Journal of Global Optimization (2005) 32:181–205
DOI 10.1007/s10898-004-2701-z

Quartic Formulation of Standard Quadratic Optimization Problems

IMMANUEL M. BOMZE[1] and LAURA PALAGI[2]
[1]*Deptartment of Statistics and Decision Support Systems, University of Vienna, Austria (e-mail: immanuel.bomze@univie.ac.at)*
[2]*Dipartimento di Informatica e Sistemistica "A. Ruberti", Università di Roma "La Sapienza", Italy*

(Received 1 April 2004; accepted 7 April 2004)

Abstract. A standard quadratic optimization problem (StQP) consists of finding the largest or smallest value of a (possibly indefinite) quadratic form over the standard simplex which is the intersection of a hyperplane with the positive orthant. This NP-hard problem has several immediate real-world applications like the Maximum-Clique Problem, and it also occurs in a natural way as a subproblem in quadratic programming with linear constraints. To get rid of the (sign) constraints, we propose a quartic reformulation of StQPs, which is a special case (degree four) of a homogeneous program over the unit sphere. It turns out that while KKT points are not exactly corresponding to each other, there is a one-to-one correspondence between feasible points of the StQP satisfying second-order necessary optimality conditions, to the counterparts in the quartic homogeneous formulation. We supplement this study by showing how exact penalty approaches can be used for finding local solutions satisfying second-order necessary optimality conditions to the quartic problem: we show that the level sets of the penalty function are bounded for a finite value of the penalty parameter which can be fixed in advance, thus establishing exact equivalence of the constrained quartic problem with the unconstrained penalized version.

Key words: exact penalization, max clique, merit function, second-order optimality conditions, standard quadratic optimization

1. Introduction and Preliminaries

1.1. STANDARD QUADRATIC OPTIMIZATION PROBLEMS (StQPs)

A standard quadratic optimization problem (StQP) consists of finding the largest or smallest value of a (possibly indefinite) quadratic form over the standard simplex which is the intersection of a hyperplane with the positive orthant. This NP-hard problem has several immediate realworld applications like the Maximum-Clique Problem, and it also occurs in a natural way as a subproblem in quadratic programming with linear constraints. For more details, we refer to [5] and [6].

We consider the standard quadratic optimization problem of the form

$$\min\{\varphi(y) = \frac{1}{2} y^\top A y : y \in \Delta\} \tag{1}$$

where Δ denotes the standard simplex in n-dimensional Euclidean space $\mathbb{R}^n$, namely

$$\Delta = \{y \in \mathbb{R}^n : e^\top y = 1, y \geqslant 0\},$$

and $A = [a_{ij}] \in \mathbb{R}^{n \times n}$ is a symmetric $n \times n$ matrix; e is the n-vector of all ones and $y^\top$ denotes the transposed vector while I denotes the $n \times n$ identity matrix.

Since the constraints are linear, the constraint qualifications are met and the first-order necessary optimality conditions (KTS) for a feasible point $\bar{y}$ to be a local solution of problem (1) require that a scalar $\bar{\lambda}$ exists such that

$$\begin{cases} (A\bar{y})_i + \bar{\lambda} = 0 & \text{for } i : \bar{y}_i > 0, \\ (A\bar{y})_i + \bar{\lambda} \geqslant 0 & \text{for } i : \bar{y}_i = 0. \end{cases} \tag{2}$$

From (2) we get also that $\bar{\lambda} = -\bar{y}^\top A \bar{y} = -2\varphi(\bar{y})$. Hence the Lagrange multiplier is uniquely determined by $\bar{y}$.

In the sequel, we will invoke (weak) second-order necessary optimality conditions (WSC) (cf. e.g., [15], p. 61). They require in addition to (2) that

$$z^\top A z \geqslant 0 \text{ for all } z \in \mathcal{Z}(\bar{y}) = \left\{ z \in \mathbb{R}^n : \sum_{i \in \mathcal{I}(\bar{y})} z_i = 0, \text{ and } z_i = 0 \text{ for all } i \notin \mathcal{I}(\bar{y}) \right\} \tag{3}$$

where $\mathcal{I}(\bar{y})$ denotes the "inactive" variables, namely

$$\mathcal{I}(\bar{y}) = \{i : \bar{y}_i > 0\}.$$

2. Quartic Formulation of StQPs

2.1. OPTIMALITY CONDITIONS

To get rid of the sign constraints $y_i \geqslant 0$, we replace the variables y_i with x_i, putting $y_i = x_i^2$. Then the condition $e^\top y = 1$ reads $\|x\|^2 = 1$, where $\|\cdot\|$ denotes the Euclidean norm. So we arrive at the ball-constrained quartic optimization problem (BQP)

$$\min\left\{ f(x) = \frac{1}{2} x^\top X A X x : \|x\|^2 = 1 \right\} \tag{4}$$

where we denote by X the diagonal matrix with elements x_i. We note that

$$\nabla f(x) = 2XAXx \quad \text{and} \quad \nabla^2 f(x) = 4XAX + 2\,\mathrm{diag}\{AXx\}. \tag{5}$$

Since there is only the norm constraint, any kind of constraint qualification is satisfied. The first order necessary optimality conditions (KTQ) for a feasible point $\bar{x}$ to be a local solution of problem (4) require that a scalar $\bar{\mu}$ exists such that $\nabla f(\bar{x}) + 2\bar{\mu}\bar{x} = 0$, that is:

$$\bar{X}A\bar{X}\bar{x} + \bar{\mu}\bar{x} = 0. \tag{6}$$

Furthermore, taking into account that by (5) we can write $\nabla^2 f(\bar{x})\bar{x} = 6\bar{X}A\bar{X}\bar{x}$ we can re-write condition (6) as

$$(\nabla^2 f(\bar{x}) + 6\bar{\mu}I)\bar{x} = 0. \tag{7}$$

The second-order necessary optimality conditions (SNC) for problem (4) involve the Hessian of the Lagrangian,

$$H_{\bar{\mu}}(\bar{x}) = \nabla^2 f(\bar{x}) + 2\bar{\mu}I = 2[2\bar{X}AX + \mathrm{diag}\{A\bar{X}\bar{x}\} + \bar{\mu}I] \tag{8}$$

and require in addition to (7) that

$$z^\top H_{\bar{\mu}}(\bar{x})z \geqslant 0 \quad \text{for all } z \perp \bar{x}, \text{ i.e., } z^\top \bar{x} = 0. \tag{9}$$

Problem (4) is a homogeneous problem to minimize a fourth-order polynomial over the unit sphere. To invoke some generalized trust-region method, we may extend this problem to the unit ball rather than to the sphere, replacing the constraint $\|x\|^2 = 1$ with the inequality $\|x\|^2 \leqslant 1$. Now, if the objective is non-negative for all x, then of course $x = 0$ is the global solution over the ball. But in the opposite case we always can be sure that there is a solution for the trust region problem which lies on the sphere, and hence is also a solution to the problem over the sphere. Moreover, every *local* solution $\bar{x}$ to the trust region problem with $\|\bar{x}\| < 1$ necessarily satisfies $f(\bar{x}) = 0$, which immediately follows by considering the ray through $\bar{x}$ emanating at the origin. By complementary slackness, then, the Lagrange multiplier $\bar{\mu}$ must be also zero, of course.

The next subsection is devoted to a discussion of the difference between homogeneous optimization over the ball and the sphere, in particular for second-order necessary conditions involving $H_{\bar{\mu}}(\bar{x})$.

2.2. HOMOGENEOUS OPTIMIZATION OVER THE BALL AND THE SPHERE

Consider a general objective function $f(x)$ which is homogeneous of degree k and the problem $\min\{f(x) : \|x\|^2 = 1\}$ (later, we shall specialize to our case $k = 4$). By Euler's identity, we have

$$\nabla f(x)^\top x = kf(x) \quad \text{and} \quad \nabla^2 f(x)x = (k-1)\nabla f(x). \tag{10}$$

Now, the KKT condition $\nabla f(\bar{x}) + 2\bar{\mu}\bar{x} = 0$ implies that the Lagrange multiplier $\bar{\mu}$ is uniquely determined via (10), namely using $kf(\bar{x}) = \bar{x}^\top \nabla f(\bar{x}) = -2\bar{\mu}\|\bar{x}\|^2 = -2\bar{\mu}$. This also holds for the ball-constrained problem, since $\bar{\mu} = f(\bar{x}) = 0$ if $\|\bar{x}\| < 1$. Furthermore, using again (10), $\nabla f(\bar{x}) + 2\bar{\mu}\bar{x} = 0$ can be rewritten as

$$[H_{\bar{\mu}}(\bar{x}) + 2(k-2)\bar{\mu}I]\bar{x} = [\nabla^2 f(\bar{x}) + 2(k-1)\bar{\mu}I]\bar{x} = (k-1)[\nabla f(\bar{x}) + 2\bar{\mu}\bar{x}] = 0. \tag{11}$$

Hence, unless $\bar{x} = 0$, the matrix $H_{\bar{\mu}}(\bar{x}) + 2(k-2)\bar{\mu}I$ is singular. In [1], a second-order condition has been proven for the inequality constrained version of the homogeneous problem which establishes positive-semidefiniteness of this matrix. For reasons which will become obvious soon, we include a proof, writing here and in the sequel $A \succeq 0$ to signify that A is positive-semidefinite.

