UNIVERSITÀ DI PARMA

ARCHIVIO DELLA RICERCA

University of Parma Research Repository

On the complexity of quadratic programming with two quadratic constraints

This is the peer reviewd version of the followng article:

Original
On the complexity of quadratic programming with two quadratic constraints / Consolini, Luca; Locatelli, Marco. - In: MATHEMATICAL PROGRAMMING. - ISSN 0025-5610. - (2017), pp. 91-128. [10.1007/s10107-016-1073-8]

Availability:
This version is available at: 11381/2819435 since: 2021-10-14T14:48:47Z

Publisher:
Springer Verlag

Published
DOI:10.1007/s10107-016-1073-8

Terms of use:

Anyone can freely access the full text of works made available as "Open Access". Works made available

Publisher copyright

note finali coverpage

(Article begins on next page)

20 April 2024

Math. Program., Ser. A
DOI 10.1007/s10107-016-1073-8

FULL LENGTH PAPER

On the complexity of quadratic programming with two quadratic constraints

Luca Consolini[1] · **Marco Locatelli**[1]

Received: 5 October 2015 / Accepted: 23 September 2016
© Springer-Verlag Berlin Heidelberg and Mathematical Optimization Society 2016

Abstract The complexity of quadratic programming problems with two quadratic constraints is an open problem. In this paper we show that when one constraint is a ball constraint and the Hessian of the quadratic function defining the other constraint is positive definite, then, under quite general conditions, the problem can be solved in polynomial time in the real-number model of computation through an approach based on the analysis of the dual space of the Lagrange multipliers. However, the degree of the polynomial is rather large, thus making the result mostly of theoretical interest.

Keywords Quadratic problems · Complexity · Bivariate polynomial systems

Mathematics Subject Classification 90C20 Quadratic programming · 90C26 Nonconvex programming, global optimization · 90C30 Nonlinear programming

1 Introduction

Quadratic programming problems with two quadratic constraints have received a lot of attention in the literature. We refer, e.g., to [1] for a thorough discussion about them. In this paper we are interested in such problems and, in particular, in the cases with a ball constraint and a further convex quadratic constraint

✉ Luca Consolini
lucac@ce.unipr.it

Marco Locatelli
locatell@ce.unipr.it

[1] Dipartimento di Ingegneria dell'Informazione, Università di Parma, Via G.P. Usberti, 181/A, 43124 Parma, Italy

Published online: 01 October 2016

$$\begin{aligned} \min\ & \tfrac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{q}^T\mathbf{x} \\ & \tfrac{1}{2}\mathbf{x}^T\mathbf{x} \le \tfrac{1}{2} \\ & \tfrac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{a}^T\mathbf{x} \le u, \end{aligned} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is assumed to be positive definite. We also assume that the feasible region of the problem has a nonempty interior. In what follows we will denote by f_0, f_1, f_2 respectively the objective and constraint functions of problem (1). As an application of this problem, in [6] it is shown that it arises as a subproblem to be solved at each iteration of a trust region algorithm for equality constrained nonlinear problems (see also [19] where it is shown that for the case of a single equality constraint the subproblem can be solved in polynomial time).

In the recent paper [5] the authors state that the computational complexity of the above problem is an open question. In fact, polynomial approaches are known for some special cases. In [19] it is shown that a semidefinite relaxation allows to solve it when $\mathbf{q} = \mathbf{a} = \mathbf{0}$, i.e., when we have no linear terms. In [2] the case of two-sided indefinite quadratic constraints is shown to be solvable in polynomial time under a suitable assumption. This case is also discussed in [13]. Approximation results for the case with an arbitrary number of quadratic equality constraints with a diagonal Hessian and no linear terms, and with additional bound constraints, are discussed in [17]. In the previously mentioned paper [5] the authors introduce a family of polynomial-time separable second order cone-reformulation/linearization technique (SOC-RLT) constraints for the problem. These constraints are added to a semidefinite relaxation. However, the authors provide examples where the optimal value of the resulting relaxation is not equal to the optimal value of the original problem. In [3,11] conditions under which a rather simple convex relaxation turns out to be exact, are given for some subclasses of quadratic problems with two quadratic constraints. In [4] the authors introduce a polynomial-time algorithm for problems with a (possibly nonconvex) quadratic objective function and a feasible region defined by: (i) quadratic constraints which define spheres; (ii) quadratic constraints which define regions which are the complements of the interiors of spheres; (iii) some linear constraints. Given the polyhedron defined by the linear constraints, the authors call intersecting a face of the polyhedron which has a nonempty intersection with the intersection of the spheres defining the feasible region. Polynomial solvability is then proved under the assumption that the number of intersecting faces is polynomial, and that the number of quadratic constraints is fixed.

Statement of contribution The aim of this paper is to prove that, under quite general conditions, problem (1) can be solved in polynomial time in the real-number model of computation. However, as we will see, the degree of the polynomial is quite high, thus making the approach more of theoretical interest rather than of practical interest. In the next section we make a detailed outline of the paper, which could be used by the reader as a guide through the (admittedly) rather technical proofs spread throughout the paper.

2 Outline of the paper

In Sect. 3 we state the complexity result and we introduce a perturbed problem with rational coefficients. We will also show in Theorem 3.1 that once we have an approximate solution of the perturbed problem, we are able to derive an approximate solution of the original problem.

In Sect. 4 we discuss the simpler case when $\mathbf{A}$ and $\mathbf{Q}$ commute. Although the assumption is restrictive, we first discuss this case because it allows for a simple description of the main ideas of the proof, which are then generalized in the next section. The scheme of the proof is the following.

- First, we state the KKT conditions for the problem, which depend on the original variables and on the two Lagrange multipliers μ_1, μ_2 associated to the constraints. Noting that the global minimizer is a KKT point, we try to identify all such points.
- Next, we make a separate discussion for the case where the Hessian of the Lagrangian function is singular, and for the case it is nonsingular.
  - In the case of singular Hessian matrix it is shown that, under a suitable assumption, at most $2n$ KKT points can be identified and their (exact) computation requires a polynomial time.
  - In the case of nonsingular Hessian matrix, we show that the μ_1, μ_2 values can be identified by solving a bivariate polynomial system with two equations with degree at most $2n$ (the system is derived from the KKT conditions). Under the assumption of zero-dimensionality, the system has at most $4n^2$ and (approximate) solutions of the system can be computed in polynomial time. Note that in this section a couple of assumptions are introduced. Conditions under which these assumptions are fulfilled are given in Sect. 5. However, we remark that these conditions are almost always fulfilled. In particular, the assumption of zero-dimensionality of the bivariate polynomial system is certainly true if the problem has a finite number of critical points, which is almost always true.
- Finally, once approximate KKT points have been computed, in Theorem 4.1 it is shown how to derive from them feasible solutions of the original problem and, thus, also a solution of the complexity problem.

Section 5 is devoted to the general case. The scheme of the proof is the same as in Sect. 4, but, as previously commented, general conditions, depending on the problem structure, are given under which the assumptions already introduced in the previous section are fulfilled. Thus, from the KKT conditions, we make a separate discussion for the two cases of singular and nonsingular Hessian of the Lagrangian function. In the former case we are able to identify (approximate) KKT points in polynomial time by solving univariate polynomial equations (Proposition 5.1). In the latter case we compute (approximate) KKT points in polynomial time by solving, again, a bivariate polynomial system (Proposition 5.2). With respect to Sect. 4, the procedure to derive feasible solutions of the original problem close to the approximate KKT points previously identified, is more complicated. It is described in Sect. 5.2, where the polynomial-time algorithm is finally described. In order to improve readability, the intermediate results, needed for the proof of the main result, are given in Sect. 6.

3 Complexity statement and introduction of a perturbed problem

In this paper we employ the real-number model of computation. Given an optimal solution $\mathbf{x}^*$ of problem (1), according to this model we are able to solve the problem in polynomial time if, for each $\rho > 0$, we are able to detect a ρ-approximate solution, i.e., to detect some feasible point $\bar{\mathbf{x}}$ such that

$$f_0(\bar{\mathbf{x}}) \leq f_0(\mathbf{x}) + \rho, \tag{2}$$

in polynomial time with respect to n, to $\log(1/\rho)$, i.e., to the logarithm of the required precision, and to two real parameters associated to the problem at hand. Following, e.g., [12,18], the first parameter is denoted by Ω and is the maximum absolute value of the coefficients of the problem, while the second parameter measures how sensitive the constraints are with respect to perturbation of the data. Stated in another way, the second parameter is a measure of strict feasibility for the problem. We define it as follows. First, we solve the following convex optimization problems

$$\begin{aligned} \tfrac{1}{2} - 2\kappa_1 = \min\ & \tfrac{1}{2}\mathbf{x}^T\mathbf{x} \\ & \tfrac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{a}^T\mathbf{x} \leq u, \\ u - 2\kappa_2 = \min\ & \tfrac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{a}^T\mathbf{x} \\ & \tfrac{1}{2}\mathbf{x}^T\mathbf{x} \leq \tfrac{1}{2}. \end{aligned}$$

Let $\mathbf{x}_0^1, \mathbf{x}_0^2$ be the optimal solutions of the first and second problem, respectively. Next, we consider the point

$$\mathbf{x}_0 = \frac{1}{2}\mathbf{x}_0^1 + \frac{1}{2}\mathbf{x}_0^2. \tag{3}$$

Then, by convexity

$$\begin{aligned} & \tfrac{1}{2}\mathbf{x}_0^T\mathbf{x}_0 \leq \tfrac{1}{2} - \kappa_1 \\ & \tfrac{1}{2}\mathbf{x}_0^T\mathbf{A}\mathbf{x}_0 + \mathbf{a}^T\mathbf{x}_0 \leq u - \kappa_2, \end{aligned} \tag{4}$$

i.e., $\mathbf{x}_0$ is a strict feasible point for problem (1). Note that $\kappa_1 \leq \frac{1}{2}$ must hold. We consider the following measure of strict feasibility

$$\chi = \frac{1}{\min\{\kappa_1, \kappa_2\}}. \tag{5}$$

As previously commented, this value measures the sensitivity of the problem with respect to perturbations of the coefficients appearing in the constraints: a large value for χ means that even small perturbations of these data can make the problem infeasible.

Thus, the complexity result we aim to prove states that a feasible point $\bar{\mathbf{x}}$ satisfying (2) can be computed in polynomial time with respect to n, $\log(\Omega)$, $\log(\chi)$, and $\log(1/\rho)$. In order to compute such point we first need to define a suitable perturbation of the original problem.

3.1 Definition of a perturbed problem

Here and in what follows, for any vector $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\|$ denotes its Euclidean norm, for any matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ with $\mathbf{M} = (m_{ij})$, $\|\mathbf{M}\| = \max_{\|x\|\leq 1} \|\mathbf{M}x\|$ denotes the matrix 2-norm and $\|\mathbf{M}\|_\infty = \max_{i,j=1,\ldots,n} |m_{ij}|$ the entrywise infinity norm.

We consider a perturbation of the original data, such that all data have a fractional part rounded at the kth binary digit, i.e., we consider the perturbed problem

$$\begin{aligned} \min\ & \tfrac{1}{2}\mathbf{x}^T\bar{\mathbf{Q}}\mathbf{x} + \bar{\mathbf{q}}^T\mathbf{x} \\ & \tfrac{1}{2}\mathbf{x}^T\mathbf{x} \leq \tfrac{1}{2} \\ & \tfrac{1}{2}\mathbf{x}^T\bar{\mathbf{A}}\mathbf{x} + \bar{\mathbf{a}}^T\mathbf{x} \leq \bar{u}, \end{aligned} \tag{6}$$

where

$$\begin{aligned} &\bar{\mathbf{Q}} = (\mathbf{Q} + \Delta\mathbf{Q}) \quad \bar{\mathbf{q}} = (\mathbf{q} + \Delta\mathbf{q}) \quad \bar{\mathbf{A}} = (\mathbf{A} + \Delta\mathbf{A}) \\ &\bar{\mathbf{a}} = (\mathbf{a} + \Delta\mathbf{a}) \qquad \bar{u} = (u + \Delta u), \end{aligned}$$

and:

$$\|\Delta\mathbf{Q}\|_\infty, \|\Delta\mathbf{q}\|_\infty, \|\Delta\mathbf{A}\|_\infty, \|\Delta\mathbf{a}\|_\infty, |\Delta u| \leq 2^{-k}. \tag{7}$$

Note that all the coefficients are now rational numbers whose numerator is never larger than $2^k\Omega$, and whose denominator is never larger than 2^k. Thus, the bit size of these rational coefficients, denoted by δ in what follows, is

$$\delta = O\left(k + \log\Omega\right). \tag{8}$$

Also note that k is a measure of the 'size" of the perturbation: the larger k is, the "smaller" the perturbation is. The objective and constraint functions of the problem (6) will be denoted in what follows by $\bar{f}_0$, $\bar{f}_1$, $\bar{f}_2$ respectively. We first need the following remark, establishing a relation between the original and the perturbed functions.

Remark 3.1 For any $\mathbf{x}$ such that $\|\mathbf{x}\| \leq 1$, it holds that

$$|\bar{f}_i(\mathbf{x}) - f_i(\mathbf{x})| \leq 2^{-k}\left(\frac{1}{2}n + \sqrt{n}\right) \quad i = 0, 1, 2.$$

Proof We prove the simple result only for the objective function. We have that

$$\bar{f}_0(\mathbf{x}) - f_0(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\Delta\mathbf{Q}\mathbf{x} + [\Delta\mathbf{q}]^T\mathbf{x}.$$

Then,

$$|\bar{f}_0(\mathbf{x}) - f_0(\mathbf{x})| \leq \frac{1}{2}\|\Delta\mathbf{Q}\|\|\mathbf{x}\|^2 + \|\Delta\mathbf{q}\|\|\mathbf{x}\|.$$

We recall that for each n-dimensional vector $\mathbf{m}$ and each n-dimensional square matrix $\mathbf{M}$ it holds that

$$\|\mathbf{m}\| \leq \sqrt{n}\|\mathbf{m}\|_\infty, \quad \|\mathbf{M}\| \leq n\|\mathbf{M}\|_\infty. \tag{9}$$

Then,

$$|\bar{f}_0(\mathbf{x}) - f_0(\mathbf{x})| \leq 2^{-k}\left(\frac{1}{2}n + \sqrt{n}\right),$$

as we wanted to prove. □

Next, we show that for k large enough we are able to guarantee a strict feasibility condition similar to (4), also for the perturbed problem.

Observation 3.1 *If*

$$k \geq \log\left(\frac{n + 2\sqrt{n} + 2}{\kappa_2}\right), \tag{10}$$

then

$$\frac{1}{2}\mathbf{x}_0^T\bar{\mathbf{A}}\mathbf{x}_0 + \bar{\mathbf{a}}^T\mathbf{x}_0 \leq \bar{u} - \frac{\kappa_2}{2}, \tag{11}$$

where $\mathbf{x}_0$ *is the point defined in* (3).

