

On two relaxations of quadratically-constrained cardinality minimization

Dennis Wei

Received: October 18, 2012/ Accepted: date

Abstract This paper considers a quadratically-constrained cardinality minimization problem with applications to digital filter design, subset selection for linear regression, and portfolio selection. Two relaxations are investigated: the continuous relaxation of a mixed integer formulation, and an optimized diagonal relaxation that exploits a simple special case of the problem. For the continuous relaxation, an absolute upper bound on the optimal cost is derived, suggesting that the continuous relaxation tends to be a relatively poor approximation. In computational experiments, diagonal relaxations often provide stronger bounds than continuous relaxations and can greatly reduce the complexity of a branch-and-bound solution, even in instances that are not particularly close to diagonal. Similar gains are observed with respect to the mixed integer programming solver CPLEX. Motivated by these results, the approximation properties of the diagonal relaxation are analyzed. In particular, bounds on the approximation ratio are established in terms of the eigenvalues of the matrix defining the quadratic constraint, and also in the diagonally dominant and nearly coordinate-aligned cases.

Keywords Cardinality minimization · Mixed integer quadratic programming · Relaxation methods · Subset selection · Portfolio optimization

Mathematics Subject Classification (2000) 90C11 · 90C57 · 90C59

1 Introduction

This paper considers the problem of minimizing the cardinality of a vector $\mathbf{x} \in \mathbb{R}^N$ subject to a single convex quadratic constraint:

$$\min_{\mathbf{x}} C(\mathbf{x}) \quad \text{s.t.} \quad (\mathbf{x} - \mathbf{c})^T \mathbf{Q}(\mathbf{x} - \mathbf{c}) \leq \gamma, \quad (1.1)$$

D. Wei
Department of Electrical Engineering and Computer Science, University of Michigan, 1301 Beal Avenue,
Ann Arbor, MI 48109, USA
E-mail: dlwei@eecs.umich.edu

where $C(\mathbf{x})$ is the number of nonzero components of \mathbf{x} , \mathbf{Q} is a positive definite matrix, and γ is a positive scalar. Geometrically, problem (1.1) corresponds to finding a point of minimal cardinality in an ellipsoid, denoted as $\mathcal{E}_{\mathbf{Q}}$, centered at the point \mathbf{c} . The orientation and relative lengths of the ellipsoid axes are determined by the eigenvectors and eigenvalues of \mathbf{Q} while γ determines its absolute size.

The author's interest in (1.1) stems from the design of digital filters in signal processing (see [27, 28] and the references therein). In this context, \mathbf{x} represents a vector of filter coefficients and cardinality minimization is motivated by the fact that the cost of implementing a filter is often dominated by arithmetic operations, especially in hardware. The quadratic constraint represents a requirement on filter performance, for example a specified fidelity in approximating a desired frequency response or a bound on recovery error in the equalization of communication channels.

Problem (1.1) also has applications to subset selection for linear regression [11, 21], more specifically the overdetermined case in which \mathbf{Q} is positive definite and less so the underdetermined case in which \mathbf{Q} is rank-deficient and control of the cardinality is employed as a regularization. A similar problem arises in optimal linear-quadratic control with cardinality-constrained input [14] (see also [20] for optimal control with sparse state-feedback gains). A problem related to (1.1) has been studied extensively in cardinality-constrained financial portfolio optimization [2, 4, 5, 10, 12, 16, 22, 25]. The portfolio optimization problem however has additional linear constraints, most notably non-negativity, upper bounds on nonzero variables, and sometimes lower bounds as well. There is some computational evidence [2] to suggest that the relative lack of constraints in (1.1) increases the difficulty of the problem, at least when approached using conventional integer optimization methods.

Certain cases of (1.1) are known to be efficiently solvable, the simplest of which is the case of diagonal \mathbf{Q} . Extensions to block-diagonal, tridiagonal, and well-conditioned \mathbf{Q} are discussed in [28]. The authors of [11] present polynomial algorithms for several additional cases, including an FPTAS for the general banded case and exact algorithms for the cases of a tree-structured covariance graph, a large independent set ("arrow"-structured \mathbf{Q}), and exponential decay in the entries of \mathbf{Q} away from the diagonal. The case in which nearly all of the eigenvalues of \mathbf{Q} are identical and larger than the rest is treated in [15].

In the general case, (1.1) is a difficult combinatorial optimization problem. Several heuristics such as forward and backward greedy selection can be used, often with good results (see e.g. [28], also [29] for references on portfolio selection heuristics). Although approximation guarantees do exist for forward selection in the near-diagonal case [11] and for backward selection when a (difficult to evaluate) threshold test is met [9], more general guarantees for heuristics are not available. Thus if a certificate of optimality or a bound on the deviation from optimality is desired, branch-and-bound remains the method of choice and has therefore been considered by many researchers [2, 4, 5, 22, 25]. In particular, [4] investigates a branch-and-cut algorithm employing disjunctive cuts and finds that such cuts are ineffective when \mathbf{Q} is near full rank. In [2], Lemke's pivoting method is used to provide warm starts in solving continuous relaxations. Lagrangian relaxations have also been considered [22]. In [25], a lifted polyhedral relaxation is applied to mixed-integer second-order cone programs

(which include the problems considered here) to take advantage of the more mature techniques for solving mixed-integer linear programs.

The complexity of branch-and-bound can be significantly reduced if specialized relaxations are available that can better approximate the original optimal cost while remaining efficiently solvable. Such relaxations permit increased pruning of the branch-and-bound tree and can also suggest stronger reformulations of the original problem. In the present context, a sequence of works [10, 12, 13, 16, 29] have developed the *perspective* relaxation, so-called because of its relationship to the perspective of a convex function. The perspective relaxation can also be viewed as a particularly tractable instance of disjunctive convex optimization [7]. In [12, 16], the relaxation is derived for general convex functions (not necessarily quadratic) using a convex hull approach; [12] emphasizes the identification of linear cuts whereas [16] proposes solving the nonlinear relaxation directly, aided by second-order cone representations. In contrast, [10] focuses on portfolio optimization and derives the relaxation through Lagrangian decomposition. The authors of [10, 12, 16] also show that the perspective relaxation is tighter than the standard continuous relaxation in certain contexts. To apply the relaxation to portfolio optimization problems, a diagonal matrix must be separated from \mathbf{Q} ; a semidefinite programming method for determining the best separation was reported very recently [29] and is shown to outperform simpler methods in [12, 13]. None of the above works however have analyzed the quality of approximation of the relaxation with respect to the original problem.

In this paper, we focus on the pure quadratically-constrained problem (1.1) and investigate two relaxations. The first is the conventional continuous relaxation, obtained by formulating (1.1) as a mixed-integer optimization and relaxing binary-value constraints to unit interval constraints. An absolute upper bound is given on the optimal cost of the continuous relaxation. The bound suggests that the continuous relaxation is relatively weak for many instances of (1.1), a hypothesis borne out by numerical experiments. The second relaxation exploits the simplicity of the case of diagonal \mathbf{Q} , specifically by constructing the best diagonal approximation to (1.1), referred to as a diagonal relaxation. A computational comparison of the two relaxations shows that diagonal relaxations often yield significantly stronger bounds and can greatly decrease the complexity of a branch-and-bound solution to (1.1), by orders of magnitude in difficult instances, and even when \mathbf{Q} does not seem close to diagonal. Similar efficiency gains are seen relative to the mixed-integer programming solver CPLEX [18]. Motivated by these results, this paper undertakes a theoretical analysis of diagonal relaxations, providing approximation guarantees for certain classes of instances and general insight into when diagonal relaxations are expected to be successful. In particular, bounds on the approximation ratio are derived in terms of the eigenvalues of \mathbf{Q} and in the cases of diagonally dominant \mathbf{Q} and nearly coordinate-aligned $\mathcal{E}\mathbf{Q}$. We note that a relaxation similar to the diagonal relaxation was proposed independently in [14] with similarly positive computational experience. A principal objective of the current paper is to support such findings with more detailed analysis.

We begin in Sect. 2 by deriving some preliminary facts pertaining to problem (1.1). In Sect. 3, continuous relaxations of (1.1) are discussed and analyzed, while the same is done for diagonal relaxations in Sect. 4. In Sect. 5, the two relaxations are compared numerically in terms of their approximation ratios and effect on branch-

and-bound complexity. A comparison with CPLEX is also reported. The paper concludes in Sect. 6.

1.1 Notation

Vectors and matrices are denoted using lowercase and uppercase boldface letters with x_n representing the n th element of a vector \mathbf{x} and Q_{mn} the (m, n) element of a matrix \mathbf{Q} . The letter \mathbf{e} is reserved for a vector of unit entries. For sets of indices Y and Z , \mathbf{x}_Y represents the $|Y|$ -dimensional subvector of \mathbf{x} corresponding to Y and \mathbf{Q}_{YZ} the $|Y| \times |Z|$ submatrix of \mathbf{Q} with rows indexed by Y and columns indexed by Z . The notation $\mathbf{Q} \succeq \mathbf{0}$ ($\mathbf{Q} \succ \mathbf{0}$) indicates that \mathbf{Q} is positive semidefinite (positive definite); $\mathbf{Q} \succeq \mathbf{D}$ is equivalent to $\mathbf{Q} - \mathbf{D} \succeq \mathbf{0}$. The n th smallest eigenvalue of \mathbf{Q} is written $\lambda_n(\mathbf{Q})$ except as noted in Sect. 4.5; we also use $\lambda_{\min}(\mathbf{Q})$ and $\lambda_{\max}(\mathbf{Q})$ for the smallest and largest eigenvalues.

2 Preliminaries

In this section, some facts related to problem (1.1) are derived for later use. In Sect. 2.1, a condition is given for the feasibility of solutions of specified cardinality. In Sect. 2.2, it is shown that variables that are either constrained to a zero value or assumed to be nonzero can be eliminated to yield a lower-dimensional instance of (1.1).

2.1 Feasibility of solutions of specified cardinality

First we obtain a condition for the existence of feasible solutions to (1.1) with a specified number K of zero-valued components. Suppose that x_n is constrained to a zero value for n in a set Z of size K . With Y denoting the complement of Z , the constraint in (1.1) becomes

$$(\mathbf{x}_Y - \mathbf{c}_Y)^T \mathbf{Q}_{YY} (\mathbf{x}_Y - \mathbf{c}_Y) - 2\mathbf{c}_Z^T \mathbf{Q}_{ZY} (\mathbf{x}_Y - \mathbf{c}_Y) + \mathbf{c}_Z^T \mathbf{Q}_{ZZ} \mathbf{c}_Z \leq \gamma. \quad (2.1)$$

Consider minimizing the left-hand side of (2.1) with respect to \mathbf{x}_Y , with solution $\mathbf{x}_Y - \mathbf{c}_Y = (\mathbf{Q}_{YY})^{-1} \mathbf{Q}_{YZ} \mathbf{c}_Z$. If (2.1) is not satisfied when the left-hand side is minimized, then it cannot be satisfied for any value of \mathbf{x}_Y . Hence a feasible solution to (1.1) exists subject to $x_n = 0$ for $n \in Z$ if and only if

$$\mathbf{c}_Z^T (\mathbf{Q}/\mathbf{Q}_{YY}) \mathbf{c}_Z \leq \gamma, \quad (2.2)$$

where $\mathbf{Q}/\mathbf{Q}_{YY} = \mathbf{Q}_{ZZ} - \mathbf{Q}_{ZY} (\mathbf{Q}_{YY})^{-1} \mathbf{Q}_{YZ} = ((\mathbf{Q}^{-1})_{ZZ})^{-1}$ is the Schur complement of \mathbf{Q}_{YY} . Condition (2.2) may be generalized to encompass all subsets of cardinality K using a similar argument: If (2.2) is not satisfied when the left-hand side is minimized over all subsets Z of cardinality K , then there can be no solution to (1.1) with K zero-valued components. This yields the condition

$$E_0(K) = \min_{|Z|=K} \{ \mathbf{c}_Z^T (\mathbf{Q}/\mathbf{Q}_{YY}) \mathbf{c}_Z \} \leq \gamma \quad (2.3)$$

for the existence of a feasible solution of cardinality $N - K$. In general, computing $E_0(K)$ in (2.3) involves an intractable combinatorial optimization. However, when \mathbf{Q} has special structure, $E_0(K)$ becomes much easier to evaluate and it is in these cases that condition (2.3) will be used.

In the special case of a single zero-value constraint, i.e., $Z = \{n\}$, condition (2.2) reduces to

$$\frac{c_n^2}{(\mathbf{Q}^{-1})_{nn}} \leq \gamma. \quad (2.4)$$

If (2.4) is not satisfied, then it is not feasible for x_n to take a value of zero. It follows that an easily computed lower bound on the optimal cost in (1.1) is obtained by counting the number of indices n for which (2.4) is not satisfied. Furthermore, in Sect. 2.2, it is shown that the variables x_n corresponding to violations of (2.4) can be eliminated from the problem to reduce its dimension.

2.2 Variable elimination

We now consider restrictions of problem (1.1) in which certain variables are constrained to zero while others are assumed to be nonzero. These two types of constraints arise in branch-and-bound as (1.1) is divided recursively into subproblems. Variables that must be non-zero to maintain feasibility may also be identified through condition (2.4).

Let Z denote as before the subset of variables constrained to zero, U the subset of variables assumed to be nonzero, and F the remainder. We show that an arbitrary subproblem defined by subsets (Z, U, F) can be reduced to the following problem:

$$\min_{\mathbf{x}_F} |U| + C(\mathbf{x}_F) \quad \text{s.t.} \quad (\mathbf{x}_F - \mathbf{c}_{\text{eff}})^T \mathbf{Q}_{\text{eff}} (\mathbf{x}_F - \mathbf{c}_{\text{eff}}) \leq \gamma_{\text{eff}}, \quad (2.5)$$

with effective parameters given by

$$\mathbf{Q}_{\text{eff}} = \mathbf{Q}_{FF} - \mathbf{Q}_{FU} (\mathbf{Q}_{UU})^{-1} \mathbf{Q}_{UF}, \quad (2.6a)$$

$$\mathbf{c}_{\text{eff}} = \mathbf{c}_F + (\mathbf{Q}_{\text{eff}})^{-1} (\mathbf{Q}_{FZ} - \mathbf{Q}_{FU} (\mathbf{Q}_{UU})^{-1} \mathbf{Q}_{UZ}) \mathbf{c}_Z, \quad (2.6b)$$

$$\gamma_{\text{eff}} = \gamma - \mathbf{c}_Z^T (\mathbf{Q} / \mathbf{Q}_{YY}) \mathbf{c}_Z. \quad (2.6c)$$

Problem (2.5) is an instance of (1.1) with $|F|$ variables instead of N .