THEOREM 1. *Let $\bar{x}$ be a local minimizer for problem* $\min\{f(x) : \|x\|^2 \leqslant 1\}$, *where f is homogeneous of degree k.* Then necessarily $f(\bar{x}) \leqslant 0$, *and for* $\bar{\mu} = -\frac{k}{2} f(\bar{x}) \geqslant 0$ *we have*

$$H_{\bar{\mu}}(\bar{x}) + 2(k-2)\bar{\mu}I = \nabla^2 f(\bar{x}) + 2(k-1)\bar{\mu}I \succeq 0. \tag{12}$$

Proof. The assertion is obviously true if $\|\bar{x}\| < 1$, taking into account the remarks concluding the previous subsection. So without loss of generality we may (and do) assume that $\|\bar{x}\| = 1$, and of course $\bar{\mu} \geqslant 0$. Hence from the previous arguments $f(\bar{x}) = -\frac{2}{k}\bar{\mu} \leqslant 0$. To establish (12), note that (11) can also be written as

$$H_{\bar{\mu}}(\bar{x})\bar{x} = (k-1)\nabla f(\bar{x}) + 2\bar{\mu}\bar{x} = -2\bar{\mu}(k-2)\bar{x}, \tag{13}$$

where in the last equality we use the KKT condition $\nabla f(\bar{x}) = -2\bar{\mu}\bar{x}$. Next, let $w \in \mathbb{R}^n$ be arbitrary and consider a point $z = w - \alpha\bar{x}$ with α such that $\bar{x}^\top z = 0$, namely $\alpha = w^\top \bar{x}$. Now, if $\|\bar{x}\| = 1$, then the standard (SNC) for the ball-constrained problem coincides with (9). Expanding and collecting terms on the right-hand side we obtain, using (13),

$$\begin{aligned} 0 \leqslant z^\top H_{\bar{\mu}}(\bar{x})z &= (w - \alpha\bar{x})^\top H_{\bar{\mu}}(\bar{x})(w - \alpha\bar{x}) \\ &= w^\top H_{\bar{\mu}}(\bar{x})w - 2\alpha w^\top H_{\bar{\mu}}(\bar{x})\bar{x} + \alpha^2 \bar{x}^\top H_{\bar{\mu}}(\bar{x})\bar{x} \\ &= w^\top H_{\bar{\mu}}(\bar{x})w - 2\alpha w^\top(-2\bar{\mu}(k-2)\bar{x}) - \alpha^2 2\bar{\mu}(k-2)\|\bar{x}\|^2 \\ &= w^\top H_{\bar{\mu}}(\bar{x})w + 4\bar{\mu}(k-2)\alpha^2 - 2\bar{\mu}(k-2)\alpha^2 \\ &= w^\top H_{\bar{\mu}}(\bar{x})w + 2\bar{\mu}(k-2)\alpha^2. \end{aligned} \tag{14}$$

Hence we get, using $\alpha^2 \leqslant \|w\|^2$ and $\bar{\mu} \geqslant 0$,

$$0 \leqslant w^\top H_{\bar\mu}(\bar x)w + 2\bar\mu(k-2)\|w\|^2 = w^\top[\nabla^2 f(\bar x) + 2(k-1)\bar\mu I]w \quad \text{for all } w \in \mathbb{R}^n,$$

which shows the assertion. □

Note that in the last implication the fact that $\bar\mu \geqslant 0$ is essential. Hence the same arguments cannot be repeated in the equality constrained case unless some additional assumptions on $f(x)$ are made.

For the special case of quartic problem (4) we obtain

COROLLARY 2. *Let $\bar x$ be a local minimizer for problem* $\min\{f(x) : \|x\|^2 \leqslant 1\}$, *where f is as in (4). Then for $\bar\mu = -2f(\bar x) \geqslant 0$ we have $(2\bar X A\bar X + diag\{A\bar X\bar x\} + 3\bar\mu I)\bar x = 0$ and $\bar\mu(\|x\|^2 - 1) = 0$ as well as*

$$2\bar X A\bar X + diag\{A\bar X\bar x\} + 3\bar\mu I \succeq 0. \tag{15}$$

Proof. Follows from Theorem 1 by noting $k = 4$ and recalling (5) holds in the quartic case. □

However, by careful inspection of the proof of Theorem 1, we can sharpen the second-order necessary conditions by [1], now using the positive-semidefiniteness of a rank-one modification of $H_{\bar\mu}(\bar x)$ (rather than perturbing it by the identity matrix), arriving at the following result that holds for general homogeneous functions of degree k. Note that the Lagrange multiplier $\bar\mu$ can now be negative.

THEOREM 3. *Let $\bar x$ be a local minimizer of a homogeneous objective f (of degree k) over the unit sphere. Then for $\bar\mu = -\frac{k}{2}f(\bar x) \in \mathbb{R}$ we have $\nabla f(\bar x) + 2\bar\mu\bar x = 0$ and*

$$H_{\bar\mu}(\bar x) + 2(k-2)\bar\mu\bar x\bar x^\top = \nabla^2 f(\bar x) + 2\bar\mu I + 2\bar\mu(k-2)\bar x\bar x^\top \succeq 0. \tag{16}$$

Proof. By the same arguments that lead to the proof of Theorem 1, we arrive at (14). But this inequality can be rewritten as

$$0 \leqslant w^\top[H_{\bar\mu}(\bar x) + 2(k-2)\bar\mu\bar x\bar x^\top]w,$$

recalling the definition of $\alpha = w^\top\bar x$. Thus the assertion. □

In the case of the quartic homogeneous problem we finally arrive at:

COROLLARY 4. *Let $\bar x$ be a local minimizer for problem* (4). *Then for $\bar\mu = -2f(\bar x) \in \mathbb{R}$ we have $\bar X A\bar X\bar x + \bar\mu\bar x = 0$ and*

$$2\bar X A\bar X + diag\{A\bar X\bar x\} + \bar\mu(I + 2\bar x\bar x^\top) \succeq 0. \tag{17}$$

Returning shortly to the general case, it is straightforward to see that (16) implies (12), if $\bar{\mu} \geqslant 0$. Indeed, since $\|\bar{x}\| \leqslant 1$, we have $I - \bar{x}\bar{x}^\top \succeq 0$. Since both are necessary conditions, Theorem 3 is a sharpening of Theorem 1. The next section contains an example where (12) is fulfilled while (16) is violated, although $\bar{\mu} > 0$.

3. BQPs versus StQPs: Relationship among Solutions

3.1. TRANSFORMING OPTIMALITY CONDITIONS

Let us consider the transformation $y = T(x)$ with $y_i = x_i^2$. First we observe that for any vector x it results $f(x) = f(|x|)$ where by $|x|$ we denote the vector whose components are $|x_i|$. Also, both $\bar{x}$ and $|\bar{x}|$ satisfy the same first and second-order optimality conditions. Hence without loss of generality we can assume in the following that $x \geqslant 0$. We denote by $x = T^{-1}(y)$ the (partial) inverse transformation, namely $x_i = +\sqrt{|y_i|}$.

THEOREM 5. *A point $\bar{y}$ is a local minimizer of problem (1) if and only if $\bar{x} = T^{-1}(\bar{y})$ is a local minimizer of problem (4). Further, a point $\bar{y}$ is a global minimizer of problem (1) if and only if $\bar{x} = T^{-1}(\bar{y})$ is a global minimizer of problem (4).*

Proof. The transformation $y = T(x)$ and its (partial) inverse $x = T^{-1}(y)$ are well-defined and continuous. Moreover we have $f(x) = \varphi(T(x))$ as well as $f(T^{-1}(y)) = \varphi(y)$ and feasible points of problem (1) correspond to feasible points of problem (4). Hence the result. □

Unfortunately, the same does not hold true for KKT points. Only one direction of the implications is still valid:

THEOREM 6. *Let $\bar{y}$ be a KKT point of problem (1), then $\bar{x} = T(\bar{y})$ is a KKT point of problem (4).*

Proof. The proof follows easily by observing that we can re-write equation (6) coordinate-wise as

$$\bar{x}_i[(A\bar{y})_i + \bar{\mu}] = 0,$$

which is implied by (2). □

The converse is not true as shown in Example 1 of the following subsection. The loss of correspondence between KKT points implies that

spurious KKT points can be created in passing from problem (1) to problem (4). However, the reverse correspondence can be proved for refined KKT points of problem (4), namely those points that satisfy also the second-order necessary conditions proposed in the previous section:

THEOREM 7. *Let $\bar{y} = T(\bar{x}) \in \Delta$ with $\bar{x} \geqslant 0$. Then the following statements are equivalent:*

(a) $\bar{y}$ is a KKT point for problem (1) which satisfies the second-order necessary condition (3);

(b) $\bar{y}$ is a KKT point for problem (1), and the second-order necessary conditions (17) are satisfied for $\bar{x}$ and the problem (4);

(c) $\bar{x}$ is a KKT point of problem (4) which satisfies the second-order necessary conditions (17);

(d) $\bar{x}$ is a KKT point of problem (4), and the second-order necessary conditions (3) are satisfied for $\bar{y}$ and the problem (1).

