Computational Optimization and Applications manuscript No.
(will be inserted by the editor)

A globally convergent Levenberg–Marquardt method for equality-constrained optimization

A. F. Izmailov · M. V. Solodov · E. I. Uskov

Received: date / Accepted: date

Abstract It is well-known that the Levenberg–Marquardt method is a good choice for solving nonlinear equations, especially in the cases of singular/non-isolated solutions. We first exhibit some numerical experiments with local convergence, showing that this method for "generic" equations actually also works very well when applied to the specific case of the Lagrange optimality system, i.e., to the equation given by the first-order optimality conditions for equality-constrained optimization. In particular, it appears to outperform not only the basic Newton method applied to such systems, but also its modifications supplied with dual stabilization mechanisms, intended specially for tackling problems with nonunique Lagrange multipliers. The usual globalizations of the Levenberg–Marquardt method are based on linesearch for the squared Euclidean residual of the equation being solved. In the case of the Lagrange optimality system, this residual does not involve the objective function of the underlying optimization problem (only its derivative), and in particular, the resulting globalization scheme has no preference for converging to minima versus maxima, or to any other stationary point. We thus develop a special globalization of the Levenberg–Marquardt method when it is applied to the Lagrange optimality system, based on linesearch for a smooth exact penalty

A. F. Izmailov
Lomonosov Moscow State University, MSU, Uchebniy Korpus 2, VMK Faculty, OR Department, Leninskiye Gory, 119991 Moscow, Russia.
RUDN University, Miklukho-Maklaya Str. 6, 117198 Moscow, Russia.
E-mail: izmaf@ccas.ru

M. V. Solodov
IMPA – Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil
E-mail: solodov@impa.br

E. I. Uskov
Department of Mathematics, Physics and Computer Sciences, Derzhavin Tambov State University, TSU, Internationalnaya 33, 392000 Tambov, Russia.
E-mail: euskov@cs.msu.ru

function of the optimization problem, which in particular involves the objective function of the problem. The algorithm is shown to have appropriate global convergence properties, preserving also fast local convergence rate under weak assumptions.

Keywords Newton-type methods · Levenberg–Marquardt method · stabilized sequential quadratic programming · local convergence · global convergence · penalty function

Mathematics Subject Classification (2000) 65K05 · 65K15 · 90C30

1 Introduction

Consider first a system of general nonlinear equations without any special structure:

$$\Phi(z) = 0, \tag{1}$$

where $\Phi : \mathbb{R}^p \to \mathbb{R}^p$ is sufficiently smooth. One classical approach to solving this problem is the Levenberg–Marquardt method; see, e.g., [19, Chapter 10]. It is especially important when solutions of (1) are nonisolated, or singular in the sense that the Jacobian of Φ is singular at solutions (note that nonisolated solutions are automatically singular). Given the current iterate $z^k \in \mathbb{R}^p$ and a regularization parameter $\sigma_k > 0$, the next iterate of the Levenberg–Marquardt method is $z^{k+1} = z^k + d^k$ where $d^k \in \mathbb{R}^p$ is the (unique) solution of the following linear system in d:

$$\left((\Phi'(z^k))^\top \Phi'(z^k) + \sigma_k I\right) d = -(\Phi'(z^k))^\top \Phi(z^k). \tag{2}$$

With appropriate control of the parameter σ_k, this method can have fast local convergence despite singularity of solutions [6, 23]. In fact, with some suitable modifications, the Levenberg–Marquardt methods can even handle both nonisolated/singular solutions and nonsmoothness of the mapping Φ [5,9].

Consider now the equality-constrained optimization problem

$$\begin{array}{l} \text{minimize} \quad f(x) \\ \text{subject to } h(x) = 0, \end{array} \tag{3}$$

where the objective function $f : \mathbb{R}^n \to \mathbb{R}$ and the constraints mapping $h : \mathbb{R}^n \to \mathbb{R}^l$ are at least twice differentiable.

Let $L : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}$ be the Lagrangian of problem (3), i.e.,

$$L(x, \lambda) = f(x) + \langle \lambda, h(x) \rangle.$$

Then the Lagrange optimality system of problem (3), characterizing its stationary points and associated Lagrange multipliers, has the form

$$\frac{\partial L}{\partial x}(x, \lambda) = 0, \quad h(x) = 0. \tag{4}$$

In this work, we are specially interested in cases when the set

$$\Lambda(\bar{x}) = \left\{ \lambda \in \mathbb{R}^l \;\middle|\; \frac{\partial L}{\partial x}(\bar{x},\, \lambda) = 0 \right\}$$

of Lagrange multipliers associated to a stationary point $\bar{x}$ of (3) is not necessarily a singleton. In particular, the system (4) may have nonisolated solutions.

Denote the mapping of the Lagrange optimality system (4) by $\Phi : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}^n \times \mathbb{R}^l$, i.e.,

$$\Phi(x,\, \lambda) = \left(\frac{\partial L}{\partial x}(x,\, \lambda),\, h(x) \right). \tag{5}$$

Note that Φ is the full gradient mapping (with respect to primal and dual variables) of the Lagrangian L. Given the current iterate $(x^k,\, \lambda^k) \in \mathbb{R}^n \times \mathbb{R}^l$, the Levenberg–Marquardt method (2) for this Φ (taking into account that its Jacobian matrix $\Phi'(x^k,\, \lambda^k)$ is symmetric in this case), defines the next iterate as $(x^{k+1},\, \lambda^{k+1}) = (x^k + \xi^k,\, \lambda^k + \eta^k)$, where $d^k = (\xi^k,\, \eta^k)$ is the unique solution of the linear system

$$\Phi'(x^k,\, \lambda^k)\Phi(x^k,\, \lambda^k) + \left((\Phi'(x^k,\, \lambda^k))^2 + \sigma_k I \right) d = 0 \tag{6}$$

with respect to $d = (\xi,\, \eta)$. As usual in Levenberg–Marquardt methods, the regularization parameter $\sigma_k > 0$ will be given by some power of the Lagrange system residual $\|\Phi(x^k,\, \lambda^k)\|$.

While the Levenberg–Marquardt method is applicable to the Lagrange optimality system (4) (in fact, just like to any other system of equations), note that it is a general-purpose algorithm which does not take into account the special primal-dual structure of (4). Despite of that, perhaps somewhat surprisingly, we found out that the Levenberg–Marquardt method nevertheless performs quite well for the Lagrange system (4), even compared to special techniques designed to tackle degeneracy in optimization problems. By degeneracy, here we mean violation of the standard constraints regularity condition

$$\operatorname{rank} h'(\bar{x}) = l. \tag{7}$$

The special methods in question for degenerate optimization problems are those of stabilized sequential quadratic programming (sSQP) [8,11,13,21,22]; see also [15, Chapter 7]; and the more recent subspace-sSQP variations [18]. The statements of these methods and the related numerical results comparing the local behavour of all the options in consideration, are reported in Section 2.

Motivated by the observed good local behaviour of the Levenberg–Marquardt method applied to the Lagrange system (4), the issue of globalizing its convergence arose naturally. We note that some approaches to globalization of sSQP for optimization have been proposed in [7,10,16,17]. However, in our view, the issue of computationally efficient, both rigorously globally and fast locally convergent algorithms for degenerate optimization problems, is not resolved to full satisfaction as of yet. Thus, looking at alternatives such as the Levenberg–Marquardt method is still warranted, especially taking into account

good local performance of the latter described in Section 2. Another appealing feature of the Levenberg–Marquardt method is that the corresponding direction exists always. This is in contrast to the usual Newton method (sequential quadratic programming, SQP), as well as sSQP and subspace-sSQP, where in principle subproblems are guaranteed to be solvable only locally, under suitable assumptions. In fact, solvability of subproblems is the first difficulty one encounters when trying to design globally convergent schemes based on those approaches. For the Levenberg–Marquardt method, solvability of subproblems is not an issue.