The reduction can be carried out in the two steps $(\emptyset, \emptyset, \{1, \dots, N\}) \rightarrow (Z, \emptyset, Y = U \cup F) \rightarrow (Z, U, F)$. In the first step, the constraints $x_n = 0$ for $n \in Z$ reduce $C(\mathbf{x})$ to $C(\mathbf{x}_Y)$ and the quadratic constraint in (1.1) to (2.1). By completing the square, (2.1) can be rewritten as

$$\begin{bmatrix} \mathbf{x}_U - \mathbf{c}'_U \\ \mathbf{x}_F - \mathbf{c}'_F \end{bmatrix}^T \begin{bmatrix} \mathbf{Q}_{UU} & \mathbf{Q}_{UF} \\ \mathbf{Q}_{FU} & \mathbf{Q}_{FF} \end{bmatrix} \begin{bmatrix} \mathbf{x}_U - \mathbf{c}'_U \\ \mathbf{x}_F - \mathbf{c}'_F \end{bmatrix} \leq \gamma_{\text{eff}}, \quad (2.7)$$

where the subset Y has been partitioned into U and F , $\mathbf{c}'_U = \mathbf{c}_U + ((\mathbf{Q}_{YY})^{-1} \mathbf{Q}_{YZ} \mathbf{c}_Z)_U$, and $\mathbf{c}'_F = \mathbf{c}_F + ((\mathbf{Q}_{YY})^{-1} \mathbf{Q}_{YZ} \mathbf{c}_Z)_F$.

In the second step $(Z, \emptyset, U \cup F) \rightarrow (Z, U, F)$, the non-zero assumption on \mathbf{x}_U allows $C(\mathbf{x}_Y)$ to be rewritten as $|U| + C(\mathbf{x}_F)$. Since \mathbf{x}_U no longer has any effect on

the objective function, its value can be freely chosen, and in the interest of minimizing $C(\mathbf{x}_F)$, \mathbf{x}_U should be chosen as a function of \mathbf{x}_F to maximize the margin in constraint (2.7), thereby making the set of feasible \mathbf{x}_F as large as possible. This is equivalent to minimizing the left-hand side of (2.7) with respect to \mathbf{x}_U while holding \mathbf{x}_F constant. Similar to the minimization of (2.1) with respect to \mathbf{x}_Y , we obtain $\mathbf{x}_U^* = \mathbf{c}'_U - (\mathbf{Q}_{UU})^{-1} \mathbf{Q}_{UF}(\mathbf{x}_F - \mathbf{c}'_F)$ as the minimizer of (2.7). Substituting back into (2.7) results in the constraint in (2.5), except with \mathbf{c}'_F in place of \mathbf{c}_{eff} . By expressing $(\mathbf{Q}_{YY})^{-1}$ in terms of the block decomposition of \mathbf{Q}_{YY} in (2.7), it can be shown that \mathbf{c}'_F is equal to \mathbf{c}_{eff} as defined in (2.6b), thus completing the reduction.

In the sequel, we focus on the unrestricted root problem (1.1) with the understanding that the results apply to any subproblem by virtue of the reduction to (2.5). In addition, the following assumption will be made:

Assumption 2.1 Condition (2.4) is satisfied for all $n = 1, \dots, N$.

In other words, it is assumed that a feasible solution exists whenever a single variable is constrained to zero, since any variables for which this is not the case can be eliminated as shown in this section. Thus the focus is solely on the “difficult” part of the problem, i.e., those variables whose status is ambiguous.

3 Continuous relaxation

In the remainder of the paper, we consider two relaxations of (1.1) for the purpose of obtaining lower bounds on its optimal cost in the context of branch-and-bound. In Sect. 3.1, (1.1) is reformulated as a mixed-integer optimization problem, yielding a continuous relaxation. Best-case and worst-case instances are exhibited in Sect. 3.2 to show that continuous relaxations can provide arbitrarily tight or loose bounds on the optimal cost of (1.1). An absolute upper bound on the optimal cost of the relaxation is then derived in Sect. 3.3, suggesting that continuous relaxations are unlikely to yield good approximations to (1.1) in most instances.

3.1 Derivation

Problem (1.1) is first reformulated as a mixed integer optimization problem by associating with each continuous variable x_n a binary-valued indicator variable i_n with the property that $i_n = 0$ if $x_n = 0$ and $i_n = 1$ otherwise. Problem (1.1) can be restated in terms of indicator variables as follows:

$$\min_{\mathbf{x}, \mathbf{i}} \sum_{n=1}^N i_n \quad \text{s.t.} \quad (\mathbf{x} - \mathbf{c})^T \mathbf{Q}(\mathbf{x} - \mathbf{c}) \leq \gamma, \quad |x_n| \leq B_n i_n, \quad i_n \in \{0, 1\} \quad \forall n. \quad (3.1)$$

The constraint $|x_n| \leq B_n i_n$ is the usual forcing constraint linking i_n with x_n in the desired manner, where the positive constants B_n are chosen large enough to keep the set of feasible \mathbf{x} unchanged from that in (1.1). It will be seen shortly that B_n should be set to the smallest possible value subject to this requirement, i.e.,

$$B_n = \max \{ |x_n| : (\mathbf{x} - \mathbf{c})^T \mathbf{Q}(\mathbf{x} - \mathbf{c}) \leq \gamma \} = \max \{ B_n^+, B_n^- \},$$

where

$$B_n^\pm = \max \{ \pm x_n : (\mathbf{x} - \mathbf{c})^T \mathbf{Q}(\mathbf{x} - \mathbf{c}) \leq \gamma \} = \sqrt{\gamma(\mathbf{Q}^{-1})_{nn}} \pm c_n. \quad (3.2)$$

The closed-form expressions for B_n^+ and B_n^- can be derived straightforwardly from the corresponding KKT conditions [1, 26].

A continuous relaxation of (3.1) results from relaxing the binary-value constraints on i_n to interval constraints $0 \leq i_n \leq 1$. By minimizing the objective with respect to \mathbf{i} and substituting back into (3.1), we obtain the following minimization with respect to \mathbf{x} :

$$\min_{\mathbf{x}} \sum_{n=1}^N \frac{|x_n|}{B_n} \quad \text{s.t.} \quad (\mathbf{x} - \mathbf{c})^T \mathbf{Q}(\mathbf{x} - \mathbf{c}) \leq \gamma. \quad (3.3)$$

The continuous relaxation (3.3) is a quadratically-constrained weighted 1-norm minimization and is therefore a convex problem. The optimal cost in (3.3) is clearly a lower bound on the optimal cost in (3.1) since the feasible set has been enlarged; more precisely, since the latter must be an integer, the ceiling of the former is also a lower bound. It is also seen that the lower bound is maximized when the constants B_n are as small as possible.

A stronger lower bound on (3.1) can be obtained by first separating each variable x_n into its positive and negative parts x_n^+ and x_n^- as follows:

$$x_n = x_n^+ - x_n^-, \quad x_n^+, x_n^- \geq 0. \quad (3.4)$$

By assigning to each pair x_n^+, x_n^- corresponding indicator variables i_n^+, i_n^- and constants B_n^+, B_n^- , a mixed integer optimization problem equivalent to (3.1) may be formulated, where the values of B_n^+ and B_n^- are given by (3.2). The continuous relaxation of this alternative mixed integer formulation corresponds to the following quadratically-constrained linear program:

$$\min_{\mathbf{x}^+, \mathbf{x}^-} \sum_{n=1}^N \left(\frac{x_n^+}{B_n^+} + \frac{x_n^-}{B_n^-} \right) \quad \text{s.t.} \quad (\mathbf{x}^+ - \mathbf{x}^- - \mathbf{c})^T \mathbf{Q}(\mathbf{x}^+ - \mathbf{x}^- - \mathbf{c}) \leq \gamma, \quad \mathbf{x}^\pm \geq \mathbf{0}. \quad (3.5)$$

Using (3.4) to replace the absolute value functions in (3.3) with linear functions as done in linear programming [3], it can be seen that (3.3) is a special case of (3.5) with B_n^+ and B_n^- replaced by B_n . Since $B_n = \max\{B_n^+, B_n^-\}$, the optimal cost in (3.5) is at least as large as that in (3.3), and therefore (3.5) is at least as strong a relaxation as (3.3). The term continuous relaxation will refer henceforth to (3.5) with B_n^\pm given by (3.2).

Fig. 3.1 shows a graphical interpretation of the continuous relaxation (3.5). The asymmetric diamond represents a level contour of the cost function, which can be regarded as a weighted 1-norm with different weights for positive and negative component values. As seen from (3.2), the weights B_n^\pm correspond to the maximum extent of the ellipsoid $\mathcal{E}_{\mathbf{Q}}$ along the positive and negative coordinate directions and can be found graphically as indicated in Fig. 3.1. The solution to the weighted 1-norm minimization can be visualized by inflating the diamond until it just touches the ellipsoid. Note that Assumption 2.1 implies that $\mathcal{E}_{\mathbf{Q}}$ must intersect all of the coordinate planes. In Sect. 3.2, we will draw upon the geometric intuition in Fig. 3.1 to construct best-case and worst-case instances for continuous relaxation.

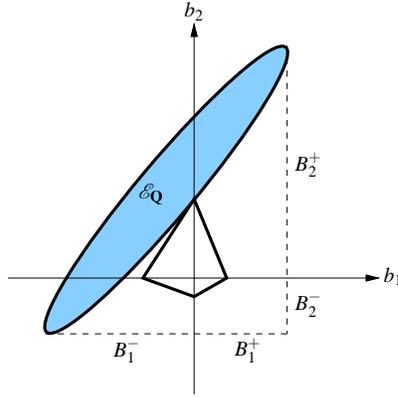


Fig. 3.1 Interpretation of the continuous relaxation as a weighted 1-norm minimization and a graphical representation of its solution.

3.2 Best-case and worst-case instances

In this subsection, instances of problem (1.1) are exhibited to show that the continuous relaxation can be a tight approximation to (1.1) as well as an arbitrarily poor one. The quality of approximation is characterized by the approximation ratio, defined as the ratio of the optimal cost of the relaxation to the optimal cost of the original problem.

In the instances to be constructed, we take $\mathbf{c} = \mathbf{e}$ and $\gamma = 1$, which can be regarded as a normalization. The matrix \mathbf{Q} is restricted to be of the form

$$\mathbf{Q} = \lambda_2 \mathbf{I} - (\lambda_2 - \lambda_1) \mathbf{v} \mathbf{v}^T, \quad (3.6)$$

where $\lambda_2 > \lambda_1$ and \mathbf{v} is vector with unit 2-norm and components equal to $\pm 1/\sqrt{N}$. It follows from (3.6) that \mathbf{v} is an eigenvector of \mathbf{Q} with eigenvalue λ_1 and the remaining $N - 1$ eigenvectors are orthogonal to \mathbf{v} with eigenvalue λ_2 . Geometrically, the ellipsoid $\mathcal{E}_{\mathbf{Q}}$ corresponding to (3.6) has a single long principal axis in the direction \mathbf{v} and shorter and equal principal axes in the other directions. We note for later use that the inverse of \mathbf{Q} and the Schur complement $\mathbf{Q}/\mathbf{Q}_{YY}$ can be computed explicitly as

$$\mathbf{Q}^{-1} = \frac{1}{\lambda_2} \mathbf{I} + \frac{\lambda_2 - \lambda_1}{\lambda_1 \lambda_2} \mathbf{v} \mathbf{v}^T, \quad (3.7)$$

$$\mathbf{Q}/\mathbf{Q}_{YY} = \lambda_2 \mathbf{I} - \frac{K \lambda_2 (\lambda_2 - \lambda_1)}{K \lambda_2 + (N - K) \lambda_1} \hat{\mathbf{v}}_Z \hat{\mathbf{v}}_Z^T, \quad (3.8)$$

where $K = |Z|$ and $\hat{\mathbf{v}}_Z$ is the unit 2-norm vector obtained by rescaling the subvector \mathbf{v}_Z .

To construct best-case instances for which the continuous relaxation is a tight approximation to (1.1), our aim is to make the optimal cost of the relaxation as large as possible. Based on Fig. 3.1 and the above structure for \mathbf{Q} , this can be done by choosing the major axis of $\mathcal{E}_{\mathbf{Q}}$ to be parallel to a level surface of the 1-norm and keeping the lengths of the minor axes to a minimum, thus allowing the ℓ_1 ball to grow relatively

unimpeded. Algebraically, we set $\lambda_1 = 1/N$, $\lambda_2 = N$, $\lceil N/2 \rceil$ of the components of \mathbf{v} equal to $+1/\sqrt{N}$, and the remaining components of \mathbf{v} equal to $-1/\sqrt{N}$.

First it is shown that the point $\mathbf{c} - \sqrt{N}\mathbf{v}$ is optimal for (1.1) with a corresponding cost of $\lfloor N/2 \rfloor$. Feasibility follows from substitution into the constraint in (1.1). To prove optimality, we verify that an additional zero-valued component is not feasible, i.e., condition (2.3) is violated for $K = N - \lfloor N/2 \rfloor + 1 = \lceil N/2 \rceil + 1$. Substituting (3.8) and $\mathbf{c} = \mathbf{e}$ into (2.3) and rearranging, we obtain

$$E_0(K) = (\lceil N/2 \rceil + 1)\lambda_2 \left(1 - \frac{(\lceil N/2 \rceil + 1)(\lambda_2 - \lambda_1)}{(\lceil N/2 \rceil + 1)\lambda_2 + (\lfloor N/2 \rfloor - 1)\lambda_1} \max_{|Z|=\lceil N/2 \rceil+1} \frac{(\mathbf{e}^T \hat{\mathbf{v}}_Z)^2}{\lceil N/2 \rceil + 1} \right). \quad (3.9)$$

The maximum in (3.9) is achieved by choosing Z to include all $\lceil N/2 \rceil$ positive components of \mathbf{v} and only one negative component, resulting in a maximum value of $(\lceil N/2 \rceil - 1)^2 / (\lceil N/2 \rceil + 1)^2$. The quantity $E_0(K)$ can then be bounded from below by removing the fraction in front of the maximization. This yields

$$(\lceil N/2 \rceil + 1)\lambda_2 \left(1 - \frac{(\lceil N/2 \rceil - 1)^2}{(\lceil N/2 \rceil + 1)^2} \right),$$

which can be seen to be strictly greater than $\gamma = 1$ as required.