Proof. First we prove that (a) implies (b). To this end, we invoke Theorem 6, whence $\bar{x}$ satisfies the KKT conditions (6) (KTQ) for problem (4). For any $d \in \mathbb{R}^n$ we can define the vector z with components

$$z_i = \begin{cases} 0 & \text{if } \bar{x}_i = 0, \\ \bar{x}_i d_i - (\bar{x}^\top d)\bar{x}_i & \text{if } \bar{x}_i > 0. \end{cases}$$

Hence, taking into account that $\bar{x}$ is feasible, we have

$$\sum_{i \in \mathcal{I}(\bar{y})} z_i = \sum_{i:\, \bar{x}_i > 0} \bar{x}_i d_i - (\bar{x}^\top d) \sum_{i:\, \bar{x}_i > 0} \bar{x}_i = 0,$$

where, as usual, $\mathcal{I}(\bar{y}) = \{i : \bar{y}_i > 0\}$.

Hence from (3) (WSC) we infer

$$0 \leqslant z^\top A z = (d - \alpha\bar{x})^\top \bar{X} A \bar{X} (d - \alpha\bar{x})$$

with $\alpha = \bar{x}^\top d$. Expanding terms we get

$$0 \leqslant d^\top \bar{X} A \bar{X} d - 2\alpha \bar{x}^\top \bar{X} A \bar{X} d + \alpha^2 \bar{x}^\top \bar{X} A \bar{X} \bar{x}.$$

Recalling that $\bar{X} A \bar{X} \bar{x} = -\bar{\mu}\bar{x}$ we can write

$$0 \leqslant d^\top \bar{X} A \bar{X} d - \alpha^2 \bar{\mu} \|x\|^2 + 2\alpha\bar{\mu}(\bar{x}^\top d) = d^\top (\bar{X} A \bar{X} + \mu \bar{x}\bar{x}^\top) d.$$

Recalling also that $A\bar{X}\bar{x} + \bar{\mu}e \geqslant 0$ which means $\operatorname{diag}\{A\bar{X}\bar{x}\} + \bar{\mu} I \succeq 0$ we get the desired result. (b) $\Rightarrow$ (c) is an immediate consequence of Theorem 6.

To establish (c) $\Rightarrow$ (d), we consider a point $z \perp \bar{x}$ and put $d = \bar{X} z$. Then $d \in \mathcal{Z}(\bar{y})$ since $i \notin \mathcal{I}(\bar{y})$ implies $\bar{x}_i = 0$ and thus $d_i = 0$ and, furthermore,

$$\sum_{i \in \mathcal{I}(\bar{y})} d_i = \sum_{i \in \mathcal{I}(\bar{y})} z_i \bar{x}_i = z^\top \bar{x} = 0.$$

Hence we obtain from (3)

$$0 \leqslant d^\top A d = z^\top \bar{X} A \bar{X} z,$$

but of course also, as above, $\operatorname{diag}\{A\bar{X}\bar{x}\} + \bar{\mu} I \succeq 0$, which entails $z^\top H_{\bar{\mu}}(\bar{x}) z \geqslant 0$, as required. Finally, to prove (d) $\Rightarrow$ (a), we have to show next that under the assumption (d), the point $\bar{y} = T(\bar{x}) = \bar{X}\bar{x} \in \Delta$ is a KKT point of problem (1). Let us write the first condition (6) coordinate-wise:

$$\bar{x}_i((A\bar{y})_i + \bar{\mu}) = 0.$$

Now $\bar{x}_i > 0$ if and only if $\bar{y}_i > 0$, in which case we infer $(A\bar{y})_i + \bar{\mu} = 0$. Else, i.e., if $\bar{x}_i = 0$, we use the second condition of (17). Then choose $z = e_i \perp \bar{x}$, to get coordinate-wise

$$0 \leqslant e_i^\top [2\bar{X}A\bar{X}] e_i + \operatorname{diag}\{A\bar{y}\}_{ii} + \bar{\mu} = 0 + [A\bar{y}]_i + \bar{\mu},$$

and (2) (KTS) is satisfied for $\bar{\lambda} = \bar{\mu}$.

Now let us prove that the point $\bar{y}$ satisfies also the second-order condition (3). Let d be any point in $\mathcal{Z}(\bar{y})$. Let us define the vector z with components:

$$z_i = \begin{cases} 0 & \text{if } \bar{x}_i = 0, \\ \frac{d_i}{\bar{x}_i} & \text{if } \bar{x}_i > 0. \end{cases}$$

Then we get $\bar{X}z = d$ and furthermore, since $d_i = 0$ for all $i \notin \mathcal{I}(\bar{y})$ by definition,

$$z^\top \bar{x}^\top = \sum_{i:\ \bar{y}_i > 0} z_i \bar{x}_i = \sum_{i:\ \bar{x}_i > 0} \bar{x}_i \frac{d_i}{\bar{x}_i} = \sum_{i \in \mathcal{I}(\bar{y})} d_i = 0.$$

But then we also know, by (17),

$$z^\top [2\bar{X}A\bar{X} + \operatorname{diag}\{A\bar{X}\bar{x}\} + \bar{\mu}(I + 2\bar{X}\bar{x}^\top)] z \geqslant 0.$$

Since we have $\{i :\ \bar{x}_i = 0\} = \{i :\ \bar{y}_i = 0\} \subseteq \{i :\ d_i = 0\}$, we obtain from

$$z^\top \operatorname{diag}\{A\bar{X}\bar{x}\} z = \sum_{i \in \mathcal{I}(\bar{y})} z_i^2 [A\bar{X}\bar{x}]_i = -\bar{\mu}\|z\|^2$$

and thus we can write

$$0 \leqslant z^\top [2\bar{X}A\bar{X} + \operatorname{diag}\{A\bar{X}\bar{x}\} + \bar{\mu}(I + 2\bar{X}\bar{x}^\top)] z = 2 d^\top A d.$$

Hence (3) holds. □

As already observed at the end of the last section, condition (16) implies (12), if $\bar{\mu} \geqslant 0$. The converse, however, does not hold, even if in addition (6) is satisfied. See, again, Example 1 below.

3.2. COUNTEREXAMPLES

EXAMPLE 1. Let us consider the following problem

$$A = \begin{bmatrix} -1 & -1 & -2 \\ -1 & -1 & -2 \\ -2 & -2 & -2 \end{bmatrix}$$

and the point $\bar{y} = \left[\frac{1}{2}, \frac{1}{2}, 0\right]^\top$, which is not a KKT point for problem (1). Indeed $A\bar{y} = \left[-\frac{1}{2}, -\frac{1}{2}, -1\right]^\top$ which violates (2) as $\bar{\lambda} = \frac{1}{2} < 1$. On the other hand, consider the transformed point $\bar{x} = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0\right]^\top$ and the transformed problem (4). Thus

$$\bar{X}A\bar{X} = -\frac{1}{2}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Now the KKT conditions require $\bar{X}A\bar{X}\bar{x} + \bar{\mu}\bar{x} = 0$ for some $\bar{\mu}$, i.e., $-1/\sqrt{2} + \bar{\mu}/\sqrt{2} = 0$ and this holds for $\bar{\mu} = 1$.

We already saw that (6) is satisfied although (2) is violated. In other words, $\bar{x}$ is a KKT point for problem (4) while $\bar{y}$ is none for problem (1). Furthermore,

$$2\bar{X}A\bar{X} + \operatorname{diag}\{A\bar{y}\} + 3\bar{\mu}I = \frac{1}{2}\begin{bmatrix} 3 & -2 & 0 \\ -2 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix} \succeq 0,$$

while, of course, the sharper second-order condition (16) is violated (otherwise, Theorem 7 would yield (2), which as shown is absurd).

