[5]	65.5	8,742	60	0.0[5]
Average			$7 \cdot 10^5$	142	0.0[15]	46.9	$6 \cdot 10^5$	21	0.0[15]	55.4	16,444	96	0.0[15]	72.1	4,696	25	0.0[15]	72.5	4,689	36	0.0[15]
80	0.1	64.4	$1 \cdot 10^7$	3600	7.6[0]	46.9	$2 \cdot 10^6$	852	0.0[5]	32.8	53,774	1,401	0.4[4]	77.4	8,979	244	0.0[5]	78.2	8,551	183	0.0[5]
	0.2	58.8	$1 \cdot 10^7$	3600	5.9[0]	48.1	$2 \cdot 10^6$	881	0.0[5]	37.8	98,151	1,997	0.6[4]	74.3	25,152	349	0.0[5]	75.4	22,630	327	0.0[5]
	0.5	51.8	$1 \cdot 10^7$	3,255	3.2[1]	49.7	$8 \cdot 10^5$	391	0.0[5]	43.7	185,839	2,462	0.4[4]	67.8	66,779	482	0.0[5]	68.5	64,512	535	0.0[5]
Average			$1 \cdot 10^7$	3,485	5.5[1]	48.2	$1 \cdot 10^6$	708	0.0[15]	38.1	112,588	1,953	0.5[12]	73.2	33,637	358	0.0[15]	74.0	31,898	349	0.0[15]
100	0.1	65.0	$9 \cdot 10^6$	3600	23.1[0]	42.3	$5 \cdot 10^6$	3600	9.1[0]	28.8	65,628	3600	6.4[0]	73.0	83,300	2,667	2.5[2]	73.8	67,074	2,904	2.6[2]
	0.2	59.4	$9 \cdot 10^6$	3600	20.9[0]	43.9	$5 \cdot 10^6$	3600	7.8[0]	32.9	72,439	3600	9.0[0]	70.6	122,553	3,031	2.8[2]	71.2	116,173	3,033	3.3[1]
	0.5	52.5	$9 \cdot 10^6$	3600	17.2[0]	46.2	$5 \cdot 10^6$	3600	5.4[0]	39.1	136,082	3600	7.7[0]	64.4	261,440	3,327	3.8[1]	64.8	270,701	3,396	3.7[1]
Average			$9 \cdot 10^6$	3600	20.4[0]	44.2	$5 \cdot 10^6$	3600	7.4[0]	33.6	91,383	3600	7.4[0]	69.3	155,764	3,008	3.0[5]	69.9	151,316	3,111	3.2[4]

Sect. 7.1, we observe little difference between `Conic` and `Conic+cuts`—consistent with Proposition 8—and that `Perspective+cuts` is not effective in closing the root gap. Approximating the nonlinear function with gradient inequalities appears to cause numerical issues as adding cuts weakens the relaxation contrary to expectations. Please see our comments at the end of Sect. 7.1.

Finally, observe that the formulations tested require adding $O(n^2)$ additional variables, one for each negative off-diagonal entry in A. Thus, solving the continuous relaxations may be computationally expensive for large values of n. Table 5 illustrates this point for matrices with $\rho = 0$ and $\delta = 1$. It shows, for the `Basic`, `Perspective` and `Conic` formulations, the value of the best feasible solution found (`sol`), the value of the lower bound after one hour of branch and bound (`ebound`), the value of the lower bound after processing the root node (`rbound`), the time used to process the root node in seconds (`rtime`), and the number of nodes explored in one hour (`nodes`). Each row represents the average over five instances, and the values of `sol`, `ebound` and `rbound` are scaled so that the best feasible solution found for a given instance has value 100. Observe that for $n \geq 150$ the lower bound found by `Conic` at the root node is stronger than the lower bounds found by other formulations after one hour of branch-and-bound. However, the continuous relaxations of `Conic` are difficult to solve for large values of n, leading to few branch-and-bound nodes explored and few or no feasible solutions found within the time limit.