Proof In view of (4), (7), and (9), the left-hand side in (11) is bounded from above by

$$u - \kappa_2 + \frac{1}{2}\|\Delta\mathbf{A}\|\|\mathbf{x}_0\|^2 + \|\Delta\mathbf{a}\|\|\mathbf{x}_0\| \leq u - \kappa_2 + 2^{-k}\left(\frac{1}{2}n + \sqrt{n}\right).$$

The right-hand side in (11) is bounded from below by

$$u - |\Delta u| - \frac{\kappa_2}{2} \geq u - 2^{-k} - \frac{\kappa_2}{2}.$$

Thus, (11) holds if

$$2^k \geq \frac{n + 2\sqrt{n} + 2}{\kappa_2},$$

from which the result immediately follows. □

We will also need the following simple result.

Remark 3.2 Let us consider the quadratic function

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{B}\mathbf{x} + \mathbf{b}^T\mathbf{x},$$

and let Γ be the maximum absolute value of the coefficients in $\mathbf{B}$ and $\mathbf{b}$. Then, for any pair of vectors $\mathbf{y}_1, \mathbf{y}_2$ such that $\|\mathbf{y}_1\|, \|\mathbf{y}_2\| \leq 1$, it holds that:

$$|f(\mathbf{y}_1) - f(\mathbf{y}_2)| \leq (n + \sqrt{n})\Gamma\|\mathbf{y}_1 - \mathbf{y}_2\| + \frac{1}{2}n\Gamma\|\mathbf{y}_1 - \mathbf{y}_2\|^2. \tag{12}$$

In particular, if $\|\mathbf{y}_1 - \mathbf{y}_2\| \leq 1$, then it also holds that

$$|f(\mathbf{y}_1) - f(\mathbf{y}_2)| \leq 3n\Gamma\|\mathbf{y}_1 - \mathbf{y}_2\|. \tag{13}$$

Proof We have that

$$f(\mathbf{y}_1) - f(\mathbf{y}_2) = \mathbf{y}_2^T \mathbf{B}(\mathbf{y}_1 - \mathbf{y}_2) + \mathbf{b}^T(\mathbf{y}_1 - \mathbf{y}_2) + \frac{1}{2}(\mathbf{y}_1 - \mathbf{y}_2)^T \mathbf{B}(\mathbf{y}_1 - \mathbf{y}_2),$$

and, consequently,

$$|f(\mathbf{y}_1) - f(\mathbf{y}_2)| \leq \|\mathbf{y}_2\|\|\mathbf{B}\|\|\mathbf{y}_1 - \mathbf{y}_2\| + \|\mathbf{b}\|\|\mathbf{y}_1 - \mathbf{y}_2\| + \frac{1}{2}\|\mathbf{B}\|\|\mathbf{y}_1 - \mathbf{y}_2\|^2.$$

Now, the result immediately follows by observing that $\|\mathbf{y}_2\| \leq 1$ by assumption, and recalling (9). □

The following observation states that, given a feasible solution of the perturbed problem, we are able to identify a feasible solution of the original problem with a similar objective function value.

Observation 3.2 *If k satisfies (10), for each feasible solution $\mathbf{z}$ of the perturbed problem (6), we are able to compute a feasible solution $\mathbf{w_z}$ of the original problem (1) such that*

$$|f_0(\mathbf{w_z}) - f_0(\mathbf{z})| \leq 6n\Omega p(n, \kappa_2, k),$$

where:

$$p(n, \kappa_2, k) = \frac{2^{-k}\left(\frac{1}{2}n + \sqrt{n} + 1\right)}{2^{-k}\left(\frac{1}{2}n + \sqrt{n} + 1\right) + \kappa_2}.$$

Proof If $\mathbf{z}$ is feasible for the perturbed problem, it satisfies

$$\frac{1}{2}\mathbf{z}^T(\mathbf{A} + \Delta\mathbf{A})\mathbf{z} + (\mathbf{a} + \Delta\mathbf{a})^T\mathbf{z} \leq u + \Delta u.$$

Thus, in view of (7) and of (9), we also have

$$\frac{1}{2}\mathbf{z}^T\mathbf{A}\mathbf{z} + \mathbf{a}^T\mathbf{z} \leq u + 2^{-k}\left(\frac{1}{2}n + \sqrt{n} + 1\right).$$

Now, let $\mathbf{w_z}$ be equal to $\mathbf{z}$, if $\mathbf{z}$ is feasible for (1) (in which case the result is trivial). Otherwise, let $\mathbf{w_z}$ be a point along the segment between $\mathbf{z}$ and the strictly feasible point $\mathbf{x}_0$ defined in (3):

$$\mathbf{w_z} = (1 - \lambda)\mathbf{z} + \lambda\mathbf{x}_0.$$

Note that both $\mathbf{z}$ and $\mathbf{w_z}$ fulfill the ball constraint. For what concerns the second constraint, in view of the convexity of the constraint function we have that, for any $\lambda \in [0, 1]$,

$$f_2(\mathbf{w_z}) \leq (1-\lambda) f_2(\mathbf{z}) + \lambda f_2(\mathbf{x}_0) \leq (1-\lambda)\left(u + 2^{-k}\left(\tfrac{1}{2}n + \sqrt{n} + 1\right)\right) + \lambda(u - \kappa_2)$$
$$= u + (1-\lambda)2^{-k}\left(\tfrac{1}{2}n + \sqrt{n} + 1\right) - \kappa_2\lambda.$$

Recalling the definition of p, feasibility of $\mathbf{w_z}$ is guaranteed for all $\lambda \geq p(n, \kappa_2, k)$. Then, by taking λ equal to $p(n, \kappa_2, k)$, it holds that

$$\|\mathbf{z} - \mathbf{w_z}\| = \lambda\|\mathbf{w_z} - \mathbf{x}_0\| \leq p(n, \kappa_2, k)(\|\mathbf{w_z}\| + \|\mathbf{x}_0\|) \leq 2p(n, \kappa_2, k).$$

For k defined as in (10), $2p(n, \kappa_2, k) \leq 1$ holds, so that, in view of Remark 3.2,

$$|f_0(\mathbf{w_z}) - f_0(\mathbf{z})| \leq 6n\Omega p(n, \kappa_2, k). \tag{14}$$

We are now ready to prove the following theorem.

Theorem 3.1 *If a $\frac{\rho}{2}$-approximate solution $\mathbf{z}$ of the perturbed problem (6) is available, then the feasible solution $\mathbf{w_z}$ of the original problem (1) is a ρ-approximate solution for such problem, provided that*

$$k \geq \log\left(\frac{1}{\rho}\right) + \max\{\log(4n + 8\sqrt{n}), \log(12n^2 + 24n\sqrt{n} + 24n) + \log(\Omega) + \log(\chi)\}. \tag{15}$$

Proof Let $\bar{\mathbf{x}}^*$ be an optimal solution of the perturbed problem (6) and $\mathbf{x}^*$ be an optimal solution of the original problem (1). Let $\mathbf{z}$ be a $\frac{\rho}{2}$-approximate solution of the perturbed problem, i.e.,

$$\bar{f}_0(\mathbf{z}) - \bar{f}_0(\bar{\mathbf{x}}^*) \leq \frac{\rho}{2}.$$

We first show that

$$f_0(\mathbf{z}) - f_0(\mathbf{x}^*) \leq \frac{\rho}{2} + 2\eta(k, n), \tag{16}$$

where $\eta(k, n) = 2^{-k}\left(\frac{1}{2}n + \sqrt{n}\right)$. In view of Remark 3.1 we have that

$$f_0(\mathbf{z}) \leq \bar{f}_0(\mathbf{z}) + \eta(k, n), \quad f_0(\mathbf{x}^*) \geq \bar{f}_0(\mathbf{x}^*) - \eta(k, n),$$

form which (16) immediately follows. Thus, in view of (14) we also have that $\mathbf{w_z}$ is a feasible point for the original problem such that

$$f_0(\mathbf{w_z}) - f_0(\mathbf{x}^*) \leq \frac{\rho}{2} + 2\eta(k, n) + 6n\Omega p(n, \kappa_2, k).$$

Thus,

$$2\eta(k, n) + 6n\Omega p(n, \kappa_2, k) \leq \frac{\rho}{2}, \tag{17}$$

implies

$$f_0(\mathbf{w_z}) - f_0(\mathbf{x}^*) \leq \rho.$$

Recalling the definitions of p and η, we must have

$$2^{-k}\left(n+2\sqrt{n}\right)+6n\Omega\frac{2^{-k}\left(\frac{1}{2}n+\sqrt{n}+1\right)}{2^{-k}\left(\frac{1}{2}n+\sqrt{n}+1\right)+\kappa_2}\leq\frac{\rho}{2},$$

The above inequality holds, e.g., if

$$2^{-k}\left(n+2\sqrt{n}\right)\leq\frac{\rho}{4},$$

and

$$6n\Omega\frac{2^{-k}\left(\frac{1}{2}n+\sqrt{n}+1\right)}{\kappa_2}\leq\frac{\rho}{4},$$

which both hold if

$$k\geq\log\left(\frac{1}{\rho}\right)+\max\{\log(4n+8\sqrt{n}),\log(12n^2+24n\sqrt{n}+24n)+\log(\Omega)+\log(\chi)\},$$

as we wanted to prove. □

According to the above theorem, if the perturbation of the original problem is "small" enough, i.e., if k is large enough, we are able to compute a ρ-approximate solution of the original problem if we are able to compute a $\frac{\rho}{2}$-approximate solution of the perturbed problem. Thus, what we need now in order to prove polynomial solvability, is that a $\frac{\rho}{2}$-approximate solution of the perturbed problem can be computed in polynomial time with respect to the previously discussed parameters. We will first consider the simpler case of commuting matrices and, later on, we will deal with the general case.

4 The case of commuting matrices

In this section we study the case where $\mathbf{A}$ and $\mathbf{Q}$ commute. In this case it turns out that $\mathbf{A}$ and $\mathbf{Q}$ are simultaneously diagonalizable (see [3]) by the orthonormal matrix $\mathbf{S}$ such that

$$\mathbf{S}^T\mathbf{Q}\mathbf{S}=\mathbf{D}_1=Diag(\gamma_1,\ldots,\gamma_n),\quad \mathbf{S}^T\mathbf{A}\mathbf{S}=\mathbf{D}_2=Diag(\eta_1,\ldots,\eta_n).$$

Note that $\mathbf{D}_2$ is positive definite, i.e., $\eta_i>0, i=1,\ldots,n$, in view of the positive definiteness of $\mathbf{A}$. Then, after the change of variable $\mathbf{x}=\mathbf{S}\mathbf{y}$, problem (1) can be rewritten in the following form where the objective and constraint functions are separable

$$\begin{aligned}\min\ & f_0(\mathbf{y}):=\tfrac{1}{2}\textstyle\sum_{i=1}^n\gamma_i y_i^2+\sum_{i=1}^n c_i y_i\\ & f_1(\mathbf{y}):=\tfrac{1}{2}\textstyle\sum_{i=1}^n y_i^2\leq\tfrac{1}{2}\\ & f_2(\mathbf{y}):=\tfrac{1}{2}\textstyle\sum_{i=1}^n\eta_i y_i^2+\sum_{i=1}^n b_i y_i\leq u.\end{aligned}\tag{18}$$

We will assume that all the coefficients of the above problem are rational numbers. This is not necessarily true after the diagonalization, but in case it is not we proceed with the already discussed perturbation procedure. Moreover, we introduce an assumption which simplifies the following development.

Assumption 4.1 We assume that:

(a) for each $i \in \{1, \ldots, n\}$, $c_i \neq 0$;
(b) for each $i, j \in \{1, \ldots, n\}$, $i \neq j$,

$$\frac{c_i}{b_i} \neq \frac{c_j}{b_j}.$$

Note that the assumption is not restrictive, since when we perturb the coefficients c_i, $i = 1, \ldots, n$, we can do that in such a way that the assumption is fulfilled. Now, the approach we are going to propose is based on the identification of the KKT points for this problem. Indeed, any local minimizer of the problem (and, thus, also the global minimizer) satisfies the KKT conditions. Note that the feasible region is a convex set and, by assumption, Slater's condition holds (a strictly feasible point exists). Thus, all feasible points satisfy a constraint qualification and we can restrict the search for the global minimizer to KKT points.

The KKT system is the following

$$\begin{array}{l} \gamma_i y_i + c_i + \mu_1 y_i + \mu_2 \eta_i y_i + \mu_2 b_i = 0 \quad i = 1, \ldots, n \\ \mu_1 \left(\sum_{i=1}^n y_i^2 - 1 \right) = 0 \\ \mu_2 \left(\frac{1}{2} \sum_{i=1}^n \eta_i y_i^2 + \sum_{i=1}^n b_i y_i - u \right) = 0 \\ \mu_1, \mu_2 \geq 0, \end{array} \tag{19}$$

where μ_1 and μ_2 are the Lagrange multipliers of the first and second constraint, respectively. We first discuss some special cases. After remarking that the values y_i, $i = 1, \ldots, n$, are uniquely identified if $\gamma_i + \mu_1 + \mu_2 \eta_i \neq 0$, we discuss the cases when $\gamma_i + \mu_1 + \mu_2 \eta_i = 0$ for some index i.

Observation 4.1 *Under Assumption 4.1, for each $i = 1, \ldots, n$, the number of KKT points for which $\gamma_i + \mu_1 + \mu_2 \eta_i = 0$ is (at most) two.*

Proof First of all we notice from (19) that $\gamma_i + \mu_1 + \mu_2 \eta_i = 0$ implies $c_i + \mu_2 b_i = 0$. In view of Assumption 4.1a, we must have $b_i \neq 0$. Then, the values of the Lagrange multipliers must be

$$\begin{array}{l} \mu_1^* = -\gamma_i + \eta_i \frac{c_i}{b_i} \\ \mu_2^* = -\frac{c_i}{b_i}. \end{array} \tag{20}$$

Note that if $\mu_1^* < 0$ or $\mu_2^* < 0$, then there is no KKT point for which $\gamma_i + \mu_1 + \mu_2 \eta_i = 0$. If for some $j \neq i$

$$\gamma_j + \mu_1^* + \mu_2^* \eta_j = 0,$$

in view of Assumption 4.1b we must have $c_j + \mu_2^* b_j \neq 0$, i.e., there is no KKT point corresponding to the two Lagrange multipliers μ_1^*, μ_2^*. Then, we can impose

$$\gamma_j + \mu_1^* + \mu_2^* \eta_j \neq 0,$$

for all $j \neq i$. Then it follows from (19) that for $j \neq i$

$$\bar{y}_j = -\frac{c_j + \mu_2^* b_j}{\gamma_j + \mu_1^* + \mu_2^* \eta_j}. \tag{21}$$

Thus, only the value y_i needs to be fixed. Since $c_i \neq 0$ in view of Assumption 4.1a, then $\mu_2^* \neq 0$, so that the second constraint is active at the KKT point and

$$\frac{1}{2}\eta_i y_i^2 + b_i y_i = u - \frac{1}{2}\sum\nolimits_{j \neq i} \eta_j \bar{y}_j^2 - \sum\nolimits_{j \neq i} b_j \bar{y}_j;$$

thus, at most two values for y_i are possible and there are at most two KKT points. □

If we discard the solutions discussed in the previous observation and we restrict the attention to KKT points where both constraints are active (the other KKT points, which are also KKT points of trust-region problems, can be detected in a similar way), we are able to rewrite the KKT conditions as follows

$$\begin{array}{ll} y_i = \frac{p_{i1}(\mu_1,\mu_2)}{p_{i2}(\mu_1,\mu_2)} & i = 1, \ldots, n \\ \sum_{i=1}^n \left(\frac{p_{i1}(\mu_1,\mu_2)}{p_{i2}(\mu_1,\mu_2)}\right)^2 = 1 & \\ \frac{1}{2}\sum_{i=1}^n \eta_i \left(\frac{p_{i1}(\mu_1,\mu_2)}{p_{i2}(\mu_1,\mu_2)}\right)^2 + \sum_{i=1}^n b_i \frac{p_{i1}(\mu_1,\mu_2)}{p_{i2}(\mu_1,\mu_2)} = u, & \end{array} \tag{22}$$

where, for each $i = 1, \ldots, n$:

$$\begin{array}{l} p_{i1}(\mu_1, \mu_2) = -(c_i + \mu_2 b_i) \\ p_{i2}(\mu_1, \mu_2) = \gamma_i + \mu_1 + \mu_2 \eta_i. \end{array} \tag{23}$$

According to (22), we need to solve the following system, after multiplying both sides of the last two equations by the least common multiple of the denominators,

$$\begin{array}{l} \sum_{i=1}^n (p_{i1}(\mu_1, \mu_2))^2 \prod_{j \neq i} (p_{j2}(\mu_1, \mu_2))^2 - \prod_{j=1}^n (p_{j2}(\mu_1, \mu_2))^2 = 0 \\ \sum_{i=1}^n \left\{ \left[\frac{1}{2}\eta_i (p_{i1}(\mu_1, \mu_2))^2 + b_i p_{i1}(\mu_1, \mu_2) p_{i2}(\mu_1, \mu_2) \right] \prod_{j \neq i} (p_{j2}(\mu_1, \mu_2))^2 \right\} \\ \quad - u \prod_{j=1}^n (p_{j2}(\mu_1, \mu_2))^2 = 0. \end{array} \tag{24}$$

This is a bivariate polynomial system with degrees of the polynomial at most equal to $2n$. Before discussing its solution, we make a few remarks.