We now prove that the lower bound provided by the continuous relaxation is equal to the optimal cost of $\lfloor N/2 \rfloor$ for the unrelaxed problem. Toward this end, we make use of the Lagrangian dual of the continuous relaxation, given by

$$\max_{\boldsymbol{\mu}} \quad \mathbf{c}^T \boldsymbol{\mu} - \sqrt{\gamma \boldsymbol{\mu}^T \mathbf{Q}^{-1} \boldsymbol{\mu}} \quad \text{s.t.} \quad -\mathbf{g}^- \leq \boldsymbol{\mu} \leq \mathbf{g}^+, \quad (3.10)$$

where $g_n^\pm = 1/B_n^\pm$ for $n = 1, \dots, N$. A derivation of the dual problem can be found in [26]. It is shown that the optimal cost of the dual is strictly bounded from below by $\lfloor N/2 \rfloor - 1$, implying through duality that the optimal cost of the primal (3.5) is between $\lfloor N/2 \rfloor - 1$ and $\lfloor N/2 \rfloor$ and is equal to $\lfloor N/2 \rfloor$ after rounding up to the next integer. From (3.2) and (3.7), we find that $B_n^+ = 1 + \sqrt{1 + (N-1)/N^2} = B^+$ for all n . Substituting the dual feasible solution $\boldsymbol{\mu} = \mathbf{g}^+ = (1/B^+)\mathbf{e}$ into the dual objective function and simplifying, we obtain

$$\begin{cases} \frac{1}{B^+}(N-1), & N \text{ even,} \\ \frac{1}{B^+} \left(N - \sqrt{2 - \frac{1}{N^2}} \right), & N \text{ odd,} \end{cases} \quad (3.11)$$

as a lower bound on the dual optimal cost. Straightforward algebraic manipulations show that the quantities in (3.11) are strictly greater than $\lfloor N/2 \rfloor - 1$ in the two cases of N even and N odd. This completes the demonstration of the potential tightness of the continuous relaxation lower bound.

Next we construct instances for which the lower bound resulting from the continuous relaxation is as loose as possible. The worst-case scenario corresponds to the optimal cost in (1.1) being equal to $N - 1$ and the optimal cost of the relaxation being less than 1. The former cannot equal N given Assumption 2.1 while the latter cannot equal zero exactly since that would require $\mathbf{x} = \mathbf{0}$ to be a feasible solution, in which case the optimal cost in (1.1) is also zero. Referring again to Fig. 3.1 and the form of

\mathbf{Q} in (3.6), the optimal cost of the continuous relaxation can be minimized by orienting the major axis of the ellipsoid $\mathcal{E}_{\mathbf{Q}}$ so that it points toward the origin and obstructs the growth of the ℓ_1 ball. Algebraically, we set $\mathbf{v} = (1/\sqrt{N})\mathbf{e}$, $\lambda_1 = 1/(N-1)$, and $\lambda_2 = (N-1)/2$. We verify that the unrelaxed optimal cost is equal to $N-1$. From (3.7), we have $(\mathbf{Q}^{-1})_{mm} = (N+1)/N$, which ensures that (2.4) is satisfied for all n . Using (3.8), $E_0(K)$ in (2.3) for $K=2$ evaluates to $N(N-1)/(N(N-1)-1)$. Since this quantity is greater than $\gamma=1$, condition (2.3) is violated for $K=2$ and the optimal cost in (1.1) must be equal to $N-1$.

To show that the optimal cost of the continuous relaxation is less than 1, we consider the feasible and strictly positive solution $\mathbf{x}^+ = \mathbf{c} - (1/\sqrt{\lambda_1})\mathbf{v} = (1 - \sqrt{(N-1)/N})\mathbf{e}$, $\mathbf{x}^- = \mathbf{0}$. From (3.5), the corresponding cost is

$$\frac{N - \sqrt{N(N-1)}}{B^+}, \quad (3.12)$$

where $B^+ = 1 + \sqrt{(N+1)/N}$ is the common value for the constants B_n^+ given by (3.2). Since $B^+ > 2$ while the numerator in (3.12) can be seen to be less than 1, we conclude that the optimal cost in (3.5) is less than 1 as claimed. The approximation ratio in these instances is thus equal to $1/(N-1)$, which approaches zero as N increases.

3.3 An absolute upper bound

The constructions in Sect. 3.2 imply that the approximation ratio for the continuous relaxation can range anywhere between 0 and 1, and thus it is not possible to place a non-trivial bound on the ratio that holds for all instances of (1.1). It is possible however to obtain an absolute upper bound on the optimal cost of the continuous relaxation in terms of the problem dimension N .

Proposition 3.1 *Under Assumption 2.1, the optimal cost of the continuous relaxation (3.5) is bounded from above by $\theta N/2$, where $\theta = 1 - \sqrt{\gamma/\mathbf{c}^T \mathbf{Q} \mathbf{c}}$.*

Proof Consider the solution $\mathbf{b}^+ - \mathbf{b}^- = \theta \mathbf{c}$, i.e., $b_n^+ = \theta c_n$, $b_n^- = 0$ for $c_n \geq 0$ and $b_n^+ = 0$, $b_n^- = \theta |c_n|$ for $c_n < 0$. It can be verified that this is a feasible solution for the continuous relaxation (the solution lies on the boundary of the ellipsoid $\mathcal{E}_{\mathbf{Q}}$), and hence the optimal cost of the relaxation is bounded from above by

$$\theta \sum_{n:c_n>0} \frac{c_n}{B_n^+} + \theta \sum_{n:c_n<0} \frac{|c_n|}{B_n^-} = \theta \sum_{n=1}^N \frac{|c_n|}{\sqrt{\gamma(\mathbf{Q}^{-1})_{nn}} + |c_n|}, \quad (3.13)$$

using (3.2). Assumption 2.1 then implies that each of the fractions on the right-hand side of (3.13) is no greater than $1/2$, completing the proof. \square

Proposition 3.1 indicates that the continuous relaxation cannot be a tight approximation if the optimal cost in (1.1) is greater than $\lceil \theta N/2 \rceil$. This suggests that it is unlikely for the continuous relaxation to yield a strong bound on (1.1) in most instances, since if it did, this would imply that the optimal cost in (1.1) is not much

greater than $\theta N/2$ in most cases, a fact considered unlikely. The situation is exacerbated if the factor θ is small. This negative result motivates the consideration of an alternative relaxation as described in Section 4.

We note in closing that Lemaréchal and Oustry [19] have shown that a common semidefinite relaxation technique is equivalent to continuous relaxation when applied to cardinality minimization problems such as (1.1). As a consequence, the properties of the continuous relaxation (3.5) noted in this section also apply to this type of semidefinite relaxation.

4 Diagonal relaxation

As an alternative to continuous relaxations, in this section we discuss relaxations of problem (1.1) in which the matrix \mathbf{Q} is replaced by a diagonal matrix, an approach referred to as diagonal relaxation. As will be seen in Sect. 4.1, problem (1.1) is easily solved in the diagonal case, thus making it attractive as a relaxation of the problem when \mathbf{Q} is non-diagonal. It is shown in Sect. 4.2 that diagonal relaxations can yield exact as well as arbitrarily poor approximations to (1.1), as was the case for the continuous relaxation in Sect. 3. However, numerical evidence in Sect. 5 and elsewhere [27] indicates that the lower bounds provided by diagonal relaxations are often significantly stronger than those from continuous relaxations. This computational experience motivates a better theoretical understanding of situations to which diagonal relaxations are particularly well-suited. Within this context, approximation guarantees are derived in Sect. 4.3–4.5 for the three specific cases of well-conditioned \mathbf{Q} matrices, diagonally dominant \mathbf{Q} , and nearly coordinate-aligned ellipsoids $\mathcal{E}_{\mathbf{Q}}$.

4.1 Derivation

To obtain a diagonal relaxation of problem (1.1), the matrix \mathbf{Q} is replaced with a positive definite diagonal matrix \mathbf{D} to yield a similar constraint:

$$(\mathbf{x} - \mathbf{c})^T \mathbf{D} (\mathbf{x} - \mathbf{c}) = \sum_{n=1}^N D_{nn} (x_n - c_n)^2 \leq \gamma. \quad (4.1)$$

Geometrically, (4.1) specifies an ellipsoid, denoted as $\mathcal{E}_{\mathbf{D}}$, with axes that are aligned with the coordinate axes. Since the relaxation is intended to provide a lower bound for the original problem, we require that the coordinate-aligned ellipsoid $\mathcal{E}_{\mathbf{D}}$ enclose the original ellipsoid $\mathcal{E}_{\mathbf{Q}}$ so that minimizing over $\mathcal{E}_{\mathbf{D}}$ yields a lower bound on the minimum over $\mathcal{E}_{\mathbf{Q}}$. For simplicity, the two ellipsoids are taken to be concentric, in which case the nesting of the ellipsoids is equivalent to the condition $\mathbf{D} \preceq \mathbf{Q}$. Sufficiency follows from the inequality

$$(\mathbf{x} - \mathbf{c})^T \mathbf{D} (\mathbf{x} - \mathbf{c}) \leq (\mathbf{x} - \mathbf{c})^T \mathbf{Q} (\mathbf{x} - \mathbf{c}) \quad \forall \mathbf{x}, \quad (4.2)$$

so if $\mathbf{x} \in \mathcal{E}_{\mathbf{Q}}$, then both sides of (4.2) are bounded by γ and $\mathbf{x} \in \mathcal{E}_{\mathbf{D}}$. Conversely, if $\mathbf{D} \not\preceq \mathbf{Q}$, then there exists a vector \mathbf{x} that violates (4.2), and by scaling $\mathbf{x} - \mathbf{c}$ so that

the right-hand side of (4.2) is equal to γ , we have $\mathbf{x} \in \mathcal{E}_{\mathbf{Q}}$ but $\mathbf{x} \notin \mathcal{E}_{\mathbf{D}}$ since \mathbf{x} does not satisfy (4.1).

Problem (1.1) is greatly simplified in the diagonal case. Replacing \mathbf{Q} by \mathbf{D} , condition (2.3) simplifies to

$$\min_{|Z|=K} \sum_{n \in Z} D_{nn} c_n^2 \leq \gamma$$

since $\mathbf{D}/\mathbf{D}_{YY} = \mathbf{D}_{ZZ}$. The minimum is attained by choosing Z to correspond to the K smallest elements of the sequence $D_{nn} c_n^2, n = 1, \dots, N$. It follows that (4.1) admits a solution with K zero-valued components if and only if

$$S_K(\{D_{nn} c_n^2\}) \leq \gamma, \quad (4.3)$$

where S_K denotes the sum of the K smallest elements of a sequence. The minimum cardinality corresponds to the largest value of K such that (4.3) holds.

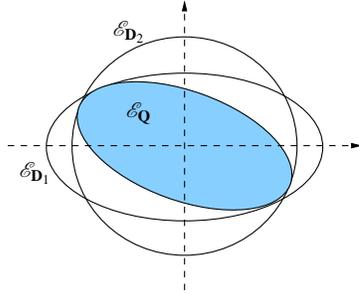


Fig. 4.1 Two different diagonal relaxations.

For every \mathbf{D} satisfying $\mathbf{0} \preceq \mathbf{D} \preceq \mathbf{Q}$, minimizing $C(\mathbf{x})$ subject to (4.1) results in a lower bound on the optimal cost in (1.1). Thus the set of diagonal relaxations is parameterized by \mathbf{D} as illustrated in Fig. 4.1. We are naturally interested in obtaining a diagonal relaxation that is as tight as possible, i.e., a matrix \mathbf{D}_d such that the minimum cardinality associated with \mathbf{D}_d is maximal among all valid choices of \mathbf{D} . Such a relaxation can be determined based on condition (4.3), specifically by solving the following optimization problem:

$$E_d(K) = \max_{\mathbf{D}} S_K(\{D_{nn} c_n^2\}) \quad \text{s.t.} \quad \mathbf{0} \preceq \mathbf{D} \preceq \mathbf{Q}, \quad \mathbf{D} \text{ diagonal}, \quad (4.4)$$

for selected values of K . If $E_d(K)$ in (4.4) is less than or equal to γ , then (4.3) holds for every \mathbf{D} satisfying the constraints in (4.4), and consequently a feasible solution \mathbf{x} with K zero-valued components exists for every such \mathbf{D} . We conclude that the optimal cost of any diagonal relaxation is at most $N - K$. On the other hand, if $E_d(K) > \gamma$, then according to (4.3) there exists a \mathbf{D} for which a vector \mathbf{x} with K zero-valued components is not feasible, and for this \mathbf{D} the optimal cost of the corresponding diagonal relaxation is at least $N - K + 1$. By selecting values of K to perform a bisection search over $1, \dots, N$ and solving (4.4) each time, we eventually arrive at the highest possible optimal cost under any diagonal relaxation, i.e., the tightest lower bound on (1.1)

achievable with a diagonal relaxation. Henceforth the term diagonal relaxation will be understood to refer to the tightest such relaxation.

The above procedure determines both the tightest possible diagonal relaxation and its optimal cost at the same time, and amounts to solving (4.4) for a maximum of $\lceil \log_2 N \rceil + 1$ values of K . Since the function S_K is concave in \mathbf{D} [6] and the constraints in (4.4) are convex, the maximization in (4.4) is a convex problem. Furthermore, (4.4) can be recast as a standard semidefinite program following [6] by expressing the function S_K as the optimal cost of a linear program and then substituting the Lagrangian dual of the linear program. Thus (4.4) can be solved efficiently using standard interior-point algorithms. Further efficiency enhancements can be made as detailed in [26, Sec. 3.5].

4.2 Worst-case instances

As with the continuous relaxation in Sect. 3.2, we consider extreme instances in which the diagonal relaxation is either a tight approximation to the original problem or an arbitrarily poor one. It is clear that if \mathbf{Q} is already diagonal, the diagonal relaxation and the original problem coincide and the approximation ratio defined in Sect. 3.2 is equal to 1. It is shown that the approximation ratio can also equal zero, i.e., the optimal cost of the diagonal relaxation can be zero while the original problem has a non-zero optimal cost. Based on Fig. 4.1, the diagonal relaxation is expected to result in a poor approximation when the original ellipsoid $\mathcal{E}_{\mathbf{Q}}$ is far from coordinate-aligned, thus forcing the coordinate-aligned enclosing ellipsoid $\mathcal{E}_{\mathbf{D}}$ to be much larger than $\mathcal{E}_{\mathbf{Q}}$. This situation is exemplified by the first class of instances in Sect. 3.2 in which $\mathcal{E}_{\mathbf{Q}}$ is dominated by a single long axis with equal components in all coordinate directions. To show that the diagonal relaxation has an optimal cost of zero in these instances, we make use of the following lemma.

