Optimization Letters (2019) 13:1269–1283
https://doi.org/10.1007/s11590-018-1316-0

ORIGINAL PAPER

On global minimizers of quadratic functions with cubic regularization

Andrea Cristofari[1] · Tayebeh Dehghan Niri[2] · Stefano Lucidi[3]

Received: 3 November 2017 / Accepted: 15 August 2018 / Published online: 23 August 2018

Abstract
In this paper, we analyze some theoretical properties of the problem of minimizing a quadratic function with a cubic regularization term, arising in many methods for unconstrained and constrained optimization that have been proposed in the last years. First we show that, given any stationary point that is not a global solution, it is possible to compute, in closed form, a new point with a smaller objective function value. Then, we prove that a global minimizer can be obtained by computing a finite number of stationary points. Finally, we extend these results to the case where stationary conditions are approximately satisfied, discussing some possible algorithmic applications.

Keywords Unconstrained optimization · Cubic regularization · Global minima

1 Introduction

In this paper, we address the solutions of the following (possibly non-convex) optimization problem:

$$\min_{s\in\mathbb{R}^n} \; m(s) := c^T s + \frac{1}{2}s^T Qs + \frac{1}{3}\sigma\|s\|^3, \qquad (1)$$

✉ Andrea Cristofari
andrea.cristofari@unipd.it

Tayebeh Dehghan Niri
t.dehghan@stu.yazd.ac.ir

Stefano Lucidi
lucidi@diag.uniroma1.it

[1] Department of Mathematics, University of Padua, Via Trieste, 63, 35121 Padua, Italy

[2] Department of Mathematics, Yazd University, P.O. Box 89195-74, Yazd, Iran

[3] Department of Computer, Control and Management Engineering, Sapienza University of Rome, Via Ariosto, 25, 00185 Rome, Italy

where $c \in \mathbb{R}^n$, Q is a symmetric $n \times n$ matrix, σ is a positive real number and, here and in the rest of the article, $\|\cdot\|$ is the Euclidean norm.

In recent years, there has been a growing interest in studying the properties of problem (1), since functions of the form of $m(s)$ are used as local models (to be minimized) in many algorithmic frameworks for unconstrained optimization [1–7, 11,12,14,17–19], which have been even extended to the constrained case [2,8,16]. To be more specific, let us consider the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x),$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable function. The class of methods proposed in the above cited papers is mostly characterized by the iteration $x^{k+1} = x^k + s^k$, being s^k a (possibly approximate) minimizer of the cubic model

$$m^k(s) := f\left(x^k\right) + \nabla f\left(x^k\right)^T s + \frac{1}{2} s^T \nabla^2 f\left(x^k\right) s + \frac{1}{3}\sigma^k \|s\|^3,$$

where σ^k is a suitably chosen positive real number. Interestingly, it can be shown that, under suitable assumptions, this algorithmic scheme is able to achieve quadratic convergence rate and a worst-case iteration complexity better than the gradient method. In particular, if $\nabla^2 f(x)$ is Lipschitz continuous and s^k is a global minimizer of $m^k(s)$, Nesterov and Polyak [18] proved a worst-case iteration count of order $O(\varepsilon^{-3/2})$ to obtain $\|\nabla f(x^k)\| \le \varepsilon$. Cartis et al. [6,7] generalized this result, obtaining the same complexity bound, but allowing for a symmetric approximation of $\nabla^2 f(x^k)$ to be used in $m^k(s)$ and relaxing the condition that s^k is a global minimizer of $m^k(s)$. Moreover, superlinear and quadratic convergence rate were proved under appropriate assumptions, but without requiring $\nabla^2 f(x^k)$ to be globally Lipschitz continuous.

The intuition behind the algorithm proposed by Cartis, Gould and Toint is that the parameter σ^k plays the same role as the (reciprocal of the) trust-region radius in trust-region methods. Moreover, some theoretical properties of trust-region models can be extended to (1), such as the existence of necessary and sufficient conditions for global minimizers even when $m(s)$ is non-convex [6,14,18]. In this fashion, Cartis, Gould and Toint proposed the Adaptive Regularization algorithm using Cubics (ARC) that, besides having the theoretical convergence properties mentioned above, is in practice comparable with state-of-the-art trust-region methods.

In this respect, in the above cited papers different strategies were proposed to minimize $m^k(s)$. In particular, in [6,18] some iterative techniques were devised to compute global minimizers, that are based on solving a one-dimensional non-linear equation.

Starting from these considerations, here we focus on the solutions of problem (1), pointing out some theoretical properties that, besides their own interest, may be useful from an algorithmic point of view. In particular, we first extend the results obtained in [15] for trust-region models and we show that, given any stationary point of (1) that is not a global minimizer, we can compute, in closed form, a new point that reduces $m(s)$. So, a global minimizer of (1) can be obtained by repeating this step a finite number of times, that is, computing at most $2(k+1)$ stationary points, where k is the number of distinct negative eigenvalues of the matrix Q. Further, we show how this strategy can be generalized to the case where stationary conditions are approximately satisfied, opening to a possible practical usage of the proposed results.

The rest of the paper is organized as follows. Section 2 is the core of the paper, where we point out some theoretical properties of the stationary points of (1) and analyze how to compute global minima by escaping from stationary points that are not global minimizers. In Sect. 3 we generalize these properties, considering approximate stationary points, and we briefly discuss how these results can used in a more general framework. Finally, we draw some conclusions in Sect. 4.