On the other hand, recall that usual globalizations of the Levenberg–Marquardt method (2) for the equation (1) would proceed, after computing the direction d^k, with linesearch to decrease the squared Euclidian residual $\|\Phi(\cdot)\|^2$ of the equation; see, e.g., [6, 23]. We emphasize that in the context of the Lagrange system (4), when Φ given by (5) comes from the optimization problem (3), this is (at least) not ideal. Indeed, the Lagrange system characterizes both minima and maxima of f on the set defined by $h(x) = 0$. Thus, if one merely aims at minimizing the residual of this system, such a scheme would not have any particular preference for minima versus maxima (or any other stationary point), i.e., for solving the optimization problem (3). This calls for some other approach tailored specifically to optimization. In other words, linesearch should better be performed for some *penalty function*, which involves the objective function of the optimization problem, and is thus at least not indifferent to minimization versus maximization. To this end, we shall employ the smooth primal-dual two-parameter penalty function for problem (3), first proposed in [3]. Specifically, this function is $\varphi_{c_1,\,c_2} : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}$,

$$\varphi_{c_1,\,c_2}(x,\,\lambda) = L(x,\,\lambda) + \frac{c_1}{2}\|h(x)\|^2 + \frac{c_2}{2}\left\|\frac{\partial L}{\partial x}(x,\,\lambda)\right\|^2, \tag{8}$$

where $c_1 > 0$ and $c_2 > 0$ are penalty parameters. Observe that the objective function f enters (8) in an additive way, through the Lagrangian L. According to [1, Section 4.3] and [2], but without going into full technical details, the properties of the penalty function in question can be summarized as follows. Observe first that for any c_1 and c_2, any solution of the Lagrange optimality system (4) is evidently a stationary point of the unconstrained optimization problem

$$\text{minimize } \varphi_{c_1,\,c_2}(x,\,\lambda), \quad (x,\,\lambda) \in \mathbb{R}^n \times \mathbb{R}^l. \tag{9}$$

Conversely, if $c_2 > 0$ is small enough and $c_1 > 0$ is large enough, then any stationary point $(\bar{x},\,\bar{\lambda})$ of problem (9), satisfying the constraints regularity condition (7), is also a solution of the Lagrange optimality system (4). Moreover, if a solution of $(\bar{x},\,\bar{\lambda})$ of (4) satisfies (7) and the second-order sufficient optimality condition (SOSC)

$$\left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x},\,\bar{\lambda})\xi,\,\xi \right\rangle > 0 \quad \forall\, \xi \in \ker h'(\bar{x}) \setminus \{0\} \tag{10}$$

for problem (3), then $(\bar{x}, \bar{\lambda})$ is a strict local minimizer in (9), and problem (9) has no other stationary points near $(\bar{x}, \bar{\lambda})$. Finally, if a solution of $(\bar{x}, \bar{\lambda})$ of (4) violates the second-order necessary optimality condition

$$\left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\lambda})\xi, \xi \right\rangle \geq 0 \quad \forall\, \xi \in \ker h'(\bar{x})$$

for problem (3), then $(\bar{x}, \bar{\lambda})$ cannot be a local minimizer in (9).

In addition, as we shall show in Section 3, the Levenberg–Marquardt direction given by (6) is *always* a descent direction for the penalty function φ_{c_1, c_2}, for an appropriate explicit choice of the parameters c_1 and c_2. This property would be the key for designing a globally convergent algorithm based on the Levenberg–Marquardt subproblems, presented in Section 3. Global convergence properties of the algorithm are obtained in Section 4. In Section 5 we show that local superlinear convergence rate of the basic Levenberg–Marquardt method is preserved, under the same weak assumptions. Finally, in Section 6 we present some numerical results for the proposed globalized Levenberg–Marquardt method for optimization.

We conclude this section recalling some notions to be used in the sequel. We say that the local Lipschitzian error bound holds at a stationary point $\bar{x}$ of problem (3) for a Lagrange multiplier $\bar{\lambda} \in \Lambda(\bar{x})$, if

$$\|x - \bar{x}\| + \operatorname{dist}(\lambda, \Lambda(\bar{x})) = O(\|\Phi(x, \lambda)\|) \tag{11}$$

as $(x, \lambda) \to (\bar{x}, \bar{\lambda})$. Note that the error bound (11) holds if, and only if, the multiplier $\bar{\lambda}$ is *noncritical* [15, Definition 1.41]:

$$\nexists\, \xi \in \ker h'(\bar{x}) \setminus \{0\} \text{ such that } \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\lambda})\xi \in \operatorname{im}(h'(\bar{x}))^{\mathrm{T}}.$$

Otherwise $\bar{\lambda}$ is called *critical*. It can be easily seen that SOSC (10) is a stronger assumption than noncriticality; thus, (10) implies the error bound (11).

2 Local numerical behavior of the Levenberg–Marquardt method compared to some alternatives

To motivate the interest in the Levenberg–Marquardt method for degenerate optimization problems, we first provide some numerical results for the basic local versions of the following algorithms (specific iteration schemes are stated a bit further):

- SQP; e.g., [15, Chapter 4].
- sSQP, as described in [13, Section 2].
- Subspace-sSQP with asymptotically vanishing stabilization, as defined in [18], with all the parameters as specified there.
- The Levenberg–Marquardt method applied to the Lagrange optimality system (4), i.e., iterations given by (6), with the regularization parameters $\sigma_k = \|\Phi(x^k, \lambda^k)\|^2$, where Φ is defined in (5).

In Figures 1 and 2 below, these methods are represented by black, dark gray, light gray, and white color, respectively. Let us first recall the iteration systems of the methods in question, also discussing briefly the differences.

The iteration system of SQP has the form

$$\frac{\partial^2 L}{\partial x^2}(x^k,\, \lambda^k)\xi + (h'(x^k))^{\mathrm{T}}\eta = -\frac{\partial L}{\partial x}(x^k,\, \lambda^k), \quad h'(x^k)\xi = -h(x^k). \tag{12}$$

This is just the iteration of the Newton method applied to the Lagrange optimality system (4) of problem (3).

The iteration system of sSQP is given by

$$\frac{\partial^2 L}{\partial x^2}(x^k,\, \lambda^k)\xi + (h'(x^k))^{\mathrm{T}}\eta = -\frac{\partial L}{\partial x}(x^k,\, \lambda^k), \quad h'(x^k)\xi - \sigma_k\eta = -h(x^k). \tag{13}$$

It differs from (12) by an extra stabilizing term in the left-hand side of the second equation, involving the parameter $\sigma_k > 0$ (for explantions of the stabilizing affects, see [15, Chapter 7]).

The iteration system of subspace-sSQP has the form

$$\frac{\partial^2 L}{\partial x^2}(x^k,\, \lambda^k)\xi + (h'(x^k))^{\mathrm{T}}\eta = -\frac{\partial L}{\partial x}(x^k,\, \lambda^k), \quad h'(x^k)\xi - \sigma_k P_k\eta = -h(x^k),$$

where the difference with (13) is solely in the presence of an $l \times l$ matrix P_k which is supposed to approximate the projector onto some complement of the subspace $\operatorname{im} h'(\bar{x})$ at a stationary point $\bar{x}$, as x approaches $\bar{x}$, and λ approaches an associated Lagrange multiplier. Using this matrix makes stabilization "less aggressive", with the iteration system being closer to that of SQP in (12), but retaining all nice local convergence features of sSQP; see [18].

For comparisons, we used randomly generated *degenerate* problems with quadratic objective functions and quadratic equality constraints. The generator is the one described in [12], with entries of all arrays taking values in $[-100,\, 100]$. For each triple of integers $(n,\, l,\, r)$, we generated 100 problems with n variables and l constraints, such that $\bar{x} = 0$ is a stationary point of each problem, with the rank of constraints' Jacobian at $\bar{x}$ equal to r (with $r < l$). The generator also provides some Lagrange multiplier $\bar{\lambda}$ associated to $\bar{x}$. These problems are nonconvex, can be unbounded below, and may have stationary points other than the one of interest, namely other than $\bar{x} = 0$.

For each problem generated this way, we ran all the algorithms from 10 primal-dual starting points taken randomly from the box centered at the point $(\bar{x},\, \bar{\lambda})$, with the half of the edge equal to 0.1. Therefore, for each triple, we perform 1000 runs in total. The methods stop once the residual of the Lagrange optimality system becomes less than $1e-8$. If this happens in no more than 500 iterations, the run is counted as successful. Otherwise, failure is declared. Convergence to the primal solution of interest ($\bar{x} = 0$) is declared if the distance from primal iterate at termination to $\bar{x}$ is less than $1e-3$. In these local experiments, we used triples (n, l, r) with $n = 5,\, 10,\, 25,\, 20,\, 100$, all $l \in \{1,\, \ldots,\, n-1\}$, and all $r \in \{0,\, \ldots,\, l-1\}$. All computations here, and in Section 6 below, were performed in Matlab environment.