Remark 4.1 The solutions of this system also include those discussed in Observation 4.1. Thus, although we have made a separate discussion for the solutions of those special cases, in fact such solutions can also be derived from the solution of the system (24).

Remark 4.2 The multiplication by the least common multiple may also introduce unwanted solutions. These occurs when

$$p_{i2}(\mu_1, \mu_2) = \gamma_i + \mu_1 + \mu_2\eta_i = \gamma_j + \mu_1 + \mu_2\eta_j = p_{j2}(\mu_1, \mu_2) = 0, \quad i \neq j. \tag{25}$$

Note that we do not need to care about pairs (i, j) such that $\gamma_i = \gamma_j$ and $\eta_i = \eta_j$. Indeed, in this case the least common multiple includes only one of the two (identical) expressions $(\gamma_i + \mu_1 + \mu_2\eta_i)^2$ and $(\gamma_j + \mu_1 + \mu_2\eta_j)^2$. In all the other cases, the bivariate linear system (25) either has no solution or identifies a unique solution $(\bar{\mu}_1, \bar{\mu}_2)$. These solutions of system (24) can be discarded since they do not correspond to KKT points.

Remark 4.3 Although we are only interested in solutions of the system with $\mu_1, \mu_2 \geq 0$, the system may also have solutions with a negative value for μ_1 or μ_2 or both. We compute also these solutions because these lead to KKT points for problem (18) where the inequality constraints are replaced by equality ones.

Now we can exploit all the theory and the practical knowledge about the solution of bivariate polynomial systems. We introduce the following assumption.

Assumption 4.2 System (24) is zero-dimensional, i.e., it has only a finite number of solutions.

The assumption is a rather general one and sufficient conditions under which it is fulfilled are discussed in Sect. 5.

Observation 4.2 *If system* (24) *is zero-dimensional, then it has at most* $4n^2$ *solutions, including the complex ones.*

Proof This is an immediate application of Bezout's theorem (see, e.g., [14]). This theorem states that two polynomial functions of degree d_1 and d_2, which do not have common factors (or whose greatest common factor is a constant, which is equivalent to say that the corresponding polynomial system is zero-dimensional) can have at most $d_1 d_2$ common roots. Now, the result follows immediately by observing that the degrees of the two polynomial functions in (24) is at most $2n$. □

In this case all the solutions of the system can be approximated in polynomial time.

Observation 4.3 *Let* 2^δ *be the size of the maximum coefficient in* (18). *If system* (24) *is zero-dimensional, its solutions can be derived with precision* ε *after*

$$\tilde{O}\left(n^6\delta + n^2 \log\left(\frac{1}{\varepsilon}\right)\right)$$

operations, where $\tilde{O}$ *means that polylogarithmic terms are omitted.*

Proof Theorem 11 in [10] states that for a bivariate polynomial system with degree d of the polynomials and with maximum size 2^τ of the coefficients, the number of operations required to accomplish precision ε in the identification of all the solutions is

$$\tilde{O}\left(d^6 + d^5\tau + d^2 \log\left(\frac{1}{\varepsilon}\right)\right).$$

Then, the results follows by observing that our polynomial system has degree $2n$ and maximum size of the coefficients $2^{n\delta}$. □

Once we are able to detect the solutions of the polynomial system with precision ε, we would like to show that we are also able to identify a point $\bar{\mathbf{y}}$ which satisfies (2). We first need to prove a lemma about the distance between distinct solutions of the system.

Lemma 4.1 *A lower bound for the minimum distance between distinct solutions of the polynomial system, denoted in what follows by* sep, *is* $O(2^{-n^4\delta})$.

Proof As a consequence of Corollary 5 in [9], we have that for a bivariate polynomial system with polynomials of degree d and an upper bound 2^τ on the size of the coefficients, the minimal distance between distinct solutions of the system is bounded from below by

$$2^{-2d^4 - 2(4\log(d)+\tau)d^3}.$$

(in fact Corollary 5 in [9] states a more general result for n-dimensional polynomial systems). In our polynomial system we have $d = 2n$ and $\tau = n\delta$ (recall 2^δ is the maximum size of the coefficients defining our problem). Thus:

$$\texttt{sep} \geq O(2^{-n^4\delta}).$$

□

Now we notice that we are able to detect exactly some of the solutions of the polynomial system, namely the solutions (μ_1^j, μ_2^j), $j = 1, \ldots, n$, such that

$$\begin{aligned} c_j + \mu_2^j b_j &= 0 \\ \gamma_j + \mu_1^j + \mu_2^j \eta_j &= 0, \end{aligned} \tag{26}$$

i.e, those discussed in Observation 4.1. Consequently, we are also able to compute exactly the corresponding coordinates y_r, $r = 1, \ldots, n$. Instead, let us consider a solution $(\bar{\mu}_1, \bar{\mu}_2)$ of the system different from (μ_1^j, μ_2^j), $j = 1, \ldots, n$. In this case we are able to compute the values $(\bar{\mu}_1, \bar{\mu}_2)$ only within some precision ε. We would like to establish how the error in the computation of these values affects the error in the computation of each $\bar{y}_j$ value. We have that

$$\bar{y}_j = -\frac{c_j + \bar{\mu}_2 b_j}{\gamma_j + \bar{\mu}_1 + \bar{\mu}_2 \eta_j} \in [-1, 1],$$

where $\bar{y}_j \in [-1, 1]$ follows from the first constraint. Since we are able to identify the solution only within the precision ε, we detect an approximate solution $(\bar{\mu}_1 + \xi_1, \bar{\mu}_2 + \xi_2)$, where $|\xi_1|, |\xi_2| \leq \varepsilon$. Thus, we will have the following approximation of $\bar{y}_j$

$$\tilde{y}_j = -\frac{c_j + (\bar{\mu}_2 + \xi_2) b_j}{\gamma_j + (\bar{\mu}_1 + \xi_1) + (\bar{\mu}_2 + \xi_2)\eta_j}.$$

We would like to compute the error of the approximation, i.e.,

$$|\tilde{y}_j - \bar{y}_j| = \left| -\frac{c_j + (\bar{\mu}_2 + \xi_2) b_j}{\gamma_j + (\bar{\mu}_1 + \xi_1) + (\bar{\mu}_2 + \xi_2)\eta_j} + \frac{c_j + \bar{\mu}_2 b_j}{\gamma_j + \bar{\mu}_1 + \bar{\mu}_2 \eta_j} \right|. \tag{27}$$

We need the following lemma, which gives an upper bound on the approximation error for each coordinate.

Lemma 4.2 *For each $j = 1, \ldots, n$, it holds that*

$$|\tilde{y}_j - \bar{y}_j| \leq \varepsilon O\left(\frac{2^{2\delta}}{\mathtt{sep}}\right).$$

Proof First we notice that we can rewrite (27) as follows

$$|\tilde{y}_j - \bar{y}_j| = \left| \left(\frac{c_j + \bar{\mu}_2 b_j}{\gamma_j + \bar{\mu}_1 + \bar{\mu}_2 \eta_j} \right) \left(\frac{\xi_1 + \xi_2 \eta_j}{\gamma_j + (\bar{\mu}_1 + \xi_1) + (\bar{\mu}_2 + \xi_2)\eta_j} \right) - \frac{\xi_2 b_j}{\gamma_j + (\bar{\mu}_1 + \xi_1) + (\bar{\mu}_2 + \xi_2)\eta_j} \right|.$$

Then,

$$|\tilde{y}_j - \bar{y}_j| \leq \left| \frac{c_j + \bar{\mu}_2 b_j}{\gamma_j + \bar{\mu}_1 + \bar{\mu}_2 \eta_j} \right| \left| \frac{\xi_1 + \xi_2 \eta_j}{\gamma_j + (\bar{\mu}_1 + \xi_1) + (\bar{\mu}_2 + \xi_2)\eta_j} \right| + \left| \frac{\xi_2 b_j}{\gamma_j + (\bar{\mu}_1 + \xi_1) + (\bar{\mu}_2 + \xi_2)\eta_j} \right|.$$

In view of

$$\left| -\frac{c_j + \bar{\mu}_2 b_j}{\gamma_j + \bar{\mu}_1 + \bar{\mu}_2 \eta_j} \right| \leq 1,$$

and $|\xi_1|, |\xi_2| \leq \varepsilon$, we have

$$|\tilde{y}_j - \bar{y}_j| \leq \varepsilon \frac{1 + |\eta_j| + |b_j|}{|\gamma_j + (\bar{\mu}_1 + \xi_1) + (\bar{\mu}_2 + \xi_2)\eta_j|}.$$

Now, for some A, B it holds that

$$\begin{aligned} c_j + \bar{\mu}_2 b_j &= A \\ \gamma_j + \bar{\mu}_1 + \bar{\mu}_2 \eta_j &= B, \end{aligned} \tag{28}$$

where $|B| \geq |A|$. If $b_j = 0$, then $A = c_j$ and $|B| \geq |c_j|$, so that for a small enough ε we have

$$|\gamma_j + (\bar{\mu}_1 + \xi_1) + (\bar{\mu}_2 + \xi_2)\eta_j| = |B + \xi_1 + \xi_2 \eta_j| \geq \frac{|c_j|}{2},$$

and, thus

$$|\tilde{y}_j - \bar{y}_j| \leq \varepsilon \frac{2(1 + |\eta_j|)}{|c_j|} \leq \varepsilon O\left(2^{\delta}\right).$$

Otherwise, if $b_j \neq 0$, if we subtract system (28) to system (26), we get

$$\begin{bmatrix} 0 & b_j \\ 1 & \eta_j \end{bmatrix} \left\{ \begin{pmatrix} \bar{\mu}_1 \\ \bar{\mu}_2 \end{pmatrix} - \begin{pmatrix} \mu_1^j \\ \mu_2^j \end{pmatrix} \right\} = \begin{pmatrix} A \\ B \end{pmatrix}.$$

If we denote by

$$M_j = \begin{bmatrix} 0 & b_j \\ 1 & \eta_j \end{bmatrix},$$

$b_j \neq 0$ implies that M_j is nonsingular. Moreover

$$\left\| \begin{pmatrix} \bar{\mu}_1 \\ \bar{\mu}_2 \end{pmatrix} - \begin{pmatrix} \mu_1^j \\ \mu_2^j \end{pmatrix} \right\| \leq \left\| M_j^{-1} \right\| \left\| \begin{pmatrix} A \\ B \end{pmatrix} \right\|.$$

Note that the left-hand side is bounded from below by $\mathtt{sep}$, which is the minimum distance between distinct solutions of the system. Then

$$\left\| \begin{pmatrix} A \\ B \end{pmatrix} \right\| \geq \frac{\mathtt{sep}}{\left\| M_j^{-1} \right\|}.$$

Thus, recalling $|B| \geq |A|$ and noticing that $\|M_j^{-1}\| = O(2^{\delta})$,

$$|B| \geq O\left(\frac{\mathtt{sep}}{2^{\delta}}\right).$$

Thus, for a small enough ε we have

$$|\gamma_j + (\bar{\mu}_1 + \xi_1) + (\bar{\mu}_2 + \xi_2)\eta_j| = |B + \xi_1 + \xi_2 \eta_j| \geq O\left(\frac{\mathtt{sep}}{2^{\delta}}\right),$$

and, thus

$$|\tilde{y}_j - \bar{y}_j| \leq \varepsilon O\left(\frac{2^{2\delta}}{\mathtt{sep}}\right).$$

□

We recall that our aim is to identify a point for which (2) holds. The following theorem states that this can be done in polynomial time.

Theorem 4.1 *A feasible solution which satisfies* (2) *can be identified in polynomial time with respect to* n, $\log(\Omega)$, $\log(\chi)$, *and* $\log(1/\rho)$.

Proof If we were able to compute exactly all the KKT points, we would be done. Unfortunately, for some points we are only able to detect approximations, which may not be feasible points. However, we would like to show that with a similar computational effort we are able to derive from the approximate solutions at least one feasible solution which satisfies (2).