2 Properties of stationary points

In this section, we present the main results of the paper. First, let us report the definition of stationary points of problem (1) and recall a known result on necessary and sufficient conditions for global optimality, whose proof can be found in [6]. From now on, we indicate with I the $n \times n$ identity matrix.

Definition 1 We say that $s^* \in \mathbb{R}^n$ is a stationary point of problem (1) if

$$\nabla m(s^*) = c + Qs^* + \sigma \|s^*\| s^* = 0,$$

or equivalently,

$$c + Qs^* + \lambda s^* = 0, \tag{2}$$

$$\lambda = \sigma \|s^*\|. \tag{3}$$

Theorem 1 *A point $s^* \in \mathbb{R}^n$ is a global minimizer of problem (1) if and only if it satisfies stationary conditions (2)–(3) and the matrix $(Q + \sigma\|s^*\| I)$ is positive semidefinite. Moreover, s^* is unique if $(Q + \sigma\|s^*\| I)$ is positive definite.*

Now, exploiting the close relation between problem (1) and the trust-region model (see [9] for an overview on trust-region methods), we extend the results obtained in [15] to show that

(i) given a stationary point $\bar{s}$ of (1) that is not a global minimizer, we can compute, in closed form, a new point $\hat{s}$ such that $m(\hat{s}) < m(\bar{s})$;
(ii) a global minimizer of (1) can be obtained by computing at most $2(k+1)$ stationary points, where k is the number of distinct negative eigenvalues of the matrix Q.

We start by proving the first point, as stated in the following theorem.

Theorem 2 *Let $\bar{s}$ be a stationary point of problem (1). We define the point $\hat{s}$ as follows:*

(a) *if $c^T \bar{s} > 0$, then*

$$\hat{s} := -\bar{s};$$

(b) *if $c^T \bar{s} \le 0$ and a vector $d \in \mathbb{R}^n$ exists such that $d^T (Q + \sigma\|\bar{s}\| I) d < 0$,*

(i) *if* $\bar{s} = 0$, *then*

$$\hat{s} := \bar{s} + \alpha d,$$

with

$$0 < \alpha < -\frac{3\, d^T Q d}{2\, \sigma \|d\|^3};$$

(ii) *if* $\bar{s} \neq 0$ *and* $\bar{s}^T d \neq 0$, *then*

$$\hat{s} := \bar{s} - 2\frac{\bar{s}^T d}{\|d\|^2} d;$$

(iii) *if* $\bar{s} \neq 0$ *and* $\bar{s}^T d = 0$, *then*

$$\hat{s} := \bar{s} - 2\frac{\bar{s}^T z}{\|z\|^2} z,$$

where $z := \bar{s} + \alpha d$ *and*

$$\alpha > \frac{c^T d - \sqrt{\left(c^T d\right)^2 + \left(c^T \bar{s}\right)\left[d^T (Q + \sigma \|\bar{s}\| I) d\right]}}{d^T (Q + \sigma \|\bar{s}\| I) d}.$$

We have that

$$m(\hat{s}) < m(\bar{s}).$$

Proof In case (a), we can write

$$\begin{aligned} m(\hat{s}) = m(-\bar{s}) &= c^T (-\bar{s}) + \frac{1}{2}\bar{s}^T Q \bar{s} + \frac{1}{3}\sigma \|\bar{s}\|^3 \\ &< c^T \bar{s} + \frac{1}{2}\bar{s}^T Q \bar{s} + \frac{1}{3}\sigma \|\bar{s}\|^3 = m(\bar{s}). \end{aligned}$$

Now, we consider case (b) and distinguish the three subcases.

(i) From (2)–(3), we have that $c = 0$. Thus, we can write

$$m(\bar{s} + \alpha d) = m(\alpha d) = \frac{1}{2}\alpha^2 d^T Q d + \frac{1}{3}\sigma \alpha^3 \|d\|^3, \quad \forall \alpha \in \mathbb{R}^n.$$

Consequently,

$$m(\bar{s} + \alpha d) = \frac{1}{6}\alpha^2 \left(3 d^T Q d + 2\sigma\alpha \|d\|^3\right) < 0 = m(\bar{s}),$$

for all $0 < \alpha < -\dfrac{3\, d^T Qd}{2\, \sigma \|d\|^3}$.

(ii) First, we observe that

$$\left\| \bar{s} - 2\frac{\bar{s}^T d}{\|d\|^2} d \right\|^2 = \|\bar{s}\|^2 + \left(2\frac{\bar{s}^T d}{\|d\|^2}\right)^2 \|d\|^2 - 4\frac{\bar{s}^T d}{\|d\|^2}\left(\bar{s}^T d\right) = \|\bar{s}\|^2. \tag{4}$$

Moreover, the function $m(s)$ can be written as

$$m(s) = c^T s + \frac{1}{2} s^T (Q + \sigma \|s\| I) s - \frac{1}{6}\sigma \|s\|^3. \tag{5}$$

Using (4) and (5), we can write $m\left(\bar{s} - 2\dfrac{\bar{s}^T d}{\|d\|^2} d\right)$ as

$$c^T\left(\bar{s} - 2\frac{\bar{s}^T d}{\|d\|^2} d\right) + \frac{1}{2}\left(\bar{s} - 2\frac{\bar{s}^T d}{\|d\|^2} d\right)^T (Q + \sigma \|\bar{s}\| I)\left(\bar{s} - 2\frac{\bar{s}^T d}{\|d\|^2} d\right) - \frac{1}{6}\sigma \|\bar{s}\|^3.$$