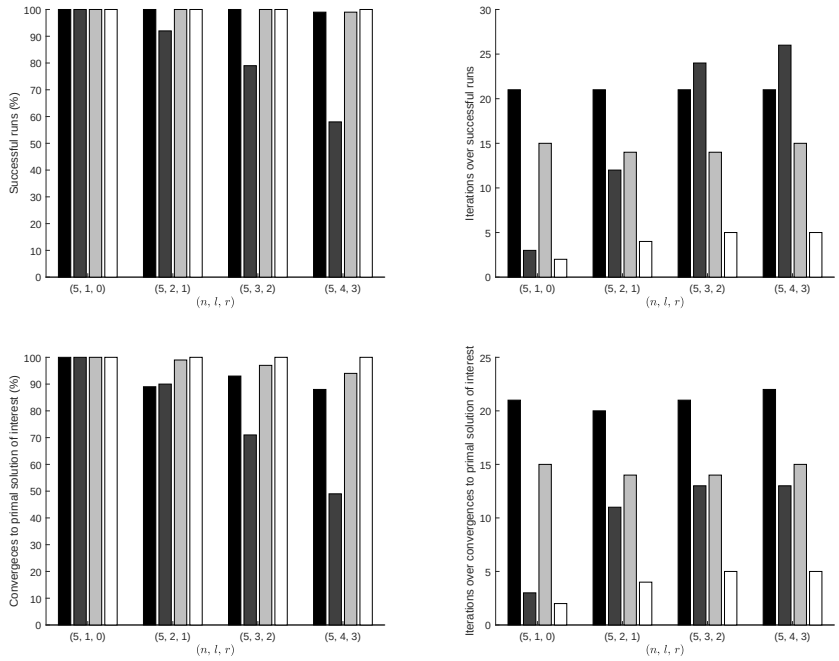


Fig. 1: Local behavior for $n = 5$.

The results for some representative triples (n, l, r) are given in Figures 1 and 2, in the form of bar diagrams. Iteration counts reported are averages over successful runs and over the cases of convergence to the primal point of interest, respectively, out of 1000 runs performed for each triple. One can observe that the Levenberg–Marquardt method is by far the more robust one, and is also more efficient. Also, unlike for the other methods, its robustness is almost not affected by the dimensions' growth. The difficulties experienced by SQP are explained by slow convergence caused by convergence of the dual sequence to critical multipliers; see [12,14] and [15, Section 7.1]. The behavior of sSQP is less deteriorated by this effect [14], but apart from the cases of full degeneracy (when $h'(\bar{x}) = 0$), this method has a tendency to generate long sequences of short steps before eventual superlinear convergence can be observed. This is a result of "over-stabilization"; see the related discussion in [17]. The subspace-sSQP variant, developed in [18] precisely with the intention to suppress this tendency, has better performance than sSQP but, nevertheless, it is still outperformed by the Levenberg–Marquardt method on this test set.

The conclusion from these experiments is that the Levenberg–Marquardt method for optimization, i.e., the method applied to the Lagrange optimality system, exhibits very good local behavior on degenerate problems. Therefore,

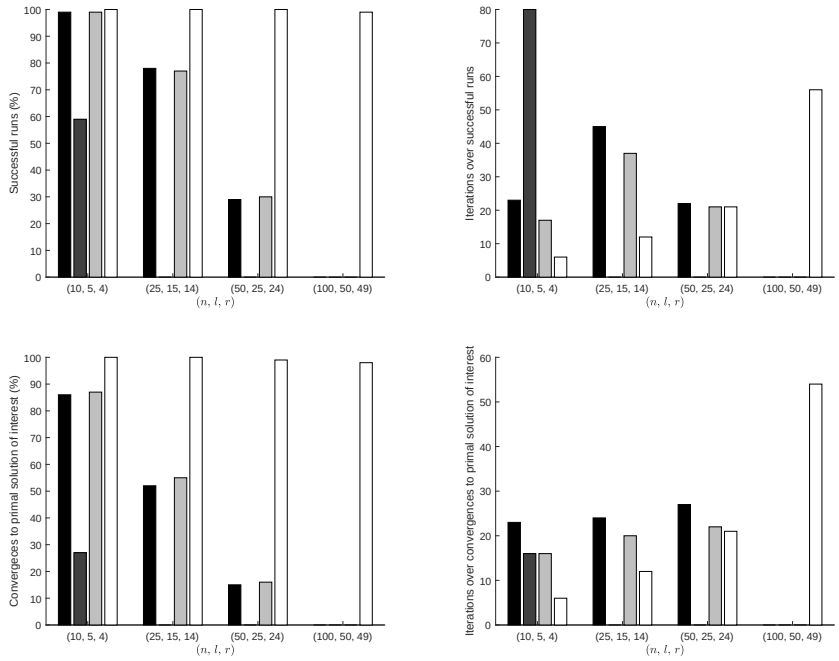


Fig. 2: Local behavior for $n \geq 10$.

developing on its basis an optimization-tailored globally convergent algorithm is of interest.

3 A globally convergent Levenberg–Marquardt algorithm for optimization

For the subsequent analysis, it will be convenient to define the following two mappings, which split the two parts of the Lagrange optimality system (4). Let $\Phi_1 : \mathbb{R}^n \to \mathbb{R}^l$ and $\Phi_2 : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}^n$ be given by

$$\Phi_1(x) = h(x), \quad \Phi_2(x, \lambda) = \frac{\partial L}{\partial x}(x, \lambda).$$

With these definitions, $\Phi(x, \lambda) = (\Phi_2(x, \lambda), \Phi_1(x))$, where the mapping Φ is defined in (5). Furthermore, the penalty function (8) can be written as

$$\varphi_{c_1, c_2}(x, \lambda) = L(x, \lambda) + \frac{c_1}{2}\|\Phi_1(x)\|^2 + \frac{c_2}{2}\|\Phi_2(x, \lambda)\|^2, \tag{14}$$

the form which we shall employ from now on.

We first show that there *always* exist values of the penalty parameters, for which the Levenberg–Marquardt direction given by (6) is a descent direction

for the penalty function (14), provided the direction is nonzero. Moreover, those penalty parameters are easily computable by explicit formulas.

By direct calculation of the gradient of the penalty function, for any point $(x^k, \lambda^k) \in \mathbb{R}^n \times \mathbb{R}^l$ and any direction $d^k = (\xi^k, \eta^k) \in \mathbb{R}^n \times \mathbb{R}^l$, we have that

$$\begin{aligned}\langle \varphi'_{c_1, c_2}(x^k, \lambda^k), d^k \rangle = \big\langle \Phi(x^k, \lambda^k), d^k \big\rangle \\ + c_1 \langle \Phi_1(x^k), \Phi'_1(x^k)\xi^k \rangle + c_2 \langle \Phi_2(x^k, \lambda^k), \Phi'_2(x^k, \lambda^k) d^k \rangle.\end{aligned} \tag{15}$$

Next, if d^k is obtained by solving the Levenberg–Marquardt equation (6), it holds that

$$\begin{aligned}\langle \Phi_1(x^k), \Phi'_1(x^k)\xi^k \rangle + \langle \Phi_2(x^k, \lambda^k), \Phi'_2(x^k, \lambda^k) d^k \rangle &= \langle \Phi(x^k, \lambda^k), \Phi'(x^k, \lambda^k) d^k \rangle \\ &= \langle \Phi'(x^k, \lambda^k) \Phi(x^k, \lambda^k), d^k \rangle \\ &= -\|\Phi'(x^k, \lambda^k) d^k\|^2 \\ &\quad -\sigma_k \|d^k\|^2.\end{aligned} \tag{16}$$

Observe now that the first term in the right-hand side of (15) does not depend on the penalty parameters. Hence, whatever its sign is, taking (for example) $c_1 = c_2 > 0$ sufficiently large, (16) shows that we can make the right-hand side of (15) negative (if $d^k \neq 0$). In fact, we can make it as negative as we wish, and the corresponding values of the penalty parameters are also explicitly computable from (15) and (16). We shall not give specifically those formulas here, because our proposal for choosing the penalty parameters would be a bit more sophisticated, for the following reason. According to the exact penalization properties of the penalty function (14), already mentioned in Section 1, it is desirable to keep c_2 small(er). Thus, by assessing all the terms in the right-hand side of (15), we shall first attempt to make it negative by increasing c_1 only (if this is necessary at all). We thus arrive at the following algorithm. Some further comments on its features would be provided immediately after its statement.

Algorithm 1 Choose parameters $\bar{\sigma} > 0$, $q > 0$, $\rho > 0$, $\tau > 1$, $\bar{c}_1 > 0$, $\bar{c}_2 > 0$, $\delta > 0$, ν, ε, $\theta \in (0, 1)$, and set $c_1 = \bar{c}_1$, $c_2 = \bar{c}_2$. Choose $(x^0, \lambda^0) \in \mathbb{R}^n \times \mathbb{R}^l$ and set $k = 0$.

1. If
$$\Phi'(x^k, \lambda^k) \Phi(x^k, \lambda^k) = 0, \tag{17}$$
where $\Phi(\cdot)$ is defined in (5), stop.
2. Set $\sigma_k = \min\{\bar{\sigma}, \|\Phi(x^k, \lambda^k)\|^q\}$. Compute $d^k = (\xi^k, \eta^k)$ as the unique solution of the Levenberg–Marquardt equation (6).
3. If
$$\langle \varphi'_{c_1, c_2}(x^k, \lambda^k), d^k \rangle \leq -\rho \|d^k\|^\tau, \tag{18}$$
go to step 4.