The preceding example already exhibits the weakness of (12) for our purposes. Still, one could hope that the stronger condition (2), together with (12), implies (16) or equivalent conditions from Theorem 7. The following example shows that this hope, too, is in vain:

EXAMPLE 2. For any $n \geqslant 2$, consider $A = -I$ and $\bar{y} = \frac{1}{n}e$ with $\bar{x} = \frac{1}{\sqrt{n}}e$. Then both KKT conditions (2) and (6) are satisfied as $\mathcal{I}(\bar{y}) = \{1, \ldots, n\}$ is maximal. Moreover, $2\bar{X}A\bar{X} + \operatorname{diag}\{A\bar{y}\} + 3\bar{\mu}I = 0 \succeq 0$ and hence (12) holds. On the other hand, $\mathcal{Z}(\bar{y}) = e^\perp \neq \{0\}$ so that (3) cannot hold.

We proceed with a third example which shows that, despite the weakness of (12), there are still KKT points which violate even this condition:

EXAMPLE 3. Let

$$A = \begin{bmatrix} -2 & 3 \\ 3 & -8 \end{bmatrix} \text{ and } \bar{y} = \begin{bmatrix} 11/16 \\ 5/16 \end{bmatrix} \text{ with } \bar{x} = \begin{bmatrix} \sqrt{11}/4 \\ \sqrt{5}/4 \end{bmatrix}.$$

As $A\bar{y} = -\frac{7}{16}[1, 1]^\top$, the KKT condition (2) holds as in Example 2, with $\bar{\mu} = \frac{7}{16} > 0$. However, here $2\bar{X}A\bar{X} + \operatorname{diag}\{A\bar{y}\} + 3\bar{\mu}I = 2[\bar{X}A\bar{X} + 2\bar{\mu}I]$ is indefinite as

$$\bar{X}A\bar{X} + 2\bar{\mu}I = \frac{1}{16}\begin{bmatrix} -8 & 3\sqrt{55} \\ 3\sqrt{55} & -26 \end{bmatrix}.$$

Let us conclude this subsection by a last example showing that the second-order condition (3) is indeed only necessary but not sufficient for local optimality. This may be surprising at first sight as the objective in (1) is quadratic. The reason for this gap is that copositivity has to replace definiteness conditions in order to make second-order conditions tight for quadratic problems. For details, we refer to [3].

EXAMPLE 4. Let

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & -1 \\ 1 & -1 & 0 \end{bmatrix} \text{ and } \bar{y} = \begin{bmatrix} 1/2 \\ 1/2 \\ 0 \end{bmatrix}.$$

Then $\mathcal{Z}(\bar{y}) = \{[v, -v, 0]^\top : v \in \mathbb{R}\}$, and hence condition (3) holds: indeed, $z^\top Az = 0$ for all $z \in \mathcal{Z}(\bar{y})$. On the other hand, we have $y^\top Ay = 2y_3(y_1 - y_2)$ for all $y \in \Delta$, so that $\bar{y}$ with its objective value of zero cannot be a local minimizer as the objective function can change sign arbitrarily close to $\bar{y}$.

4. Finding a Local Solution of the BQP

Theorem 7 states the correspondence among points satisfying the second order necessary conditions of problems (3) and (4). Hence, if we want to use the quartic formulation to obtain a solution of the StQP, we need an algorithm that converges to second-order KKT points of problem (4).

In the constrained optimization field, very few algorithms have been proposed that achieve convergence to points satisfying second order necessary conditions (see e.g. [9, 12] and references therein). However, up to the authors' knowledge, there are no available implementations of such algorithms. Furthermore, the special structure of the constraint of problem (4)

allows us to define "ad-hoc" algorithms that may be simplified with respect to those for more general problems.

In this section we propose an approach based on the use of a continuously differentiable exact penalty function, which, following the same lines of [18, 19], has the distinguishing feature of exploiting the particular structure of the objective function and of the constraint. By means of the penalty function P, the problem of locating a constrained global minimizer of problem (4) is recast as the problem of locating an unconstrained global minimizer of P. This allows us to use an unconstrained method for the minimization of the penalty function P, which makes life simpler. Indeed, in the unconstrained optimization field there are different approaches that allows to find points satisfying the second order necessary conditions for unconstrained minimization (see e.g. [20, 14]) and software is available, too.

This exact penalty approach can be used for finding local solutions to the quartic problem, because it is possible to show that the level sets of P are bounded for a finite value of the penalty parameter which can be fixed in advance. Thus exact equivalence of the constrained quartic problem with the unconstrained penalized version can be established in the sense that first- and second order optimality conditions can be related to each other in an exact way.

One may wonder why we go through the quartic formulation (4) and not apply the penalty transformation directly to the StQP (3). Actually, although it is possible to define a continuously differentiable exact penalty function for the StQP itself [10, 13], and although the feasible set has a special structure, their expression involves the presence of barrier terms to avoid unboundedness of the level sets so that theoretical study and also practical implementation of minimization algorithms are more complicated.

In the next subsection we describe the structure of the penalty function and then we discuss more in details its algorithmic use.

4.1. A CONTINUOUSLY DIFFERENTIABLE PENALTY FUNCTION

In this subsection, we show that problem (4) is equivalent to the unconstrained minimization of a twice continuously differentiable merit function. The transformation of a constrained optimization problem into an unconstrained one by means of a continuously differentiable exact penalty function has been addressed in many papers (see for example [2, 10] and references therein). Almost all the expressions of continuously differentiable penalty function are derived from the original augmented Lagrangean function proposed by Hestenes–Powell–Rockafellar [24, 16] by substituting

the multiplier vector by a suitable multiplier function. Indeed the properties and the expression of the penalty function depend on the particular choice of the multiplier function. In the literature different multiplier functions have been proposed (see e.g. [2, 10] and references therein) that, however, require assumptions on the constraints (such as linear independence). In particular, applying the standard definition we should have $\mu(x) = -2\frac{f(x)}{\|x\|}$ which is not defined in the origin.

However, following the guidelines of [18, 19] we define here a multiplier function $\mu(x): \mathbb{R}^n \to \mathbb{R}$, which yields an estimate of the multiplier associated to problem (4) as a function of the variables x, that fully exploits the particular structure of the constraint, namely

$$\mu(x) = -x^\top XAXx = -2f(x). \tag{18}$$

We note that the expression (18) equals the standard definition of $\mu(x)$ only on the feasible set. It is easily seen that $\mu(x)$ given by (18) is twice continuously differentiable on $\mathbb{R}^n$ with derivatives

$$\nabla\mu(x) = -2\nabla f(x) = -4XAXx \quad \text{and}$$
$$\nabla^2\mu(x) = -2\nabla^2 f(x) = -4(2XAX + \operatorname{diag}\{AXx\}).$$

The main property of the multiplier function is summarized in the following proposition.

PROPOSITION 8. *If $(\bar{x}, \bar{\mu})$ is a KKT point for problem (4) then we have $\mu(\bar{x}) = \bar{\mu}$.*

Now following a standard approach [19] for the definition of a penalty function, we obtain the following expression for $P(x;\varepsilon)$:

$$P(x;\varepsilon) = f(x) + \frac{1}{\varepsilon}(\|x\|^2 - 1)^2 + \mu(x)(\|x\|^2 - 1),$$

which, using the expression (18), becomes

$$P(x;\varepsilon) = f(x)(3 - 2\|x\|^2) + \frac{1}{\varepsilon}(\|x\|^2 - 1)^2. \tag{19}$$

Usually, exactness results for the penalty functions are stated for sufficiently small values of the penalty parameter, that has to be adjusted iteratively during the minimization process ([13] and references therein). However, due to the special structure of the objective function and constraint, it is possible to define a priori the threshold value $\bar{\varepsilon}$ (depending on the problem data A) for the penalty parameter ε in the penalty function $P(x;\varepsilon)$:

$$\bar{\varepsilon} = \min\left\{\frac{1}{2}\|A\|\frac{(1+3\Theta^4)}{(\Theta^2-1)^2}, \frac{1}{2\|A\|}\frac{(\gamma^2-1)^2}{\gamma^4}, \frac{2\Theta^2}{3\|A\|\gamma^4}\right\}, \tag{20}$$

where $\Theta \in (0,1)$ and $\gamma > 1$ are user-selected constants (see Theorem 9 below).