Rearranging and taking into account that $\nabla m(\bar{s}) = Q\bar{s} + \sigma \|\bar{s}\| \bar{s} + c$, we obtain

$$m\left(\bar{s} - 2\frac{\bar{s}^T d}{\|d\|^2} d\right) = m(\bar{s}) + \frac{1}{2}\left(2\frac{\bar{s}^T d}{\|d\|^2}\right)^2 d^T (Q + \sigma \|\bar{s}\| I) d - 2\frac{\bar{s}^T d}{\|d\|^2} \nabla m(\bar{s})^T d. \tag{6}$$

Stationary conditions (2)–(3) imply that $\nabla m(\bar{s}) = 0$. Exploiting the fact that $d^T (Q + \sigma \|\bar{s}\| I) d < 0$, we get $m\left(\bar{s} - 2\dfrac{\bar{s}^T d}{\|d\|^2} d\right) < m(\bar{s})$.

(iii) Using the definition of z, we can write

$$\begin{aligned} z^T (Q + \sigma \|\bar{s}\| I) z &= (\bar{s} + \alpha d)^T (Q + \sigma \|\bar{s}\| I)(\bar{s} + \alpha d) \\ &= \bar{s}^T (Q + \sigma \|\bar{s}\| I)\bar{s} + \alpha^2 d^T (Q + \sigma \|\bar{s}\| I) d \\ &\quad + 2\alpha d^T (Q + \sigma \|\bar{s}\| I)\bar{s}. \end{aligned}$$

From stationary conditions (2)–(3), we have that $Q\bar{s} + \sigma \|\bar{s}\| \bar{s} = -c$. So, we obtain

$$z^T (Q + \sigma \|\bar{s}\| I) z = \alpha^2 d^T (Q + \sigma \|\bar{s}\| I) d - 2\alpha c^T d - c^T \bar{s}.$$

It is straightforward to verify that the right-hand side of the above equality is negative for all $\alpha > \tilde{\alpha}$, where

$$\tilde{\alpha} = \frac{c^T d - \sqrt{\left(c^T d\right)^2 + \left(c^T \bar{s}\right)\left[d^T (Q + \sigma \|\bar{s}\| I) d\right]}}{d^T (Q + \sigma \|\bar{s}\| I) d}.$$

Consequently, since $z = \bar{s} + \alpha d$ with $\alpha > \tilde{\alpha}$, it follows that $z^T(Q + \sigma\|\bar{s}\|I)z < 0$. We can thus proceed as in case (ii) by defining the point $\hat{s} = \bar{s} - 2\frac{\bar{s}^T z}{\|z\|^2}z$ and we get the result. □

Remark 1 Conditions of Theorem 2 are satisfied if and only if the stationary point $\bar{s}$ is not a global minimizer. It follows from the fact that, if (a) or (b) hold at $\bar{s}$, then $\bar{s}$ is not a global minimizer; vice versa, if $\bar{s}$ is not a global minimizer, then $(Q + \sigma\|\bar{s}\|I)$ is not positive semidefinite (see Theorem 1) and then (b) holds.

Now, we show how the above result can be exploited to obtain a global minimizer of (1) by computing a finite number of stationary points. We first need the following lemma, stating that two stationary points of problem (1) with the same norm produce the same objective value.

Lemma 1 *Let $\hat{s}$ and $\bar{s}$ be two points satisfying stationary conditions (2)–(3) with the same λ. Then,*

$$m(\hat{s}) = m(\bar{s}).$$

Proof For every pair (s, λ) satisfying (2)–(3), we can write

$$\begin{aligned} m(s) &= c^T s + \frac{1}{2}s^T(-c - \lambda s) + \frac{1}{3}\sigma\|s\|^3 \\ &= \frac{1}{2}c^T s - \frac{1}{2}\lambda\|s\|^2 + \frac{1}{3}\sigma\|s\|^3 = \frac{1}{2}c^T s - \frac{1}{6}\sigma\|s\|^3. \end{aligned}$$

Then,

$$\begin{aligned} m(\hat{s}) &= \frac{1}{2}c^T \hat{s} - \frac{1}{6}\sigma\|\hat{s}\|^3 = -\frac{1}{2}\bar{s}^T(Q + \lambda I)\hat{s} - \frac{1}{6}\sigma\|\bar{s}\|^3 \\ &= \frac{1}{2}c^T \bar{s} - \frac{1}{6}\sigma\|\bar{s}\|^3 = m(\bar{s}). \end{aligned}$$

□

The following proposition establishes a bound on the maximum number of stationary points with different norm. The proof follows the same line of arguments used in [6] to characterize global minimizers of the cubic model. It is entirely reported here for the sake of completeness.