If

$$\langle \Phi_1(x^k),\, \Phi_1'(x^k)\xi^k\rangle \le -\nu\left(\|\Phi'(x^k,\,\lambda^k)d^k\|^2 + \sigma_k\|d^k\|^2\right), \tag{19}$$

define $\bar{c}_{1,\,k}$ as the minimal value of c_1 such that (18) holds, i.e.,

$$\bar{c}_{1,\,k} = -\,\frac{\langle \Phi(x^k,\,\lambda^k),\, d^k\rangle + c_2\langle \Phi_2(x^k,\,\lambda^k),\, \Phi_2'(x^k,\,\lambda^k)d^k\rangle + \rho\|d^k\|^\tau}{\langle \Phi_1(x^k),\, \Phi_1'(x^k)\xi^k\rangle}, \tag{20}$$

set $c_1 = \bar{c}_{1,\,k} + \delta$, and go to step 4.
Else, define $\bar{c}_{2,\,k}$ as the minimal value of c_2 such that (18) holds, i.e.,

$$\bar{c}_{2,\,k} = -\,\frac{\langle \Phi(x^k,\,\lambda^k),\, d^k\rangle + c_1\langle \Phi_1(x^k),\, \Phi_1'(x^k)\xi^k\rangle + \rho\|d^k\|^\tau}{\langle \Phi_2(x^k,\,\lambda^k),\, \Phi_2'(x^k,\,\lambda^k)d^k\rangle}, \tag{21}$$

and set $c_2 = \bar{c}_{2,\,k} + \delta$.

4. Compute $\alpha_k = \theta^j$, where j is the smallest nonnegative integer satisfying the Armijo inequality

$$\varphi_{c_1,\,c_2}((x^k,\,\lambda^k) + \theta^j d^k) \le \varphi_{c_1,\,c_2}(x^k,\,\lambda^k) + \varepsilon\theta^j\left\langle \varphi_{c_1,\,c_2}'(x^k,\,\lambda^k),\, d^k\right\rangle. \tag{22}$$

5. Set $(x^{k+1},\,\lambda^{k+1}) = (x^k,\,\lambda^k) + \alpha_k d^k$, increase k by 1, and go to step 1.

Some comments are in order. We first note that the stopping outcome (17), as well as the corresponding possibility for the global convergence in Theorem 1 below, mean stationarity for the squared residual of the Lagrange optimality system. While this is not what we are looking for (recalling our motivation for using a penalty function instead of the residual for globalization), such an outcome is clearly unavoidable as a possibility: if the starting point satisfies (17), then the Levenberg–Marquardt equation (6) shows that the zero direction is generated. However, while such a situation cannot be ruled out, the expectation is that it makes it way less likely to occur "in the limit" if linesearch minimizes the penalty function and not the residual.

Other observations concern the penalty parameters. Note that if the computed direction is of sufficient descent (i.e., (18) holds), no parameters are changed. If that is not the case but descent can be achieved by increasing c_1 only, i.e., (19) holds, then the value of c_1 is increased. The value of c_2 is increased only as the last option. Note also that at each iteration either the penalty parameters do not change, or one (and only one) of them increases, by a quantity no less than δ. The latter is because $\bar{c}_{1,\,k}$ and $\bar{c}_{2,\,k}$ cannot be smaller than the current values of c_1 and c_2, respectively, as otherwise there would be a contradiction with violation of (18). For this, recall the discussion following (15) and (16), which makes it clear that (18) always holds for the values of parameters large enough; hence, if the current values are not suitable, then they must increase. It follows that the sequences of penalty parameters are non-decreasing.

The final comment is that, since the Levenberg–Marquardt equation (6) is always solvable, and since the procedure to choose appropriate values of parameters that guarantee the descent property for the computed directions is well-defined (as explained above), the whole algorithm is well-defined at every

step. We emphasize, in particular, that no safeguarding descent directions (e.g., the gradient directions) are ever needed, as there are no "unfavorable" situations when the computed direction cannot be used. This is in contrast, for example, with the globalization strategy for sSQP in [17].

4 Global convergence properties

We are now in position to establish global convergence properties of Algorithm 1. In particular, we show that in any situation (regarding what happens with the sequences of the penalty parameters), a stationary point of the penalty function and/or of the squared residual of the Lagrange optimality system is obtained. Recall that the value of each parameter either stays constant from some iteration on, or it goes to infinity.

Theorem 1 *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ be twice continuously differentiable on $\mathbb{R}^n$.*

Then Algorithm 1 is well-defined, and either finitely terminates at a point (x^k, λ^k) satisfying (17), or generates an infinite sequence $\{(x^k, \lambda^k)\}$ possessing the following properties:

(i) If the values of c_1 and c_2 do not change for all k sufficiently large, then every accumulation point $(\bar{x}, \bar{\lambda})$ of $\{(x^k, \lambda^k)\}$ satisfies at least one of the equalities

$$\Phi'(\bar{x}, \bar{\lambda})\Phi(\bar{x}, \bar{\lambda}) = 0 \tag{23}$$

or

$$\varphi'_{c_1, c_2}(\bar{x}, \bar{\lambda}) = 0, \tag{24}$$

with those asymptotically constant values of c_1 and c_2.

(ii) If the value of c_2 is asymptotically constant and there exists an infinite set K of iteration indices such that c_1 changes at each iteration $k \in K$, then every accumulation point $(\bar{x}, \bar{\lambda})$ of $\{(x^k, \lambda^k) \mid k \in K\}$ satisfies (23).

(iii) Assertion (ii) is also valid if the value of c_1 does not change for all k sufficiently large, and there exists an infinite set K of iteration indices such that the value of c_2 changes at each iteration $k \in K$.

(iv) If the sequence $\{(x^k, \lambda^k)\}$ converges to some $(\bar{x}, \bar{\lambda})$, and the values of both c_1 and c_2 are changed infinitely many times, then $(\bar{x}, \bar{\lambda})$ satisfies (23).

Proof Let k be any iteration index, and let (17) not hold, so that the algorithm does not terminate. Then it holds that $\sigma_k > 0$ and, according to (6), $d^k \neq 0$.

It follows that if (19) is satisfied, then

$$\langle \Phi_1(x^k), \Phi'_1(x^k)\xi^k \rangle < 0, \tag{25}$$

and hence, by (15), the value of $\bar{c}_{1,k}$ in (20) is well-defined in this case.

On the other hand, if (19) is not satisfied, then by (16) we obtain that

$$\begin{aligned}\langle \Phi_2(x^k,\lambda^k),\Phi_2'(x^k,\lambda^k)d^k\rangle &= -\|\Phi'(x^k,\lambda^k)d^k\|^2-\sigma_k\|d^k\|^2\\ &\quad -\langle \Phi_1(x^k),\Phi_1'(x^k)\xi^k\rangle\\ &\le -(1-\nu)\left(\|\Phi'(x^k,\lambda^k)d^k\|^2+\sigma_k\|d^k\|^2\right)\\ &< 0, \end{aligned}\tag{26}$$

and therefore, by (15), $\bar{c}_{2,\,k}$ in (21) is well-defined in this case.

We conclude that our choice of penalty parameters is well-defined, and guarantees that the computed direction d^k is of descent for the associated penalty function. Hence, linesearch in Algorithm 1 always terminates with some stepsize $\alpha_k > 0$. This implies that the whole algorithm is well-defined.

Observe next that if for some infinite set K of iteration indices, the subsequence $\{(x^k,\lambda^k) \mid k\in K\}$ converges to some point $(\bar{x},\bar{\lambda})$ violating (4), i.e., $\Phi(\bar{x},\bar{\lambda})\ne 0$, then the sequence $\{\sigma_k \mid k\in K\}$ is separated from zero by a positive constant (recall that $\sigma_k=\min\{\bar{\sigma},\|\Phi(x^k,\lambda^k)\|^q\}$). This implies that the matrices in the sequence $\{(\Phi'(x^k,\lambda^k))^2+\sigma_k I \mid k\in K\}$ are uniformly positive definite. Hence, the inverses of these matrices exist, and depend continuously on (x^k,λ^k) at $(\bar{x},\bar{\lambda})$. Therefore, from (6) it follows that in this case the sequence $\{d^k \mid k\in K\}$ necessarily converges to some d; moreover, if (23) is violated as well, then $d\ne 0$.

We proceed to the possible cases of the asymptotic behaviour of the penalty parameters.