It easily seen that $P(x;\varepsilon)$ is twice continuously differentiable on the *whole space* $\mathbb{R}^n$, with the gradient and Hessian given by

$$\nabla P(x;\varepsilon) = \nabla f(x)(3 - 2\|x\|^2) - 4f(x)x + \frac{4}{\varepsilon}(\|x\|^2 - 1)x \tag{21}$$

$$\nabla^2 P(x;\varepsilon) = \nabla^2 f(x)(3 - 2\|x\|^2) - 4f(x)I - 4[\nabla f(x)x^\top + x\nabla f(x)^\top] + \frac{8}{\varepsilon}xx^\top + \frac{4}{\varepsilon}(\|x\|^2 - 1)I. \tag{22}$$

Moreover, for every feasible x we have

$$P(x;\varepsilon) = f(x). \tag{23}$$

Furthermore, the penalty function $P(x;\varepsilon)$ differs from other continuously differentiable exact penalty functions also in the important fact that it admits *compact level sets,* without the need of barrier terms. This implies the existence of a global minimizer and also the satisfaction of a minimal assumption in unconstrained optimization which implies boundedness of the sequence generated by an unconstrained method.

In particular, given any arbitrary but fixed point $x^0 \in \mathbb{R}^n$, let us define the level set

$$\mathcal{L}_0 = \{x \in \mathbb{R}^n : P(x;\varepsilon) \leqslant P(x^0;\varepsilon)\}.$$

Then we have the following result.

THEOREM 9. *Let* $x^0 \in \mathbb{R}^n$ *be a point such that* $\|x^0\| = 1$. *If* $\Theta \in (0,1)$ *and* $\gamma > 1$ *are the constants appearing in (20), then for* $0 < \varepsilon < \bar{\varepsilon}$

$$\mathcal{L}_0 \subseteq \{x \in \mathbb{R}^n : \Theta \leqslant \|x\| \leqslant \gamma\}.$$

Proof. Denote by $\|A\| = \sup_{\|y\| \leqslant 1} \|Ay\|$ the usual operator norm. Then it is straightforward to see that for any $x \in \mathbb{R}^n$ we have $\|X\| \leqslant \|x\|$, and hence $\|f(x)\| \leqslant \frac{1}{2}\|A\|\|x\|^4$ holds. By consequence, we make use of the inequalities:

$$P(x^0;\varepsilon) = f(x^0) = \frac{1}{2}x^{0\top}X^0AX^0x^0 \leqslant \frac{1}{2}\|A\|\,\|x^0\|^4 = \frac{1}{2}\|A\| \tag{24}$$

and, because of $P(x;\varepsilon) \geqslant -\frac{1}{2}\|A\|\,\|x\|^4(3 - 2\|x\|^2) + \frac{1}{\varepsilon}(\|x\|^2 - 1)^2 = -\frac{3}{2}\|A\|\,\|x\|^4 + \|A\|\,\|x\|^6 + \frac{1}{\varepsilon}(\|x\|^2 - 1)^2$,

$$P(x;\varepsilon) \geqslant -\frac{3}{2}\|A\|\,\|x\|^4 + \frac{1}{\varepsilon}(1 - \|x\|^2)^2. \tag{25}$$

First we prove that $\|x\| < \Theta$ implies $P(x;\varepsilon) > P(x^0;\varepsilon)$. Indeed, since $\|x\| < \Theta < 1$, we can deduce from (25)

$$P(x;\varepsilon) \geqslant -\frac{3}{2}\|A\|\Theta^4 + \frac{1}{\varepsilon}(\Theta^2-1)^2. \tag{26}$$

On the other hand, $\varepsilon \leqslant \bar{\varepsilon} < \frac{1}{2}\|A\|\frac{(1+3\Theta^4)}{(\Theta^2-1)^2}$ implies $\frac{1}{\varepsilon}(\Theta^2-1)^2 > \frac{1}{2}\|A\|(1+3\Theta^4)$, yielding

$$\frac{1}{\varepsilon}(\Theta^2-1)^2 - \frac{3}{2}\|A\|\Theta^4 > \frac{1}{2}\|A\|.$$

The result now follows from (26) and (24).

Now we prove that $\|x\| > \gamma$ implies $P(x;\varepsilon) > P(x^0;\varepsilon)$. If $\|x\| > \gamma > 1$ we can write

$$P(x;\varepsilon) \geqslant -\frac{3}{2}\|A\|\|x\|^4 + \frac{\|x\|^4}{\varepsilon}\left(1-\frac{1}{\|x\|^2}\right)^2 \geqslant \|x\|^4\left[-\frac{3}{2}\|A\| + \frac{1}{\varepsilon\gamma^4}(\gamma^2-1)^2\right] \tag{27}$$

and similarly

$$P(x^0;\varepsilon) \leqslant \frac{1}{2}\|A\| \leqslant \frac{1}{2}\|A\|\gamma \leqslant \frac{1}{2}\|A\|\|x\|^4. \tag{28}$$

Hence we get from $\varepsilon < \bar{\varepsilon}\frac{1}{2\|A\|}\frac{(\gamma^2-1)^2}{\gamma^4}$ the relation $\frac{1}{\varepsilon\gamma^4}(\gamma^2-1)^2 > 2\|A\|$ and finally

$$-\frac{3}{2}\|A\| + \frac{1}{\varepsilon\gamma^4}(\gamma^2-1)^2 > \frac{1}{2}\|A\|,$$

which via (28) and (27) establishes the result. □

The next four theorems establish the main exactness properties of the penalty function $P(x;\varepsilon)$ that we need for the purpose of solving StQP.

THEOREM 10. (First-order exactness property). *For $0 < \varepsilon < \bar{\varepsilon}$ as in (20), a point $\bar{x} \in \mathcal{L}_0$ is a stationary point of $P(x;\varepsilon)$ if and only if $(\bar{x}, \mu(\bar{x}))$ is a KKT point for problem (4).*

Proof. ($\Leftarrow$) Since $\bar{x}$ is a KKT point for (4), it is in particular feasible, so that we have, by the first equation in (21), $\nabla P(\bar{x};\varepsilon) = \nabla f(\bar{x}) + 2\mu(\bar{x})\bar{x} = 0$.

($\Rightarrow$) If $\nabla P(\bar{x};\varepsilon) = 0$ then, by (21) and using $x^\top \nabla f(x) = 4f(x)$,

$$\begin{aligned} 0 &= \varepsilon x^\top \nabla P(x;\varepsilon) = \varepsilon x^\top \nabla f(x)[1-2(\|x\|^2-1)] - 4\varepsilon f(x)\|x\|^2 + 4(\|x\|^2-1)\|x\|^2 \\ &= (\|x\|^2-1)(4\|x\|^2 - 12\varepsilon f(x)) \end{aligned}$$

We can write for $\bar{x} \in \mathcal{L}_0$

$$4\|\bar{x}\|^2 - 12\varepsilon f(\bar{x}) \geqslant 4\|\bar{x}\|^2 - 6\varepsilon\|A\|\|\bar{x}\|^4 \geqslant 4\Theta^2 - 6\varepsilon\|A\|\gamma^4.$$

Since $\varepsilon < \frac{4\Theta^2}{6\|A\|\gamma^4}$ we get that the term $4\|\bar{x}\|^2 - 12\varepsilon f(\bar{x}) > 0$. Hence $\nabla P(\bar{x};\varepsilon) = 0$ implies $\|\bar{x}\|^2 = 1$ and again $\nabla f(\bar{x}) + 2\bar{\mu}\bar{x} = \nabla P(\bar{x};\varepsilon) = 0$ for $\bar{\mu} = \mu(\bar{x}) = -2f(\bar{x})$. □

It is easy to show that there is a one-to-one correspondence between global minimizers of problem (4) and global minimizers of the penalty function P. The proof is quite standard in penalty approach and we report it here for sake of completeness.

THEOREM 11. (Correspondence of global minimizers). *For $0 < \varepsilon < \bar{\varepsilon}$ as in (20), every global minimizer of problem (4) is a global minimizer of $P(x;\varepsilon)$ and conversely.*

Proof. By Theorem 9, the penalty function $P(.;\varepsilon)$ admits a global minimizer $\hat{x}$, which is obviously a stationary point of $P(.;\varepsilon)$ and hence, by Theorem 10, a KKT point of Problem (4), so that we have:

$$P(\hat{x};\varepsilon) = f(\hat{x}).$$

On the other hand, if x^* is a global minimizer of problem (4), it is also a KKT point and hence by the preceding proposition it is a stationary point of $P(.;\varepsilon)$ which implies again that $P(x^*.;\varepsilon) = f(x^*)$. Now, we proceed by contradiction. Assume that a global minimizer $\hat{x}$ of $P(.;\varepsilon)$ is not a global minimizer of problem (4), then there should exists a point x^*, global minimizer of problem (4), such that

$$P(\hat{x};\varepsilon) = f(\hat{x}) > f(x^*) = P(x^*;\varepsilon)$$

that contradicts the assumption that $\hat{x}$ is a global minimizer of $P(x.;\varepsilon)$. The converse is true by analogous considerations. □

Regarding local minimizers, we have the following result, whose proof is standard in penalty approach and is reported only for sake of completeness.