Proposition 1 *At most $2(k+1)$ points that satisfy (2)–(3) with distinct values of λ exist, where k is the number of distinct negative eigenvalues of Q.*

Proof First, we observe that if $\lambda = 0$, then $s = 0$ is the only point that satisfies (2)–(3). So, in the following we consider the case in which $\lambda > 0$ (i.e., $s \neq 0$). Let $V \in \mathbb{R}^{n\times n}$ be an orthonormal matrix such that

$$V^T Q V = M,$$

where $M := \text{diag}_{i=1,\dots,n}\{\mu_i\}$ and $\mu_1 \leq \cdots \leq \mu_n$ are the eigenvalues of Q. Now, we can introduce the vector $a \in \mathbb{R}^n$ and consider the transformation

$$s = Va.$$

Pre-multiplying (2) by V^T, we get

$$V^T(Q + \lambda I)s = -V^T c,$$

and then

$$(M + \lambda I)a = -\beta,$$

where $\beta = -V^T c$.
The above expression can be equivalently written as

$$a_i = -\frac{\beta_i}{\mu_i + \lambda}, \quad i = 1, \dots, n. \tag{7}$$

Moreover, from (3) we get

$$\lambda^2 = \sigma^2 \|s\|^2 = \sigma^2 \|Va\|^2 = \sigma^2 \|a\|^2. \tag{8}$$

Using (7) and (8), we can rewrite the stationary conditions as follows:

$$\begin{cases} g(\lambda) = \dfrac{1}{\sigma^2}, \\ \lambda > 0, \end{cases} \tag{9}$$

where

$$g(\lambda) := \frac{1}{\lambda^2} \sum_{i=1}^{n} \frac{\beta_i^2}{(\mu_i + \lambda)^2}.$$

Now, we have two cases.

(i) $\beta_i = 0$ for all $i = 1, \dots, n$ (i. e., $c = 0$). It follows that $g(\lambda) = 0$ in all the domain and system (9) does not admit solutions. In this case, only $s = 0$ satisfies stationary conditions (2)–(3).

(ii) An index $i \in \{1, \dots, n\}$ exists such that $\beta_i \neq 0$ (i. e., $c \neq 0$). Without loss of generality, we assume that $\mu_1, \dots, \mu_p \leq 0$, with $p \leq n$. Then $g(\lambda)$ is defined in the following $n + 2$ subintervals:

$$\begin{aligned} &(-\infty, -\mu_n) \cup (-\mu_n, -\mu_{n-1}) \cup \cdots \cup \left(-\mu_{p+1}, 0\right) \\ &\quad \cup \left(0, -\mu_p\right) \cup \cdots \cup (-\mu_2, -\mu_1) \cup (-\mu_1, +\infty) . \end{aligned}$$

Computing the derivatives of $g(\lambda)$, we obtain

$$\frac{d}{d\lambda} g(\lambda) = -2 \sum_{i=1}^{n} \beta_i^2 \left[\lambda(\mu_i + \lambda)\right]^{-3} (\mu_i + 2\lambda),$$

$$\frac{d^2}{d\lambda^2} g(\lambda) = 2 \sum_{i=1}^{n} \beta_i^2 \left[\lambda(\mu_i + \lambda)\right]^{-4} \left[10\lambda^2 + 10\mu_i\lambda + 3\mu_i^2\right].$$

It is straightforward to verify that $\frac{d^2}{d\lambda^2} g(\lambda) > 0$ in all the points where $g(\lambda)$ is defined, that is, $g(\lambda)$ is strictly convex in all the non-empty subintervals that define its domain.
Taking into account that $\lim_{\lambda \to 0} g(\lambda) = +\infty$, $\lim_{\lambda \to -\mu_i} g(\lambda) = +\infty$ for all $\beta_i \neq 0$ and $\lim_{\lambda \to \pm\infty} g(\lambda) = 0$, we get that $g(\lambda)$ has at most $2(n + 1)$ roots: at most one in each extreme subinterval and at most two in all the other subintervals.
Now, let $k \leq p$ be the number of distinct negative eigenvalues μ_i. It follows that system (9) has at most $2k + 1$ solutions: at most two in each subinterval $(0, -\mu_k)$, $(-\mu_k, -\mu_{k-1}), \ldots, (-\mu_2, -\mu_1)$, and at most one in the subinterval $(-\mu_1, +\infty)$. Taking into account the case $\lambda = 0$, we conclude that there exist at most $2(k + 1)$ distinct values of λ satisfying stationary conditions (2)–(3). □

From Lemma 1 and Proposition 1, we easily get the following corollary, establishing a bound on the maximum number of distinct values assumed by the objective function $m(s)$ at stationary points.

Corollary 1 *The maximum number of distinct values of the objective function $m(s)$ at stationary points is $2(k + 1)$, where k is the number of distinct negative eigenvalues of Q.*

At least from a theoretical point of view, Theorem 2 and Corollary 1 suggest a possible iterative strategy to obtain a global minimizer of problem (1). Namely, we can compute a stationary point $\bar{s}$ by some local algorithm and check the conditions of Theorem 2: if none of them is satisfied, then $\bar{s}$ is a global minimizer (see Remark 1); otherwise, we get a new point $\hat{s}$ such that $m(\hat{s}) < m(\bar{s})$ and, starting from $\hat{s}$, we can compute a new stationary point and iterate. Corollary 1 ensures that this procedure is finite and returns a global minimizer of problem (1).

To be rigorous, the above strategy is well defined under the assumption that stationary points can be computed in a finite number of iterations by a local algorithm. Unfortunately, optimization methods only ensure asymptotic convergence and, in practice, a point $\bar{s}$ is returned such that $\|\nabla m(\bar{s})\| \leq \varepsilon$, being ε a desired tolerance. In the next section, we show how Theorem 2 can be generalized to cope with this case and discuss possible algorithmic applications.

3 Extension to approximate stationary points

In this section, first we extend Theorem 2 to the case where stationary conditions are approximately satisfied, and then we briefly discuss how these results may be used in an algorithmic framework, showing some numerical examples.

Assuming that $\bar{s} \in \mathbb{R}^n$ is a non-stationary point of problem (1), of course we have $\|\nabla m(\bar{s})\| > 0$, or equivalently, $|\nabla m(\bar{s})^T d| > 0$ for some $d \in \mathbb{R}^n$. The next theorem states some conditions to compute a point $\hat{s}$ such that $m(\hat{s}) < m(\bar{s})$.