A possible approach that achieves a compromise between the strength and the size of the formulation is to apply the proposed conic inequalities for a subset of the matrix: given an M-matrix Q, choose $I \subset \left\{(i,j) \in N \times N : Q_{ij} < 0\right\}$ and use the formulation

$$\begin{aligned}\min \; & \sum_{i \in P} \bar{Q}_i z_i + \sum_{i \in \bar{P}} \bar{Q}_i y_i - \sum_{(i,j) \in I} Q_{ij} t_{ij} - \sum_{(i,j) \notin I} Q_{ij} (y_i - y_j)^2 \\ \text{s.t.} \; & y_i^2 \leq z_i x_i, \;\; \forall i \in P, \quad (x_i, x_j, y_i, y_j, t_{ij}) \in X, \;\; \forall (i,j) \in I.\end{aligned}$$

In particular, if $|I| \approx 4n$, then the results in Sect. 7.1 suggest that the formulations would scale well. Additionally, the component corresponding to the remainder, $-\sum_{(i,j)\notin I} Q_{ij}(y_i - y_j)^2$, could be further strengthened by linear inequalities (37) (and other subgradient inequalities corresponding to points where $\bar{y} \neq \bar{x}$) in the original space of variables instead of extended reformulations. An effective implementation of such a partial strengthening is beyond the scope of the current paper.

8 Conclusions

In this paper we show, under mild assumptions, that minimization of a quadratic function with an M-matrix with indicator variables is a submodular minimization problem, hence, solvable in polynomial time. We derive strong formulations using the convex hull description of non-separable quadratic terms with two indicator variables arising from a decomposition of the quadratic function. Additionally, we provide strong conic quadratic valid inequalities approximating the convex hulls. The derived formulations generalize previous results in the binary case and separable case, and the inequalities

Table 5 Experiments with $n \geq 100$ and $k = n/10$

n	Basic					Perspective					Conic				
	Sol	Ebound	Rbound	Rtime	Nodes	Sol	Ebound	Rbound	Rtime	Nodes	Sol	Ebound	Rbound	Rtime	Nodes
100	100.0	94.9	11.0	0.09	12,375,694	100.0	100.0	42.2	0.05	1,968,600	100.0	100.0	77.9	2.13	2,176
150	100.0	61.3	11.7	0.07	9,739,922	100.3	81.3	45.6	0.08	3,788,060	100.6	96.2	83.8	141.46	3,174
200	100.0	46.5	12.4	0.11	6,382,960	100.3	72.5	48.4	0.13	2,644,816	–	90.8	86.4	1090.73	1,531
250	100.0	34.7	11.6	0.22	4,092,948	100.3	72.5	48.4	0.21	1,692,204	–	82.7	82.7	1732.13	3
300	100.0	29.5	12.0	0.41	2,763,780	100.9	61.0	47.1	0.32	1,166,534	–	86.1	86.1	2333.81	1

dominate valid inequalities given in the literature. Computational experiments indicate that the proposed conic formulations may be significantly more effective compared to the natural convex relaxation and the perspective reformulation.