Let us assume that we have identified a number R of approximate KKT points $\tilde{\mathbf{y}}^r$, $r = 1, \ldots, R$, and let $\bar{\mathbf{y}}^r$ be the corresponding exact KKT points. Let

$$\bar{\mathbf{y}}^s \in \arg\min_{r=1,\ldots,R} \bar{f}_0(\bar{\mathbf{y}}^r),$$

be the best of these KKT points in terms of objective function value, and

$$\tilde{\mathbf{y}}^t \in \arg\min_{r=1,\ldots,R} \bar{f}_0(\tilde{\mathbf{y}}^r),$$

be the best approximated KKT point (notice that $t \neq s$ may hold). In view of Lemma 4.2, we have that for each $r = 1, \ldots, R$

$$\|\tilde{\mathbf{y}}^r - \bar{\mathbf{y}}^r\| \leq \sqrt{n}\varepsilon O\left(\frac{2^{2\delta}}{\mathtt{sep}}\right) \leq \sqrt{n}\varepsilon O\left(2^{2\delta+n^4\delta}\right), \tag{29}$$

where the last inequality follows from the lower bound for $\mathtt{sep}$ established in Lemma 4.1. Now, let us set

$$r_1(n, \delta, \varepsilon) = \sqrt{n}\varepsilon O\left(2^{2\delta+n^4\delta}\right).$$

We choose

$$\varepsilon \leq \frac{1}{O\left(\sqrt{n}2^{2\delta+n^4\delta}\right)},$$

so that $r_1(n, \delta, \varepsilon) \leq 1$. In view of Remark 3.2

$$|\bar{f}_i(\tilde{\mathbf{y}}^r) - \bar{f}_i(\bar{\mathbf{y}}^r)| \leq 3n\Omega r_1(n, \delta, \varepsilon), \quad i = 0, 1, 2. \tag{30}$$

Now we set

$$r_2(n, \delta, \varepsilon, \Omega) = 3n\Omega r_1(n, \delta, \varepsilon).$$

We further restrict the possible values for ε by imposing

$$r_2(n, \delta, \varepsilon, \Omega) \leq \frac{1}{3} \min\left\{\kappa_1, \frac{\kappa_2}{2}\right\} \leq \frac{1}{3} \min\{\kappa_1, \kappa_2\} = \frac{1}{3\chi} \leq \frac{1}{3}, \tag{31}$$

where the last inequality follows from $\kappa_1 \leq 1$. This holds if

$$\varepsilon \geq \frac{1}{O\left(\Omega\chi n^{\frac{3}{2}} 2^{2\delta+n^4\delta}\right)}. \tag{32}$$

In view of the feasibility of $\bar{\mathbf{y}}^t$, we have that

$$\begin{aligned} \bar{f}_1(\tilde{\mathbf{y}}^t) &\leq \tfrac{1}{2} + r_2(n, \delta, \varepsilon, \Omega) \\ \bar{f}_2(\tilde{\mathbf{y}}^t) &\leq \bar{u} + r_2(n, \delta, \varepsilon, \Omega). \end{aligned}$$

Let $\mathbf{z}^t$ be equal to $\tilde{\mathbf{y}}^t$, in case $\tilde{\mathbf{y}}^t$ is feasible for the perturbed problem (6), otherwise let $\mathbf{z}^t$ be a point along the segment between $\tilde{\mathbf{y}}^t$ and the strictly feasible point $\mathbf{x}_0$ defined in (3):

$$\mathbf{z}^t = (1-\lambda)\tilde{\mathbf{y}}^t + \lambda\mathbf{x}_0.$$

In view of the convexity of the constraint functions we have, for any $\lambda \in [0, 1]$,

$$\begin{aligned} \bar{f}_1(\mathbf{z}^t) &\leq (1-\lambda)\bar{f}_1(\tilde{\mathbf{y}}^t) + \lambda\bar{f}_1(\mathbf{x}_0) \leq (1-\lambda)\bar{f}_1(\tilde{\mathbf{y}}^t) + \lambda\left(\tfrac{1}{2} - \kappa_1\right), \\ \bar{f}_2(\mathbf{z}^t) &\leq (1-\lambda)\bar{f}_2(\tilde{\mathbf{y}}^t) + \lambda\bar{f}_2(\mathbf{x}_0) \leq (1-\lambda)\bar{f}_2(\tilde{\mathbf{y}}^t) + \lambda\left(\bar{u} - \tfrac{\kappa_2}{2}\right). \end{aligned}$$

Thus, for λ such that

$$\begin{aligned} (1-\lambda)\left(\tfrac{1}{2} + r_2\right) + \lambda\left(\tfrac{1}{2} - \kappa_1\right) &\leq \tfrac{1}{2} \\ (1-\lambda)(\bar{u} + r_2) + \lambda\left(\bar{u} - \tfrac{\kappa_2}{2}\right) &\leq \bar{u}, \end{aligned}$$

(dependency of r_2 on $n, \delta, \varepsilon, \Omega$ has been omitted), i.e.,

$$\lambda \geq \frac{r_2}{r_2 + \min\left\{\kappa_1, \frac{\kappa_2}{2}\right\}},$$

point $\mathbf{z}^t$ is feasible. If we set λ equal to the right-hand side of the above inequality, then

$$\|\mathbf{z}^t - \tilde{\mathbf{y}}^t\| = \lambda\|\tilde{\mathbf{y}}^t - \mathbf{x}_0\| \leq \frac{r_2}{r_2 + \min\left\{\kappa_1, \frac{\kappa_2}{2}\right\}}\left(\|\tilde{\mathbf{y}}^t\| + \|\mathbf{x}_0\|\right)$$
$$\leq \frac{r_2}{r_2 + \min\left\{\kappa_1, \frac{\kappa_2}{2}\right\}}(2 + 2r_2).$$

Now, let us set

$$r_3(n, \delta, \varepsilon, \Omega, \chi) = \frac{r_2}{r_2 + \min\left\{\kappa_1, \frac{\kappa_2}{2}\right\}}(2 + 2r_2).$$

In view of (31) the right-hand side is not larger than 1, so that in view of Remark 3.2

$$|\bar{f}_0(\tilde{\mathbf{y}}^t) - \bar{f}_0(\mathbf{z}^t)| \leq 3n\Omega\frac{r_2}{r_2 + \min\left\{\kappa_1, \frac{\kappa_2}{2}\right\}}(2 + 2r_2).$$

In view of (30) and of the definition of $\tilde{\mathbf{y}}^t$ as the best approximate KKT point, it holds that

$$\bar{f}_0(\bar{\mathbf{y}}^s) \geq \bar{f}_0(\tilde{\mathbf{y}}^s) - r_2 \geq \bar{f}_0(\tilde{\mathbf{y}}^t) - r_2 \geq \bar{f}_0(\mathbf{z}^t)$$
$$-r_2 - 3n\Omega\frac{r_2}{r_2 + \min\left\{\kappa_1, \frac{\kappa_2}{2}\right\}}(2 + 2r_2).$$

Thus, $\mathbf{z}^t$ is a $\frac{\rho}{2}$-approximate solution of the perturbed problem provided that

$$r_2 + 3n\Omega\frac{r_2}{r_2 + \min\left\{\kappa_1, \frac{\kappa_2}{2}\right\}}(2 + 2r_2) \leq \frac{\rho}{2}. \tag{33}$$

Observing that $r_2 \leq \frac{1}{2}$, we have that the left-hand side in (33) is bounded from above by

$$\frac{9n\Omega r_2 + r_2^2 + r_2 \min\left\{\kappa_1, \frac{\kappa_2}{2}\right\}}{r_2 + \frac{1}{2}\min\{\kappa_1, \kappa_2\}}.$$

This can be further bounded from above, after observing that $r_2^2 \leq r_2$ and $\min\left\{\kappa_1, \frac{\kappa_2}{2}\right\} \leq 1$, with

$$\frac{r_2(18n\Omega + 4)}{\frac{1}{\chi}}.$$

Thus, (33) is fulfilled if

$$2r_2\chi(18n\Omega + 4) \leq \rho,$$

which, recalling the definition of r_2 holds if

$$\varepsilon \leq \frac{\rho}{2\Omega n^{\frac{3}{2}} 2^{2\delta + n^4\delta}\chi(18n\Omega + 4)}. \tag{34}$$

Now, according to Observation 4.3, we are able to detect an approximated KKT point in time

$$\tilde{O}\left(n^6\delta + n^2 \log\left(\frac{1}{\varepsilon}\right)\right).$$

Then, taking into account (34), the time to compute an approximated KKT point is

$$\tilde{O}\left(n^6\delta + n^2\left[\log(1/\rho) + \log(\Omega) + \log(n) + n^4\delta + \log(\chi)\right]\right).$$

Recalling that $\delta = O(k + \log(\Omega))$ and the lower bound (15) for k, we further have that the computing time is

$$\tilde{O}\left(n^6\left[\log(n) + \log(1/\rho) + \log(\Omega) + \log(\chi)\right]\right).$$

Since all the R approximate KKT points need to be computed, and since R is bounded from above by the maximum number $4n^2$ of solutio6ns of the system (24) (see Observation 4.2), we are able to detect a feasible solution which satisfies (2) in polynomial time

$$\tilde{O}\left(n^8\left[\log(n) + \log(1/\rho) + \log(\Omega) + \log(\chi)\right]\right).$$

□

For the sake of illustration, we present a simple example which illustrates how the approximation algorithm works.

Example 4.1 Let us consider the following problem

$$\begin{aligned} \min\ & x^2 - y^2 - 2y + 2x \\ & x^2 + y^2 \leq 1 \\ & x^2 + y^2 - 2x - 2y \leq -1. \end{aligned}$$

It can be easily checked that the problem admits a strictly feasible solution, e.g., the point $(1/2, 1/2)$. The KKT system is the following

$$\begin{aligned} & 2x + 2 + 2\mu_1 x + 2\mu_2 x - 2\mu_2 = 0 \\ & -2y - 2 + 2\mu_1 y + 2\mu_2 y - 2\mu_2 = 0 \\ & \mu_1(x^2 + y^2 - 1) = 0 \\ & \mu_2(x^2 + y^2 - 2x - 2y + 1) = 0 \\ & \mu_1, \mu_2 \geq 0. \end{aligned}$$

We first deal with the case where both constraints are active. The (possible) KKT solutions discussed in Observation 4.1 are those for which the Lagrange multipliers satisfy the two following linear systems

$$\begin{cases} \mu_1 + \mu_2 + 1 = 0 \\ \mu_2 - 1 = 0 \end{cases} \qquad \begin{cases} \mu_1 + \mu_2 - 1 = 0 \\ \mu_2 + 1 = 0. \end{cases}$$

The solutions of these systems are, respectively, $(-2, 1)$ and $(2, -1)$. Both have one negative value and, thus, can be discarded in advance. Once we discard these solutions, we can derive x and y from the first two equations of the KKT system,

$$x = \frac{\mu_2 - 1}{1 + \mu_1 + \mu_2}, \quad y = \frac{\mu_2 + 1}{-1 + \mu_1 + \mu_2}, \tag{35}$$

so that, after imposing that both constraints are active, we end up with the following system

$$\begin{aligned}
&\left(\tfrac{\mu_2-1}{1+\mu_1+\mu_2}\right)^2 + \left(\tfrac{\mu_2+1}{-1+\mu_1+\mu_2}\right)^2 = 1 \\
&\left(\tfrac{\mu_2-1}{1+\mu_1+\mu_2}\right)^2 + \left(\tfrac{\mu_2+1}{-1+\mu_1+\mu_2}\right)^2 - 2\left(\tfrac{\mu_2-1}{1+\mu_1+\mu_2}\right) - 2\left(\tfrac{\mu_2+1}{-1+\mu_1+\mu_2}\right) = -1,
\end{aligned}$$

or, equivalently

$$\begin{aligned}
&(\mu_2 - 1)^2(\mu_1 + \mu_2 - 1)^2 + (\mu_2 + 1)^2(1 + \mu_1 + \mu_2)^2 \\
&\quad -(1 + \mu_1 + \mu_2)^2(-1 + \mu_1 + \mu_2)^2 = 0 \\
&(\mu_1 + 2)^2(\mu_1 + \mu_2 - 1)^2 + (2 - \mu_1)^2(1 + \mu_1 + \mu_2)^2 \\
&\quad -(1 + \mu_1 + \mu_2)^2(-1 + \mu_1 + \mu_2)^2 = 0.
\end{aligned}$$

The solutions of this system are $\{(2, 1), (-2, 1), (2, -1), (-2, -1)\}$. We only need to consider $(2, 1)$ since all other solutions have at least one negative value. Note that, as previously commented, we detected again the solutions derived in Observation 4.1. If we substitute $\mu_1 = 2,\ \mu_2 = 1$ in (35), we find the KKT point $x^* = 0,\ y^* = 1$. We remark that in this simple case exact solutions of the system could be detected. Usually, only approximated Lagrange multipliers can be detected, and, thus, approximated KKT points may need to be projected back within the feasible region.

Next, we need to consider possible KKT points with a single active constraint. Note that, since the objective function is not convex, optimal solutions must have at least one active constraint. We only discuss the case where the first constraint is active (the discussion for the other case is analogous). Since the second constraint is not required to be active, we can impose $\mu_2 = 0$. Thus, the first two equations in the KKT system lead to

$$x = \frac{-1}{1 + \mu_1}, \quad y = \frac{1}{-1 + \mu_1} \tag{36}$$

(note that $\mu_1 = -1, 1$ cannot hold). By imposing the equality in the first constraint, we have

$$\left(\frac{-1}{1 + \mu_1}\right)^2 + \left(\frac{1}{-1 + \mu_1}\right)^2 = 1,$$

or, equivalently

$$\mu_1^4 - 4\mu_1^2 - 1 = 0.$$

The unique real and positive solution is $\sqrt{\sqrt{5}+2}$, from which, substituting in (36), we get the point

$$x = -\frac{1}{1+\sqrt{\sqrt{5}+2}}, \quad y = \frac{1}{-1+\sqrt{\sqrt{5}+2}},$$

which, however, does not satisfy the second constraint. A similar result holds for the case where only the second constraint is active. Thus, we can conclude that the unique KKT point and, thus, the global minimizer of our problem is $x^* = 0,\ y^* = 1$, with global minimum value equal to -3.

5 The case of non commuting matrices

In this section, we no longer assume that $\mathbf{A}$ and $\mathbf{Q}$ commute. This case can still be solved with the ideas presented in Sect. 4. In order to make the paper more readable, in this section we present the main results, while all the intermediate (and more technical) results are presented in Sect. 6. In Sect. 5.1 we discuss the complexity of computing approximate KKT points under suitable generic assumptions. In Sect. 5.2 we introduce the approximation algorithm and prove that it is a polynomial-time algorithm. We remark that, with a slight abuse, in this section we assume that the data of problem (1) have a fractional part rounded at the kth binary digit as in the perturbed problem (6). This is not true in general but is justified by the fact that, as we proved in Sect. 3.1, we can always find an approximate solution of (1) once we have an approximate solution for the perturbed problem (6).

5.1 Solving the KKT system

As before, we discuss only the case when both constraints are active (the other cases can be dealt with in a completely analogous way). Thus, the KKT points for which both constraints are active are the stationary points of the Lagrangian

$$L(\mathbf{x}, \mu_1, \mu_2) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{q}^T\mathbf{x} + \mu_1\left(\frac{1}{2}\mathbf{x}^T\mathbf{x} - 1\right) + \mu_2\left(\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{a}^T\mathbf{x} - u\right), \tag{37}$$

given by the solution of the following system of equations

$$(\mathbf{Q} + \mu_1\mathbf{I} + \mu_2\mathbf{A})\mathbf{x} + \mathbf{q} + \mu_2\mathbf{a} = 0 \tag{38}$$

$$\frac{1}{2}\mathbf{x}^T\mathbf{x} - \frac{1}{2} = 0$$

$$\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{a}^T\mathbf{x} - u = 0. \tag{39}$$

We first need to introduce two hypotheses on the coefficients appearing in (1) that are generically satisfied. The first one is the controllability of the couple $(\mathbf{A}, \mathbf{a})$.