Theorem 3 *Given $\bar{s} \in \mathbb{R}^n$, let us define the point $\hat{s}$ as follows:*

(a) *if $c^T \bar{s} > 0$, then*

$$\hat{s} := -\bar{s};$$

(b) *if $c^T \bar{s} \leq 0$ and a vector $d \in \mathbb{R}^n$ exists such that $d^T (Q + \sigma \|\bar{s}\| I) d < -\varepsilon_2 \|d\|^2$,*

(i) *if $\bar{s} = 0$ and $\varepsilon_2 \geq 0$, then, assuming without loss of generality that $c^T d \leq 0$,*

$$\hat{s} := \bar{s} + \alpha d,$$

with $0 < \alpha < -\dfrac{3\, d^T Q d}{2\, \sigma \|d\|^3}$;

(ii) *if $\bar{s} \neq 0$, $\bar{s}^T d \neq 0$ and $\varepsilon_2 \geq \left| \dfrac{\nabla m(\bar{s})^T d}{\bar{s}^T d} \right|$, then*

$$\hat{s} := \bar{s} - 2 \frac{\bar{s}^T d}{\|d\|^2} d;$$

(iii) *if $\bar{s} \neq 0$, $\bar{s}^T d = 0$ and $\varepsilon_2 > \dfrac{|\nabla m(\bar{s})^T \bar{s}|}{\|\bar{s}\|^2}$, then, assuming without loss of generality that $\nabla m(\bar{s})^T d \geq 0$,*

$$\hat{s} := \bar{s} - 2 \frac{\bar{s}^T z}{\|z\|^2} z,$$

where $z := \bar{s} + \alpha d$ and $\alpha > 0$ is sufficiently large to satisfy

$$z^T (Q + \sigma \|\bar{s}\| I) z < -\varepsilon_2 \|z\|^2.$$

We have that

$$m(\hat{s}) < m(\bar{s}).$$

Proof The proof of case (a) is the same as for Theorem 2. Now, we consider case (b) and distinguish the three subcases.

(i) Since we are assuming that $c^T d \leq 0$, we can write

$$\begin{aligned} m(\bar{s} + \alpha d) = m(\alpha d) &= \alpha c^T d + \frac{1}{2}\alpha^2 d^T Q d + \frac{1}{3}\sigma\alpha^3 \|d\|^3 \\ &\leq \frac{1}{2}\alpha^2 d^T Q d + \frac{1}{3}\sigma\alpha^3 \|d\|^3 \end{aligned}$$

and we obtain the result by the same arguments used in the proof of point (b)-(i) of Theorem 2.

(ii) Using (6), and exploiting the fact that $d^T(Q + \sigma\|\bar{s}\|I)d < -\varepsilon_2\|d\|^2$, we get

$$\begin{aligned} m\left(\bar{s} - 2\frac{\bar{s}^T d}{\|d\|^2} d\right) &< m(\bar{s}) - \frac{1}{2}\left(2\frac{\bar{s}^T d}{\|d\|^2}\right)^2 \varepsilon_2 \|d\|^2 - 2\frac{\bar{s}^T d}{\|d\|^2}\nabla m(\bar{s})^T d \\ &\leq m(\bar{s}) - \frac{1}{2}\left(2\frac{\bar{s}^T d}{\|d\|^2}\right)^2 \varepsilon_2 \|d\|^2 + 2\frac{\left|\bar{s}^T d\right|}{\|d\|^2}\left|\nabla m(\bar{s})^T d\right| \\ &= m(\bar{s}) - 2\frac{\left|\bar{s}^T d\right|}{\|d\|^2}\left(\left|\bar{s}^T d\right|\varepsilon_2 - \left|\nabla m(\bar{s})^T d\right|\right) \leq m(\bar{s}), \end{aligned}$$

where the last inequality follows from the fact that $\varepsilon_2 \geq \left|\dfrac{\nabla m(\bar{s})^T d}{\bar{s}^T d}\right|$.

(iii) Since $d \neq 0$, we can first assume that $\alpha > 0$ is sufficiently large to satisfy $z \neq 0$. Replacing d with z in (6), we obtain

$$m\left(\bar{s} - 2\frac{\bar{s}^T z}{\|z\|^2} z\right) = m(\bar{s}) + \frac{1}{2}\left(2\frac{\bar{s}^T z}{\|z\|^2}\right)^2 z^T(Q + \sigma\|\bar{s}\|I)z - 2\frac{\bar{s}^T z}{\|z\|^2}\nabla m(\bar{s})^T z.$$