Lemma 4.1 *Assume that the vector \mathbf{c} has unit-magnitude components. Then the optimal cost $E_d(K)$ in (4.4) is bounded from below by $K\lambda_{\min}(\mathbf{Q})$. This lower bound is tight if the eigenvector \mathbf{v} corresponding to $\lambda_{\min}(\mathbf{Q})$ has components of equal magnitude.*

Proof The diagonal matrix $\mathbf{D} = \lambda_{\min}(\mathbf{Q})\mathbf{I}$ satisfies $\mathbf{D} \preceq \mathbf{Q}$ and is therefore a feasible solution to (4.4). Hence the corresponding objective value $K\lambda_{\min}(\mathbf{Q})$ (with $c_n^2 = 1$ for all n) is a lower bound on $E_d(K)$. If the eigenvector \mathbf{v} has equal-magnitude components and is normalized to have unit 2-norm, then the inequality $\mathbf{D} \preceq \mathbf{Q}$ implies that

$$\mathbf{v}^T \mathbf{D} \mathbf{v} = \frac{1}{N} \sum_{n=1}^N D_{nn} \leq \mathbf{v}^T \mathbf{Q} \mathbf{v} = \lambda_{\min}(\mathbf{Q}) \quad (4.5)$$

for any feasible \mathbf{D} in (4.4). The solution $\mathbf{D} = \lambda_{\min}(\mathbf{Q})\mathbf{I}$ satisfies (4.5) with equality and is therefore an optimal solution to (4.4) for $K = N$ under the assumptions of the lemma, yielding $E_d(N) = N\lambda_{\min}(\mathbf{Q})$. Using the fact that the mean of the K smallest D_{nn} for $K < N$ is no greater than the mean of all N diagonal entries, it follows from (4.5) that

$$S_K(\{D_{nn}\}) \leq K\lambda_{\min}(\mathbf{Q}), \quad K = 1, 2, \dots, N-1, \quad (4.6)$$

again for any feasible \mathbf{D} in (4.4). Since the solution $\mathbf{D} = \lambda_{\min}(\mathbf{Q})\mathbf{I}$ also satisfies (4.6) with equality, it is an optimal solution to (4.4) for all K under the assumptions of the lemma and we have $E_d(K) = K\lambda_{\min}(\mathbf{Q})$. \square

In the first class of instances in Sect. 3.2, $\mathbf{c} = \mathbf{e}$, $\lambda_{\min}(\mathbf{Q}) = \lambda_1 = 1/N$ and the corresponding eigenvector \mathbf{v} has equal-magnitude components. It follows from Lemma 4.1 that $E_d(K) = K\lambda_1 = K/N$, which does not exceed $\gamma = 1$ for any K . Hence the optimal cost of the diagonal relaxation is zero while the optimal cost in the unrelaxed problem (1.1) is $\lfloor N/2 \rfloor$. This implies that it is not possible to bound the approximation ratio away from zero for all instances of (1.1), as with the continuous relaxation. Furthermore, since the continuous relaxation yields a tight approximation for the same class of instances, neither relaxation strictly dominates the other (diagonal relaxations are clearly dominant in the case of diagonal \mathbf{Q}). These conclusions however are based on extreme instances. It will be seen in Sect. 5 that in more typical instances the diagonal relaxation can offer a significantly better quality of approximation than the continuous relaxation. In addition, non-trivial lower bounds on the diagonal relaxation approximation ratio can be obtained as in Sect. 4.3–4.5 when the class of instances of (1.1) is restricted.

4.3 Eigenvalue-based approximation guarantees

In this subsection, the quality of approximation of the diagonal relaxation is characterized in terms of the eigenvalues of the matrix \mathbf{Q} . The resulting bounds on the approximation ratio are strongest in the case of well-conditioned \mathbf{Q} , i.e., when the eigenvalues of \mathbf{Q} have a low spread. Geometrically, the well-conditioned case corresponds to a nearly spherical ellipsoid $\mathcal{E}_{\mathbf{Q}}$, which can be enclosed by a coordinate-aligned ellipsoid $\mathcal{E}_{\mathbf{D}}$ of comparable size as illustrated in Fig. 4.2. Given the close approximation of $\mathcal{E}_{\mathbf{Q}}$ by $\mathcal{E}_{\mathbf{D}}$ in terms of volume, one would expect a close approximation in terms of the cardinality cost as well. This geometric intuition is confirmed by the analysis.

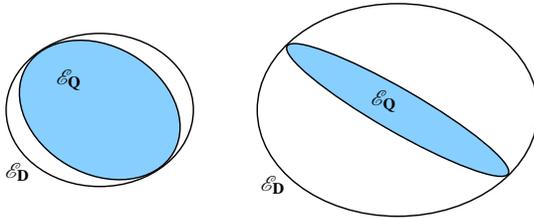


Fig. 4.2 Diagonal relaxations for two ellipsoids $\mathcal{E}_{\mathbf{Q}}$ with contrasting condition numbers.

The results presented in the remainder of the section are more conveniently stated in terms of the number of zero-valued components rather than the number of non-zero components. Define K^* to be the maximum number of zero-valued components in (1.1) and K_d to be the maximum number of zero-valued components in the diagonal

relaxation of (1.1). The enclosing condition $\mathcal{E}_{\mathbf{Q}} \subseteq \mathcal{E}_{\mathbf{D}}$ ensures that $K^* \leq K_d$, and a good approximation corresponds to the ratio K_d/K^* being not much larger than 1. It is shown that K^* and K_d can be bounded by the following quantities related to the eigenvalues of \mathbf{Q} and its Schur complements:

$$\underline{K} = \max \{K : \lambda_{\max}(\mathbf{Q}/\mathbf{Q}_{Y(K)Y(K)})S_K(\{c_n^2\}) \leq \gamma\}, \quad (4.7a)$$

$$\overline{K} = \max \{K : \lambda_{\min}(\mathbf{Q})S_K(\{c_n^2\}) \leq \gamma\}, \quad (4.7b)$$

where $Y(K)$ denotes the index set corresponding to the $N - K$ largest-magnitude components of \mathbf{c} (its complement $Z(K)$ corresponds to the K smallest components). The relationships among K^* , K_d , \underline{K} and \overline{K} are specified below.

Theorem 4.2 *The maximum numbers of zero-valued components in problem (1.1) and its diagonal relaxation, K^* and K_d respectively, satisfy the ordering $\underline{K} \leq K^* \leq K_d \leq \overline{K}$, where \underline{K} and \overline{K} are defined in (4.7). Furthermore, the approximation ratio K_d/K^* is bounded as follows:*

$$\frac{K_d}{K^*} \leq \frac{\overline{K}}{\underline{K}} \leq \frac{[(\underline{K} + 1)\lambda_{\max}(\mathbf{Q}/\mathbf{Q}_{Y(\underline{K}+1)Y(\underline{K}+1)})/\lambda_{\min}(\mathbf{Q})] - 1}{\underline{K}}. \quad (4.8)$$

Proof The quantity K^* is equivalently the largest value of K such that condition (2.3) is satisfied, and hence K^* can be bounded from below through an upper bound on $E_0(K)$ in (2.3). By choosing a specific subset $Z(K)$ corresponding to the K smallest-magnitude components of \mathbf{c} , we obtain

$$\begin{aligned} E_0(K) &= \min_{|Z|=K} \{\mathbf{c}_Z^T(\mathbf{Q}/\mathbf{Q}_{YY})\mathbf{c}_Z\} \leq \mathbf{c}_{Z(K)}^T(\mathbf{Q}/\mathbf{Q}_{Y(K)Y(K)})\mathbf{c}_{Z(K)} \\ &\leq \lambda_{\max}(\mathbf{Q}/\mathbf{Q}_{Y(K)Y(K)})S_K(\{c_n^2\}), \end{aligned} \quad (4.9)$$

where the second inequality is due to a property of quadratic forms [17]. It follows from (4.9) and the definition of \underline{K} in (4.7a) that $K^* \geq \underline{K}$. Similarly, K_d is the largest value of K such that $E_d(K)$ in (4.4) is no greater than γ and can therefore be bounded from above through a lower bound on $E_d(K)$. Since $\mathbf{D} = \lambda_{\min}(\mathbf{Q})\mathbf{I}$ is a feasible solution to (4.4), we have $E_d(K) \geq \lambda_{\min}(\mathbf{Q})S_K(\{c_n^2\})$ and $K_d \leq \overline{K}$ from the definition of \overline{K} in (4.7b).

To obtain the bound on the ratio $\overline{K}/\underline{K}$, we infer from the definition of \underline{K} in (4.7a) that $\lambda_{\max}(\mathbf{Q}/\mathbf{Q}_{Y(\underline{K}+1)Y(\underline{K}+1)})S_{\underline{K}+1}(\{c_n^2\}) > \gamma$. The left-hand side of this inequality can be bounded from above as follows:

$$\lambda_{\max}(\mathbf{Q}/\mathbf{Q}_{Y(\underline{K}+1)Y(\underline{K}+1)})S_{\underline{K}+1}(\{c_n^2\}) \leq \lceil k \rceil \lambda_{\min}(\mathbf{Q}) \frac{S_{\underline{K}+1}(\{c_n^2\})}{\underline{K} + 1} \leq \lambda_{\min}(\mathbf{Q}) S_{\lceil k \rceil}(\{c_n^2\}), \quad (4.10)$$

where $k = (\underline{K} + 1)\lambda_{\max}(\mathbf{Q}/\mathbf{Q}_{Y(\underline{K}+1)Y(\underline{K}+1)})/\lambda_{\min}(\mathbf{Q}) \geq \underline{K} + 1$. The last inequality in (4.10) is due to the fact that the mean of the smallest elements in a sequence is non-decreasing when a larger number of elements is included. From the inequality $\lambda_{\min}(\mathbf{Q})S_{\lceil k \rceil}(\{c_n^2\}) > \gamma$ and the definition of \overline{K} in (4.7b), we conclude that $\overline{K} \leq \lceil k \rceil - 1$. \square

In the limit of large \underline{K} , the bound on the approximation ratio K_d/K^* in Theorem 4.2 is approximately equal to the eigenvalue ratio $\lambda_{\max}(\mathbf{Q}/\mathbf{Q}_{Y(\underline{K}+1)Y(\underline{K}+1)})/\lambda_{\min}(\mathbf{Q})$, which can be regarded as a type of condition number. This eigenvalue ratio is in turn bounded from above by the conventional condition number $\kappa(\mathbf{Q}) = \lambda_{\max}(\mathbf{Q})/\lambda_{\min}(\mathbf{Q})$ [17], thus linking approximation quality in terms of cardinality to the geometric approximation quality illustrated in Fig. 4.2.

Theorem 4.2 can be strengthened somewhat by exploiting an invariance property of problem (1.1) and its diagonal relaxation. It is straightforward to see that the optimal cost in (1.1) (and hence K^*) is invariant to diagonal scaling transformations of the feasible set, i.e., transformations parameterized by an invertible diagonal matrix \mathbf{S} mapping \mathbf{c} to $\mathbf{S}\mathbf{c}$ and \mathbf{Q} to $\mathbf{S}^{-1}\mathbf{Q}\mathbf{S}^{-1}$. Likewise, the optimal cost $E_d(K)$ in (4.4) can be shown to be invariant to the same transformations, and thus K_d is invariant [26]. By generalizing the definitions of \underline{K} and \overline{K} , Theorem 4.2 can be generalized as follows:

Corollary 4.3 *For any invertible diagonal matrix \mathbf{S} , define $Y_{\mathbf{S}}(K)$ to be the index set corresponding to the $N - K$ largest $S_{nn}c_n^2$ and*

$$\begin{aligned}\underline{K}_{\mathbf{S}} &= \max \{K : \lambda_{\max}((\mathbf{S}^{-1}\mathbf{Q}\mathbf{S}^{-1})/(\mathbf{S}^{-1}\mathbf{Q}\mathbf{S}^{-1})_{Y_{\mathbf{S}}(K)Y_{\mathbf{S}}(K)})S_K(\{S_{nn}c_n^2\}) \leq \gamma\}, \\ \overline{K}_{\mathbf{S}} &= \max \{K : \lambda_{\min}(\mathbf{S}^{-1}\mathbf{Q}\mathbf{S}^{-1})S_K(\{S_{nn}c_n^2\}) \leq \gamma\}.\end{aligned}$$

Then Theorem 4.2 holds with \mathbf{Q} , \underline{K} , \overline{K} , and $Y(K)$ replaced by $\mathbf{S}^{-1}\mathbf{Q}\mathbf{S}^{-1}$, $\underline{K}_{\mathbf{S}}$, $\overline{K}_{\mathbf{S}}$, and $Y_{\mathbf{S}}(K)$ respectively.

The scaling matrix \mathbf{S} can be chosen to minimize the eigenvalue ratio in Theorem 4.2, i.e., as a type of optimal diagonal preconditioner for \mathbf{Q} , thus minimizing the bound on the approximation ratio.

The bounds in Theorem 4.2 are essentially tight. Specifically, it is shown that for $N \geq 5$, the inequalities $\underline{K} \leq K^*$ and $K_d \leq \overline{K}$ can be simultaneously tight so that the left-hand inequality in (4.8) is met with equality, while the right-hand inequality reduces to $\overline{K}/\underline{K} \leq (\overline{K} + 1)/\underline{K}$ and is asymptotically tight as $\underline{K} \rightarrow \infty$. We consider again the first class of instances constructed in Sect. 3.2 in which $\mathbf{c} = \mathbf{e}$, $\gamma = 1$, and the eigenvector \mathbf{v} corresponding to the smallest eigenvalue of \mathbf{Q} has $\lceil N/2 \rceil$ components equal to $+1/\sqrt{N}$ and $\lfloor N/2 \rfloor$ components equal to $-1/\sqrt{N}$. We keep $\lambda_1 = 1/N$ and change λ_2 to $\lambda_2 = 1/(2\lceil N/2 \rceil - \lfloor \sqrt{N} \rfloor - 1)$. Given these choices, (4.7b) yields $\overline{K} = N = 1/\lambda_1$, while from (3.8) we have $\lambda_{\max}(\mathbf{Q}/\mathbf{Q}_{Y(K)Y(K)}) = \lambda_2$ and hence $\underline{K} = 1/\lambda_2 = 2\lceil N/2 \rceil - \lfloor \sqrt{N} \rfloor - 1$ from (4.7a). It can then be verified through substitution that the rightmost quantity in (4.8) is equal to $(\overline{K} + 1)/\underline{K}$ for $N \geq 5$ as claimed. Furthermore, the construction satisfies the assumptions of Lemma 4.1 and thus $E_d(K) = K\lambda_{\min}(\mathbf{Q}) = K/N$, from which it follows that $K_d = N = \overline{K}$.