References

1. Ahuja, R.K., Hochbaum, D.S., Orlin, J.B.: A cut-based algorithm for the nonlinear dual of the minimum cost network flow problem. Algorithmica **39**, 189–208 (2004)
2. Aktürk, M.S., Atamtürk, A., Gürel, S.: A strong conic quadratic reformulation for machine-job assignment with controllable processing times. Oper. Res. Lett. **37**, 187–191 (2009)
3. Anstreicher, K.M.: On convex relaxations for quadratically constrained quadratic programming. Math. Program. **136**, 233–251 (2012)
4. Atamtürk, A., Bhardwaj, A.: Network design with probabilistic capacities. Networks **71**, 16–30 (2018)
5. Atamtürk, A., Gomez, A.: Submodularity in conic quadratic mixed 0–1 optimization. BCOL Research Report 16.02, IEOR, UC Berkeley. arXiv preprint arXiv:1705.05918 (2017)
6. Atamtürk, A., Jeon, H.: Lifted polymatroid for mean-risk optimization with indicator variables. BCOL Research Report 17.01, IEOR, UC Berkeley. arXiv preprint arXiv:1705.05915 (2017)
7. Atamtürk, A., Narayanan, V.: Cuts for conic mixed integer programming. In: Fischetti, M., Williamson, D.P. (eds.) Proceedings of the 12th International IPCO Conference, pp. 16–29 (2007)
8. Balas, E.: Disjunctive programming and a hierarchy of relaxations for discrete optimization problems. SIAM J. Algebr. Discrete Methods **6**, 466–486 (1985)
9. Belotti, P., Góez, J.C., Pólik, I., Ralphs, T.K., Terlaky, T.: A conic representation of the convex hull of disjunctive sets and conic cuts for integer second order cone optimization. In: Numerical Analysis and Optimization, pp. 1–35. Springer (2015)
10. Bertsimas, D., King, A., Mazumder, R.: Best subset selection via a modern optimization lens. Ann. Stat. **44**, 813–852 (2016)
11. Bienstock, D.: Computational study of a family of mixed-integer quadratic programming problems. Math. Program. **74**, 121–140 (1996)
12. Bienstock, D., Michalka, A.: Cutting-planes for optimization of convex functions over nonconvex sets. SIAM J. Optim. **24**, 643–677 (2014)
13. Boland, N., Dey, S.S., Kalinowski, T., Molinaro, M., Rigterink, F.: Bounding the gap between the McCormick relaxation and the convex hull for bilinear functions. Math. Program. **162**, 523–535 (2017a)
14. Boland, N., Gupte, A., Kalinowski, T., Rigterink, F., Waterer, H.: Extended formulations for convex hulls of graphs of bilinear functions. arXiv preprint arXiv:1702.04813 (2017b)
15. Bonami, P., Lodi, A., Tramontani, A., Wiese, S.: On mathematical programming with indicator constraints. Math. Program. **151**, 191–223 (2015)
16. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. **23**, 1222–1239 (2001)
17. Ceria, S., Soares, J.: Convex programming for disjunctive convex optimization. Math. Program. **86**, 595–614 (1999)
18. Cornuejols, G., Tütüncü, R.: Optimization Methods in Finance, vol. 5. Cambridge University Press, Cambridge (2006)
19. Dong, H., Linderoth, J.: On valid inequalities for quadratic programming with continuous variables and binary indicators. In: Goemans, M., Correa, J. (eds.) Proceedings of IPCO 2013, pp. 169–180. Springer, Berlin (2013)
20. Edmonds, J.: Submodular functions, matroids, and certain polyhedra. In: Guy, R., Hanani, H., Sauer, N., Schönenheim, J. (eds.) Combinatorial Structures and Their Applications, pp. 69–87. Gordon and Breach, Philadelphia (1970)
21. Frangioni, A., Gentile, C.: Perspective cuts for a class of convex 0–1 mixed integer programs. Math. Program. **106**, 225–236 (2006)
22. Frangioni, A., Gentile, C., Hungerford, J.: Decompositions of semidefinite matrices and the perspective reformulation of nonseparable quadratic programs. Report R-16-10, IASI, Rome (2016)
23. Gao, J., Li, D.: Cardinality constrained linear-quadratic optimal control. IEEE Trans. Autom. Control **56**, 1936–1941 (2011)