Taking into account that $z = \bar{s} + \alpha d$ and $\bar{s}^T z = \bar{s}^T(\bar{s} + \alpha d) = \|\bar{s}\|^2$, we can write

$$\begin{aligned} m\left(\bar{s} - 2\frac{\bar{s}^T z}{\|z\|^2} z\right) &= m(\bar{s}) + \frac{1}{2}\left(2\frac{\|\bar{s}\|^2}{\|z\|^2}\right)^2 z^T(Q + \sigma\|\bar{s}\|I)z - 2\frac{\|\bar{s}\|^2}{\|z\|^2}\nabla m(\bar{s})^T(\bar{s} + \alpha d) \\ &\leq m(\bar{s}) + \frac{1}{2}\left(2\frac{\|\bar{s}\|^2}{\|z\|^2}\right)^2 z^T(Q + \sigma\|\bar{s}\|I)z - 2\frac{\|\bar{s}\|^2}{\|z\|^2}\nabla m(\bar{s})^T \bar{s} \\ &\leq m(\bar{s}) + 2\frac{\|\bar{s}\|^2}{\|z\|^2}\left(\|\bar{s}\|^2 \frac{z^T(Q + \sigma\|\bar{s}\|I)z}{\|z\|^2} + \left|\nabla m(\bar{s})^T \bar{s}\right|\right), \end{aligned} \tag{10}$$

where the first inequality follows from the fact that $\nabla m(\bar{s})^T d \geq 0$ and $\alpha > 0$. Now, let us define $\theta \in (0, 1)$ such that $\varepsilon_2 = \dfrac{1}{\theta}\dfrac{|\nabla m(\bar{s})^T \bar{s}|}{\|\bar{s}\|^2}$. Exploiting the fact that $\theta \in (0, 1)$ and $d^T(Q + \sigma\|\bar{s}\|I)d < -\varepsilon_2\|d\|^2$, for sufficiently large $\alpha > 0$ we have

$$\frac{\left(\frac{\bar{s}}{\alpha}+d\right)^T(Q+\sigma\|\bar{s}\|I)\left(\frac{\bar{s}}{\alpha}+d\right)}{\frac{\|\bar{s}\|^2}{\alpha^2}+\|d\|^2}=\frac{(\bar{s}+\alpha d)^T(Q+\sigma\|\bar{s}\|I)(\bar{s}+\alpha d)}{\|\bar{s}\|^2+\alpha^2\|d\|^2}<-\theta\varepsilon_2.$$

Taking into account that $z=\bar{s}+\alpha d$ and $\|z\|^2=\|\bar{s}\|^2+\alpha^2\|d\|^2$, it follows that, for sufficiently large $\alpha>0$,

$$\frac{z^T(Q+\sigma\|\bar{s}\|I)z}{\|z\|^2}<-\theta\varepsilon_2.$$

Combining this inequality with (10), for sufficiently large $\alpha>0$ we can write

$$m\left(\bar{s}-2\frac{\bar{s}^Tz}{\|z\|^2}z\right)<m(\bar{s})+2\frac{\|\bar{s}\|^2}{\|z\|^2}\left(-\theta\varepsilon_2\|\bar{s}\|^2+\left|\nabla m(\bar{s})^T\bar{s}\right|\right)=m(\bar{s}),$$

where the equality follows from the fact that $\varepsilon_2=\frac{1}{\theta}\frac{|\nabla m(\bar{s})^T\bar{s}|}{\|\bar{s}\|^2}$. □

Remark 2 It is straightforward to verify that, when $\bar{s}$ is a stationary point, Theorem 3 coincides with Theorem 2.

Remark 3 Using (6), Theorem 3 can be strengthened by replacing the condition b-(ii) with the condition that a direction d exists such that $\bar{s}\neq 0$, $\bar{s}^Td\neq 0$ and

$$\frac{1}{2}\left(2\frac{\bar{s}^Td}{\|d\|^2}\right)^2d^T(Q+\sigma\|\bar{s}\|I)d-2\frac{\bar{s}^Td}{\|d\|^2}\nabla m(\bar{s})^Td<0.$$

Remark 4 From a computational point of view, condition (a) of Theorem 3 can be easily checked with a negligible cost. To check condition (b), we have to verify if there exists a negative curvature direction with respect to the matrix $(Q+\sigma\|\bar{s}\|I)$. This can be done, for example, by calculating the smallest eigenvalue and the associate eigenvector of that matrix. If such a direction exists, we see that, for case (b)-(i), this is enough to ensure that $m(\hat{s})<m(\bar{s})$. For case (b)-(ii) and (b)-(iii), we have to check if ε_2 is sufficiently large. It is easy to verify that, if $\|\nabla m(\bar{s})\|\leq\varepsilon$, then condition (b)-(ii) is verified whenever $\varepsilon_2\geq\varepsilon\|d\|/|\bar{s}^Td|$, and condition (b)-(iii) is verified whenever $\varepsilon_2>\varepsilon/\|\bar{s}\|$. Therefore, the threshold value of ε_2 for satisfying conditions b-(ii) and b-(iii) is related to $\|\nabla m(\bar{s})\|$, that is, the tolerance we have chosen to solve problem (1).

Let us concluding this section by discussing some possible algorithmic applications of our results, even if defining a proper optimization method is beyond the scope of the paper. A first naive strategy to exploit Theorem 3 is checking if one of its conditions holds after that an approximate stationary point $\bar{s}$ of problem (1) is computed with the desired tolerance by a local algorithm. If this is the case, then we can compute the point $\hat{s}$ and restart the local algorithm from $\hat{s}$. To provide some numerical examples, we have inserted this strategy within the ARC algorithm described in [6,7] to minimize the cubic model at each iteration, giving rise to an algorithm that we name ARC$^+$. In particular, at every iteration of ARC$^+$ and ARC, a truncated-Newton method has been used as local solver for the minimization of the cubic model, starting from a randomly chosen point. The codes have been written in Matlab, using built-in functions to compute eigenvalues and eigenvectors needed to check the conditions of Theorem 3. We have considered a set of 130 unconstrained test problems of the form $\min_{x \in \mathbb{R}^n} \ f(x)$ from the CUTEst collection [13] and, among them, we have then selected the 39 for which the two algorithms performed differently and both converged to a point x^* such that $\|\nabla f(x^*)\|_\infty \leq 10^{-5}$ within a maximum number of iterations, set equal to 10^5. The results on this subset of problems are reported in Table 1, where *obj* and *iter* denote the final objective value and the number of iterations, respectively. We see that, in 28 out 39 cases, ARC$^+$ converged in fewer iterations. Taking a look to the performance profile [10] reported in Fig. 1, we also observe that, on the considered subset of problems, ARC$^+$ is more robust than ARC in terms of number of iterations. We have then repeated the same experiments by using the Cauchy point as starting point for the minimization of the cubic model, but no significative difference emerged between ARC$^+$ and ARC. This opens a question about possible relations between the Cauchy point and the global minimizers, which can be subject of future research.