It remains to show that $\underline{K} = K^*$ for this class of instances. This is equivalent to showing that condition (2.3) is violated for $K = \underline{K} + 1$. Substituting (3.8) and the chosen parameter values into (2.3) and performing some simplifications, the required condition $E_0(\underline{K} + 1) > \gamma$ is equivalent to

$$(\underline{K} + 1)(\kappa - 1) \max_{|Z|=\underline{K}+1} (\mathbf{e}^T \hat{\mathbf{v}}_Z)^2 < (\underline{K} + 1)(\kappa - 1) + N, \quad (4.11)$$

where $\kappa = \lambda_2/\lambda_1 = \overline{K}/\underline{K}$. As was the case in (3.9), the maximum in (4.11) is achieved by including in Z all $\lceil N/2 \rceil$ positive components of \mathbf{v} , with the remaining components

being negative. Noting that $\underline{K} + 1 = 2\lceil N/2 \rceil - \lfloor \sqrt{N} \rfloor \geq \lceil N/2 \rceil$ for $N \geq 5$, the maximum value can be seen to be $\lfloor \sqrt{N} \rfloor^2 / (\underline{K} + 1)$. Condition (4.11) then becomes

$$\left(2\lceil N/2 \rceil - \lfloor \sqrt{N} \rfloor\right)^2 (\kappa - 1) + N - \lfloor \sqrt{N} \rfloor (\kappa - 1) > 0,$$

which is true given that $1 < \kappa \leq 2$ for $N \geq 5$.

Theorem 4.2 and Corollary 4.3 characterize the approximation quality of the diagonal relaxation in terms of extreme eigenvalues, specifically the smallest eigenvalue of \mathbf{Q} and the largest eigenvalue of a Schur complement of \mathbf{Q} . A second characterization involving intermediate eigenvalues can be obtained under the stochastic assumption that the eigenvectors of \mathbf{Q} are chosen as an orthonormal set uniformly at random from the unit sphere. This assumption allows the bound on $E_0(K)$ in (4.9) to be improved, essentially replacing the largest eigenvalue of $\mathbf{Q}/\mathbf{Q}_{Y(\underline{K}+1)Y(\underline{K}+1)}$ with the mean eigenvalue of \mathbf{Q} , $\bar{\lambda}(\mathbf{Q}) = \frac{1}{N} \sum_{n=1}^N \lambda_n(\mathbf{Q})$. By retaining the other elements in the proof of Theorem 4.2, we obtain the following bound on the approximation ratio, which holds with high probability as N becomes large.

Theorem 4.4 *Let the matrix \mathbf{V} of eigenvectors of \mathbf{Q} be drawn uniformly at random from the set of $N \times N$ orthogonal matrices. Then the approximation ratio K_d/K^* is bounded from above by*

$$\frac{\lceil (\underline{K} + 1)(1 + \varepsilon) \bar{\lambda}(\mathbf{Q}) / \lambda_{\min}(\mathbf{Q}) \rceil - 1}{\underline{K}}$$

with probability at least

$$\begin{cases} 1 - \exp\left(-\frac{N}{8} \frac{\varepsilon^2 \bar{\lambda}(\mathbf{Q})^2}{\varepsilon^2 \bar{\lambda}(\mathbf{Q})^2 + \text{var}(\lambda(\mathbf{Q}))}\right), & \varepsilon \in (0, \varepsilon_{\max}) \setminus \mathcal{I}, \\ 1 - \exp\left(-\frac{N}{8} \frac{\varepsilon \bar{\lambda}(\mathbf{Q})}{\varepsilon \bar{\lambda}(\mathbf{Q}) + (\lambda_{\max}(\mathbf{Q}) - \bar{\lambda}(\mathbf{Q}))}\right), & \varepsilon \in \mathcal{I}, \\ 1, & \varepsilon \geq \varepsilon_{\max}, \end{cases} \quad (4.12)$$

where $\text{var}(\lambda(\mathbf{Q})) = \frac{1}{N} \sum_{n=1}^N (\lambda_n(\mathbf{Q}) - \bar{\lambda}(\mathbf{Q}))^2$ is the variance of the eigenvalues of \mathbf{Q} , $\varepsilon_{\max} = \lambda_{\max}(\mathbf{Q}) / \bar{\lambda}(\mathbf{Q}) - 1$,

$$\mathcal{I} = \begin{cases} (\varepsilon_-, \varepsilon_+), & (\lambda_{\max}(\mathbf{Q}) - \bar{\lambda}(\mathbf{Q}))^2 > 8 \text{var}(\lambda(\mathbf{Q})), \\ \emptyset, & (\lambda_{\max}(\mathbf{Q}) - \bar{\lambda}(\mathbf{Q}))^2 \leq 8 \text{var}(\lambda(\mathbf{Q})), \end{cases} \quad (4.13)$$

and

$$\varepsilon_{\pm} = \frac{1}{4} \left(\varepsilon_{\max} \pm \sqrt{\varepsilon_{\max}^2 - 8 \frac{\text{var}(\lambda(\mathbf{Q}))}{\bar{\lambda}(\mathbf{Q})^2}} \right).$$

Proof As noted above, it suffices to replace the bound in (4.9) with

$$E_0(K) \leq (1 + \varepsilon) \bar{\lambda}(\mathbf{Q}) S_K(\{c_n^2\}) \quad (4.14)$$

and show that (4.14) holds with the probabilities indicated in the theorem statement. The remainder of the proof proceeds as in Theorem 4.2. First note that for $\varepsilon \geq \varepsilon_{\max}$, (4.14) is implied by (4.9) and is therefore true with probability 1. For $\varepsilon \in (0, \varepsilon_{\max})$,

we use an upper bound on $E_0(\mathbf{K})$ to bound the probability that (4.14) is violated. Choosing the same subset $Z(\mathbf{K})$ as in Theorem 4.2 and using the definition of the Schur complement, we have

$$E_0(\mathbf{K}) \leq \mathbf{c}_{Z(\mathbf{K})}^T \mathbf{Q}_{Z(\mathbf{K})Z(\mathbf{K})} \mathbf{c}_{Z(\mathbf{K})} = \tilde{\mathbf{c}}^T \mathbf{\Lambda} \tilde{\mathbf{c}}, \quad \tilde{\mathbf{c}} = \mathbf{V}^T \begin{bmatrix} \mathbf{c}_{Z(\mathbf{K})} \\ \mathbf{0} \end{bmatrix}, \quad (4.15)$$

where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues of \mathbf{Q} . The assumption on \mathbf{V} implies that $\tilde{\mathbf{c}}$ is distributed uniformly over the sphere of radius $\sqrt{S_K(\{c_n^2\})}$ centered at the origin. Hence the quantity $\tilde{\mathbf{c}}^T \mathbf{\Lambda} \tilde{\mathbf{c}}$ can be equivalently expressed as $S_K(\{c_n^2\})(\mathbf{z}^T \mathbf{\Lambda} \mathbf{z} / \mathbf{z}^T \mathbf{z})$, where the components of \mathbf{z} are independent standard normal random variables.

We now bound the probability that $\tilde{\mathbf{c}}^T \mathbf{\Lambda} \tilde{\mathbf{c}} > (1 + \varepsilon) \bar{\lambda}(\mathbf{Q}) S_K(\{c_n^2\})$, which in turn bounds the probability that (4.14) is not satisfied. The event in question can be rewritten as

$$S = \sum_{n=1}^N \left[\lambda_n(\mathbf{Q}) - (1 + \varepsilon) \bar{\lambda}(\mathbf{Q}) \right] z_n^2 \equiv \sum_{n=1}^N \delta_n z_n^2 > 0.$$

It can be seen that the expected value of S is equal to $-\varepsilon N \bar{\lambda}(\mathbf{Q})$, and hence we are bounding the probability that a linear combination of independent chi-squared random variables exceeds its mean by $\varepsilon N \bar{\lambda}(\mathbf{Q})$. A straightforward application of the Chernoff bound [8] yields

$$\log \Pr(S > 0) \leq \min_{0 \leq t < 1/(2\delta_{\max})} -\frac{1}{2} \sum_{n=1}^N \log(1 - 2\delta_n t),$$

where $\delta_{\max} = \lambda_{\max}(\mathbf{Q}) - (1 + \varepsilon) \bar{\lambda}(\mathbf{Q})$. To derive a closed-form expression for the Chernoff exponent, the function $-(1/2) \log(1 - 2\delta_n t)$ is bounded from above by the quadratic function $2\delta_n^2 t^2 + \delta_n t$ over the interval $[0, 1/(4\delta_{\max})]$ (this upper bound can be verified by comparing derivatives over $[0, 1/(4\delta_{\max})]$). It follows that

$$\log \Pr(S > 0) \leq N \min_{0 \leq t \leq 1/(4\delta_{\max})} 2 \left(\text{var}(\lambda(\mathbf{Q})) + \varepsilon^2 \bar{\lambda}(\mathbf{Q})^2 \right) t^2 - \varepsilon \bar{\lambda}(\mathbf{Q}) t, \quad (4.16)$$

using the definition of $\text{var}(\lambda(\mathbf{Q}))$. We consider the two cases in which the unconstrained minimizer $t^* = (1/4) \varepsilon \bar{\lambda}(\mathbf{Q}) / (\text{var}(\lambda(\mathbf{Q})) + \varepsilon^2 \bar{\lambda}(\mathbf{Q})^2)$ is either less than or greater than $1/(4\delta_{\max})$. These correspond to the first two cases in (4.12). In the first case, substituting $t = t^*$ into (4.16) yields the exponent in (4.12) directly, while in the second case, the exponent in (4.12) results from substituting $t = 1/(4\delta_{\max})$ in (4.16) and then using the assumed inequality $t^* > 1/(4\delta_{\max})$. Solving the boundary condition $t^* = 1/(4\delta_{\max})$ for ε yields the expression in (4.13) for the interval \mathcal{I} . \square

Theorem 4.4 can be significantly less conservative than Theorem 4.2, in particular when most of the eigenvalues are small and comparable so that the mean eigenvalue of \mathbf{Q} is much closer to the minimum eigenvalue than to the maximum eigenvalue. This preference for eigenvalue distributions weighted toward small values is seen in the numerical results in Sect. 5. Furthermore, it agrees with the following geometric intuition: Assuming that the ellipsoid $\mathcal{E}_{\mathbf{Q}}$ is not close to spherical ($\kappa(\mathbf{Q})$ is large),

it is preferable for most of the ellipsoid axes to be comparatively long (corresponding to small eigenvalues) and of the same order. Such an ellipsoid tends to require a smaller coordinate-aligned enclosing ellipsoid, and consequently the diagonal relaxation tends to be a better approximation. For example, in three dimensions, a severely oblate spheroid can be enclosed on average in a smaller coordinate-aligned ellipsoid than an equally severely prolate spheroid. Note also that the exponents in (4.12) depend on the eigenvalue distribution and are larger (i.e., the decay is sharper) when the spread of the eigenvalues is small as measured by $\text{var}(\lambda(\mathbf{Q}))$ or $\lambda_{\max}(\mathbf{Q}) - \bar{\lambda}(\mathbf{Q})$.

4.4 The diagonally dominant case

We now consider the case in which the matrix \mathbf{Q} is diagonally dominant, specifically in the sense that

$$\max_m \sum_{n \neq m} \frac{|Q_{mn}|}{\sqrt{Q_{mm}Q_{nn}}} < 1, \quad (4.17)$$

i.e., the absolute sum of the normalized off-diagonal entries in any row or column is small. It is expected in this case that the original problem (1.1) can be well-approximated by its diagonal relaxation, and that the quality of approximation depends on the degree of diagonal dominance. Indeed, it can be shown that the maximum numbers of zero-valued components in (1.1) and its diagonal relaxation, K^* and K_d respectively, are bounded by the following quantities related to diagonal dominance:

$$\underline{K}_{\text{dd}} = \max \left\{ K : \left(1 + \max_{m \in Z_{\text{dd}}(K)} \sum_{\substack{n \in Z_{\text{dd}}(K) \\ n \neq m}} \frac{|Q_{mn}|}{\sqrt{Q_{mm}Q_{nn}}} \right) S_K(\{Q_{nn}c_n^2\}) \leq \gamma \right\}, \quad (4.18a)$$

$$\bar{K}_{\text{dd}} = \max \left\{ K : \left(1 - \max_m \sum_{n \neq m} \frac{|Q_{mn}|}{\sqrt{Q_{mm}Q_{nn}}} \right) S_K(\{Q_{nn}c_n^2\}) \leq \gamma \right\}, \quad (4.18b)$$

where $Z_{\text{dd}}(K)$ in (4.18a) denotes the index set corresponding to the K smallest $Q_{nn}c_n^2$. A bound on the approximation ratio K_d/K^* follows.

Theorem 4.5 *Assume that the matrix \mathbf{Q} is diagonally dominant in the sense of (4.17). Then the maximum numbers of zero-valued components in problem (1.1) and its diagonal relaxation, K^* and K_d respectively, satisfy the ordering $\underline{K}_{\text{dd}} \leq K^* \leq K_d \leq \bar{K}_{\text{dd}}$, where $\underline{K}_{\text{dd}}$ and \bar{K}_{dd} are defined in (4.18). The approximation ratio K_d/K^* is bounded as follows:*

$$\frac{K_d}{K^*} \leq \frac{\bar{K}_{\text{dd}}}{\underline{K}_{\text{dd}}} \leq \frac{\lceil (\underline{K}_{\text{dd}} + 1)r_{\text{dd}} \rceil - 1}{\underline{K}_{\text{dd}}}, \quad (4.19)$$

where

$$r_{\text{dd}} = \left(1 + \max_{m \in Z_{\text{dd}}(\underline{K}_{\text{dd}}+1)} \sum_{\substack{n \in Z_{\text{dd}}(\underline{K}_{\text{dd}}+1) \\ n \neq m}} \frac{|Q_{mn}|}{\sqrt{Q_{mm}Q_{nn}}} \right) / \left(1 - \max_m \sum_{n \neq m} \frac{|Q_{mn}|}{\sqrt{Q_{mm}Q_{nn}}} \right).$$

The ratio r_{dd} in Theorem 4.5 plays the same role as the eigenvalue ratio in Theorem 4.2. As \mathbf{Q} becomes more diagonally dominant, r_{dd} approaches 1 from above. Unlike with Theorem 4.2, there is no benefit to allowing diagonal scaling transformations because the measure of diagonal dominance used here remains unchanged when \mathbf{Q} is replaced by $\mathbf{S}^{-1}\mathbf{Q}\mathbf{S}^{-1}$.

To prove the inequality $K_d \leq \bar{K}_{\text{dd}}$, we use the following lemma, which specifies the optimal cost of (4.4) under the additional constraint that \mathbf{D} is a multiple of a fixed diagonal matrix.

Lemma 4.6 *For any positive definite diagonal matrix \mathbf{D}_0 , the optimal cost $E_d(K)$ in (4.4) is bounded from below by $\lambda_{\min}(\mathbf{D}_0^{-1/2}\mathbf{Q}\mathbf{D}_0^{-1/2})S_K(\{(\mathbf{D}_0)_{mn}c_n^2\})$.*

Proof We restrict \mathbf{D} in (4.4) to be a multiple of \mathbf{D}_0 , thus obtaining a lower bound on $E_d(K)$. With $\mathbf{D} = \alpha\mathbf{D}_0$, (4.4) reduces to

$$\max_{\alpha} \quad \alpha S_K(\{(\mathbf{D}_0)_{mn}c_n^2\}) \quad \text{s.t.} \quad \mathbf{0} \preceq \alpha\mathbf{D}_0 \preceq \mathbf{Q}.$$

Since \mathbf{D}_0 is invertible, the constraint can be rewritten as $\mathbf{0} \preceq \alpha\mathbf{I} \preceq \mathbf{D}_0^{-1/2}\mathbf{Q}\mathbf{D}_0^{-1/2}$, from which it follows that α should be chosen as the smallest eigenvalue of $\mathbf{D}_0^{-1/2}\mathbf{Q}\mathbf{D}_0^{-1/2}$. \square

We now proceed with the proof of Theorem 4.5.