Definition 5.1 The pair $(\mathbf{A}, \mathbf{a})$ is controllable if

$$\det(\mathbf{a}, \mathbf{A}\mathbf{a}, \ldots, \mathbf{A}^{n-1}\mathbf{a}) \neq 0. \tag{40}$$

Controllability (40) is a property considered in the field of control theory. It is a generic property. Indeed (40) is a non-zero polynomial on the entries of $\mathbf{A}$ and $\mathbf{a}$ that defines an algebraic variety of codimension 1. Now, define the constraint functions $f_1, f_2 : \mathbb{C}^n \to \mathbb{C}$ as

$$f_1(\mathbf{x}) = \tfrac{1}{2}(\mathbf{x}^T\mathbf{x} - 1)$$
$$f_2(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{a}^T\mathbf{x} - u.$$

and, for $i = 1, 2$, define the complex algebraic varieties

$$V_i = \{\mathbf{x} \in \mathbb{C}^n : f_i(\mathbf{x}) = 0\}.$$

The second hypothesis is transversality of sets V_1, V_2, defined as follows.

Definition 5.2 Manifolds V_1, V_2 are transversal if, for every $\mathbf{x} \in V_1 \cap V_2$, $\nabla f_1(\mathbf{x})$ and $\nabla f_2(\mathbf{x})$ are linearly independent.

Geometrically, V_1 and V_2 are transversal if they are not tangent. It is well known that transversality is a generic property. The main result of this section, about the detection of the set of solutions of (38)–(39), is the following.

Theorem 5.1 *If the couple* $(\mathbf{A}, \mathbf{a})$ *is controllable and if* V_1, V_2 *are transversal, then the set of solutions of* (38)–(39) *can be approximated within precision* ε_1 *in polynomial time with respect to the number of variables n, to the bit size of the rational coefficients* δ, *and to* $\log(1/\varepsilon_1)$.

The proof follows a reasoning similar to the case of commuting matrices in Sect. 4 and is an immediate consequence of two propositions. Setting $\mathbf{M}(\mu_1, \mu_2) = (\mathbf{Q} + \mu_1\mathbf{I} + \mu_2\mathbf{A})$, the first proposition parallels Observation 4.1 and characterizes the solutions of the first equation when $\mathbf{M}(\mu_1, \mu_2)$ is singular.

Proposition 5.1 *If* $(\mathbf{A}, \mathbf{a})$ *is controllable, the subset of solutions of* (38)–(39) *for which* $\mathbf{M}(\mu_1, \mu_2)$ *is singular can be approximated within precision* ε_1 *in polynomial time with respect to n,* δ, *and* $\log(1/\varepsilon_1)$.

The second proposition characterizes the detection of KKT points for which $\mathbf{M}(\mu_1, \mu_2)$ is non singular, generalizing Observation 4.3.

Proposition 5.2 *If the couple* $(\mathbf{A}, \mathbf{a})$ *is controllable and if* V_1, V_2 *are transversal, the subset of solutions of* (38)–(39) *for which* $\mathbf{M}(\mu_1, \mu_2)$ *is nonsingular can be approximated within precision* ε_1 *in polynomial time with respect to n,* δ, *and* $\log(1/\varepsilon_1)$.

The proofs of these propositions will be presented in Sect. 6.

5.2 The polynomial-time algorithm

Each KKT point $\mathbf{x}^*$ is associated to exact solutions μ_1^*, μ_2^*, $\mathbf{x}^*$ of (38)–(39). However, root-finding algorithms allow to find only an approximation μ_1, μ_2 of the exact solution. In this section, we present a method for computing a feasible solution $\mathbf{x}$ associated to an approximated solution μ_1, μ_2 that guarantees that the value of the objective function of problem (1) on $\mathbf{x}$ is arbitrarily close to its value on $\mathbf{x}^*$. This result will be the basis for the definition of the polynomial-time algorithm for solving problem (1). We first need to introduce some notation and a couple of definitions. Let $\mathcal{F} = \left\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq 1, \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{a}^T\mathbf{x} \leq u\right\}$ be the feasible set of problem (1). Recall that $\mathcal{F}$ is convex and has a non-empty interior. Define function $\nu : \mathbb{R}^n \to \mathcal{F}$ as $\nu(\mathbf{x}) = (1 - \bar{\lambda})\mathbf{x} + \bar{\lambda}\mathbf{x}_0$ where $\bar{\lambda} = \min\{\lambda : 0 < \lambda \leq 1,\ (1 - \lambda)\mathbf{x} + \lambda\mathbf{x}_0 \in \mathcal{F}\}$, where $\mathbf{x}_0$ is the strictly feasible point defined in (3). In this way, function $\nu(\mathbf{x})$ represents the projection of vector $\mathbf{x}$ into the feasible set $\mathcal{F}$. The following decomposition of a symmetric matrix will be needed in what follows.

Definition 5.3 Consider a symmetric matrix $\mathbf{T}$, let λ_i denote its eigenvalues ordered in increasing norm (i.e., $|\lambda_1| \leq |\lambda_2| \leq \cdots \leq |\lambda_n|$), and let $\mathbf{v}_i$ be the associated normalized eigenvectors. For a positive real ϵ, set $l = \min\{i : |\lambda_{i+1}| > \epsilon\}$ and define the projection matrices

$$\mathbf{\Pi}_l = \sum\nolimits_{i=1,\ldots,l} \mathbf{v}_i\mathbf{v}_i^T, \quad \mathbf{\Pi}_g = \mathbf{I} - \mathbf{\Pi}_l.$$

Decompose matrix $\mathbf{T}$ as

$$\mathbf{T}_g = \mathbf{\Pi}_g\mathbf{T}\mathbf{\Pi}_g, \quad \mathbf{T}_l = \mathbf{T} - \mathbf{T}_g.$$

We call the pair $(\mathbf{T}_l, \mathbf{T}_g)$ the ϵ-decomposition of $\mathbf{T}$. Moreover, we decompose each vector $\mathbf{x} \in \mathbb{R}^n$ as $\mathbf{x} = \mathbf{x}_l + \mathbf{x}_g$, where

$$\mathbf{x}_g = \mathbf{\Pi}_g\mathbf{x}, \quad \mathbf{x}_l = \mathbf{\Pi}_l\mathbf{x},$$

and call $(\mathbf{x}_l, \mathbf{x}_g)$ the T_l^g-decomposition of $\mathbf{x}$.

We also need to introduce the following class of functions.

Definition 5.4 Let $\theta = (\mathbf{Q}, \mathbf{A}, \mathbf{q}, \mathbf{a}, u, n, \delta) \in \Lambda$ be the set of parameters appearing in problem (1), together with δ which is the bit size used in their float representation (recall we are assuming that all data have a fractional part rounded at the kth binary digit). A function $\phi : \mathbb{R} \times \Lambda \to \mathbb{R}$ is of class P if there exists a polynomial function p such that

$$|\phi(\epsilon, \theta)| \leq \epsilon 2^{p(n,\delta)}, \quad \forall\theta \in \Lambda.$$

Note that the sum of two functions of class P is also of class P. The following proposition allows to associate to every approximate solution μ_1, μ_2 of (48) a feasible solution $\mathbf{x}$ such that the value of the objective function on $\mathbf{x}$ is very close to the value on $\mathbf{x}^*$.

Proposition 5.3 *Let μ_1^*, μ_2^*, $\mathbf{x}^*$ be a solution of (38) and (39). Let $0 < \epsilon < 1$, μ_1, μ_2 be real constants such that*

$$\|\mathbf{T} - \mathbf{T}^*\|, \|\mathbf{v} - \mathbf{v}^*\| \leq \frac{1}{2}\epsilon^2, \tag{41}$$

where $\mathbf{T} = \mathbf{M}(\mu_1, \mu_2)$, $\mathbf{T}^* = \mathbf{M}(\mu_1^*, \mu_2^*)$, $\mathbf{v} = -\mathbf{q} - \mu_2\mathbf{a}$, $\mathbf{v}^* = -\mathbf{q} - \mu_2^*\mathbf{a}$. Let $(\mathbf{T}_l, \mathbf{T}_g)$ define the ϵ-decomposition of $\mathbf{T}$, and $(\mathbf{v}_l, \mathbf{v}_g)$ define the corresponding T_l^g-decomposition of $\mathbf{v}$. Then, the following statements hold.

(i) Problem

$$\begin{aligned} &\|\mathbf{x}\| \in [1-\epsilon, 1+\epsilon] \\ &\tfrac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{a}^T\mathbf{x} \in \left[u - \left(\tfrac{3}{2}\|\mathbf{A}\| + \|\mathbf{a}\|\right)\epsilon,\, u + \left(\tfrac{3}{2}\|\mathbf{A}\| + \|\mathbf{a}\|\right)\epsilon\right] \\ &\mathbf{T}_g\mathbf{x} = \mathbf{v}_g, \end{aligned} \tag{42}$$

is feasible.

(ii) There exists a class P function $\phi(\epsilon, \theta)$ such that for each feasible solution $\bar{\mathbf{x}}$ of problem

$$\begin{aligned} &\|\mathbf{x}\| \in [1-2\epsilon, 1+2\epsilon] \\ &\tfrac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{a}^T\mathbf{x} \in \left[u - 2\left(\tfrac{3}{2}\|\mathbf{A}\| + \|\mathbf{a}\|\right)\epsilon,\, u + 2\left(\tfrac{3}{2}\|\mathbf{A}\| + \|\mathbf{a}\|\right)\epsilon\right] \\ &\mathbf{T}_g\mathbf{x} = \mathbf{v}_g, \end{aligned} \tag{43}$$

$\nu(\bar{\mathbf{x}})$ is a feasible solution of (1) and $|f(\nu(\bar{\mathbf{x}})) - f(\mathbf{x}^*)| \leq \phi(\epsilon, \theta)$, where $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{q}^T\mathbf{x}$.

The proof of this result is given in Sect. 6.2. First we discuss its main implication, which is also the main result of the paper. For some $\rho > 0$, we search for a ρ-optimal solution of problem (1), i.e., a feasible solution $\bar{\mathbf{x}}$ such that

$$f(\bar{\mathbf{x}}) \leq f(\mathbf{x}^*) + \rho,$$

where $\mathbf{x}^*$ is an optimal solution of (1). We first need a proposition whose proof is given in Sect. 6.3.

Proposition 5.4 *A feasible solution for (43) can be computed in polynomial time by solving a convex optimization problem.*

Next we introduce Algorithm 1 and claim it is a polynomial time algorithm to compute a ρ-optimal solution for problem (1) (to be more precise, the algorithm should include also the approximate computation of KKT points where only one constraint is active, which are also KKT points of a trust-region problem where the non active constraint is ignored, but such computation can be carried on in a way analogous to the case of two active constraints) .

Data: ρ: the required tolerance on the minimum value of the objective function;
ε_1: a positive tolerance value for the computation of the approximate Lagrange multipliers;
Compute all the solutions $(\mu_1, \mu_2)_i$, $i = 1, \ldots, M$ of system (38)–(39), within the tolerance ε_1;
foreach $(\mu_1, \mu_2)_i$, $i = 1, \ldots, M$ **do**
  Compute a solution $\mathbf{x}_i$ of the feasibility problem (43).
end
Set $j \in \arg\min_{i=1,\ldots,M}\{f(v(\mathbf{x}_i)\}$;
return $\mathbf{x}^\star = v(\mathbf{x}_j)$.

Algorithm 1: Polynomial-time approximation algorithm.

The claim is proved in the following theorem.

Theorem 5.2 *Let ε_1 be defined as follows*

$$\varepsilon_1 \geq \frac{\rho^2}{2n^2 2^{2p(n,\delta)+\delta}}, \tag{44}$$

and let ϵ be derived from

$$n^2 2^\delta \varepsilon_1 = \frac{1}{2}\epsilon^2. \tag{45}$$

Then, Algorithm 1 *returns a ρ-optimal solution for problem* (1) *in polynomial time with respect to n, δ, and* $\log(1/\rho)$.

Proof For each approximate pair of Lagrange multipliers (μ_1, μ_2) we need to consider the time $\mathcal{T}_1$ needed for its computation, plus the time $\mathcal{T}_2$ needed for the solution of the feasibility problem (43). According to Proposition 5.4, time $\mathcal{T}_2$ is the time needed to solve a convex optimization problem and is, thus, polynomial. Then, we only need to compute time $\mathcal{T}_1$. Since ε_1 is the precision with which we compute the approximate Lagrange multipliers μ_1, μ_2,

$$|\mu_i^* - \mu_i| \leq \varepsilon_1, \quad i = 1, 2.$$

In view of the definition of $\mathbf{T}$, $\mathbf{T}^*$, $\mathbf{v}$, $\mathbf{v}^*$, it holds that

$$\|\mathbf{T} - \mathbf{T}^*\|, \|\mathbf{v} - \mathbf{v}^*\| \leq n^2 2^\delta \varepsilon_1.$$

Thus, in order to satisfy (41) we need to impose (45). According to Part (ii) of Proposition 5.3, imposing

$$\rho = |\phi(\epsilon, \theta)| \leq \epsilon 2^{p(n,\delta)}, \tag{46}$$

we are able to detect a feasible solution whose function value differs by at most ρ from the function value at the KKT point $\mathbf{x}^*$ corresponding to the exact Lagrange multipliers μ_1^*, μ_2^*. Formula (44) for the selection of the precision ε_1 is then a consequence of (45) and (46). According to Observation 4.3, the approximate Lagrange multipliers can be computed in time

$$\mathcal{T}_1 = \tilde{O}\left(n^6\delta + n^2 \log\left(\frac{1}{\varepsilon_1}\right)\right).$$

Recalling that a lower bound for ε_1 is given in (44), and deriving a lower bound for ϵ from (46), we have that the overall time $\mathcal{T}_1 + \mathcal{T}_2$ is polynomial with respect to n, δ and $\log(1/\rho)$. Since the number of KKT points is polynomial with respect to n, and the global optimizer is a KKT point, we can conclude that a ρ-optimal solution can be attained in polynomial time with respect to n, δ, and $\log(1/\rho)$, as we wanted to prove. □

6 Proofs of the intermediate results

This section is devoted to the proof of all the intermediate results needed for the proofs of the main results.

6.1 Proofs of Propositions 5.1 and 5.2

Before proving these two propositions, we need some technical results. First, controllability is characterized by the following result (known as PBH test in control theory, see for instance Theorem 4.8 of [15]).

Proposition 6.1 *The pair* $(\mathbf{A}, \mathbf{a})$ *is controllable if and only if for every eigenvalue* λ *of* $\mathbf{A}$

$$rank\ (\mathbf{A} - \lambda\mathbf{I}, \mathbf{a}) = n.$$

The following lemma presents a consequence of the controllability of the couple $(\mathbf{A}, \mathbf{a})$.

Lemma 6.1 *If the pair* $(\mathbf{A}, \mathbf{a})$ *is controllable, the set of values* μ_1, μ_2 *for which matrix*

$$[\mathbf{Q} + \mu_1\mathbf{I} + \mu_2\mathbf{A}, \mathbf{q} + \mu_2\mathbf{a}]$$

is not full rank is zero dimensional.