24. Günlük, O., Linderoth, J.: Perspective reformulations of mixed integer nonlinear programs with indicator variables. Math. Program. **124**, 183–205 (2010)
25. Hijazi, H., Bonami, P., Cornuéjols, G., Ouorou, A.: Mixed-integer nonlinear programs featuring "on/off" constraints. Comput. Optim. Appl. **52**, 537–558 (2012)
26. Hiriart-Urruty, J.B., Lemaréchal, C.: Convex Analysis and Minimization Algorithms I: Fundamentals, vol. 305. Springer, Berlin (2013)
27. Hochbaum, D.S.: Multi-label markov random fields as an efficient and effective tool for image segmentation, total variations and regularization. Numer. Math. Theory Methods Appl. **6**, 169–198 (2013)
28. Ivănescu, P.L.: Some network flow problems solved with pseudo-boolean programming. Oper. Res. **13**, 388–399 (1965)
29. Jeon, H., Linderoth, J., Miller, A.: Quadratic cone cutting surfaces for quadratic programs with on–off constraints. Discrete Optim. **24**, 32–50 (2017)
30. Keilson, J., Styan, G.P.H.: Markov chains and M-matrices: inequalities and equalities. J. Math. Anal. Appl. **41**, 439–459 (1973)
31. Kılınç-Karzan, F., Yıldız, S.: Two-term disjunctions on the second-order cone. Math. Program. **154**, 463–491 (2015)
32. Kolmogorov, V., Zabin, R.: What energy functions can be minimized via graph cuts? IEEE Trans. Pattern Anal. Mach. Intell. **26**, 147–159 (2004)
33. Lobo, M.S., Fazel, M., Boyd, S.: Portfolio optimization with linear and fixed transaction costs. Ann. Oper. Res. **152**, 341–365 (2007)
34. Lovász, L.: Submodular functions and convexity. In: Bachem, A., Korte, B., Grötschel, M. (eds.) Mathematical Programming The State of the Art: Bonn 1982, pp. 235–257. Springer, Berlin (1983)
35. Luedtke, J., Namazifar, M., Linderoth, J.: Some results on the strength of relaxations of multilinear functions. Math. Program. **136**, 325–351 (2012)
36. Luedtke, J., D'Ambrosio, C., Linderoth, J., Schweiger, J.: Strong convex nonlinear relaxations of the pooling problem. arXiv preprint arXiv:1803.02955 (2018)
37. Luk, F.T., Pagano, M.: Quadratic programming with M-matrices. Linear Algebra Appl. **33**, 15–40 (1980)
38. Mahajan, A., Leyffer, S., Linderoth, J., Luedtke, J., Munson, T.: Minotaur: A mixed-integer nonlinear optimization toolkit. ANL/MCS-P8010-0817, Argonne National Lab (2017)
39. Modaresi, S., Kılınç, M.R., Vielma, J.P.: Intersection cuts for nonlinear integer programming: convexification techniques for structured sets. Math. Program. **155**, 575–611 (2016)
40. Nemhauser, G.L., Wolsey, L.A., Fisher, M.L.: An analysis of approximations for maximizing submodular set functions I. Math. Program. **14**, 265–294 (1978)
41. Orlin, J.B.: A faster strongly polynomial time algorithm for submodular function minimization. Math. Program. **118**, 237–251 (2009)
42. Picard, J.C., Ratliff, H.D.: Minimum cuts and related problems. Networks **5**, 357–370 (1975)
43. Plemmons, R.J.: M-matrix characterizations. I—nonsingular M-matrices. Linear Algebra Appl. **18**, 175–188 (1977)
44. Poljak, S., Wolkowicz, H.: Convex relaxations of (0,1)-quadratic programming. Math. Oper. Res. **20**, 550–561 (1995)
45. Stubbs, R.A., Mehrotra, S.: A branch-and-cut method for 0–1 mixed convex programming. Math. Program. **86**, 515–532 (1999)
46. Vielma, J.P.: Small and strong formulations for unions of convex sets from the cayley embedding. To appear in Mathematical Programming, arXiv preprint arXiv:1704.03954 (2018)
47. Wei, D., Sestok, C.K., Oppenheim, A.V.: Sparse filter design under a quadratic constraint: low-complexity algorithms. IEEE Trans. Signal Process. **61**, 857–870 (2013)
48. Wu, B., Sun, X., Li, D., Zheng, X.: Quadratic convex reformulations for semicontinuous quadratic programming. SIAM J. Optim. **27**, 1531–1553 (2017)
49. Young, N.: The rate of convergence of a matrix power series. Linear Algebra Appl. **35**, 261–278 (1981)