Consider first the case when the values of c_1 and c_2 do not change for all indices k sufficiently large. In this case, the "tail" of the sequence $\{(x^k,\lambda^k)\}$ is generated by a descent method with Armijo linesearch for a fixed smooth function $\varphi_{c_1,\,c_2}$. Suppose that $(\bar{x},\bar{\lambda})$ is an accumulation point of $\{(x^k,\lambda^k)\}$, violating (23). Let K be an infinite set of iteration indices such that the subsequence $\{(x^k,\lambda^k) \mid k\in K\}$ converges to $(\bar{x},\bar{\lambda})$. Then according to the observation above, $\{\sigma_k \mid k\in K\}$ is separated from zero by a positive constant, while $\{d^k \mid k\in K\}$ is bounded and its norm is separated from zero. Therefore, (18) implies that $\{\langle \varphi'_{c_1,\,c_2}(x^k,\lambda^k),d^k\rangle \mid k\in K\}$ is separated from zero by some negative constant. It remains to repeat the argument from the proof of [1, Theorem 1.8] (on linesearch methods using "uniformly gradient-related directions") to conclude that $(\bar{x},\bar{\lambda})$ is a stationary point in (9) i.e., it satisfies (24). This proves item (i).

Consider now the case when c_2 is constant from some iteration on, but there exists an infinite index set K such that c_1 increases at each iteration $k\in K$ (note that, at any iteration, c_1 can only increase or remain unchanged). Possibly taking a further subsequence, suppose that $\{(x^k,\lambda^k) \mid k\in K\}$ converges to some $(\bar{x},\bar{\lambda})$ violating (23). Recalling the observation above, then $\{d^k \mid k\in K\}$ converges to some $d\ne 0$, and $\{\sigma_k \mid k\in K\}$ is separated from zero by a positive constant. As the value of c_1 increases at each iteration $k\in K$, condition (19) must hold for all $k\in K$, where the right-hand side, and hence also the left-hand side, are separated from zero by a negative constant. This shows that the quantity $\bar{c}_{1,\,k}$ defined in (20) is bounded above, as it depends

continuously on (x^k, λ^k, d^k) at $(\bar{x}, \bar{\lambda}, d)$, and c_2 is fixed. Hence, the sequence $\{\bar{c}_{1,k} \mid k \in K\}$ is bounded. On the other hand, for each $k \in K$, the value of c_1 increases at least by $\delta > 0$, which implies that c_1 goes to infinity. Clearly, this contradicts boundedness of $\{\bar{c}_{1,k} \mid k \in K\}$, because for $k \in K$ the value of c_1 is modified to $\bar{c}_{1,k} + \delta$. The proof of item (ii) is complete.

Assertion (iii) is proven in a similar way as (ii). In particular, if c_1 does not change from some point on, this means that (19) never holds for k sufficiently large. Thus (26) holds. If (23) were not true, the left-hand and right-hand sides in (26) stay separated from zero by a negative constant. Then the quantity $\bar{c}_{2,k}$ defined in (21) is bounded above, and the sequence $\{\bar{c}_{2,k} \mid k \in K\}$ is bounded. As any change of c_2 for k large enough means modifying it to $\bar{c}_{2,k} + \delta$, in particular increasing it at least by $\delta > 0$, unboundedness of c_2 contradicts boundedness of $\{\bar{c}_{2,k} \mid k \in K\}$.

Suppose finally that the sequence $\{(x^k, \lambda^k)\}$ converges to some $(\bar{x}, \bar{\lambda})$, and the values of both c_1 and c_2 are changed infinitely many times (and thus tend to infinity). Suppose that $(\bar{x}, \bar{\lambda})$ does not satisfy (23). Then, by the observation above, $\{d^k\}$ converges to some $d = (\xi, \eta) \neq 0$ and $\{\sigma_k\}$ is separated from zero by a positive constant. Let K_1 (respectively, K_2) be an infinite index set such that c_1 (c_2) increases for each iteration $k \in K_1$ ($k \in K_2$). As above, condition (19) must hold for all $k \in K_1$, where the right-hand side, and hence also the left-hand side are separated from zero by a negative constant. Passing to the limit, we then obtain that $\langle \Phi_1(\bar{x}), \Phi_1'(\bar{x})\xi \rangle < 0$. Then by the convergence of $\{(x^k, \lambda^k)\}$ to $(\bar{x}, \bar{\lambda})$, and of $\{d^k\}$ to d, it follows that (25) holds for all $k \in K_2$ large enough. Also, (26) holds, because the test (19) failed for $k \in K_2$. Employing (21), we then obtain that $\bar{c}_{2,k}$ computed for such k is bounded from above by a quantity not depending on c_1 (as according to (25), the latter appears in (21) multiplied by a negative quantity), and depending continuously on (x^k, λ^k, d^k) at $(\bar{x}, \bar{\lambda}, d)$ (as according to (26), c_2 in (15) is multiplied by a quantity smaller than $-(1-\nu)\sigma_k\|d^k\|^2$, which is separated from zero by a negative constant). Therefore, the sequence $\{\bar{c}_{2,k} \mid k \in K_2\}$ is bounded. On the other hand, for each $k \in K_2$, the value of c_2 increases at least by $\delta > 0$, which implies that c_2 goes to infinity. This contradicts boundedness of $\{\bar{c}_{2,k} \mid k \in K_2\}$, because for $k \in K_2$ the value of c_2 is modified to $\bar{c}_{2,k} + \delta$. This completes the proof of item (iv). □

5 Superlinear convergence

To avoid certain (essentially technical) complications, it is convenient to assume formally in the statements below that if (x^k, λ^k) satisfies the Lagrange system (4) (and hence, $\sigma_k = 0$), then $d^k = 0$ is "computed". Note that the algorithm actually terminates in this case at step 1, and no d^k is computed in reality. On the other hand, $d^k = 0$ is always a solution of (6) in this case, and for the purposes of convergence analysis it is convenient to assume that precisely this solution is meant (note that in the degenerate case, there may

exist other solutions of (6) with $\sigma_k = 0$, for (x^k, λ^k) satisfying the Lagrange system (4)).

We start our local rate of convergence analysis with the following proposition.

Proposition 1 *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ be twice differentiable near $\bar{x} \in \mathbb{R}^n$, with their second derivatives being continuous at $\bar{x}$. Let $\bar{x}$ be a stationary point of problem* (3) *with an associated noncritical Lagrange multiplier $\bar{\lambda} \in \mathbb{R}^l$.*

Then for any $q \in [1, 2]$ there exists $\gamma > 0$ such that for the solution d^k of (6)*, where $\sigma_k = \|\Phi(x^k, \lambda^k)\|^q$, it holds that*

$$\|\Phi'(x^k, \lambda^k)d^k\| \geq \gamma\|d^k\|$$

for all $(x^k, \lambda^k) \in \mathbb{R}^n \times \mathbb{R}^l$ close enough to $(\bar{x}, \bar{\lambda})$.

Proof Observe first that under the error bound (11) (which is equivalent to noncriticality of $\bar{\lambda}$), the solution set $\Phi^{-1}(0)$ of the Lagrange system (4), near $(\bar{x}, \bar{\lambda})$, has the form $\{\bar{x}\} \times \Lambda(\bar{x})$. Therefore, (11) implies that

$$\operatorname{dist}((x^k, \lambda^k), \Phi^{-1}(0)) = O(\|\Phi(x^k, \lambda^k)\|). \tag{27}$$

Hence, according to [6, Lemmas 2.1, 2.2],

$$\|d^k\| = O(\operatorname{dist}((x^k, \lambda^k), \Phi^{-1}(0))), \tag{28}$$

and

$$\operatorname{dist}((x^k, \lambda^k) + d^k, \Phi^{-1}(0)) = O\left((\operatorname{dist}((x^k, \lambda^k), \Phi^{-1}(0)))^{(2+q)/2}\right) \tag{29}$$

as $(x^k, \lambda^k) \to (\bar{x}, \bar{\lambda})$. (More precisely, in [6, Lemma 2.2] it is additionally *assumed* that $(x^k, \lambda^k) + d^k$ stays close enough to $(\bar{x}, \bar{\lambda})$. However, in the proof of [6, Theorem 2.1] it is established that any needed proximity of $(x^k, \lambda^k) + d^k$ to $(\bar{x}, \bar{\lambda})$ is guaranteed by sufficient proximity of (x^k, λ^k) to $(\bar{x}, \bar{\lambda})$, and hence, this additional assumption can actually be removed.)