Proof If matrix

$$[\mathbf{Q} + \mu_1\mathbf{I} + \mu_2\mathbf{A}, \mathbf{q} + \mu_2\mathbf{a}]$$

is not full rank, its rows are not linearly independent and there is a non-zero row vector $\mathbf{w} \neq \mathbf{0}$ such that $\mathbf{w}[\mathbf{Q} + \mu_1\mathbf{I} + \mu_2\mathbf{A}, \mathbf{q} + \mu_2\mathbf{a}] = 0$, hence

$$\mathbf{w}(\mathbf{Q} + \mu_2\mathbf{A}) = -\mu_1\mathbf{w}, \quad \mathbf{w}(\mathbf{q} + \mu_2\mathbf{a}) = 0. \tag{47}$$

Set $\mathbf{v}_i = (\mathbf{Q} + \mu_2\mathbf{A})^i(\mathbf{q} + \mu_2\mathbf{a})$, then (47) implies

$$\mathbf{w}[\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_{n-1}] = 0,$$

so that

$$\det[\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_{n-1}] = 0.$$

(Note that this part of the proof follows the ideas of the PBH test for reachability in Proposition 6.1). Function $\det[\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_{n-1}]$ is a polynomial in μ_2 of maximum degree

$$d = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}.$$

Moreover, the coefficient of μ_2^d (the coefficient of maximum degree) is given by

$$\det[\mathbf{a}, \mathbf{A}\mathbf{a}, \mathbf{A}^2\mathbf{a}, \ldots, \mathbf{A}^{n-1}\mathbf{a}],$$

and is not null since $(\mathbf{A}, \mathbf{a})$ is controllable. Hence, the set of μ_2 for which (47) has a solution is finite, being constituted by the roots of a polynomial of degree d. Moreover, for each of these values of μ_2, the set of μ_1 for which the first equation in (47) is satisfied corresponds to the eigenvalues of $\mathbf{Q} + \mu_2\mathbf{A}$ and, again, is finite since it is given by the roots of a polynomial of degree n. □

If $\mathbf{M}(\mu_1, \mu_2)$ is not singular, $\mathbf{x}$ can be computed as a function of μ_1 and μ_2 from (38). Its substitution in (39) gives a system of two equations in the two variables μ_1, μ_2 with the following form

$$\begin{aligned}
&(\mathbf{q} + \mu_2\mathbf{a})^T \operatorname{adj} \mathbf{M}(\mu_1, \mu_2)^T \operatorname{adj} \mathbf{M}(\mu_1, \mu_2)(\mathbf{q} + \mu_2\mathbf{a}) - \det \mathbf{M}(\mu_1, \mu_2)^2 = 0 \\
&(\mathbf{q} + \mu_2\mathbf{a})^T \operatorname{adj} \mathbf{M}(\mu_1, \mu_2)^T \mathbf{A} \operatorname{adj} \mathbf{M}(\mu_1, \mu_2)(\mathbf{q} + \mu_2\mathbf{a}) \\
&\quad + 2\mathbf{a}^T \operatorname{adj} \mathbf{M}(\mu_1, \mu_2)(\mathbf{q} + \mu_2\mathbf{a}) \det \mathbf{M}(\mu_1, \mu_2) - 2u \det \mathbf{M}(\mu_1, \mu_2)^2 \quad = 0,
\end{aligned} \tag{48}$$

where adj denotes the adjugate of a matrix. Its solution set $V \subset \mathbb{C}^2$ is an algebraic variety. We will show that, under the given hypotheses, generically satisfied, V is zero-dimensional, hence, it is composed of a finite set of couples (μ_1, μ_2) that can be approximated in polynomial time. The main idea of the proof is based on the observation that, if the system were positive-dimensional, then it should have complex solutions diverging to infinity, and then to show that, under the given hypotheses, this cannot occur.

The following remark parallels Remark 4.1 of the commuting matrix case. It shows that the set of solutions (μ_1, μ_2) of system (48) includes the couples discussed in Lemma 6.1.

Remark 6.1 If μ_1, μ_2 are such that matrix

$$[\mathbf{Q} + \mu_1\mathbf{I} + \mu_2\mathbf{A}, \mathbf{q} + \mu_2\mathbf{a}] \tag{49}$$

is not full rank, then Eq. (48) is satisfied.

Proof The hypothesis implies that $\mathbf{M}(\mu_1, \mu_2)$ is not full rank and $\det \mathbf{M}(\mu_1, \mu_2) = 0$. Moreover, by the definition of the adjoint matrix, the ith component of $\operatorname{adj} \mathbf{M}(\mu_1, \mu_2)(\mathbf{q} + \mu_2 \mathbf{a})$ is the determinant of the matrix obtain by substituting the ith column of $\mathbf{M}(\mu_1, \mu_2)$ with vector $\mathbf{q} + \mu_2 \mathbf{a}$ and it is 0 since matrix (49) is not full rank. □

Consider now the equation obtained from (48) by homogenization, that is, the system of homogeneous equations obtained by multiplying each monomial in (48) of non maximum degree by a suitable power of the new variable $\mu_0 \in \mathbb{C}$

$$\begin{aligned}
&(\mu_0 \mathbf{q} + \mu_2 \mathbf{a})^T \operatorname{adj} \mathbf{M^h}(\mu_0, \mu_1, \mu_2)^T \operatorname{adj} \mathbf{M^h}(\mu_0, \mu_1, \mu_2)(\mu_0 \mathbf{q} + \mu_2 \mathbf{a}) \\
&\quad - \det \mathbf{M^h}(\mu_0, \mu_1, \mu_2)^2 = 0 \\
&(\mu_0 \mathbf{q} + \mu_2 \mathbf{a})^T \operatorname{adj} \mathbf{M^h}(\mu_0, \mu_1, \mu_2)^T \mathbf{A} \operatorname{adj} \mathbf{M^h}(\mu_0, \mu_1, \mu_2)(\mu_0 \mathbf{q} + \mu_2 \mathbf{a}) \\
&\quad + 2\mu_0 \mathbf{a}^T \operatorname{adj} \mathbf{M^h}(\mu_0, \mu_1, \mu_2)(\mu_0 \mathbf{q} + \mu_2 \mathbf{a}) \det \mathbf{M^h}(\mu_1, \mu_2) \\
&\quad - 2u \det \mathbf{M^h}(\mu_0, \mu_1, \mu_2)^2 = 0,
\end{aligned} \tag{50}$$

where $\mathbf{M}^h(\mu_0, \mu_1, \mu_2) = (\mu_0 \mathbf{Q} + \mu_1 \mathbf{I} + \mu_2 \mathbf{A})$ is the homogenization of $\mathbf{M}$. It is easy to verify that system (50) is homogeneous. The solution set W of (50) is a projective variety (see Definition 5, Chapter 8.2 of [7]). Note that (48) can be reobtained from (50) by setting $\mu_0 = 1$ (dehomogenization). Since $W \supset V$, by Proposition 7, Chapter 8.4 of [7], then $W \supset \bar{V}$, where $\bar{V}$ is the projective closure of V, defined as the homogenization of the ideal associated to V (see Definition 6, Chapter 8.4 of [7]). The subset $W_\infty = W \cap \{(\mu_0, \mu_1, \mu_2) \in \mathbb{P}(\mathbb{C}^3) | \mu_0 = 0\}$ is called the subset of points at infinity of W. This set is obtained by setting $\mu_0 = 0$ in (50) and is given by the solution of the homogeneous system

$$\begin{aligned}
&(\mu_2 \mathbf{a})^T \operatorname{adj} \mathbf{N}(\mu_1, \mu_2)^T \operatorname{adj} \mathbf{N}(\mu_1, \mu_2)(\mu_2 \mathbf{a}) - \det \mathbf{N}(\mu_1, \mu_2)^2 \quad = 0 \\
&(\mu_2 \mathbf{a})^T \operatorname{adj} \mathbf{N}(\mu_1, \mu_2)^T \mathbf{A} \operatorname{adj} \mathbf{N}(\mu_1, \mu_2)(\mu_2 \mathbf{a}) \\
&\quad + 2\mathbf{a}^T \operatorname{adj} \mathbf{N}(\mu_1, \mu_2)(\mu_2 \mathbf{a}) \det \mathbf{M}(\mu_1, \mu_2) - 2u \det \mathbf{N}(\mu_1, \mu_2)^2 = 0,
\end{aligned} \tag{51}$$

where $\mathbf{N}(\mu_1, \mu_2) = (\mu_1 \mathbf{I} + \mu_2 \mathbf{A})$. System (51) plays an important role in the discussion of the solution set of (48). The following lemma shows that transversality of V_1 and V_2 and controllability of $(\mathbf{A}, \mathbf{a})$ guarantee that (51) has no nontrivial (i.e., non null) solution.

Lemma 6.2 *If the couple* $(\mathbf{A}, \mathbf{a})$ *is controllable and* V_1 *and* V_2 *are transversal, system* (51) *has no nontrivial solutions.*

Proof By contradiction, assume that (51) has a nontrivial solution (μ_1, μ_2). Assume first that $\det \mathbf{N}(\mu_1, \mu_2) = 0$, then $\mu_2 \neq 0$, since otherwise it would be $\mu_1 = 0$. Being the system homogeneous, it can be assumed that $\mu_2 = 1$, thus μ_1 is an eigenvalue of $\mathbf{A}$. If matrix $\mathbf{N}(\mu_1, 1)$ had rank lower than $n - 1$, the couple $(\mathbf{A}, \mathbf{a})$ would not be controllable (by Proposition 6.1). Hence, $\mathbf{N}(\mu_1, 1)$ has rank $n-1$ and, by Theorem 4.7C of [16], adj $\mathbf{N}$ has rank 1 and $\operatorname{adj} \mathbf{N}(\mu_1, 1)\mathbf{N}(\mu_1, 1) = 0$. Then, the image of $\mathbf{N}(\mu_1, 1)$

coincides with the kernel of adj $\mathbf{N}(\mu_1, 1)$. Moreover, the first equation of (51) implies that $\mathbf{a}$ belongs to the kernel of adj $\mathbf{N}(\mu_1, 1)$ and, consequently, to the image of $\mathbf{N}(\mu_1, 1)$. Then, there exists a vector $\mathbf{v}$ such that

$$\mathbf{v}^T[\mathbf{N}(\mu_1, 1), \mathbf{a}] = 0,$$

which, by Proposition 6.1, implies that the couple $(\mathbf{A}, \mathbf{a})$ is not controllable. At this point, necessarily $\det \mathbf{N}(\mu_1, \mu_2) \neq 0$. In this case, in view of the first equation of (51), we must have $\mu_2 \neq 0$. Then,

$$\mathbf{x} = -\mu_2 \mathbf{N}(\mu_1, \mu_2)^{-1}\mathbf{a} = -\mu_2 \frac{\text{adj } \mathbf{N}(\mu_1, \mu_2)}{\det \mathbf{N}(\mu_1, \mu_2)}\mathbf{a},$$

is well-defined, and $f_1(\mathbf{x}) = f_2(\mathbf{x}) = 0$, i.e., $\mathbf{x} \in V_1 \cap V_2$, holds. Moreover, by the definition of $\mathbf{x}$

$$\mu_1 \nabla f_1(\mathbf{x}) + \mu_2 \nabla f_2(\mathbf{x}) = (\mu_1 \mathbf{I} + \mu_2 \mathbf{A})\mathbf{x} + \mu_2 \mathbf{a} = \mathbf{N}(\mu_1, \mu_2)\mathbf{x} + \mu_2 \mathbf{a} = 0,$$

which means that V_1 and V_2 are not transversal at $\mathbf{x}$. □

The following lemma presents a property of the solutions of the homogeneous system (50).

Lemma 6.3 *If V_1 and V_2 are transversal and the couple $(\mathbf{A}, \mathbf{a})$ is controllable, then the set of solutions of system (48) is zero-dimensional.*

Proof Assume by contradiction that the set of solutions $V \subset \mathbb{C}^2$ of system (48) is positive dimensional. Let $\bar{V} \in \mathbb{P}(\mathbb{C}^2)$ be the projective closure of V. By definition of a projective space (see, e.g., Section 2, Chapter 8 in [7]) the origin does not belong to $\bar{V}$. By Proposition 7 of Chapter 9.4 of [7], $\bar{V} \neq V$, hence the solution set of the homogeneous system (50) contains points at infinity. Note that $(\bar{V} \setminus V) \subset W_\infty$, where W_∞ is the set of solutions at infinity of (50), given by the solution set of (51). Hence W_∞ contains a nontrivial solution, which contradicts the result of Lemma 6.2. □

Using the previous results, we can prove Propositions 5.1 and 5.2.

Proof of Proposition 5.1 If $\mathbf{M}(\mu_1, \mu_2)$ is singular, (38) has a solution only if $\mathbf{q} + \mu_2 \mathbf{a} \in \text{Im } \mathbf{M}(\mu_1, \mu_2)$ (the column space of matrix $\mathbf{M}(\mu_1, \mu_2)$), which implies that matrix

$$[\mathbf{Q} + \mu_1 \mathbf{I} + \mu_2 \mathbf{A}, \mathbf{q} + \mu_2 \mathbf{a}]$$

is not full rank. Then, the result follows from Lemma 6.1.

Proof of Proposition 5.2 If $\mathbf{M}(\mu_1, \mu_2)$ is nonsingular, $\mathbf{x}$ can be explicitly computed as

$$\mathbf{x} = -(\mathbf{Q} + \mu_1 \mathbf{I} + \mu_2 \mathbf{A})^{-1}(\mathbf{q} + \mu_2 \mathbf{a}) = -\frac{\text{adj } \mathbf{M}(\mu_1, \mu_2)}{\det \mathbf{M}(\mu_1, \mu_2)}(\mathbf{q} + \mu_2 \mathbf{a}).$$

Its substitution in (39) and multiplication by the non-null term $\det \mathbf{M}(\mu_1, \mu_2)^2$ gives the polynomial system (48). Note that multiplication by $\det \mathbf{M}(\mu_1, \mu_2)^2$ adds to the zero set the values of μ_1, μ_2 for which $\det \mathbf{M}(\mu_1, \mu_2) = 0$. The zero set of (48) is zero-dimensional by Lemma 6.2. Since (48) is a system of two bivariate polynomial equations of degree $2n$ with a zero-dimensional solution, its set of solutions can be approximated in polynomial time (see [8]).