Proof (Theorem 4.5) To prove that $K_d \leq \bar{K}_{\text{dd}}$, we let $\mathbf{D}_0 = \text{Diag}(\mathbf{Q})$ in Lemma 4.6, where $\text{Diag}(\mathbf{Q})$ denotes a diagonal matrix with the same diagonal entries as \mathbf{Q} . Using the Gershgorin circle theorem [17] to bound the smallest eigenvalue of $\tilde{\mathbf{Q}} = \text{Diag}(\mathbf{Q})^{-1/2}\mathbf{Q}\text{Diag}(\mathbf{Q})^{-1/2}$, we then obtain

$$E_d(K) \geq \left(1 - \max_m \sum_{n \neq m} \frac{|Q_{mn}|}{\sqrt{Q_{mm}Q_{nn}}} \right) S_K(\{Q_{nn}c_n^2\}),$$

from which we infer that $K_d \leq \bar{K}_{\text{dd}}$ based on (4.18b).

To prove that $K^* \geq \underline{K}_{\text{dd}}$, the quantity $E_0(K)$ in (2.3) is bounded from above as follows, starting with the specific choice of subset $Z = Z_{\text{dd}}(K)$:

$$\begin{aligned} E_0(K) &\leq \mathbf{c}_{Z_{\text{dd}}(K)}^T (\mathbf{Q}/\mathbf{Q}_{Y_{\text{dd}}(K)} Y_{\text{dd}}(K)) \mathbf{c}_{Z_{\text{dd}}(K)} \\ &\leq \mathbf{c}_{Z_{\text{dd}}(K)}^T \mathbf{Q}_{Z_{\text{dd}}(K)Z_{\text{dd}}(K)} \mathbf{c}_{Z_{\text{dd}}(K)} \\ &= (\text{Diag}(\mathbf{Q})^{1/2} \mathbf{c})_{Z_{\text{dd}}(K)}^T \tilde{\mathbf{Q}}_{Z_{\text{dd}}(K)Z_{\text{dd}}(K)} (\text{Diag}(\mathbf{Q})^{1/2} \mathbf{c})_{Z_{\text{dd}}(K)} \\ &\leq \lambda_{\max}(\tilde{\mathbf{Q}}_{Z_{\text{dd}}(K)Z_{\text{dd}}(K)}) S_K(\{Q_{nn}c_n^2\}) \\ &\leq \left(1 + \max_{m \in Z_{\text{dd}}(K)} \sum_{\substack{n \in Z_{\text{dd}}(K) \\ n \neq m}} \frac{|Q_{mn}|}{\sqrt{Q_{mm}Q_{nn}}} \right) S_K(\{Q_{nn}c_n^2\}). \end{aligned}$$

The second line follows from the definition of the Schur complement, the third from a rescaling, the fourth from eigenvalue properties and the definition of $Z_{\text{dd}}(K)$, and the last from the Gershgorin circle theorem. Comparing with (4.18a), we conclude that $K^* \geq \underline{K}_{\text{dd}}$. The proof of the bound on $\bar{K}_{\text{dd}}/\underline{K}_{\text{dd}}$ is similar to that in Theorem 4.2. \square

As with Theorem 4.2, there exist instances for which the left-hand bound in (4.19) is tight and the right-hand bound is asymptotically tight. We consider the same class of instances as in Sect. 3.2 with $\mathbf{c} = \mathbf{e}$, $\gamma = 1$, and \mathbf{v} having $\lceil N/2 \rceil$ components equal to $+1/\sqrt{N}$ and $\lfloor N/2 \rfloor$ components equal to $-1/\sqrt{N}$. From (3.6) we obtain $Q_{nn} = (N-1)\lambda_2/N + \lambda_1/N$ for all n and $|Q_{mm}| = (\lambda_2 - \lambda_1)/N$ for all $m \neq n$, from which it follows that

$$\left(1 - \max_m \sum_{n \neq m} \frac{|Q_{mn}|}{\sqrt{Q_{mm}Q_{nn}}}\right) S_K(\{Q_{nn}c_n^2\}) = K\lambda_1,$$

$$\left(1 + \max_{\substack{m \in Z \\ n \in Z \\ n \neq m}} \sum \frac{|Q_{mn}|}{\sqrt{Q_{mm}Q_{nn}}}\right) S_K(\{Q_{nn}c_n^2\}) = K \left(\lambda_2 + \frac{K-2}{N}(\lambda_2 - \lambda_1)\right)$$

for any Z of cardinality K . Choosing $\lambda_1 = 1/N$ and $\lambda_2 = 1/N + 1/((N-1)(2N-3))$, some straightforward calculations yield $\underline{K}_{\text{dd}} = N-1$ and $\overline{K}_{\text{dd}} = N$ from (4.18), and $r_{\text{dd}} = 1 + 2/(2N-3)$ for the ratio defined in Theorem 4.5. It can then be seen that the right-hand inequality in (4.19) reads $N/(N-1) \leq (N+1)/(N-1)$ for $N \geq 3$, which is asymptotically tight as $N \rightarrow \infty$.

To show that the left-hand inequality in (4.19) is tight, we note that the construction satisfies the assumptions of Lemma 4.1 so we again have $E_d(K) = K\lambda_{\min}(\mathbf{Q}) = K/N$ and $K_d = N = \overline{K}_{\text{dd}}$. The remaining required equality $K^* = \underline{K}_{\text{dd}} = N-1$ is equivalent to the all-zero solution being infeasible for (1.1), i.e., $\mathbf{e}^T \mathbf{Q} \mathbf{c} > \gamma = 1$. Using (3.6) and substituting the selected parameter values, we find $\mathbf{e}^T \mathbf{Q} \mathbf{c} = 1 + N/((N-1)(2N-3)) > 1$ for N even and $\mathbf{e}^T \mathbf{Q} \mathbf{c} = 1 + (N+1)/(N(2N-3)) > 1$ for N odd, completing the demonstration.

4.5 The nearly coordinate-aligned case

A geometric analogue to diagonal dominance is the case in which the axes of the ellipsoid $\mathcal{E}_{\mathbf{Q}}$ are nearly aligned with the coordinate axes. Algebraically, this corresponds to the eigenvectors of \mathbf{Q} being close to the standard basis vectors. We assume that \mathbf{Q} is diagonalized as $\mathbf{Q} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, where the eigenvalues $\lambda_n(\mathbf{Q})$ and the eigenvector matrix \mathbf{V} are ordered in such a way that $\mathbf{\Delta} = \mathbf{V} - \mathbf{I}$ is small, specifically in the sense that its spectral radius $\rho(\mathbf{\Delta})$ satisfies $\kappa(\mathbf{Q})\rho(\mathbf{\Delta}) < 1$. It is expected in this case that the diagonal relaxation would give a better approximation for smaller $\mathbf{\Delta}$, i.e., for closer alignments. Following the approach in Sect. 4.3–4.4, it is shown that K^* and K_d may be bounded by

$$\underline{K}_{\text{na}} = \max \{K : (1 + \kappa(\mathbf{Q})(\rho(\mathbf{\Delta}) + \rho(\mathbf{\Delta})^2))S_K(\{\lambda_n(\mathbf{Q})c_n^2\}) \leq \gamma\}, \quad (4.20a)$$

$$\overline{K}_{\text{na}} = \max \{K : (1 - \kappa(\mathbf{Q})\rho(\mathbf{\Delta}))S_K(\{\lambda_n(\mathbf{Q})c_n^2\}) \leq \gamma\}. \quad (4.20b)$$

The approximation ratio K_d/K^* may be bounded accordingly.

Theorem 4.7 *Assume that the matrix \mathbf{Q} can be diagonalized as $\mathbf{Q} = (\mathbf{I} + \mathbf{\Delta})\mathbf{\Lambda}(\mathbf{I} + \mathbf{\Delta})^T$, where $\mathbf{\Delta}$ is such that $\kappa(\mathbf{Q})\rho(\mathbf{\Delta}) < 1$. Then the maximum numbers of zero-valued*

components in (1.1) and its diagonal relaxation, K^* and K_d respectively, satisfy the ordering $\underline{K}_{\text{na}} \leq K^* \leq K_d \leq \overline{K}_{\text{na}}$, where $\underline{K}_{\text{na}}$ and \overline{K}_{na} are defined in (4.20). The approximation ratio K_d/K^* is bounded as follows:

$$\frac{K_d}{K^*} \leq \frac{\overline{K}_{\text{na}}}{\underline{K}_{\text{na}}} \leq \frac{\lceil (\underline{K}_{\text{na}} + 1)r_{\text{na}} \rceil - 1}{\underline{K}_{\text{na}}}, \quad (4.21)$$

where

$$r_{\text{na}} = \frac{1 + \kappa(\mathbf{Q})(\rho(\mathbf{\Delta}) + \rho(\mathbf{\Delta})^2)}{1 - \kappa(\mathbf{Q})\rho(\mathbf{\Delta})}.$$

Theorem 4.7 characterizes the quality of approximation in terms of the ratio r_{na} . As $\mathbf{\Delta}$ approaches $\mathbf{0}$, r_{na} approaches 1 as expected. Similar to Theorem 4.2, Theorem 4.7 may be strengthened using diagonal scaling transformations since both $\rho(\mathbf{\Delta})$ and the condition number $\kappa(\mathbf{Q})$ may decrease as \mathbf{Q} is transformed into $\mathbf{S}^{-1}\mathbf{Q}\mathbf{S}^{-1}$ for different choices of \mathbf{S} . The dependence on the condition number can be explained geometrically as illustrated in Fig. 4.3. On the left, the original ellipsoid $\mathcal{E}_{\mathbf{Q}}$ is both nearly coordinate-aligned and nearly spherical (i.e., $\kappa(\mathbf{Q})$ is close to 1), and can therefore be enclosed by a coordinate-aligned ellipsoid that is only slightly larger. Indeed in the limit $\kappa(\mathbf{Q}) = 1$, $\mathcal{E}_{\mathbf{Q}}$ is spherical and thus already coordinate-aligned, and the eigenvector matrix \mathbf{V} can be chosen equal to \mathbf{I} resulting in $\mathbf{\Delta} = \mathbf{0}$. On the other hand, if $\kappa(\mathbf{Q})$ is large, even a small misalignment between the ellipsoid and coordinate axes results in a much larger enclosing ellipsoid, as seen on the right in Fig. 4.3.

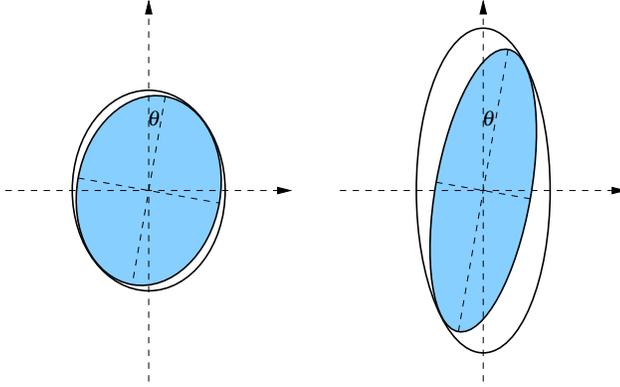


Fig. 4.3 The effect of the condition number $\kappa(\mathbf{Q})$ on the approximation quality in the nearly coordinate-aligned case. For the same angular offset θ between the axes of the original ellipsoid and the coordinate axes, the coordinate-aligned enclosing ellipsoid on the right is comparatively larger.

In the proof of Theorem 4.7 below, we make reference to the scaled matrix $\mathbf{\Lambda}^{-1/2}\mathbf{Q}\mathbf{\Lambda}^{-1/2}$. When $\mathbf{\Delta}$ is small, $\mathbf{\Lambda}^{-1/2}\mathbf{Q}\mathbf{\Lambda}^{-1/2}$ is close to the identity matrix and the deviation of its eigenvalues from 1 is specified by the following lemma.

Lemma 4.8 Assume that the matrix \mathbf{Q} can be diagonalized as $\mathbf{Q} = (\mathbf{I} + \mathbf{\Delta})\mathbf{\Lambda}(\mathbf{I} + \mathbf{\Delta})^T$, where $\mathbf{\Delta}$ is such that $\kappa(\mathbf{Q})\rho(\mathbf{\Delta}) < 1$. Then

$$\begin{aligned}\lambda_{\min}(\mathbf{\Lambda}^{-1/2}\mathbf{Q}\mathbf{\Lambda}^{-1/2}) &\geq 1 - \kappa(\mathbf{Q})\rho(\mathbf{\Delta}), \\ \lambda_{\max}(\mathbf{\Lambda}^{-1/2}\mathbf{Q}\mathbf{\Lambda}^{-1/2}) &\leq 1 + \kappa(\mathbf{Q})(\rho(\mathbf{\Delta}) + \rho(\mathbf{\Delta})^2).\end{aligned}$$

Proof Expanding $\mathbf{\Lambda}^{-1/2}\mathbf{Q}\mathbf{\Lambda}^{-1/2}$ yields $\mathbf{I} + \tilde{\mathbf{\Delta}} + \tilde{\mathbf{\Delta}}^T + \tilde{\mathbf{\Delta}}\tilde{\mathbf{\Delta}}^T$, where $\tilde{\mathbf{\Delta}} = \mathbf{\Lambda}^{-1/2}\mathbf{\Delta}\mathbf{\Lambda}^{1/2}$. The eigenvalues of $\mathbf{\Lambda}^{-1/2}\mathbf{Q}\mathbf{\Lambda}^{-1/2}$ can then be bounded by

$$\lambda_{\min}(\mathbf{\Lambda}^{-1/2}\mathbf{Q}\mathbf{\Lambda}^{-1/2}) \geq 1 + \lambda_{\min}(\tilde{\mathbf{\Delta}} + \tilde{\mathbf{\Delta}}^T), \quad (4.22a)$$

$$\lambda_{\max}(\mathbf{\Lambda}^{-1/2}\mathbf{Q}\mathbf{\Lambda}^{-1/2}) \leq 1 + \lambda_{\max}(\tilde{\mathbf{\Delta}} + \tilde{\mathbf{\Delta}}^T) + \lambda_{\max}(\tilde{\mathbf{\Delta}}\tilde{\mathbf{\Delta}}^T), \quad (4.22b)$$

noting that $\tilde{\mathbf{\Delta}}\tilde{\mathbf{\Delta}}^T$ is positive semidefinite in (4.22a). The rightmost term in (4.22b) can be bounded using the sub-multiplicative property of the spectral norm [17]:

$$\lambda_{\max}(\tilde{\mathbf{\Delta}}\tilde{\mathbf{\Delta}}^T) = \|\tilde{\mathbf{\Delta}}\|_2^2 \leq \|\mathbf{\Lambda}^{1/2}\|_2^2 \|\mathbf{\Delta}\|_2^2 \|\mathbf{\Lambda}^{-1/2}\|_2^2 = \lambda_{\max}(\mathbf{Q})\rho(\mathbf{\Delta})^2 \lambda_{\min}^{-1}(\mathbf{Q}) = \kappa(\mathbf{Q})\rho(\mathbf{\Delta})^2.$$

To bound the eigenvalues of $\tilde{\mathbf{\Delta}} + \tilde{\mathbf{\Delta}}^T$, we make use of a diagonalization of $\tilde{\mathbf{\Delta}}$. Given that \mathbf{V} is orthogonal, it has unit-modulus eigenvalues and can be diagonalized by a unitary matrix \mathbf{U} . From the relations $\mathbf{\Delta} = \mathbf{V} - \mathbf{I}$ and $\rho(\mathbf{\Delta}) < 1/\kappa(\mathbf{Q})$, we see that $\mathbf{\Delta}$ can be diagonalized as $\mathbf{\Delta} = \mathbf{U}\Psi\mathbf{U}^H$, where the eigenvalues ψ_n of $\mathbf{\Delta}$ lie on the highlighted arc in Fig. 4.4. It follows that $\tilde{\mathbf{\Delta}} = \tilde{\mathbf{U}}\Psi\tilde{\mathbf{U}}^{-1}$ with $\tilde{\mathbf{U}} = \mathbf{\Lambda}^{-1/2}\mathbf{U}$.