THEOREM 12. (Correspondence of local minimizers). *For $0 < \varepsilon < \bar{\varepsilon}$ as in (20), let $\bar{x} \in \mathcal{L}_0$ be a local minimizer of $P(x;\varepsilon)$. Then $\bar{x}$ is a local solution to problem (4), and $\mu(\bar{x})$ is the associated KKT multiplier.*

Proof. We first recall that if $\bar{x}$ is a local minimizer of $P(x;\varepsilon)$ then the pair $(\bar{x}, \mu(\bar{x}))$ satisfies the KKT conditions for problem (4) and we have $P(\bar{x};\varepsilon) = f(\bar{x})$. Since $\bar{x}$ is a local minimizer of P, there exists a neighborhood Ω of $\bar{x}$ such that

$$f(\bar{x}) = P(\bar{x};\varepsilon) \leqslant P(x;\varepsilon) \quad \text{for all } x \in \Omega\ .$$

Since for every feasible point x we have $P(x;\varepsilon) = f(x)$, we can also write

$$f(\bar{x}) \leqslant P(x;\varepsilon) = f(x) \quad \text{for all } x \in \Omega \cap \mathcal{F} \tag{29}$$

and hence $\bar{x}$ is a local minimizer for problem (4). □

As we use the exact penalty approach to locate a solution of the StQP, we need also to exploit the correspondence among points satisfying second-order necessary conditions. In particular, the following result is needed.

THEOREM 13. (Second-order exactness property). *For $0 < \varepsilon < \bar{\varepsilon}$ as in (20), let $\bar{x} \in \mathcal{L}_0$ be a stationary point of $P(x;\varepsilon)$ satisfying the standard second-order necessary conditions for unconstrained optimality. Then $(\bar{x}, \mu(\bar{x}))$ satisfies the second-order necessary conditions for problem (4).*

Proof. From Theorem 10 we know that first order conditions hold. Hence we have that $\|\bar{x}\| = 1$ and that $\nabla f(\bar{x}) + 2\bar{\mu}\bar{x} = 0$. Now, recalling (22) we obtain

$$\begin{aligned}\nabla^2 P(\bar{x};\varepsilon) &= \nabla^2 f(\bar{x}) - 4f(\bar{x})I - 4[\nabla f(\bar{x})\bar{x}^\top + \bar{x}\nabla f(\bar{x})^\top] + \frac{8}{\varepsilon}\bar{x}\bar{x}^\top \\ &= \nabla^2 f(\bar{x}) + 2\bar{\mu}I + 8\left(\frac{1}{\varepsilon} + 2\bar{\mu}\right)\bar{x}\bar{x}^\top,\end{aligned}$$

since $\nabla f(\bar{x}) = -2\bar{\mu}\bar{x}$ by the KKT conditions. Next, for every d such that $d^\top \bar{x} = 0$ we get

$$0 \leqslant d^\top \nabla^2 P(\bar{x};\varepsilon) d = d^\top (\nabla^2 f(\bar{x}) + 2\bar{\mu}I) d$$

and the proof is completed. □

4.2. ALGORITHMIC ASPECTS

On the basis of the definition of the penalty function above, we have recast the problem of locating a constrained solution of problem (4) as the problem of locating an unconstrained solution of P. As we mentioned at the

beginning of this section, this allows us to use an unconstrained method for the minimization of the penalty function P converging to points satisfying the second order necessary conditions. Indeed, by Theorem 13 stationary points of P satisfying the second order necessary conditions, are points satisfying the second-order necessary conditions (17) for problem (4) which, in turn, by Theorem 7 are points satisfying the second-order necessary condition (3) for the StQP (1).

We observe that given a feasible starting point x^0, any of these algorithms is able to locate a KKT point with a lower value of the objective function. In fact, any unconstrained algorithm obtains a stationary point $\bar{x}$ for P such that

$$P(\bar{x};\varepsilon) < P(x^0;\varepsilon).$$

Then, Theorem 10 ensures that $\bar{x}$ is a KKT point of problem (4). On the other hand, if x^0 is a feasible point, recalling (23), we get that

$$f(\bar{x}) = P(\bar{x};\varepsilon) < P(x^0;\varepsilon) = f(x^0).$$

In conclusion, by using an unconstrained optimization algorithm, we get a KKT point $\bar{x}$ of problem (4) with a value of the objective function lower than the value at the starting point x^0.

We have performed numerical experiments based on the Penalty transformation of the **BQP**. We refer to the implementation of the method based on these successive reformulations as **BQP**.

As a method for unconstrained minimization of P we have used the curvilinear nonmonotone line search algorithm **NMonNC** described in [20]. The algorithm generates a sequence $\{x^k\}$ as

$$x^{k+1} = x^k + \alpha^k s^k + \alpha^{k^2} d^k$$

where d^k is a Newton-type direction, s^k is a particular negative curvature direction which has some resemblance to an eigenvector corresponding to the smallest eigenvalue of the hessian matrix $\nabla^2 P$ and α^k is a stepsize. **NMonNC** uses a Lanczos bases iterative scheme to compute both the directions s^k, d^k. It has been proved in [20] to be globally convergent to points satisfying second order necessary conditions with superlinear rate of convergence.

Of course, we could have also defined ad-hoc algorithms for finding a local minimizer of problem (4) that exploit its structure (following similar approaches of [18, 19]). However, this is out of the scope of the paper.

We remark, however, that the **BQP** method is a local method, in the sense that it guarantees convergence only to local solutions and there is no guarantee that the points obtained by the algorithm **NMonNC** are global minimizers of $P(x;\varepsilon)$ and hence of problem (1).

Hence, if we want to determine a global solution, we ought to include some global procedure to 'escape' from local solutions. However, we have performed the numerical experiments, reported in the next section, without any 'escape step', namely without implementing any heuristic or global procedure to 'escape' from local-nonglobal points whenever they are found during the minimization. The only global aspects in the implementation stays in the fact that we adopt a multi-start approach, namely we perform many minimization processes starting from different starting point chosen randomly; then we select the best value obtained.

5. Numerical Experiments

Standard quadratic optimization problems arise in several applications (see [5] for a full review). We consider a special StQP problem that arises from a continuous formulation of a classical problem in graph theory, namely the maximum clique problem.

5.1. THE MAXIMUM CLIQUE PROBLEM (MCP)

Given an undirected graph $G = (V, E)$ with vertex V and edge set $E \subset V \times V$, the max clique problem consists on finding a complete subgraph of G of maximum cardinality ω^*.

This problem has many different continuous formulation as a nonconvex optimization problem. For a survey we refer to [7]. Here we use the continuous formulation given by Bomze [4] as a regularization of the Motzkin–Straus [22] formulation. In the original Motzkin–Straus formulation, the value of the maximum clique ω^* is obtained as $(1 - f^*)^{-1}$ where f^* denotes the optimal value of the indefinite quadratic program

$$\max \{y^\top A_G y : y \in \Delta\},$$

where A_G denotes the adjacency matrix of the graph, namely $a_{ij} = 1$ if $(i,j) \in E$ and Δ is the standard simplex in the n-dimensional Euclidean space. The regularized version of Bomze is obtained by adding to the objective function the term $\frac{1}{2}\|y\|^2$, so that the maximum clique problem can be written as:

$$\max \{y^\top \left(A_G + \frac{1}{2} I\right) y : y \in \Delta\}, \tag{30}$$

an StQP of the type (1) with matrix $A = -2(A_G + \frac{1}{2} I)$.

The regularized version (30) avoids the drawback of the original Motzkin–Straus formulation of having spurious solutions, namely of solutions that are not in a one-to-one correspondence with solutions of the original combinatorial problems. The main result proved in [4] is reported here:

THEOREM 14. *Let G be an undirected graph and consider problem (30). Then the following assertion are equivalent:*

(a) $\bar{y}$ is a strict local maximum for problem (30);

(b) $\bar{y}$ is a local maximum for problem (30);

(c) $\bar{y} = \frac{1}{\bar{\omega}} \sum_{i \in \sigma} e_i$ where σ is a maximal clique of cardinality $\bar{\omega}$.

If one of the above conditions (and therefore all) is met, the objective $\bar{y}^\top (A_{\mathcal{G}} + \frac{1}{2} I)\bar{y}$ equals the value $1 - \frac{1}{2\bar{\omega}}$.

Assertions (a) and (b) imply that every local solution of (30) is strict, so that there is no problem in identifying a clique σ from $\bar{y}$. Indeed a vertex $i \in \sigma$ if and only if $\bar{y}_i > 0$ and $\bar{\omega} = \frac{1}{2}(1 - \bar{f})^{-1}$. Obviously σ^* is a maximum clique of G if and only if x^* is the global solution of (30).