6.2 Proof of Proposition 5.3

Part (i) Consider the T_l^g-decomposition $(\mathbf{x}_l^*, \mathbf{x}_g^*)$ of $\mathbf{x}^*$, and set $\bar{\mathbf{x}} = \mathbf{x}_l^* + \bar{\mathbf{x}}_g$, where $\bar{\mathbf{x}}_g$ is the solution of $\mathbf{T}_g \bar{\mathbf{x}}_g = \mathbf{v}_g$. Note that, by (38), $\mathbf{T}^* \mathbf{x}^* = \mathbf{v}^*$. We claim that $\bar{\mathbf{x}}$ is a feasible solution of (42). First, note that for any $\mathbf{x} \in \mathbb{R}^n$, if $\mathbf{x}_g = \mathbf{\Pi}_g \mathbf{x}$ and $\mathbf{x}_l = \mathbf{\Pi}_l \mathbf{x}$, $\mathbf{T}_g \mathbf{x} = \mathbf{T}_g(\mathbf{x}_g + \mathbf{x}_l) = \mathbf{T}_g \mathbf{x}_g$, being $\mathbf{T}_g \mathbf{x}_l = 0$. Moreover,

$$\begin{aligned}\mathbf{T}_g(\bar{\mathbf{x}}_g - \mathbf{x}_g^*) &= \mathbf{T}_g \bar{\mathbf{x}}_g - (\mathbf{\Pi}_g \mathbf{T}^* + \mathbf{T}_g - \mathbf{\Pi}_g \mathbf{T}^*)\mathbf{x}^* = \mathbf{v}_g - \mathbf{\Pi}_g \mathbf{v}^* - (\mathbf{T}_g - \mathbf{\Pi}_g \mathbf{T}^*)\mathbf{x}^* \\ &= \mathbf{\Pi}_g(\mathbf{v} - \mathbf{v}^*) - \mathbf{\Pi}_g(\mathbf{T} - \mathbf{T}^*)\mathbf{x}^*,\end{aligned}$$

where we have used the identity $\mathbf{\Pi}_g \mathbf{T} = \mathbf{T}_g$. Hence, since $\mathbf{\Pi}_g$ is a projection matrix, $\|\mathbf{\Pi}_g\| \leq 1$ and

$$\|\mathbf{T}_g(\bar{\mathbf{x}}_g - \mathbf{x}_g^*)\| \leq \|\mathbf{v} - \mathbf{v}^*\| + \|\mathbf{T} - \mathbf{T}^*\|.$$

Observe that, by the definition of $\mathbf{T}_g$, for any $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{T}_g \mathbf{x}_g\| > \epsilon \|\mathbf{x}_g\|$. Being $\mathbf{x}_l = \mathbf{x}_l^*$ and by bounds (41),

$$\|\bar{\mathbf{x}} - \mathbf{x}^*\| = \|\bar{\mathbf{x}}_g - \mathbf{x}_g^*\| \leq \epsilon^{-1} \|\mathbf{T}_g(\bar{\mathbf{x}}_g - \mathbf{x}_g^*)\| \leq \epsilon^{-1}(\|\mathbf{v} - \mathbf{v}^*\| + \|\mathbf{T} - \mathbf{T}^*\|) \leq \epsilon.$$

Set $\mathbf{h} = \bar{\mathbf{x}} - \mathbf{x}^*$, then $|\|\bar{\mathbf{x}}\| - 1| = |\|\mathbf{x}^* + \mathbf{h}\| - \|\mathbf{x}^*\|| \leq \|\mathbf{h}\| \leq \epsilon$, and

$$\begin{aligned}&\left|\frac{1}{2}\bar{\mathbf{x}}^T \mathbf{A} \bar{\mathbf{x}} + \mathbf{a}^T \bar{\mathbf{x}} - u\right| = \left|(\bar{\mathbf{x}} - \mathbf{x}^*)^T \mathbf{A} \bar{\mathbf{x}} - \frac{1}{2}(\bar{\mathbf{x}} - \mathbf{x}^*)\mathbf{A}(\bar{\mathbf{x}} - \mathbf{x}^*) + \mathbf{a}^T(\bar{\mathbf{x}} - \mathbf{x}^*)\right| \\ &\quad \leq \epsilon \left(\frac{3}{2}\|\mathbf{A}\| + \|\mathbf{a}\|\right),\end{aligned}$$

where we have used the fact that both equality constraints are active on $\mathbf{x}^*$ and that $\epsilon < 1$. Hence, $\bar{\mathbf{x}}$ is a feasible solution of (42).

Part (ii) Let $\mathbf{x}$ be a feasible solution of (43). We first give a comment on the general idea on this part of the proof. Since $\mathbf{T}_g^* \mathbf{x}^* = \mathbf{v}_g^*$, the equality condition in (43) guarantees that the component of the error $\mathbf{x}_g - \mathbf{x}_g^*$ is infinitesimal with ϵ (more precisely, we will show that it is of class P). We cannot guarantee that the other component of the error $\mathbf{x}_l - \mathbf{x}_l^*$ is small, but we will show that this component of the error does not affect significantly the value of the objective function (more precisely, we will show that $f(\mathbf{x}) - f(\mathbf{x}^*)$ is a class P function).

Note that, by the definition of $\mathbf{T}^*$ and $\mathbf{v}^*$, $\forall \mathbf{x} \in \mathbb{R}^n$

$$\frac{1}{2}\mathbf{x}^T(\mathbf{T}^*\mathbf{x}-\mathbf{v}^*) = f(\mathbf{x}) + \mu_1^*\left(\frac{1}{2}\mathbf{x}^T\mathbf{x}-\frac{1}{2}\right) + \mu_2^*\left(\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}+\mathbf{a}^T\mathbf{x}-u\right) + \frac{1}{2}\mathbf{x}^T\mathbf{v}^* + \frac{1}{2}\mu_1^* + \mu_2^* u. \tag{52}$$

Since $\mathbf{x}^*$ and $\mathbf{v}^*$ satisfy conditions

$$\begin{aligned}&\mathbf{T}^*\mathbf{x}^* - \mathbf{v}^* = 0\\ &\tfrac{1}{2}(\mathbf{x}^*)^T\mathbf{x}^* - \tfrac{1}{2} = 0\\ &\tfrac{1}{2}(\mathbf{x}^*)^T\mathbf{A}\mathbf{x}^* + \mathbf{a}^T\mathbf{x}^* - u = 0,\end{aligned}$$

Equation (52) for $\mathbf{x} = \mathbf{x}^*$ writes as

$$0 = f(\mathbf{x}^*) + \frac{1}{2}(\mathbf{x}^*)^T\mathbf{v}^* + \frac{1}{2}\mu_1^* + \mu_2^* u. \tag{53}$$

Subtracting (53) from (52),

$$f(\mathbf{x}) - f(\mathbf{x}^*) = -\mu_1^*\left(\frac{1}{2}\mathbf{x}^T\mathbf{x}-\frac{1}{2}\right) - \mu_2^*\left(\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}+\mathbf{a}^T\mathbf{x}-u\right) + \frac{1}{2}\mathbf{x}^T(\mathbf{T}^*\mathbf{x}-\mathbf{v}^*) - \frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^T\mathbf{v}^*. \tag{54}$$

To prove the statement, we show that every term appearing on the right-hand side of (54) is a class P function. Note that, by reference [9], $|\mu_1^*|, |\mu_2^*| \leq 2^{d(2d^3+6+2\tau)}$, where $d = 2n$ is the maximum degree of the polynomials in system (48), and $\tau = n\delta$ is the bit size of the coefficients of the same polynomials. Moreover, since $\mathbf{x}$ is a feasible solution of (43)

$$\left\|\mu_1^*\left(\frac{1}{2}\mathbf{x}^T\mathbf{x}-\frac{1}{2}\right)\right\| \leq 4|\mu_1^*|\epsilon, \quad \left|\mu_2^*\left(\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}+\mathbf{a}^T\mathbf{x}-u\right)\right| \leq |\mu_2^*|\,(4\|\mathbf{A}\|+2\|\mathbf{a}\|)\,\epsilon,$$

which implies that the first two terms of the right-hand side of (54) are class P functions. Since $\mathbf{T}^*\mathbf{x}^* - \mathbf{v}^* = 0$, the third term of the rhs of (54) is rewritten as

$$\frac{1}{2}\mathbf{x}^T(\mathbf{T}^*\mathbf{x}-\mathbf{v}^*) = \frac{1}{2}\mathbf{x}^T(\mathbf{T}^*(\mathbf{x}-\mathbf{x}^*+\mathbf{x}^*)-\mathbf{v}^*) = \frac{1}{2}\mathbf{x}^T\mathbf{T}^*(\mathbf{x}-\mathbf{x}^*),$$

and

$$\mathbf{T}^*(\mathbf{x}-\mathbf{x}^*) = ((\mathbf{T}^*-\mathbf{T})+\mathbf{T})(\mathbf{x}-\mathbf{x}^*) = (\mathbf{T}^*-\mathbf{T})(\mathbf{x}-\mathbf{x}^*)+\mathbf{T}(\mathbf{x}_g-\mathbf{x}_g^*)+\mathbf{T}(\mathbf{x}_l-\mathbf{x}_l^*). \tag{55}$$

Thus, each term of the right-hand side of (55) is of class P. Indeed,

$$\begin{aligned}\|(\mathbf{T}^* - \mathbf{T})(\mathbf{x} - \mathbf{x}^*)\| &\leq \|\mathbf{T}^* - \mathbf{T}\|\,\|\mathbf{x} - \mathbf{x}^*\| \leq \|\mathbf{T}^* - \mathbf{T}\|\,(\|\mathbf{x}\| + \|\mathbf{x}^*\|) \\ &\leq \frac{1}{2}\epsilon^2(2 + 2\epsilon),\end{aligned}$$

where we have used the facts that $\|\mathbf{x}\| \leq 1 + 2\epsilon$ and $\|\mathbf{x}^*\| = 1$. Moreover, by the definition of $\mathbf{T}_l$,

$$\left\|\mathbf{T}\left(\mathbf{x}_l - \mathbf{x}_l^*\right)\right\| = \left\|\mathbf{T}_l\left(\mathbf{x}_l - \mathbf{x}_l^*\right)\right\| \leq \epsilon(2 + 2\epsilon).$$

Next, we notice that

$$\left\|\mathbf{T}\left(\mathbf{x}_g - \mathbf{x}_g^*\right)\right\| = \left\|\mathbf{T}\mathbf{x}_g - \left(\mathbf{T}^* - \mathbf{T}^* + \mathbf{T}\right)\mathbf{x}_g^*\right\| \leq \left\|\mathbf{v}_g - \mathbf{v}_g^*\right\| + \left\|(\mathbf{T} - \mathbf{T}^*)\mathbf{x}_g^*\right\| \leq \epsilon^2,$$

where the last inequality follows from (41). The last term on the right-hand side of (54) is bounded as follows

$$\begin{aligned}\left\|\frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T\mathbf{v}^*\right\| &= \left\|\frac{1}{2}\left(\mathbf{x}_g - \mathbf{x}_g^*\right)^T\mathbf{v}_g^*\right\| + \left\|\frac{1}{2}\left(\mathbf{x}_l - \mathbf{x}_l^*\right)^T\mathbf{v}_l^*\right\| \\ &\leq \frac{1}{2}(\|\mathbf{x}_g - \mathbf{x}_g^*\|\|\mathbf{v}_g^*\| + (2 + 2\epsilon)\|\mathbf{v}_l^*\|. \qquad (56)\end{aligned}$$

Note that $\|\mathbf{v}_g^*\|, \|\mathbf{v}_l^*\| \leq \|\mathbf{v}^*\|$ and $\|\mathbf{v}^*\| \leq \|\mathbf{q}\| + \|\mathbf{a}\||\mu_2^*|$. Conditions $\mathbf{T}_g\bar{\mathbf{x}}_g = \mathbf{v}_g$ and $\mathbf{T}_g\mathbf{x}_g = \mathbf{v}_g$ imply that $\mathbf{x}_g = \bar{\mathbf{x}}_g$, hence, by part (i) of the proof $\|\mathbf{x}_g - \mathbf{x}_g^*\| \leq \epsilon$, which shows that the first term of the right-hand side of (56) is of class P. Moreover,

$$\|\mathbf{v}_l^*\| = \|(\mathbf{T}^*\mathbf{x}^*)_l\| = \|((\mathbf{T}^* - \mathbf{T})\mathbf{x}^*)_l + \mathbf{T}\mathbf{x}_l^*\| \leq \frac{1}{2}\epsilon^2 + \epsilon.$$

Finally, $\mathbf{x}$ may not be a feasible solution for (1), since the constraints have been relaxed. A feasible solution is given by $\nu(\mathbf{x})$ and, as a last step of the proof, we need to prove that the distance between $\mathbf{x}$ and the projection $\nu(\mathbf{x})$ is of class P. Indeed, setting $r_2 = \max\{(3\|\mathbf{A}\| + 2\|\mathbf{a}\|)\epsilon, 4\epsilon\}$ and following the same reasoning used in the proof of Theorem 4.1, we obtain the bound

$$\|\nu(\mathbf{x}) - \mathbf{x}\| \leq \frac{r_2}{r_2 + \min\{\kappa_1, \frac{\kappa_2}{2}\}}(2 + 2r_2),$$

which concludes the proof. □

6.3 Proof of Proposition 5.4

For each (approximate) solution pair (μ_1, μ_2) of the system (38)–(39), set $\mathbf{T} = \mathbf{M}(\mu_1, \mu_2)$, $\mathbf{v} = -\mathbf{q} - \mu_2\mathbf{a}$ and let $\mathbf{x}_g$ be the minimum norm solution of $\mathbf{T}_g\mathbf{x}_g = \mathbf{v}_g$,

i.e., $\mathbf{x}_g = \mathbf{T}_g^+ \mathbf{v}_g$, where $\mathbf{T}_g^+$ is the Moore–Penrose inverse of $\mathbf{T}_g$. Let $\mathbf{P}$ be an (orthonormal) basis matrix of the null space of $\mathbf{T}_g$. Then, a feasible point for (43) can be computed by solving the following problem in variable $\mathbf{y}$ for some non null vector $\mathbf{w} \in \mathbb{R}^n$

$$\begin{aligned}
&\min \mathbf{w}^T \mathbf{y} \\
&\|\mathbf{y}\|^2 + \|\mathbf{x}_g\|^2 \in (1 - 2\epsilon, 1 + 2\epsilon) \\
&\frac{1}{2}(\mathbf{x}_g + \mathbf{P}\mathbf{y})^T \mathbf{A}(\mathbf{x}_g + \mathbf{P}\mathbf{y}) + \mathbf{a}^T(\mathbf{x}_g + \mathbf{P}\mathbf{y}) \in \left[u - \left(\frac{3}{2}\|\mathbf{A}\| + \|\mathbf{a}\|\right) 2\epsilon, u \right. \\
&\quad \left. + \left(\frac{3}{2}\|\mathbf{A}\| + \|\mathbf{a}\|\right) 2\epsilon \right].
\end{aligned} \tag{57}$$

Note that, in view of Proposition 5.3, which states the existence of a feasible solution for (42), a strictly feasible solution for (43) exists for this problem, i.e., Slater's condition holds. Before proceeding, we make the following remark.

Remark 6.2 Exact computation of $\mathbf{x}_g$ and $\mathbf{P}$ might not be possible, so that we need to employ approximations $\bar{\mathbf{x}}_g = \mathbf{x}_g + \Delta\mathbf{x}_g$ and $\bar{\mathbf{P}} = \mathbf{P} + \Delta\mathbf{P}$. In this case, rather than solving (57), we should solve

$$\begin{aligned}
&\min \mathbf{w}^T \mathbf{y} \\
&\|\mathbf{y}\|^2 + \|\bar{\mathbf{x}}_g\|^2 \in \left(1 - \frac{3}{2}\epsilon, 1 + \frac{3}{2}\epsilon\right) \\
&\frac{1}{2}(\bar{\mathbf{x}}_g + \bar{\mathbf{P}}\mathbf{y})^T \mathbf{A}(\bar{\mathbf{x}}_g + \bar{\mathbf{P}}\mathbf{y}) + \mathbf{a}^T(\bar{\mathbf{x}}_g + \bar{\mathbf{P}}\mathbf{y}) \in \left[u - \left(\frac{3}{2}\|\mathbf{A}\| + \|\mathbf{a}\|\right) \frac{3}{2}\epsilon, u \right. \\
&\quad \left. + \left(\frac{3}{2}\|\mathbf{A}\| + \|\mathbf{a}\|\right) \frac{3}{2}\epsilon \right].
\end{aligned}$$

Provided that $\|\Delta\mathbf{x}_g\|$ and $\|\Delta\mathbf{P}\|$ are small enough (say, not larger than $\frac{\epsilon}{8}$), any feasible solution of this problem is also a feasible solution of problem (57). Keeping this in mind, in what follows we will assume that exact computation of $\mathbf{x}_g$ and $\mathbf{P}$ is possible.