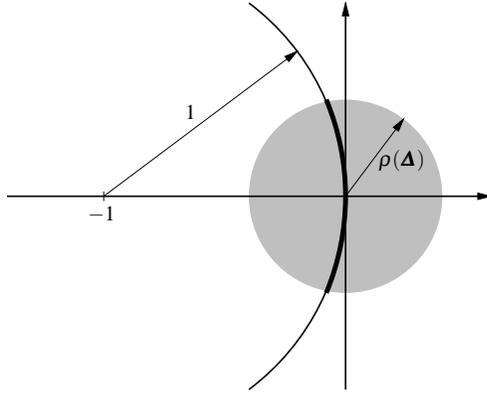


Fig. 4.4 The dark segment of the arc indicates the set of possible locations in the complex plane for the eigenvalues of $\mathbf{\Delta}$ and $\tilde{\mathbf{\Delta}}$.

We now invoke a theorem from [17], which states that for any eigenvalue of $\tilde{\mathbf{\Delta}} + \tilde{\mathbf{\Delta}}^T$, there exists an eigenvalue of $\tilde{\mathbf{\Delta}}$ such that $|\lambda(\tilde{\mathbf{\Delta}} + \tilde{\mathbf{\Delta}}^T) - \lambda(\tilde{\mathbf{\Delta}})| \leq \|\tilde{\mathbf{U}}^{-1}\tilde{\mathbf{\Delta}}^T\tilde{\mathbf{U}}\|_2$.

Expanding the right-hand side of this inequality and using the sub-multiplicative property of spectral norms, we obtain

$$\begin{aligned} \left| \lambda(\tilde{\mathbf{A}} + \tilde{\mathbf{A}}^T) - \lambda(\tilde{\mathbf{A}}) \right| &\leq \left\| \mathbf{U}^H \mathbf{\Lambda}^{1/2} \mathbf{\Lambda}^{1/2} \mathbf{\Delta}^T \mathbf{\Lambda}^{-1/2} \mathbf{\Lambda}^{-1/2} \mathbf{U} \right\|_2 \\ &\leq \|\mathbf{U}^H\|_2 \|\mathbf{\Lambda}\|_2 \|\mathbf{\Delta}^T\|_2 \|\mathbf{\Lambda}^{-1}\|_2 \|\mathbf{U}\|_2 \\ &= \kappa(\mathbf{Q})\rho(\mathbf{\Delta}). \end{aligned} \quad (4.23)$$

The bound in (4.23) constrains the eigenvalues of $\tilde{\mathbf{A}} + \tilde{\mathbf{A}}^T$ to lie within a Euclidean distance of $\kappa(\mathbf{Q})\rho(\mathbf{\Delta})$ from the arc in Fig. 4.4. Furthermore, the symmetry of $\tilde{\mathbf{A}} + \tilde{\mathbf{A}}^T$ implies that its eigenvalues are real-valued. It is clear then that $\lambda_{\max}(\tilde{\mathbf{A}} + \tilde{\mathbf{A}}^T) \leq \kappa(\mathbf{Q})\rho(\mathbf{\Delta})$. From Fig. 4.4 and the assumption that $\kappa(\mathbf{Q})\rho(\mathbf{\Delta}) < 1$, it can also be seen that $\lambda_{\min}(\tilde{\mathbf{A}} + \tilde{\mathbf{A}}^T)$ is minimized by setting $\lambda(\tilde{\mathbf{A}}) = 0$ in (4.23) since all other choices for $\lambda(\tilde{\mathbf{A}})$ would yield more positive values for $\lambda_{\min}(\tilde{\mathbf{A}} + \tilde{\mathbf{A}}^T)$. Substituting the resulting bound $\lambda_{\min}(\tilde{\mathbf{A}} + \tilde{\mathbf{A}}^T) \geq -\kappa(\mathbf{Q})\rho(\mathbf{\Delta})$ into (4.22a) completes the proof. \square

Theorem 4.7 can now be proved straightforwardly using previous results.

Proof (Theorem 4.7) As in the proof of Theorem 4.5, we use Lemma 4.6 to show that $K_d \leq \bar{K}_{\text{na}}$, this time choosing $\mathbf{D}_0 = \mathbf{\Lambda}$. Combining Lemma 4.6 with Lemma 4.8 then yields $E_d(K) \geq (1 - \kappa(\mathbf{Q})\rho(\mathbf{\Delta}))S_K(\{\lambda_n(\mathbf{Q})c_n^2\})$, which implies that $K_d \leq \bar{K}_{\text{na}}$ in light of (4.20b).

To prove that $K^* \geq \underline{K}_{\text{na}}$, we proceed as in the proof of Theorem 4.5 by fixing a specific subset $Z_{\text{na}}(K)$ corresponding to the K smallest $\lambda_n(\mathbf{Q})c_n^2$. This yields

$$\begin{aligned} E_0(K) &\leq \mathbf{c}_{Z_{\text{na}}(K)}^T \mathbf{Q}_{Z_{\text{na}}(K)Z_{\text{na}}(K)} \mathbf{c}_{Z_{\text{na}}(K)} \\ &= \left(\mathbf{\Lambda}^{1/2} \begin{bmatrix} \mathbf{c}_{Z_{\text{na}}(K)} \\ \mathbf{0} \end{bmatrix} \right)^T \mathbf{\Lambda}^{-1/2} \mathbf{Q} \mathbf{\Lambda}^{-1/2} \left(\mathbf{\Lambda}^{1/2} \begin{bmatrix} \mathbf{c}_{Z_{\text{na}}(K)} \\ \mathbf{0} \end{bmatrix} \right) \\ &\leq \lambda_{\max}(\mathbf{\Lambda}^{-1/2} \mathbf{Q} \mathbf{\Lambda}^{-1/2}) S_K(\{\lambda_n(\mathbf{Q})c_n^2\}) \\ &\leq (1 + \kappa(\mathbf{Q})(\rho(\mathbf{\Delta}) + \rho(\mathbf{\Delta})^2)) S_K(\{\lambda_n(\mathbf{Q})c_n^2\}). \end{aligned}$$

In the second line above, the quadratic form has been rewritten in terms of the full matrix \mathbf{Q} and then rescaled. The last two lines result from the definition of $Z_{\text{na}}(K)$ and Lemma 4.8. Combining the last inequality with (4.20a) yields $K^* \geq \underline{K}_{\text{na}}$ as desired. The proof of the bound on $\bar{K}_{\text{na}}/\underline{K}_{\text{na}}$ is similar to that in Theorem 4.2. \square

5 Numerical evaluation

In this section, numerical results are presented to illustrate the performance of the two relaxations discussed in Sect. 3 and Sect. 4. In Sect. 5.1, the relaxations are compared on the basis of their approximation ratios under different conditions. In Sect. 5.2, the relaxations are incorporated in a branch-and-bound algorithm to gauge their effectiveness in reducing the complexity of solving problem (1.1).

5.1 Approximation ratios

Randomly generated instances of problem (1.1) are used in this section to evaluate the approximation quality of the two relaxations. While it was seen in Sect. 4.2 that neither relaxation dominates the other over all possible instances, the present comparison using random instances indicates that diagonal relaxations yield significantly stronger bounds in many situations, including but not limited to those analyzed in Sect. 4.3–4.5.

In these experiments, the problem dimension N is varied between 10 and 150 and the parameter γ is normalized to 1 throughout. The continuous relaxation of each instance, and more specifically the dual problem (3.10), is solved using the MATLAB function `fmincon`. A customized solver described in [26, Sect. 3.5] is used for the diagonal relaxation; a general-purpose semidefinite optimization solver such as SDPT3 [24] or SeDuMi [23] could also be used. In addition, a feasible solution is obtained for each instance using the backward greedy selection method in [28]. The cost of this feasible solution is used as a substitute for the true optimal cost, which is difficult to compute given the large number of instances. Numerical experience in [28] however suggests that backward greedy selection is often optimal. The approximation quality of each relaxation is measured by the ratio of the optimal cost of the relaxation to the cost of the feasible solution. These ratios are denoted R_c and R_d for continuous and diagonal relaxations respectively; they are lower bounds on the true approximation ratios. Note that we are returning to the original definition of approximation ratio in terms of the number of non-zero components and not the number of zero-valued components as in Sect. 4.3–4.5.

In the first three experiments, the eigenvector matrix \mathbf{V} of \mathbf{Q} is chosen uniformly from the set of $N \times N$ orthogonal matrices (as assumed in Theorem 4.4). The eigenvalues are drawn from different power-law distributions and then rescaled to match a specified condition number $\kappa(\mathbf{Q})$ chosen from the values \sqrt{N} , N , $10N$, and $100N$. Once \mathbf{Q} is fixed, each component of the ellipsoid center \mathbf{c} is drawn uniformly from the interval $[-\sqrt{(\mathbf{Q}^{-1})_{mm}}, \sqrt{(\mathbf{Q}^{-1})_{mm}}]$, in keeping with Assumption 2.1.

Fig. 5.1(a) plots the approximation ratios R_c and R_d as functions of N and $\kappa(\mathbf{Q})$ for an eigenvalue distribution proportional to $1/\lambda$, which corresponds to a uniform distribution for $\log \lambda$. Each point represents the average of 1000 instances. A $1/\lambda$ eigenvalue distribution is unbiased in the sense that it is invariant under matrix inversion (up to a possible overall scaling), an operation that maps the positive definite cone to itself. The continuous relaxation approximation ratio R_c does not vary much with N or $\kappa(\mathbf{Q})$. In contrast, the diagonal relaxation approximation ratio R_d is markedly higher for lower $\kappa(\mathbf{Q})$, in agreement with Theorem 4.2 and the geometric intuition in Fig. 4.2. Moreover, R_d improves with increasing N so that even for $\kappa(\mathbf{Q}) = 100N$ the diagonal relaxation outperforms the continuous relaxation for $N \geq 20$, with the difference being substantial at large N .

Figs. 5.1(b) and 5.1(c) show average approximation ratios for a uniform eigenvalue distribution and a $1/\lambda^2$ distribution, the latter corresponding to a uniform distribution for the eigenvalues of \mathbf{Q}^{-1} . Compared to a $1/\lambda$ distribution, a $1/\lambda^2$ distribution is more heavily weighted toward small values whereas a uniform distribution is less so. Accordingly, each R_d curve in Fig. 5.1(b) is lower than its counterpart in

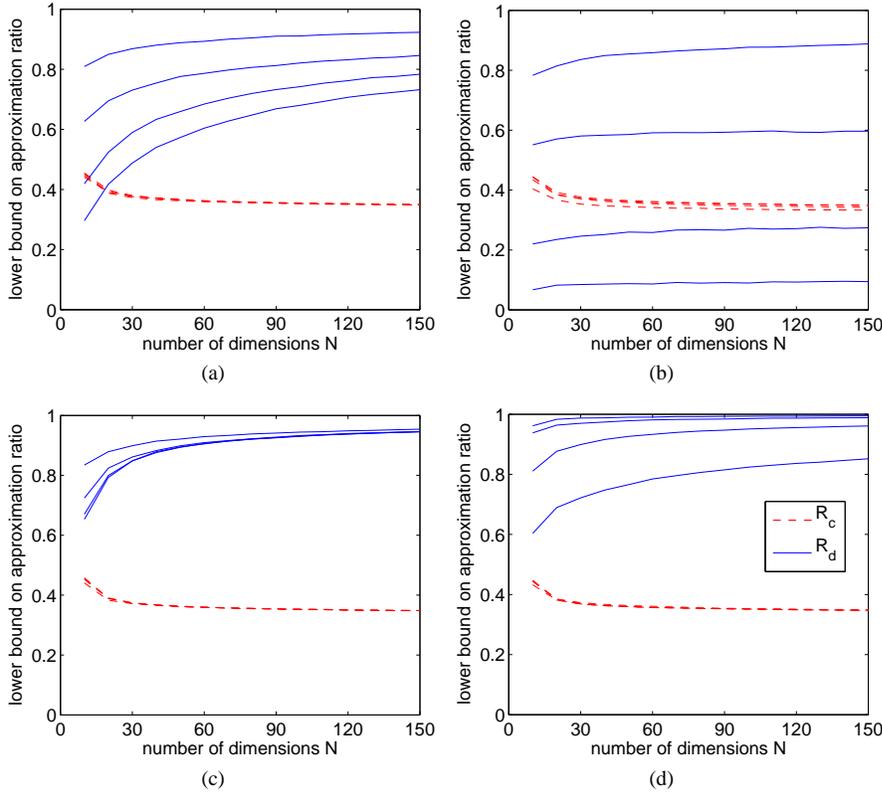


Fig. 5.1 Average approximation ratios R_c and R_d for (a) a $1/\lambda$ eigenvalue distribution, (b) a uniform eigenvalue distribution, (c) a $1/\lambda^2$ eigenvalue distribution, and (d) unit diagonal entries and off-diagonal entries drawn uniformly from $[-a, a]/\sqrt{N}$. In (a)–(c), $\kappa(\mathbf{Q}) = \sqrt{N}, N, 10N, 100N$ from top to bottom within each set of curves. In (d), $a = 0.1, 0.2, 0.5, 0.8$ from top to bottom within each set of curves.