It is worth pointing out that the above described ARC$^+$ method could be too expensive in terms of CPU time, since it requires the computation of eigenvalues and eigenvectors at the end of each local minimization. Nevertheless, a more refined way to exploit Theorem 3 for algorithmic purposes can be based on checking if one of its conditions is satisfied during the iterations of the local method, instead of at the end. This can be done efficiently when the local method is able to detect negative curvature directions. Assuming that a sequence of points $\{s^k\}$ and a sequence of directions $\{d^k\}$ are produced by the local algorithm, since $\nabla^2 m(s^k) = Q + \sigma \|s^k\| I + \sigma \dfrac{s^k (s^k)^T}{\|s^k\|}$, we have $(d^k)^T (Q + \sigma \|s^k\| I) d^k = (d^k)^T \nabla^2 m(s^k) d^k - \sigma \dfrac{((s^k)^T d^k)^2}{\|s^k\|}$. Therefore, if d^k is a negative curvature direction with respect to $\nabla^2 m(s^k)$, condition (b) of Theorem 3 is verified for some $\varepsilon_2 \geq 0$, provided $c^T \bar{s} \leq 0$. Then, a new point that ensures a decrease in the objective function may be easily computed. In this case, condition (b) of Theorem 3 can therefore be checked without the need of computing eigenvalues and eigenvectors. Finally, other checks can be included in the scheme to ensure convergence of such modification of the local algorithm.

4 Conclusions

In this paper, we have highlighted some theoretical properties of the stationary points of problem (1), whose solutions are of interest for many optimization methods. We have shown that, given a stationary point of problem (1) that is not a global minimizer, it is possible to compute, in closed form, a new point that reduces the objective function value. Then, we have pointed out how a global minimum point of problem (1) can be obtained by computing at most $2(k+1)$ stationary points, where k is the number of

Table 1 Numerical results of ARC$^+$ and ARC on CUTEst problems. ARC$^+$ differs from ARC in that a globalization strategy, outlined in Theorem 3, is used to minimize the cubic model at each iteration

Problem	n	ARC$^+$		ARC	
		obj	iter	obj	iter
BROWNAL	200	1.00e−07	**103**	1.00e−07	472
BROWNBS	2	7.40e−12	**27552**	0.00e+00	27560
CURLY10	100	− 1.00e+04	**82**	− 1.00e+04	281
CURLY20	100	− 1.00e+04	**53**	− 1.00e+04	288
CURLY30	100	− 1.00e+04	**39**	− 1.00e+04	590
DECONVU	63	9.10e−07	**162**	8.52e−07	167
DENSCHND	3	2.63e−07	**2154**	2.82e−07	2293
DIXMAANH	300	1.00e+00	424	1.00e+00	**423**
DIXMAANJ	300	1.00e+00	4762	1.00e+00	**4739**
DIXMAANK	300	1.00e+00	5335	1.00e+00	**5265**
DIXMAANL	300	1.00e+00	5008	1.00e+00	**4941**
EIGENCLS	462	4.70e−09	**254**	4.37e−09	258
ENGVAL2	3	8.49e−16	**30**	2.04e−20	50
FLETCHBV	10	− 2.04e+06	551	− 2.09e+06	**460**
GENHUMPS	10	4.49e−12	**8968**	2.77e−11	9283
GENROSE	100	1.00e+00	**119**	1.00e+00	120
GENROSEB	500	1.00e+00	**505**	1.00e+00	511
GROWTHLS	3	1.00e+00	**271**	1.00e+00	4557
GULF	3	3.51e−06	4642	3.51e−06	**4640**
HAIRY	2	2.00e+01	**108**	2.00e+01	158
HEART8LS	8	4.91e−12	**86**	6.97e−17	130
HUMPS	2	1.91e−10	**1611**	8.40e−11	1858
JENSMP	2	1.24e+02	**28**	1.24e+02	47
LIARWHD	100	1.39e−19	**12**	2.97e−20	14
LOGHAIRY	2	1.82e−01	**5177**	1.82e−01	5316
MEXHAT	2	− 4.00e−02	523	− 4.00e−02	**68**
NONCVXU2	100	2.33e+02	**571**	2.33e+02	572
NONDIA	100	1.57e−18	**7**	9.66e−26	9
OSCIPATH	10	1.00e+00	39	1.00e+00	**22**
PALMER6C	8	1.64e−02	21678	1.64e−02	**17418**
PALMER7C	8	6.02e−01	31863	6.02e−01	**24683**
PALMER8C	8	1.60e−01	33434	1.60e−01	**14945**
PARKCH	15	1.62e+03	**65**	1.62e+03	250
PFIT1LS	3	2.10e−10	**501**	4.75e−04	2810
SINEVAL	2	2.13e−17	**101**	5.40e−12	137
SPARSINE	100	1.83e−14	**38**	1.13e−10	39
SROSENBR	100	4.02e−14	**12**	2.13e−17	500
VARDIM	200	6.90e−31	**36**	7.29e−27	37
WATSON	12	3.57e−06	**82**	2.84e−06	91