After diagonalization, where $\mathbf{P}^T\mathbf{A}\mathbf{P}$ is replaced by $\mathbf{Q}\mathbf{D}\mathbf{Q}^T$, and the change of variable $\mathbf{z} = \mathbf{Q}^T\mathbf{y} \in \mathbb{R}^{n'}$, where $n' = n - l$, the problem to be solved is the following

$$\begin{aligned}
\min\ & \textstyle\sum_{i=1}^{n'} w_i z_i \\
& \tfrac{1}{2}\textstyle\sum_{i=1}^{n'} z_i^2 \le \alpha_0 + 2\epsilon \\
& \tfrac{1}{2}\textstyle\sum_{i=1}^{n'} \theta_i z_i^2 + \sum_{i=1}^{n'} d_i z_i \le \beta_0 + \tilde{\epsilon} \\
& -\tfrac{1}{2}\textstyle\sum_{i=1}^{n'} z_i^2 \le -\alpha_0 + 2\epsilon \\
& -\tfrac{1}{2}\textstyle\sum_{i=1}^{n'} \theta_i z_i^2 - \sum_{i=1}^{n'} d_i z_i \le -\beta_0 + \tilde{\epsilon},
\end{aligned} \tag{58}$$

with $\theta_i > 0$, $i = 1, \ldots, n'$, and:

$$\begin{aligned}
\alpha_0 &= 1 - \|\mathbf{x}_g\|^2 \\
\beta_0 &= u - \tfrac{1}{2}\mathbf{x}_g^T \mathbf{A}\mathbf{x}_g - \mathbf{a}^T \mathbf{x}_g \\
\tilde{\epsilon} &= \left(\tfrac{3}{2}\|\mathbf{A}\| + \|\mathbf{a}\|\right) 2\epsilon.
\end{aligned}$$

During the diagonalization we might introduce some approximation of the exact matrices $\mathbf{Q}$ and $\mathbf{D}$. We refer again to Remark 6.2 to deal with this issue. Note that we can choose the values w_i, $i = 1, \ldots, n'$, arbitrarily. We claim that, for suitable choices of the w_i values, a solution of this problem can be detected in polynomial time. Note that this is a result of independent interest. From the geometrical point of view, the problem is equivalent to the detection of an approximate point at the intersection of the borders of two axis-aligned ellipsoids.

After introducing the variables t_i, $i = 1, \ldots, n'$, problem (58), can be rewritten as

$$\begin{aligned}
\min\ & \textstyle\sum_{i=1}^{n'} w_i z_i \\
& \textstyle\sum_{i=1}^{n'} t_i \leq \alpha_0 + 2\epsilon \\
& \textstyle\sum_{i=1}^{n'} \theta_i t_i + \sum_{i=1}^{n'} d_i z_i \leq \beta_0 + \tilde{\epsilon} \\
& \textstyle -\sum_{i=1}^{n'} t_i \leq -\alpha_0 + 2\epsilon \\
& \textstyle -\sum_{i=1}^{n'} \theta_i t_i - \sum_{i=1}^{n'} d_i z_i \leq -\beta_0 + \tilde{\epsilon} \\
& \tfrac{1}{2} z_i^2 = t_i \qquad\qquad i = 1, \ldots, n'.
\end{aligned} \tag{59}$$

If we replace $=$ with $\leq$ in the last constraints, we are led to the following convex relaxation of (58)

$$\begin{aligned}
\min\ & \textstyle\sum_{i=1}^{n'} w_i z_i \\
& \textstyle\sum_{i=1}^{n'} t_i \leq \alpha_0 + 2\epsilon \\
& \textstyle\sum_{i=1}^{n'} \theta_i t_i + \sum_{i=1}^{n'} d_i z_i \leq \beta_0 + \tilde{\epsilon} \\
& \textstyle -\sum_{i=1}^{n'} t_i \leq -\alpha_0 + 2\epsilon \\
& \textstyle -\sum_{i=1}^{n'} \theta_i t_i - \sum_{i=1}^{n'} d_i z_i \leq -\beta_0 + \tilde{\epsilon} \\
& \tfrac{1}{2} z_i^2 - t_i \leq 0 \qquad\qquad i = 1, \ldots, n'.
\end{aligned} \tag{60}$$

As already remarked, a feasible solution for this problem exists even if we replace 2ϵ in the right-hand side of the constraints with ϵ, and $\tilde{\epsilon}$ with $\frac{\tilde{\epsilon}}{2}$. Thus: (i) the convex problem (60) admits a strictly feasible solution and Slater's condition holds; (ii) the feasible region of the problem contains a $2n'$-dimensional sphere whose radius is at least $O\left(\frac{\epsilon}{n\Omega}\right)$.

The KKT conditions for the convex problem are

$$\begin{array}{ll}
\lambda_1 + \lambda_2\theta_i - \lambda_3 - \lambda_4\theta_i - \nu_i = 0 & i = 1, \ldots, n' \\
w_i + \lambda_2 d_i - \lambda_4 d_i + \nu_i z_i = 0 & i = 1, \ldots, n' \\
\sum_{i=1}^{n'} t_i \leq \alpha_0 + 2\epsilon & \\
\sum_{i=1}^{n'} \theta_i t_i + \sum_{i=1}^{n'} d_i z_i \leq \beta_0 + \tilde{\epsilon} & \\
-\sum_{i=1}^{n'} t_i \leq -\alpha_0 + 2\epsilon & \\
-\sum_{i=1}^{n'} \theta_i t_i - \sum_{i=1}^{n'} d_i z_i \leq -\beta_0 + \tilde{\epsilon} & \\
\frac{1}{2} z_i^2 \leq t_i & i = 1, \ldots, n \\
\lambda_1 \left(\sum_{i=1}^{n'} t_i - \alpha_0 - 2\epsilon \right) = 0 & \\
\lambda_2 \left(\sum_{i=1}^{n'} \theta_i t_i + \sum_{i=1}^{n'} d_i z_i - \beta_0 - \tilde{\epsilon} \right) = 0 & \\
\lambda_3 \left(\sum_{i=1}^{n'} t_i - \alpha_0 + 2\epsilon \right) = 0 & \\
\lambda_4 \left(\sum_{i=1}^{n'} \theta_i t_i + \sum_{i=1}^{n'} d_i z_i - \beta_0 + \tilde{\epsilon} \right) = 0 & \\
\nu_i \left(\frac{1}{2} z_i^2 - t_i \right) = 0 & i = 1, \ldots, n' \\
\lambda_1, \lambda_2, \lambda_3, \lambda_4, \nu_i \geq 0 & i = 1, \ldots, n',
\end{array} \tag{61}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are the Lagrange multipliers of the first four constraint, while ν_i, $i = 1, \ldots, n'$, are the Lagrange multipliers of the constraints $\frac{1}{2} z_i^2 \leq t_i$. If $\nu_i > 0 \; \forall i$, then the convex relaxation is exact. In particular, if this is the case, a feasible point for (43) can be computed by solving a convex problem in polynomial time, which is exactly what we want to prove. To see that $\nu_i > 0 \; \forall i$, let us assume by contradiction that $\nu_i = 0$ for some i. We show that, under suitable choices for the values w_i, we are led to a contradiction. Let

$$J_{min} = \{i \; : \; \theta_i \leq \theta_j, \; j = 1, \ldots, n\} \quad J_{max} = \{i \; : \; \theta_i \geq \theta_j, \; j = 1, \ldots, n\}.$$

We impose:

- $w_i \neq 0 \; \forall i$;
- $\frac{w_i}{d_i} \neq \frac{w_j}{d_j} \; \forall i \neq j$ such that $d_i, d_j \neq 0$;
- if $|J_{min}| = 1$ or $|J_{max}| = 1$, select an index j according to the following rules

$$j \notin \begin{cases} J_{min} \cup J_{max} & \text{if } |J_{min}|, \; |J_{max}| = 1 \\ J_{min} & \text{if } |J_{min}| = 1, \; |J_{max}| > 1 \\ J_{max} & \text{if } |J_{min}| > 1, \; |J_{max}| = 1 \end{cases} \tag{62}$$

and set w_j large enough so that for each $i \in J_{min} \cup J_{max}$ if $|J_{min}|, \; |J_{max}| = 1$ (respectively, $i \in J_{min}$ if only $|J_{min}| = 1$, and $i \in J_{max}$ if only $|J_{max}| = 1$)

$$\frac{1}{2}\left(\frac{-w_j + \frac{w_i}{d_i}d_j}{\frac{w_i}{d_i}(\theta_i - \theta_j)}\right)^2 > \alpha_0 + 2\epsilon. \tag{63}$$

Note that $\theta_i - \theta_j \neq 0$ holds in view of the rule (62) used to select j.

We first remark that we cannot have $v_i = 0$ when $d_i = 0$. Indeed, in this case $w_i + v_i z_i = 0$ holds with $w_i \neq 0$. Next, we remark that we cannot have two distinct i, j such that $v_i = v_j = 0$. Indeed, in view of $w_i + \lambda_2 d_i - \lambda_4 d_i = 0$ and $w_j + \lambda_2 d_j - \lambda_4 d_j = 0$, we should either have

$$\lambda_2 = -\frac{w_i}{d_i} = -\frac{w_j}{d_j},$$

or

$$\lambda_4 = \frac{w_i}{d_i} = \frac{w_j}{d_j},$$

which is not possible. Then, only a single value v_i can be equal to 0. But if $v_i = 0$, then

$$\begin{aligned} &\lambda_1 + \lambda_2\theta_i - \lambda_3 - \lambda_4\theta_i = 0 \\ &w_i + \lambda_2 d_i - \lambda_4 d_i = 0. \end{aligned} \tag{64}$$

Since $w_i \neq 0$, then either $\lambda_2 > 0$ or $\lambda_4 > 0$ (not both, since we cannot have both constraints active). Similarly, exactly one among λ_1 and λ_3 must be strictly positive. Let us consider the possible combinations.

$\lambda_1, \lambda_2 > 0$ In this case it follows from (64) that

$$\lambda_2 = -\frac{w_i}{d_i}, \quad \lambda_1 = \theta_i \frac{w_i}{d_i},$$

so that, in view of $\theta_i > 0$, either λ_1 or λ_2 is negative, which is a contradiction.

$\lambda_3, \lambda_4 > 0$ Similar to the previous case.

$\lambda_2, \lambda_3 > 0$ In this case we have from (64)

$$\lambda_2 = -\frac{w_i}{d_i}, \quad \lambda_3 = -\theta_i \frac{w_i}{d_i}.$$

From $\lambda_2\theta_j - \lambda_3 - v_j = 0$, $j \neq i$, we must have

$$v_j = \frac{w_i}{d_i}(\theta_i - \theta_j) \quad \forall j \neq i.$$

Note that $\theta_j \neq \theta_i$ must hold, otherwise $v_j = 0$, which is not possible, as already commented. In view of $v_j > 0, \forall j \neq i$, and since $\lambda_2 = -\frac{w_i}{d_i} > 0$, we must have that

$$v_i = 0 \Rightarrow \theta_i < \theta_j \quad \forall j \neq i.$$

In other words we must have $i \in J_{min}$ and $|J_{min}| = 1$. Now, let j be selected as in (62). From $w_j + \lambda_2 d_j + \nu_j z_j = 0$, $j \neq i$, we have

$$z_j = \frac{-w_j + \frac{w_i}{d_i} d_j}{\frac{w_i}{d_i}(\theta_i - \theta_j)}.$$

In view of the definition (63) of w_j we have $t_j = \frac{1}{2} z_j^2 > \alpha_0 + 2\epsilon$, so that $\sum_{i=1}^{n} t_i \leq \alpha_0 + 2\epsilon$ is violated;

$\lambda_1, \lambda_4 > 0$ similar to the previous case with J_{max} replacing J_{min}.

Acknowledgments The authors are extremely grateful to the Associate Editor and the anonymous reviewers. The comments and critics to previous submissions of our work allowed us to significatively improve the final result.

References

1. Ai, W., Zhang, S.: Strong duality for the CDT subproblem: a necessary and sufficient condition. SIAM J. Optim. **19**, 1735–1756 (2008)
2. Ben-Tal, A., Teboulle, M.: Hidden convexity in some nonconvex quadratically constrained quadratic programming. Math. Program. **72**, 51–63 (1996)
3. Ben-Tal, A., den Her tog, D.: Hidden conic quadratic representation of some nonconvex quadratic optimization problems. Math. Program. **143**, 1–29 (2014)
4. Bienstock, D., Michalka, A.: Polynomial solvability of variants of the trust-region subproblem. In: SODA '14 Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 380–390 (2014)
5. Burer, S., Anstreicher, K.: Second-order cone constraints for extendedtrust-region subproblems. SIAM J. Optim. **23**, 432451 (2013)
6. Celis, M.R., Dennis, J.E., Tapia, R.A.: A trust region algorithm for nonlinear equality constrained optimization. In: Boggs, R.T., Byrd, R.H., Schnabel, R.B. (eds.) Numerical Optimization, pp. 71–82. SIAM, Philadelphia (1985)
7. Cox, D.A., Little, J., O'Shea, D.: Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra. Undergraduate Texts in Mathematics. Springer, Berlin (2015)
8. Emeliyanenko, P., Sagraloff, M.: On the complexity of solving a bivariate polynomial system. In: International Symposium on Symbolic and Algebraic Computation (ISSAC), ACM (2012)
9. Emiris, I.Z., Mourrain, B., Tsigaridas, E.P.: The DMM bound: multivariate (aggregate) separation bounds. In: Proceedings of Symbolic and Algebraic Computation, International Symposium, ISSAC 2010, Munich, Germany (2010)
10. Kobel, A., Sagraloff, M.: On the complexity of computing with planar algebraic curves. J. Complex. **31**(2), 206–236 (2015)
11. Locatelli, M.: Some results for quadratic problems with one or two quadratic constraints. Oper. Res. Lett. **43**(2), 126–131 (2015)
12. Renegar, J.: Incorporating condition measures into the complexity theory of linear programming. SIAM J. Optim. **5**(3), 506–524 (1995)
13. Stern, R.J., Wolkowicz, H.: Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations. SIAM J. Optim. **5**, 286–313 (1995)
14. Sturmfels, B.: Solving systems of polynomial equations. In: CBMS Regional Conference Series in Mathematics. American Mathematical Society (2002)
15. Terrell, W.J.: Stability and Stabilization: An Introduction. Princeton University Press, Princeton (2009)
16. Thrall, R.M., Tornheim, L.: Vector Spaces and Matrices, Dover Books on Mathematics. Dover, New York (2014)
17. Ye, Y.: Approximating quadratic programming with bound and quadratic constraints. Math. Program. **84**, 219–226 (1999)

18. Ye, Y.: Approximating global quadratic optimization with convex quadratic constraints. J. Glob. Optim. **15**(1), 1–17 (1999)
19. Ye, Y., Zhang, S.: New results on quadratic minimization. SIAM J. Optim. **14**, 245–267 (2003)