5.2. IMPLEMENTATION DETAILS

First we note that in the unconstrained algorithm Lanczos only matrix times vector products are required, so that in principle the storage of the adjacency matrix $A_{\mathcal{G}}$ is not required. However, since for the problems of the Dimacs collection $A_{\mathcal{G}}$ is given (only the nonzero elements), we store it. These may increase the cpu time required at each iteration of the **NMonNC** method. In **NMonNC**, we set to 100 the memory for the nonmonotone scheme and we set to 10 the number of Lanczos basis vectors stored.

To obtain an estimate of the threshold value of ε to be used in the penalty function P, we use in (20) the fact that $\|A\|_F \leqslant \|A\| \leqslant \sqrt{n}\|A\|_F$ where $\|A\|_F = \sum_i \sum_j a_{ij}$ denotes the Frobenius norm of the matrix A; we obtain the following estimate of the value of the penalty parameter $\varepsilon < \bar{\varepsilon}$

$$\begin{aligned} &\min\left\{\frac{1}{2}\|A\|\frac{(1+3\Theta^4)}{(\Theta^2-1)^2}, \frac{1}{2\|A\|}\frac{(\gamma^2-1)^2}{\gamma^4}, \frac{2\Theta^2}{3\|A\|\gamma^4}\right\} \\ &\quad \geqslant \min\left\{\frac{1}{2}\|A\|_F\frac{(1+3\Theta^4)}{(\Theta^2-1)^2}, \frac{1}{2\sqrt{n}\|A\|_F}\frac{(\gamma^2-1)^2}{\gamma^4}, \frac{2\Theta^2}{3\sqrt{n}\|A\|_F\gamma^4}\right\}. \end{aligned} \tag{31}$$

If we set $\Theta = 0.5$ and $\gamma = 2$ in (31), and we use the fact that, in the case of the matrix $A_{\mathcal{G}} + \frac{1}{2} I$ it results $\|A_{\mathcal{G}} + \frac{1}{2} I\|_F = |E| + \frac{1}{2} n$ where $|E|$ is the number of edges in the graph, we obtain the following estimate of the threshold value for the penalty parameter

$$\varepsilon < \frac{1}{48\sqrt{n}(2|E|+1)}.$$

5.3. BENCHMARK RESULTS AND COMPARISON

As a benchmark, we use a set of 64 graph obtained from the DIMACS challenge [17]. Each problem has been solved starting with a randomly

Table 1. Best, average, worst results over 150 random runs

Graph	n	Max	Average	Min	Time average	Time max
brock200_1	200	20	15.91	13	0.91	0.81
brock200_2	200	10	8.073	6	0.94	0.84
brock200_3	200	13	10.36	8	0.99	1.18
brock200_4	200	15	12.09	10	0.97	1.32
brock400_1	400	24	18.67	17	9.44	9.72
brock400_2	400	22	15.91	16	9.44	9.33
brock400_3	200	21	18.64	15	9.01	9.36
brock400_4	200	23	18.65	16	9.04	9.16
brock800_1	800	19	15.14	13	58.45	57.05
brock800_2	800	18	15.31	13	58.28	72.63
brock800_3	800	18	15.08	12	59.07	58.06
brock800_4	800	18	15.01	13	59.38	66.48
c-fat200-1	200	**12**	11.69	10	0.34	0.38
c-fat200-2	200	**24**	22.25	22	0.44	0.43
c-fat200-5	200	**58**	57.36	55	0.57	0.53
c-fat500-1	500	**14**	13.33	12	4.55	4.86
c-fat500-10	200	**126**	12.49	122	12.97	14.23
c-fat500-2	500	**26**	25.52	24	5.57	5.07
c-fat500-5	500	**64**	62.57	60	8.41	7.60
hamming6-2	64	**32**	23.53	16	0.06	0.07
hamming6-4	64	**4**	39.06	2	0.06	0.10
hamming8-2	256	**128**	84.01	60	2.28	2.38
hamming8-4	256	**16**	11.50	6	2.58	2.92
hamming10-2	1024	453	300.4	245	358.75	326.86
hamming10-4	1024	34	30.05	20	291.60	246.68
johnson8-2-4	28	**4**	4	4	0.01	0.02
johnson8-4-4	70	**14**	10.87	7	0.07	0.07
johnson16-2-4	120	**8**	7.993	7	0.07	0.07
johnson32-2-4	496	**16**	16	16	21.89	23.44
keller4	171	10	7.7	7	0.75	0.91
keller5	776	19	16.43	15	50.63	53.27

generated point x^0 with $\|x^0\| = 1$. We perform 150 random runs. In Tables 1 and 2 we report the best, average and worst results obtained in terms of cardinality of the clique, and average and worst results in terms cpu time. In the tables, the entries that correspond to the best known result for a given graph are in bold face.

We consider a comparison with the results presented in [8], obtained with 10 random runs. There two different heuristics (h1, h21) have been presented based on a semidefinite program where the matrix is restricted to be respectively rank-one and rank-two. These semidefinite programs are equivalent to nonlinear continuous optimization problems that have been solved with an augmented Lagrangian approach. Actually in the paper also a third heuristic (h25) based on the two-rank formulation is presented, which outperforms both h1 and h21. The heuristic h25 uses a rule to escape from canonical solutions, and since we did not implement any rule

Table 2. Best, average, worst results over 150 random runs

Graph	n	Max	Average	Min	Time average	Time max
MANN_a9	45	**16**	14.91	13	0.02	0.02
MANN_a27	378	119	117.4	117	4.16	4.86
MANN_a45	1035	330	330	330	4.81	4.81
P_hat300-1	300	**8**	6.367	5	1.88	1.80
P_hat300-2	300	**25**	21.37	18	1.82	2.45
P_hat300-3	300	33	29.99	26	2.21	2.27
P_hat500-1	500	**9**	7.273	6	10.14	6.70
P_hat500-2	500	34	30.37	25	9.64	11.01
P_hat500-3	500	48	43.69	39	10.50	12.36
P_hat700-1	700	9	7.42	6	25.40	28.92
P_hat700-2	700	43	37.94	33	27.31	26.82
P_hat700-3	700	60	54.54	47	33.00	27.50
P_hat1000-1	1000	**10**	7.92	7	63.31	49.29
P_hat1000-2	1000	45	39.83	35	64.07	43.87
P_hat1000-3	1000	63	57.17	52	76.38	85.57
P_hat1500-1	1500	10	8.467	7	186.97	235.89
P_hat1500-2	1500	62	55.63	49	183.58	199.47
P_hat1500-3	1500	91	80.17	73	191.62	179.33
San200_0.7_1	200	17	15.02	12	1.03	1.17
San200_0.7_2	200	12	12	12	0.94	0.91
San200_0.9_1	200	46	45.05	45	0.80	1.03
San200_0.9_2	200	47	35.80	28	0.83	0.71
San200_0.9_3	200	34	30.03	25	0.87	1.02
San400_0.5_1	400	7	6.987	6	10.36	9.23
San400_0.7_1	400	20	20	20	9.72	13.25
San400_0.7_2	400	16	15.01	15	10.35	12.62
San400_0.7_3	400	12	12	12	10.87	10.99
San400_0.9_1	400	69	50.13	39	9.09	10.77
Sanr200_0.7	200	17	13.71	11	0.94	0.95
Sanr200_0.9	200	37	33.56	28	0.84	0.73
Sanr400_0.5	400	12	9.467	8	8.21	9.00
Sanr400_0.7	400	20	15.87	13	9.07	12.23
San1000	1000	8	8.00	8	127.24	137.06

of this type, we do not compare with it. We also include comparison with the heuristics LDR and PBH proposed in [21].

The c-*fat* and *Johnson* graph categories are not reported because all the algorithms return the maximum clique. The same happens for most of the *Hamming* graphs and for *Mann_a9* so that we report only the problems where a different behavior appears. We do not compare on the two largest problems in the DIMACS collection (*Keller6* and *Mann_a81*), because results for these two problems are not reported for most of the other heuristics. BQP finds cliques of size 36 and 1080, respectively, in these instances, which have clique numbers of at least 59 (best known solution), and 1100 (certified), respectively.