5.1(a) while the opposite is true in Fig. 5.1(c), in agreement with the dependence on the eigenvalue distribution in Theorem 4.4. The effect of the condition number on R_d is also more pronounced under a uniform eigenvalue distribution and less so under a $1/\lambda^2$ distribution. The behavior of R_c on the other hand is largely unchanged from Fig. 5.1(a).

In a fourth experiment, the diagonal entries of \mathbf{Q} are normalized to 1 while the off-diagonal entries are drawn uniformly from the interval $[-a, a]/\sqrt{N}$, where $a = 0.1, 0.2, 0.5, 0.8$. With high probability, such matrices are diagonally dominant in the sense of (4.17) for $a = 0.1, 0.2$, and are not positive definite for $a > 0.85$. The vector \mathbf{c} is generated as before based on the diagonal entries of \mathbf{Q}^{-1} . The average approximation ratios are shown in Fig. 5.1(d). Similar to the condition number in Figs. 5.1(a)–(c), the parameter a does not appear to have much effect on R_c . For the diagonal relaxation, while Theorem 4.5 predicts a close approximation for $a = 0.1, 0.2$, the performance is still relatively good for $a = 0.8$.

The results in Fig. 5.1 demonstrate that better bounds are achieved in many instances with diagonal relaxations than with continuous relaxations. Furthermore, this can be true even when the condition number $\kappa(\mathbf{Q})$ or the off-diagonal amplitude a is high, whereas the analysis in Sect. 4.3–4.5 is more conservative.

5.2 Branch-and-bound complexity

Next we consider the effect of the two relaxations on the complexity of a branch-and-bound solution to (1.1). For this purpose, the relaxations are incorporated into a basic MATLAB implementation of branch-and-bound, referred to as BB. This algorithm is also compared to the mixed-integer programming solver CPLEX 12.4 [18] as a point of reference. The comparisons show that diagonal relaxations can significantly increase the efficiency of branch-and-bound. It is also seen that a more specialized solver can outperform a sophisticated general-purpose solver in solving (1.1).

Algorithm BB is based on the mixed integer formulation (3.1) and is summarized below. Full details can be found in [27]. The branching rule is to select the variable for which the margin in condition (2.4) is minimal. This rule is similar to the maximum absolute value rule in [2, 4] in that the $i_n = 0$ subproblem is more likely to be severely constrained. The next node is chosen according to the “best node” rule, i.e., a node with a minimal lower bound. Feasible solutions are generated by running the backward selection heuristic at every node. To obtain lower bounds, condition (2.4) is checked at every node and bounds are updated as appropriate. Variable elimination as described in Sect. 2.2 is employed to reduce subproblem dimensions. For stronger lower bounds, either continuous or diagonal relaxations are solved, corresponding to two algorithm variants BB-C and BB-D. Relaxations are solved only after constraining a variable to zero ($i_n = 0$ branch) and when the subproblem dimension is at least 20. In other cases, the increased computation does not seem to be justified by the improvement in bounds.

For CPLEX, the split-variable mixed integer formulation corresponding to (3.4) and (3.5) is passed to the CPLEX MEX executable through the provided MATLAB interface. Because of the relative inefficiency of CPLEX as seen below, BB is run first and the optimal solution is used to initialize CPLEX. Given this initialization, CPLEX is instructed to emphasize optimality rather than feasibility, while all other options are set to their default values. Preliminary experimentation with changing solver parameters did not yield any gains. The experiments are run on a 2.4 GHz quad-core Linux computer with 8 GB of memory. BB is generally not observed to use more than one core at a time; CPLEX however is able to continuously exploit all four cores.

Problem instances are generated randomly from the same four classes and in the same manner as in Sect. 5.1, thus satisfying Assumption 2.1 in particular. Table 5.1 shows the solution times and numbers of nodes for the first three classes in which the eigenvalues of \mathbf{Q} are drawn from different distributions. Each entry represents the average over 100 instances. For certain instance classes and solvers, the high computational complexity does not permit an accurate evaluation. In these cases, the

solution time is estimated by extrapolating from lower values of N ; such estimates are marked by parentheses.

Table 5.1 Average computational complexity for different eigenvalue distributions. Times in parentheses represent extrapolated values.

eig. dist.	$\kappa(\mathbf{Q})$	N	time [s]			number of nodes		
			BB-C	BB-D	CPLEX	BB-C	BB-D	CPLEX
$1/\lambda$	N	40	1.24	0.70	18.38	810	599	5979
		70	662	75	2146	2.60×10^4	0.69×10^4	2.72×10^5
		100	(4×10^5)	1.09×10^4	(2×10^5)		7.40×10^4	
	100N	40	0.84	0.67	(2×10^5)	628	611	
		70	334	213		1.85×10^4	1.46×10^4	
		100	(1×10^5)	(5×10^4)				
uniform	N	40	1.09	0.72	15.28	689	616	5500
		70	261	98	1159	1.77×10^4	1.01×10^4	2.03×10^5
		100	(7×10^4)	1.43×10^4	(7×10^4)		1.36×10^5	
	100N	40	0.18	0.19	(3×10^4)	189	189	
		70	3.64	2.95		1.69×10^3	1.69×10^3	
		100	98.6	77.1		9.71×10^3	9.95×10^4	
$1/\lambda^2$	N	40	1.93	0.51	23.65	1111	438	6929
		70	2949	12	3139	4.72×10^4	0.19×10^4	3.44×10^5
		100	(6×10^6)	633	(4×10^5)		1.40×10^4	
	100N	40	1.12	0.40	(3×10^5)	742	328	
		70	1756	14		4.19×10^4	0.23×10^4	
		100	(1×10^6)	848			1.60×10^4	

Considering first the comparison between BB-C and BB-D, it is clear from Table 5.1 that diagonal relaxations can significantly decrease complexity. The gains generally increase with the dimension N and can reach several orders of magnitude for the $1/\lambda^2$ eigenvalue distribution, which as seen in Sect. 5.1 is most favorable toward diagonal relaxations. Even for a uniform distribution and $\kappa(\mathbf{Q}) = 100N$, BB-D is slightly more efficient than BB-C, in apparent contradiction with the comparison in Fig. 5.1(b). This can be explained by noting that Fig. 5.1(b) represents the average approximation ratios for the root node whereas subproblems may have more non-uniform eigenvalue distributions and lower condition numbers. It is also interesting that instances in this class appear to be the easiest to solve.

The comparison with CPLEX in Table 5.1 shows the value of a more specialized algorithm for solving (1.1), as has been observed by others [2, 14]. This is in spite of the fact that CPLEX is run as a compiled executable with full multicore capabilities. Indeed, the advantage extends to the BB-C variant at low N , although the margin decreases at higher N . Note also that CPLEX has difficulty with the more poorly-conditioned instances. Given CPLEX's use of techniques beyond pure branch-and-bound, it is difficult to identify precisely the reasons for its relative inefficiency. One factor is the poor performance of the heuristic used by CPLEX relative to the backward selection heuristic in BB. For this reason, CPLEX is initialized with the BB solution in the experiments. As for lower bounds, it is likely that checking condition (2.4) confers significant benefits because of the ability to eliminate many infeasible

ble subproblems and improve bounds incrementally with minimal computation, and also because of the subsequent reduction in dimension. Another difference is the frequency at which relaxations are solved since in BB, some effort is made to avoid solving unprofitable relaxations.

Table 5.2 shows a complexity comparison for \mathbf{Q} matrices with unit diagonal entries and uniformly distributed off-diagonal entries, corresponding to Fig. 5.1(d) in Sect. 5.1. The difference between BB-C and BB-D in this case is as dramatic as it is for the $1/\lambda^2$ eigenvalue distribution in Table 5.1. The performance of CPLEX is similar to its performance in Table 5.1 for $\kappa(\mathbf{Q}) = N$. It is clear that BB-D remains the best option.

Table 5.2 Average computational complexity for different off-diagonal amplitudes a . Times in parentheses represent extrapolated values.

a	N	time [s]			number of nodes		
		BB-C	BB-D	CPLEX	BB-C	BB-D	CPLEX
0.2	40	1.66	0.13	26.87	1128	93	8698
	70	2941	1.1	4107	6.76×10^4	151	4.85×10^5
	100	(7×10^6)	2.6	(9×10^5)		187	
0.8	40	1.56	0.76	24.04	849	543	7896
	70	577	50	2853	3.21×10^4	0.57×10^4	3.86×10^5
	100	(4×10^5)	4.86×10^3	(4×10^5)		7.51×10^4	

6 Conclusion and future work

Two relaxations of a quadratically-constrained cardinality minimization problem (1.1) were investigated, the first being the continuous relaxation of a mixed integer formulation, the second an optimized diagonal relaxation based on a simple special case of the problem. An absolute upper bound on the optimal cost of the continuous relaxation suggests that it yields relatively weak approximations. In computational experiments, diagonal relaxations were seen to result in stronger bounds and significantly reduced complexity in solving (1.1) via branch-and-bound. Substantial gains were also observed relative to the general-purpose solver CPLEX. To support these numerical results, this paper analyzed the approximation properties of diagonal relaxations, providing general insight and establishing guarantees in terms of the eigenvalues of the matrix \mathbf{Q} and in the diagonally dominant and nearly coordinate-aligned cases.

Given the interest in generalizations of (1.1) in portfolio optimization, it is hoped that the analysis in this paper could be extended to these more general formulations and to other relaxations such as the perspective relaxation [12, 16, 29]. In addition, the positive experience with diagonal relaxations motivates further exploration of relaxations based on other efficiently solvable special cases, for example those in [11].

Acknowledgements The author thanks Pablo A. Parrilo, Alan V. Oppenheim and Vivek K. Goyal for helpful discussions that shaped this work.

References

1. Bertsekas, D.P.: *Nonlinear Programming*. Athena Scientific, Belmont, MA (1999)
2. Bertsimas, D., Shioda, R.: Algorithm for cardinality-constrained quadratic optimization. *Comput. Optim. Appl.* **43**, 1–22 (2009)
3. Bertsimas, D., Tsitsiklis, J.N.: *Introduction to Linear Optimization*. Athena Scientific, Nashua, NH (1997)
4. Bienstock, D.: Computational study of a family of mixed-integer quadratic programming problems. *Math. Program.* **74**(2), 121–140 (1996)
5. Bonami, P., Lejeune, M.A.: An exact solution approach for portfolio optimization problems under stochastic and integer constraints. *Oper. Res.* **57**(3), 650–670 (2009)
6. Boyd, S., Vandenberghe, L.: *Convex Optimization*. Cambridge University Press, Cambridge, UK (2004)
7. Ceria, S., Soares, J.: Convex programming for disjunctive convex optimization. *Math. Program.* **86**(3), 595–614 (1999)
8. Chernoff, H.: A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.* **23**(4), 493–507 (1952)
9. Cuvreur, C., Bresler, Y.: On the optimality of the backward greedy algorithm for the subset selection problem. *SIAM J. Matrix Anal. Appl.* **21**(3), 797–808 (2000)
10. Cui, X., Zheng, X., Zhu, S., Sun, X.: Convex relaxations and MIQCQP reformulations for a class of cardinality-constrained portfolio selection problems. *J. Glob. Optim.* pp. 1–15 (2012)
11. Das, A., Kempe, D.: Algorithms for subset selection in linear regression. In: *40th ACM Symposium on Theory of Computing (STOC)*, pp. 45–54. Victoria, Canada (2008)
12. Frangioni, A., Gentile, C.: Perspective cuts for a class of convex 0-1 mixed integer programs. *Math. Program., Ser. A* **106**(2), 225–236 (2006)
13. Frangioni, A., Gentile, C.: SDP diagonalizations and perspective cuts for a class of nonseparable MIQP. *Oper. Res. Lett.* **35**(2), 181–185 (2007)
14. Gao, J., Li, D.: Cardinality constrained linear-quadratic optimal control. *IEEE Trans. Autom. Control* **56**(8), 1936–1941 (2011)
15. Gao, J., Li, D.: A polynomial case of the cardinality-constrained quadratic optimization problem. *J. Glob. Optim.* pp. 1–15 (2012)
16. Günlük, O., Linderoth, J.: Perspective reformulations of mixed integer nonlinear programs with indicator variables. *Math. Program., Ser. B* **124**(1–2), 183–205 (2010)
17. Horn, R.A., Johnson, C.R.: *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, UK (1994)
18. IBM ILOG: *IBM ILOG CPLEX 12.4 User’s Manual* (2012)
19. Lemaréchal, C., Oustry, F.: Semidefinite relaxations and Lagrangian duality with application to combinatorial optimization. *Tech. Rep. RR-3710, INRIA* (1999)
20. Lin, F., Fardad, M., Jovanovic, M.R.: Design of optimal sparse feedback gains via the alternating direction method of multipliers (2012). ArXiv preprint, <http://arxiv.org/abs/1111.6188>
21. Miller, A.J.: *Subset selection in regression*, 2 edn. Chapman & Hall/CRC, Boca Raton, FL (2002)
22. Shaw, D.X., Liu, S., Kopman, L.: Lagrangian relaxation procedure for cardinality-constrained portfolio optimization. *Optim. Method. Softw.* **23**(3), 411–420 (2008)
23. Sturm, J.F.: Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones. *Optim. Method. Softw.* **11**, 625–653 (1999)
24. Toh, K.C., Todd, M.J., Tütüncü, R.H.: SDPT3 — a MATLAB software package for semidefinite programming. *Optim. Method. Softw.* **11**, 545–581 (1999). Latest version available at <http://www.math.nus.edu.sg/~mattohkc/sdpt3.html>
25. Vielma, J.P., Ahmed, S., Nemhauser, G.L.: A lifted linear programming branch-and-bound algorithm for mixed-integer conic quadratic programs. *INFORMS J. Comput.* **20**(3), 438–450 (2008)
26. Wei, D.: *Design of discrete-time filters for efficient implementation*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA (2011)
27. Wei, D., Oppenheim, A.V.: A branch-and-bound algorithm for quadratically-constrained sparse filter design. *IEEE Trans. Signal Process.* (to appear)
28. Wei, D., Sestok, C.K., Oppenheim, A.V.: Sparse filter design under a quadratic constraint: Low-complexity algorithms. *IEEE Trans. Signal Process.* (to appear)
29. Zheng, X., Sun, X., Li, D.: Improving the performance of MIQP solvers for quadratic programs with cardinality and minimum threshold constraints: A semidefinite program approach. Preprint, http://www.optimization-online.org/DB_FILE/2010/11/2797.pdf