We argue by contradiction. Suppose that there exists $\{(x^k, \lambda^k)\} \subset \mathbb{R}^n \times \mathbb{R}^l$, convergent to $(\bar{x}, \bar{\lambda})$, and such that

$$\Phi'(x^k, \lambda^k)d^k = o(\|d^k\|)$$

as $k \to \infty$. Then, according to (28), it holds that

$$\Phi'(x^k, \lambda^k)d^k = o(\operatorname{dist}((x^k, \lambda^k), \Phi^{-1}(0))). \tag{30}$$

We also have that

$$\begin{aligned}\Phi((x^k, \lambda^k) + d^k) &= O(\operatorname{dist}((x^k, \lambda^k) + d^k, \Phi^{-1}(0))) \\ &= o(\operatorname{dist}((x^k, \lambda^k), \Phi^{-1}(0))),\end{aligned} \tag{31}$$

where the first equality is by the Lipschitz-continuity of Φ, and the second is by (29). Then, employing (28) again, by the mean-value theorem we derive that

$$\begin{aligned}\Phi((x^k,\,\lambda^k)+d^k)-\Phi(x^k,\,\lambda^k)-\Phi'(x^k,\,\lambda^k)d^k &= o(\|d^k\|)\\ &= o(\mathrm{dist}((x^k,\,\lambda^k),\,\Phi^{-1}(0))).\end{aligned}$$

Hence, taking into account (30) and (31),

$$\Phi(x^k,\,\lambda^k)=o(\mathrm{dist}((x^k,\,\lambda^k),\,\Phi^{-1}(0)))$$

as $k\to\infty$, which contradicts (27). □

Recall that by the construction of the algorithm, each of the penalty parameters either stays constant from some iteration on, or it tends to infinity. Next, we show that if one parameter stays constant, then so does the other. We note that, in fact, this is the behavior we observed (for successful runs) in numerical experiments reported in Section 6 below.

Lemma 1 *Under the assumptions of Proposition 1, suppose that a sequence* $\{(x^k,\,\lambda^k)\}$ *generated by Algorithm 1 with* $q\in[1,\,2]$*, and with* $\tau\geq 2$*, converges to* $(\bar{x},\,\bar{\lambda})$*.*

Then if the value of c_1 *or* c_2 *does not change for all* k *sufficiently large, then so does the value of the other parameter.*

Proof Employing (11), (28) and (29), we obtain

$$\begin{aligned}\Phi(x^k,\,\lambda^k) &= (\Phi(x^k,\,\lambda^k)-\Phi((x^k,\,\lambda^k)+d^k))+\mathrm{dist}((x^k,\,\lambda^k)+d^k,\,\Phi^{-1}(0))\\ &= O(\|d^k\|)+o(\mathrm{dist}((x^k,\,\lambda^k),\,\Phi^{-1}(0)))\\ &= O(\|d^k\|)+o(\|\Phi(x^k,\,\lambda^k)\|),\end{aligned}$$

implying that

$$\Phi(x^k,\,\lambda^k)=O(\|d^k\|)$$

as $k\to\infty$. Thus, also

$$\Phi_2(x^k,\,\lambda^k)=O(\|d^k\|).$$

We then obtain that

$$\langle\Phi(x^k,\,\lambda^k),\,d^k\rangle\leq\|\Phi(x^k,\,\lambda^k)\|\|d^k\|=O(\|d^k\|^2),\tag{32}$$

implying that

$$\langle\Phi_2(x^k,\,\lambda^k),\Phi_2'(x^k,\,\lambda^k)d^k\rangle=O(\|d^k\|^2).\tag{33}$$

Recall that any change of c_1 is only possible when (19) holds. Therefore, for the corresponding k we obtain that

$$\langle\Phi_1(x^k),\,\Phi_1'(x^k)\xi^k\rangle\leq-\nu\|\Phi'(x^k,\lambda^k)d^k\|^2\leq-\nu\gamma\|d^k\|^2,\tag{34}$$

where the second inequality is by Proposition 1.

Suppose that c_2 is constant for k sufficiently large. Then, recalling (20) and using (32)–(34), we conclude that $\{\bar{c}_{1,\,k}\}$ is bounded. This implies that c_1 does not change from some point on. Indeed, when it does change, it increases at least by $\delta > 0$. If this were to happen an infinite number of times, then c_1 actually tends to infinity, in contradiction with boundedness of $\{\bar{c}_{1,\,k}\}$ (again, since the update is $c_1 = \bar{c}_{1,\,k} + \delta$).

The argument for the case when c_1 is asymptotically constant is similar. Any change of c_2 consists of modifying it to $\bar{c}_{2,\,k} + \delta$, when (19) does not hold. In this case, by (26) and by Proposition 1, for the corresponding k it holds that

$$\begin{aligned}\langle \Phi_2(x^k,\,\lambda^k),\,\Phi_2'(x^k,\,\lambda^k)d^k\rangle &\le -(1-\nu)\|\Phi'(x^k,\,\lambda^k)d^k\|^2\\ &\le -(1-\nu)\gamma\|d^k\|^2. \end{aligned}\tag{35}$$

Moreover, since (19) is violated for such k, it holds that either

$$\langle \Phi_1(x^k),\,\Phi_1'(x^k)\xi^k\rangle \ge 0$$

or

$$|\langle \Phi_1(x^k),\,\Phi_1'(x^k)\xi^k\rangle| \le \nu\|\Phi'(x^k,\lambda^k)d^k\|^2 = O(\|d^k\|^2).$$

Then, since c_1 is constant for sufficiently large k, recalling (21) and using (32) it follows that $\{\bar{c}_{2,\,k}\}$ is bounded. The same way as for c_1 above, this implies that c_2 does not change from some point on. □

The chain of results presented next eventually leads to Lemma 3, demonstrating the asymptotic acceptance of the unit sptepsize in Algorithm 1. This guarantees the quadratic convergence rate of the algorithm.

The following result is [17, Proposition 5.3], which improves [1, Theorem 4.16 (a)] by removing the regularity assumption on the constraints.

Proposition 2 *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ be three times differentiable at $\bar{x} \in \mathbb{R}^n$. Let $\bar{x}$ be a stationary point of problem (3) with an associated Lagrange multiplier $\bar{\lambda} \in \mathbb{R}^l$ satisfying SOSC (10).*

Then for every $\bar{c}_2 > 0$ there exists $\bar{c}_1 \ge 0$ such that for all $c_1 \ge \bar{c}_1$, $c_2 \ge \bar{c}_2$, $\varphi''_{c_1,\,c_2}(\bar{x},\,\bar{\lambda})$ is positive semidefinite, and moreover, $\langle \varphi''_{c_1,\,c_2}(\bar{x},\,\bar{\lambda})d,\,d\rangle = 0$ if and only if $d \in \ker \Phi'(\bar{x},\,\bar{\lambda})$. In addition, for any fixed $d \in \mathbb{R}^n \times \mathbb{R}^l$, increasing c_1 or c_2 can only make $\langle \varphi''_{c_1,\,c_2}(\bar{x},\,\bar{\lambda})d,\,d\rangle$ larger.

We then obtain the following.

Lemma 2 *Under the assumptions of Proposition 2, for every $q \in [1,\,2]$ and $\bar{c}_2 > 0$, there exist $\gamma > 0$, $\bar{c}_1 \ge 0$, and a neighborhood U of $(\bar{x},\,\bar{\lambda})$ such that for all $c_1 \ge \bar{c}_1$ and $c_2 \ge \bar{c}_2$, and for all $(x^k,\,\lambda^k) \in U$, for the solution d^k of (6) with $\sigma_k = \|\Phi(x^k,\,\lambda^k)\|^q$ it holds that*

$$\langle \varphi''_{c_1,\,c_2}(\bar{x},\,\bar{\lambda})d^k,\,d^k\rangle \ge \gamma\|d^k\|^2.$$

Proof Let $q \in [1, 2]$ and $\bar{c}_2 > 0$ be fixed, and let $\bar{c}_1 \geq 0$ be defined according to Proposition 2. Suppose that there exist sequences $\{c_{1,k}\} \subset \mathbb{R}$, $\{c_{2,k}\} \subset \mathbb{R}$, and $\{(x^k, \lambda^k)\} \subset \mathbb{R}^n \times \mathbb{R}^l$ such that $c_{1,k} \geq \bar{c}_1$ and $c_{2,k} \geq \bar{c}_2$ for all k, $\{(x^k, \lambda^k)\}$ converges to $(\bar{x}, \bar{\lambda})$, and it holds that

$$\langle \varphi''_{c_{1,k}, c_{2,k}}(\bar{x}, \bar{\lambda})d^k, d^k \rangle = o(\|d^k\|^2) \tag{36}$$

as $k \to \infty$ ((36) implies, in particular, that $d^k \neq 0$ for all k large enough). According to the last assertion of Proposition 2, we can always assume that $c_{1,k} = \bar{c}_1$ and $c_{2,k} = \bar{c}_2$ for all k (as decreasing $c_{1,k}$ or $c_{2,k}$ can only make the nonnegative left-hand side of (36) smaller, thus keeping the relation in question valid).

Dividing both sides of (36) by $\|d^k\|^2$, and passing to subsequences if necessary, we can assume that $\{d^k/\|d^k\|\}$ converges to some $d \in \mathbb{R}^n \times \mathbb{R}^l$ such that $\|d\| = 1$ and $\langle \varphi''_{\bar{c}_1, \bar{c}_2}(\bar{x}, \bar{\lambda})d, d\rangle = 0$. According to Proposition 2, this is only possible when $\Phi'(\bar{x}, \bar{\lambda})d = 0$. At the same time, Proposition 1 implies that $\Phi'(\bar{x}, \bar{\lambda})d \neq 0$, giving a contradiction. □

The next result shows conditions under which the unit stepsize is accepted by the algorithm; it employs the argument from [17, Corollary 5.1].