In Table 3 we report the number of wins, ties and defeats of PQB with respect to each heuristic h1, h21, LDR and PBH, in terms of the best value

Table 3. Cumulative comparison with other continuous heuristics

	BQP		
	Wins	Tie	Defeats
h1	11	22	15
h21	1	5	42
LDR	40	4	4
PBH	1	9	38

Table 4. Comparison with other continuous heuristics

Graph	BQP	h1	h21	LDR PBH
brock200_1	20	20	**21**	13 20
brock200_2	10	10	**11**	7 **11**
brock200_3	13	13	**14**	10 **14**
brock200_4	15	15	**16**	11 **16**
brock400_1	**24**	22	**24**	17 **24**
brock400_2	22	24	**25**	17 24
brock400_3	21	24	**25**	17 24
brock400_4	23	23	**24**	16 **24**
brock800_1	19	20	**21**	13 **21**
brock800_2	18	**20**	**20**	13 **20**
brock800_3	18	19	**21**	15 20
brock800_4	18	18	**21**	16 20
hamming10-2	453	**512**	**512**	**512 512**
hamming10-4	34	**40**	**40**	32 32
keller4	10	7	**11**	7 **11**
keller5	19	16	24	15 **26**
MANN_a27	119	118	**125**	**125 125**
MANN_a45	330	45	45	340 **342**

obtained over the 48 problems where the behavior is different. From Table 3, we can conclude that our method is comparable to h1, better than LDR, but worse than h21 and PBH.

The overall results are reported in Tables 4, 5 and 6. In these tables, the entries corresponding to the best value obtained by the heuristics are in bold face.

It is worthwhile to remark that many other algorithms for the max-clique based on continuous formulations have been proposed in the literature; see, again, [7]. Most of them are designed specifically for the max-clique problem and use the inherent structure of the problem in a specific way; some of them incorporate also heuristics to escape from inefficient local solutions (due to the special structure of the problem, a multitude of these is inherent to the MCP). By contrast, our algorithm BQP is not specifically designed for the MCP, but for a generic StQP. Indeed, in the two-step transformation from the StQP through the formulation as a BQP we never used

Table 5. Comparison with other continuous heuristics

Graph	BQP	h1	h21	LDR	PBH
P_hat300-1	**8**	7	**8**	6	8
P_hat300-2	**25**	**25**	**25**	16	**25**
P_hat300-3	33	35	**36**	21	35
P_hat500-1	**9**	9	9	6	**9**
P_hat500-2	34	**36**	**36**	26	**36**
P_hat500-3	48	48	**50**	30	48
P_hat700-1	9	9	11	5	10
P_hat700-2	43	**44**	44	20	**44**
P_hat700-3	60	60	**62**	29	**62**
P_hat1000-1	**10**	9	**10**	7	10
P_hat1000-2	45	45	**46**	18	**46**
P_hat1000-3	63	63	**68**	31	64
P_hat1500-1	10	10	11	9	**12**
P_hat1500-2	62	64	**65**	28	64
P_hat1500-3	91	93	**94**	43	91

Table 6. Comparison with other continuous heuristics

Graph	BQP	h1	h21	LDR	PBH
San200_0.7_1	17	15	**30**	16	**30**
San200_0.7_2	12	12	**18**	12	17
San200_0.9_1	46	**70**	**70**	38	**70**
San200_0.9_2	47	36	**70**	30	60
San200_0.9_3	34	**44**	**44**	25	**44**
San400_0.5_1	7	7	**9**	7	**13**
San400_0.7_1	20	20	**40**	20	**40**
San400_0.7_2	16	15	**19**	15	**30**
San400_0.7_3	12	12	**18**	14	17
San400_0.9_1	69	52	**100**	45	**100**
Sanr200_0.7	17	17	**18**	12	**18**
San200_0.9	37	41	**42**	32	41
Sanr400_0.5	12	12	**13**	10	**13**
Sanr400_0.7	20	20	**21**	16	20
San1000	8	8	9	8	**15**

information about the fact that the matrix is an adjacency matrix of a graph and most of this structure may be lost. Moreover, as we already mentioned, BQP is designed to obtain local solutions of the StQP which satisfies the second order necessary conditions. Of course, this method can be integrated into a global optimization scheme that incorporates some special heuristics to escape from inefficient local minimizers. However, this type of heuristic should be tied to the structure of the problem under study and this remains to be done yet. Hence comparisons in Table 3 are not completely fair, but still shed some light on the general performance of the approach proposed.

6. Acknowledgement

The authors thank Massimo Roma for having providing the code of the algorithm **NMonNC** described in [20].

References

1. Bagchi, A. and Kalantari, B. (1990), New optimality conditions and algorithms for homogeneous and polynomial optimization over spheres. Rutcor Research Report No. 40–90.
2. Bertsekas, D.P. (1982), *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, New York.
3. Bomze, I.M. (2002), Regularity vs. degeneracy in dynamics, games, and optimization: a unified approach to different aspects. *SIAM Review* 44, 394–414.
4. Bomze, I.M. (1997), Evolution towards the maximum clique. *Journal of Global Optimization*, 10, 143–164.
5. Bomze, I.M. (1998), On standard quadratic optimization problems. *Journal of Global Optimization*, 13, 369–387.
6. Bomze, I.M. (2001), Quadratic optimization: standard problems; I – theory; II – algorithms; III – applications. In: Floudas, C.A. and Pardalos, P.M. (eds.), *Encyclopedia of Optimization,* Vol. 5, Kluwer, Dordrecht pp. 266–268; 268–270; 270–272.
7. Bomze, I.M., Budinich, M., Pardalos, P. and Pelillo, M. (1999), The maximum clique problem. In: Du, D.-Z. and Pardalos, P.M. (eds.), *Handbook of Combinatorial Optimization,* supp.Vol. A, Kluwer Academic Publishers, pp. 1–74.
8. Burer, S., Monteiro, R.D.C. and Zhang, Y. (2002), Maximum stable set formulations and heuristics based on continuous optimization. *Mathematical Programming,* 94, 137–166.
9. Conn, A.R., Gould, N.I.M., Orban, D. and Toint, Ph.L. (2000), A primal-dual trust region algorithm for non-convex nonlinear programming. *Mathematical programming, Ser. B* 87, 215–249.
10. Di Pillo, G. (1994), Exact penalty methods. In: Spedicato, E. (ed.), *Algorithms for continuous optimization: The state of the art,* Kluwer Academic Publishers, pp. 209–253.
11. Di Pillo, G., Grippo, L. and Lampariello, F. (1980), A method for solving equality constrained optimization problems by unconstrained minimization. In: Malinowski, K., Iracki, K. and Walukiewicz, S. (eds.), *Optimization Techniques – Proceedings of the 9th IFIP Conference, Berlin,* Lecture Notes, Springer- Verlag.
12. Facchinei, F. and Lucidi, S. (1998), Convergence to 2nd order stationary points in inequality constrained optmization. *Mathematics of Operations Research*, 23, 746–766.
13. Di Pillo, G., Liuzzi, G., Lucidi, S. and Palagi, L. (2002), Fruitful uses of smooth exact merit functions in constrained optimization. In: Di Pillo, G. and Murli, A. (eds.), *High Performance Algorithms and Software for Nonlinear Optimization,* Kluwer Academic Publishers, pp. 198–222.
14. Ferris, M., Lucidi, S. and Roma, M. (1996), Nonmonotone curvilinear Line Search Methods for Unconstrained Optimization. *Computational Optimization and Applications*, 6, 117–136.
15. Fletcher, R. (1981), *Practical Methods of Optimization, Vol.2: Constrained Optimization*. Wiley, New York.
16. Hestenes, M. (1969), Multiplier and gradient methods. *Journal of Optimization Theory and Application*, 4, 303–320.

17. Johnson, D. and Trick, M.A. (1996), (eds.), *Cliques Coloring and Satisfiability: Second DIMACS Implementation Challenge,* Vol. 26 of DIMACS Series, AMS.
18. Lucidi, S., Palagi, L. and Roma, M. (1998), On some properties of quadratic programs with a convex quadratic constraint. *SIAM Journal on Optimization*, 8, 105–122.
19. Lucidi, S. and Palagi, L. Solution of the trust region problem via a smooth Unconstrained Reformulation. In: Pardalos, P. and Wolkowicz, H. (eds.), *Topics in Semidefinite and Interior-Point methods,* Fields Institute Communications Vol. 18, 237–250, AMS.
20. Lucidi, S., Rochetich, F. and Roma, M. (1998), Curvilinear stabilizaton techniques for truncated Newton methods in large scale unconstrained optimization. *SIAM Journal on Optimization*, 8, 916–939.
21. Massaro, A., Pelillo, M. and Bomze, I.M. (2002), A complementarity pivoting approach to the Maximum weight clique problem. *Siam Journal on Optimization,* 12, 928–948.
22. Motzkin, T.S. and Straus, E.G. (1965), Maxima for graphs and a new proof of a theorem of Turan.*Canadian Journal of Mathematics,* 17, 533–540.
23. Matlab-Optimization Toolbox 2.1, The MathWorks, Inc. (2001).
24. Powell, M.J.D. (1969), *A method for nonlinear constraints in minimization problem*. In: Fletcher, R. (ed.), Optimization, Academic Press, New York, pp. 283–298.