For each problem, the smallest number of iterations is highlighted in bold

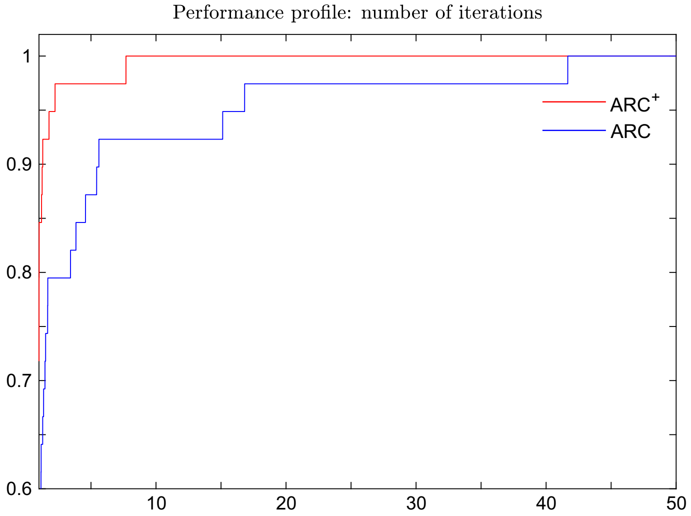

Fig. 1 Performance profile for the number of iterations related to the numerical experiments reported in Table 1

distinct negative eigenvalues of the matrix Q. Further, we have extended these results to the case where stationary conditions are approximately satisfied, sketching some possible algorithmic applications.

We think that the most natural extension of the results presented in this paper is the definition of a proper algorithm for unconstrained optimization, based on the iterative computation of the solutions of problem (1), for which some preliminary ideas have been proposed at the end of Sect. 3. This can be a challenging task for future research.

References

1. Bellavia, S., Morini, B.: Strong local convergence properties of adaptive regularized methods for nonlinear least squares. IMA J. Numer. Anal. **35**(2), 947–968 (2014)
2. Benson, H.Y., Shanno, D.F.: Interior-point methods for nonconvex nonlinear programming: cubic regularization. Comput. Optim. Appl. **58**(2), 323–346 (2014)
3. Bianconcini, T., Sciandrone, M.: A cubic regularization algorithm for unconstrained optimization using line search and nonmonotone techniques. Optim. Methods Softw. **31**(5), 1008–1035 (2016)
4. Bianconcini, T., Liuzzi, G., Morini, B., Sciandrone, M.: On the use of iterative methods in cubic regularization for unconstrained optimization. Comput. Optim. Appl. **60**(1), 35–57 (2015)
5. Birgin, E.G., Gardenghi, J.L., Martínez, J.M., Santos, S.A., Toint, P.L.: Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. Math. Program. **163**(1–2), 359–368 (2017)
6. Cartis, C., Gould, N.I.M., Toint, P.L.: Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. Math. Program. **127**(2), 245–295 (2011)
7. Cartis, C., Gould, N.I.M., Toint, P.L.: Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function-and derivative-evaluation complexity. Math. Program. **130**(2), 295–319 (2011)
8. Cartis, C., Gould, N.I.M., Toint, P.L.: An adaptive cubic regularization algorithm for nonconvex optimization with convex constraints and its function-evaluation complexity. IMA J. Numer. Anal. **32**(4), 1662–1695 (2012)

9. Conn, A.R., Gould, N.I.M., Toint, P.L.: Trust Region Methods. SIAM, Philadelphia (2000)
10. Dolan, E.D., Moré, J.J.: Benchmarking optimization software with performance profiles. Math. Program. **91**(2), 201–213 (2002)
11. Dussault, J.P.: Simple unified convergence proofs for the trust-region and a new ARC variant. Tech. rep., University of Sherbrooke, Sherbrooke, Canada (2015)
12. Gould, N.I.M., Porcelli, M., Toint, P.L.: Updating the regularization parameter in the adaptive cubic regularization algorithm. Comput. Optim. Appl. **53**(1), 1–22 (2012)
13. Gould, N.I.M., Orban, D., Toint, P.L.: CUTEst: a constrained and unconstrained testing environment with safe threads for mathematical optimization. Comput. Optim. Appl. **60**(3), 545–557 (2015)
14. Griewank, A.: The modification of Newtons method for unconstrained optimization by bounding cubic terms. Tech. Rep. NA/12 (1981)
15. Lucidi, S., Palagi, L., Roma, M.: On some properties of quadratic programs with a convex quadratic constraint. SIAM J. Optim. **8**(1), 105–122 (1998)
16. Nesterov, Y.: Cubic regularization of Newton's method for convex problems with constraints. Tech. Rep. 39, CORE (2006)
17. Nesterov, Y.: Accelerating the cubic regularization of Newtons method on convex problems. Math. Program. **112**(1), 159–181 (2008)
18. Nesterov, Y., Polyak, B.T.: Cubic regularization of Newton method and its global performance. Math. Program. **108**(1), 177–205 (2006)
19. Weiser, M., Deuflhard, P., Erdmann, B.: Affine conjugate adaptive Newton methods for nonlinear elastomechanics. Optim. Methods Softw. **22**(3), 413–431 (2007)