Lemma 3 *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ be three times differentiable near $\bar{x} \in \mathbb{R}^n$, with their third derivatives being continuous at $\bar{x}$. Let $\bar{x}$ be a stationary point of problem* (3) *with an associated Lagrange multiplier $\bar{\lambda} \in \mathbb{R}^l$ satisfying SOSC* (10).

Then for every $q \in [1, 2]$ and $\bar{c}_2 > 0$, there exist $\rho > 0$, $\bar{c}_1 \geq 0$, and a neighborhood U of $(\bar{x}, \bar{\lambda})$ such that for all $c_1 \geq \bar{c}_1$ and $c_2 \geq \bar{c}_2$, and for all $(x^k, \lambda^k) \in U$, condition (18) *holds with $\tau = 2$ for the solution d^k of* (6) *with $\sigma_k = \|\Phi(x^k, \lambda^k)\|^q$, and moreover, if $\varepsilon \in (0, 1/2)$, then* (22) *holds with $j = 0$ (i.e., $\alpha_k = 1$ is accepted by the algorithm).*

Proof Let $q \in [1, 2]$ and $\bar{c}_2 > 0$ be fixed, and let $\bar{c}_1 \geq 0$ be defined according to Lemma 2. Fix any $c_1 \geq \bar{c}_1$ and $c_2 \geq \bar{c}_2$.

For each k, let $\hat{\lambda}^k$ stand for the projection of $\lambda^k + \eta^k$ onto $\Lambda(\bar{x})$. By (29) we obtain that

$$\begin{aligned}(x^k, \lambda^k) + d^k - (\bar{x}, \hat{\lambda}^k) &= o\left((\mathrm{dist}((x^k, \lambda^k), \Phi^{-1}(0)))\right)\\ &= o(\|(x^k - \bar{x}, \lambda^k - \hat{\lambda}^k)\|),\end{aligned} \tag{37}$$

implying that

$$(x^k - \bar{x}, \lambda^k - \hat{\lambda}^k) = -d^k + o(\|d^k\|) \tag{38}$$

as (x^k, λ^k) tends to $(\bar{x}, \bar{\lambda})$.

Recall that $(\bar{x}, \hat{\lambda}^k)$ is always a stationary point in problem (9). By the mean-value theorem, and by the continuity of φ''_{c_1, c_2} at $(\bar{x}, \bar{\lambda})$, it then holds

that

$$\begin{aligned}\varphi'_{c_1,\,c_2}((x^k,\,\lambda^k)+d^k) &= \varphi'_{c_1,\,c_2}((x^k,\,\lambda^k)+d^k)-\varphi'_{c_1,\,c_2}(\bar{x},\,\hat{\lambda}^k)\\ &= O(\|(x^k,\,\lambda^k)+d^k-(\bar{x},\,\hat{\lambda}^k)\|)\\ &= o(\|(x^k-\bar{x},\,\lambda^k-\hat{\lambda}^k))\\ &= o(\|d^k\|)\end{aligned}$$

as $(x^k,\,\lambda^k)$ tends to $(\bar{x},\,\bar{\lambda})$, where the last two equalities are by (37) and (38), respectively. On the other hand,

$$\varphi'_{c_1,\,c_2}((x^k,\,\lambda^k)+d^k)=\varphi'_{c_1,\,c_2}(x^k,\,\lambda^k)+\varphi''_{c_1,\,c_2}(x^k,\,\lambda^k)d^k+o(\|d^k\|),$$

and combining this with the previous estimate we conclude that

$$\varphi'_{c_1,\,c_2}(x^k,\,\lambda^k)+\varphi''_{c_1,\,c_2}(x^k,\,\lambda^k)d^k=o(\|d^k\|). \tag{39}$$

Therefore,

$$\langle\varphi'_{c_1,\,c_2}(x^k,\,\lambda^k),\,d^k\rangle=-\langle\varphi''_{c_1,\,c_2}(x^k,\,\lambda^k)d^k,\,d^k\rangle+o(\|d^k\|^2)$$

as $(x^k,\,\lambda^k)$ tends to $(\bar{x},\,\bar{\lambda})$. Hence, by the choice of $\bar{c}_1$, for any $\rho\in(0,\,\gamma)$ condition (18) holds with $\tau=2$, assuming that $(x^k,\,\lambda^k)$ is close enough to $(\bar{x},\,\bar{\lambda})$.

Next, we obtain that

$$\begin{aligned}&\varphi_{c_1,\,c_2}((x^k,\,\lambda^k)+d^k)-\varphi_{c_1,\,c_2}(x^k,\,\lambda^k)\\ &=\langle\varphi'_{c_1,\,c_2}(x^k,\,\lambda^k),d^k\rangle+\frac{1}{2}\langle\varphi''_{c_1,\,c_2}(x^k,\,\lambda^k)d^k,\,d^k\rangle+o(\|d^k\|^2)\\ &=\frac{1}{2}\langle\varphi'_{c_1,\,c_2}(x^k,\,\lambda^k),d^k\rangle+o(\|d^k\|^2)\end{aligned}$$

as $(x^k,\,\lambda^k)$ tends to $(\bar{x},\,\bar{\lambda})$, where the last equality is by (39). Combining this with (18) with $\tau=2$, we finally derive that

$$\begin{aligned}&\varphi_{c_1,\,c_2}((x^k,\,\lambda^k)+d^k)-\varphi_{c_1,\,c_2}(x^k,\,\lambda^k)-\varepsilon\langle\varphi'_{c_1,\,c_2}(x^k,\,\lambda^k),\,d^k\rangle\\ &=\left(\frac{1}{2}-\varepsilon\right)\langle\varphi'_{c_1,\,c_2}(x^k,\,\lambda^k),\,d^k\rangle+o(\|d^k\|^2)\\ &\le-\rho\left(\frac{1}{2}-\varepsilon\right)\|d^k\|^2+o(\|d^k\|^2)\\ &\le 0\end{aligned}$$

for all $(x^k,\,\lambda^k)$ is close enough to $(\bar{x},\,\bar{\lambda})$. This means that (22) holds for $j=0$. □

Remark 1 Observe that if (18) holds with some $\rho>0$ and $\tau=2$, then it also holds with any $\rho>0$ and any $\tau>2$, for all $(x^k,\,\lambda^k)$ close enough to $(\bar{x},\,\bar{\lambda})$.

We are now ready to state our local quadratic rate of convergence result. It follows from the fact that, once the unit stepsize is accepted, the method behaves as the usual Levenberg–Marquardt scheme.

Theorem 2 *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ be three times differentiable at $\bar{x} \in \mathbb{R}^n$. Let $\bar{x}$ be a stationary point of problem* (3) *with an associated Lagrange multiplier $\bar{\lambda} \in \mathbb{R}^l$ satisfying SOSC* (10)*. Suppose that a sequence $\{(x^k, \lambda^k)\}$ generated by Algorithm 1 with $q \in [1, 2]$, $\varepsilon \in (0, 1/2)$, $\bar{c}_1 > 0$ large enough with respect to $\bar{c}_2 > 0$, and either with $\tau > 2$, or with $\tau = 2$ and a sufficiently small $\rho > 0$, converges to $(\bar{x}, \bar{\lambda})$. Suppose c_1 and c_2 are asymptotically constant.*

Then $\alpha_k = 1$ holds for all sufficiently large k, and the rate of convergence of $\{(x^k, \lambda^k)\}$ to $(\bar{x}, \bar{\lambda})$ is quadratic.

Proof Follows combining Lemma 3, Remark 1, and [6, Theorem 2.2]. □

Note that Lemma 1 states that if one penalty parameter stays asymptotically constant then so does the other, but it does not rule out the theoretical possibility of both tending to infinity. On the other hand, Theorem 2 refers to the case of the parameters asymptotically fixed. While we have never encountered both c_1 and c_2 growing to infinity in our numerical experiments in Section 6, the following simple procedure can be used (along with many others) to guarantee this. Fix some (large) constant, and once c_2 exceeds this limit, start using $c_1 = c_2 = c$, where $c > 0$ is large enough so that (18) holds. This possibility had been already commented in Section 3; it follows directly from (15) and (16). With some more-or-less obvious refinements of this procedure, it can be easily seen that both global convergence and local rate of convergence results stated above, can be adapted to cover this modification. In particular, in the local convergence analysis, $c_1 = c_2 = c$ eventually stays fixed. However, this modification is more of a theoretical guarantee: as we have never encountered both parameters being unbounded for Algorithm 1 in our numerical experiments in Section 6, we do not add this feature in our implementation.

6 Numerical results for the globalized Levenberg–Marquardt method for optimization

In this section, we present some numerical experiments with the globalized versions of the Levenberg–Marquardt method applied to *degenerate* optimization problems. Specifically, we consider here Algorithm 1 and its counterpart with linesearch for the squared Euclidian residual of the Lagrange optimality system (4), as defined in [6,23]. To avoid any possible ambiguity, we shall give the full statement of the latter algorithm. Define the function $\varphi : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}$,

$$\varphi(x, \lambda) = \frac{1}{2}\|\Phi(x, \lambda)\|^2. \tag{40}$$

Algorithm 2 Choose the parameters $\bar{\sigma} > 0$, $q > 0$, and $\beta, \varepsilon, \theta \in (0, 1)$. Choose $(x^0, \lambda^0) \in \mathbb{R}^n \times \mathbb{R}^l$ and set $k = 0$.

1. If (17) holds with $\Phi(\cdot)$ defined in (5), stop.
2. Set $\sigma_k = \min\{\bar{\sigma}, \|\Phi(x^k, \lambda^k)\|^q\}$. Compute $d^k = (\xi^k, \eta^k)$ as the unique solution of the Levenberg–Marquardt equation (6).
3. If
$$\|\Phi((x^k, \lambda^k) + d^k)\| \le \beta\|\Phi(x^k, \lambda^k)\|,$$
set $\alpha_k = 1$ and go to step 5.
4. Compute $\alpha_k = \theta^j$, where j is the smallest nonnegative integer satisfying the Armijo inequality
$$\varphi((x^k, \lambda^k) + \theta^j d^k) \le \varphi(x^k, \lambda^k) + \varepsilon\theta^j \left\langle \varphi'(x^k, \lambda^k), d^k \right\rangle.$$
5. Set $(x^{k+1}, \lambda^{k+1}) = (x^k, \lambda^k) + \alpha_k d^k$, increase k by 1, and go to step 1.

In the bar diagrams below, Algorithms 1 and 2 are represented by white and gray color, respectively. For both algorithms, we used the exponent $q = 2$ when defining the regularization parameter σ_k. Moreover, the rule for σ_k was modified as follows:

$$\sigma_k = \begin{cases} \|\Phi(x^k, \lambda^k)\|^2 & \text{if } \|\Phi(x^k, \lambda^k)\| \le \bar{\sigma}, \\ \bar{\sigma}^4/\|\Phi(x^k, \lambda^k)\|^2 & \text{otherwise.} \end{cases}$$

We note that using this (empirically determined) rule does not affect neither the behavior of the algorithms near solutions, nor any of the convergence theory. On the other hand, this modification has some stabilizing effect (making iterations closer to those of SQP when far from solutions).

The other parameters in Algorithms 1 and 2 were taken as follows: $\bar{\sigma} = 1$, $\beta = 0.9$, $\rho = 1e-9$, $\tau = 2.1$, $\nu = 0.1$, $\bar{c}_1 = 10$, $\bar{c}_2 = 0.1$, $\delta = 1$, $\varepsilon = 0.01$, and $\theta = 0.5$. Also, the algorithms terminate, in addition to the other criteria, when the stepsize parameter during linesearch becomes smaller than $1e-10$ (of course, this is counted as a failure). The other stopping rules are the same as in Section 2.

Test problems and starting points are generated the same way as in Section 2 (10 starting points for each of 100 problems generated for each triple (n, l, r), making it 1000 runs for each triple in total), but starting points for the global runs are now in the box with the half of the edge equal to 100. Triples used here are the same as in Section 2, but only for $n = 5, 10, 25$, and now excluding the case of $r = 0$, as the behavior of the algorithms in this case is way too different from the other values of r. (The value $r = 0$ was not excluded from local experiments to demonstrate specially good behavior of sSQP in this case, at the same time showing that it degrades as r grows.)

The first part of our experiments consists in comparing the algorithms' ability to compute approximate solutions with better objective function values, and to detect unbounded problems. This is meant precisely to highlight the advantage of our proposal to use a penalty function of the optimization problem (which is not indifferent to minimization versus maximization) instead of the residual of the Lagrange system. In this series of experiments, both algorithms were always run from the same starting points. For a run

successful for both algorithms, the objective function value at termination of one algorithm is regarded "better" than that of the other when it is smaller by no less than 0.1. Note that in successful runs feasibility is always satisfied up to the stopping tolerance, and so comparing objective function values is meaningful. We also count the runs for which exactly one (but not the other) of the algorithms detected an unbounded problem, by which we mean that at some iteration the constraints residual was no greater than $1e-4$, while the objective function value was no larger than $-1e+4$. This information for some triples $(n,\, l,\, r)$ is reported in Figure 3. We show only some few triples, but they reflect well enough the overall tendency. The conclusion is that the method using penalty function computes lower objective function values and detects unboundedness more often.

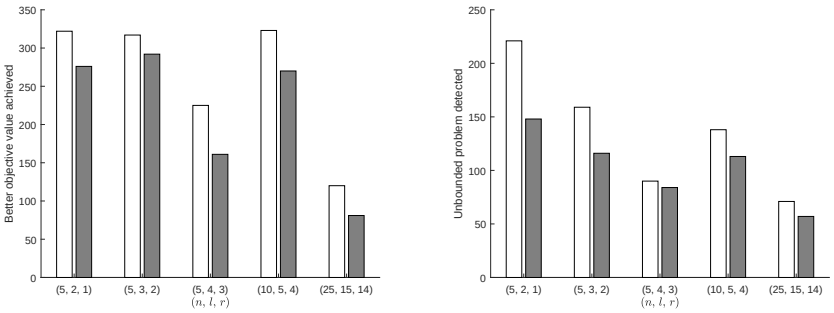
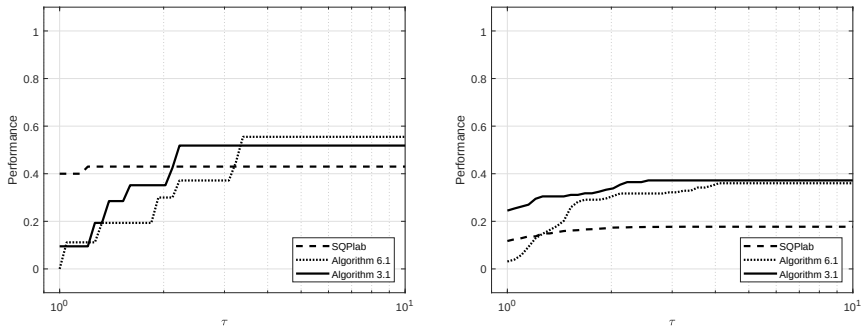


Fig. 3: Global behavior: quality of outcome.

Finally, we exhibit some results that give an impression of the overall performance of the algorithms in question on this kind of degenerate test problems. In particular, these results demonstrate that the benefits in the quality of outcome obtained by Algorithm 1, and reported in Figure 3, are achieved without degrading robustness and efficiency in any serious way. We emphasize that the test problems used here are quite difficult for computational methods, even though we restrict testing to relatively low dimensions. Partially to confirm the latter claim, we also provide results for SQPlab solver [20], which is a well-established open-source Matlab implementation of the quasi-Newton SQP method globalized by linesearch for the l_1 penalty function. SQPlab was run with all the default settings, except for stopping tolerances, iteration limit, and lower step length limit, which were put in agreement with those for the other two algorithms. In this series of experiments, the objective function was always taken strongly convex, to avoid failures caused by unbounded problems. We used the same triples $(n,\, l,\, r)$ as above, but we restricted ourselves to $n = 5,\, 10$, as already for $n = 10$ robustness of the methods (and especially of SQPlab) is quite low on this test set.



(a) $n = 5$ (b) $n = 10$

Fig. 4: Relative robustness and efficiency.

The results are reported in Figure 4, in the form of performance profiles [4]. They show that on the given test set, Algorithm 1 has similar robustness, and is more efficient (in terms of the iteration count), than the Levenberg–Marquardt method globalized using the squared residual. The results for SQPlab are included to show that the test set in question is somehow difficult, and so the results obtained for the proposed algorithm are relatively favorable indeed. Recall also that, as already reported above, Algorithm 1 produces better quality solutions than the usal Levenberg–Marquardt method.

Table 1: Behavior of penalty parameters; $n = 5$

Triples			Successful runs				Failures			
n	l	r	E1S	E2S	D1S	D2S	E1F	E2F	D1F	D2F
5	2	1	0	1	0	1	24	4	17	7
5	3	1	3	4	1	1	24	4	17	6
5	3	2	2	0	0	0	14	2	12	6
5	4	1	3	7	0	4	27	3	15	4
5	4	2	2	3	0	3	18	2	12	6
5	4	3	1	0	0	1	13	3	8	5

To complete the picture, we report in Tables 1 and 2 some interesting information regarding the behavior of penalty parameters in Algorithm 1. The meaning of the columns in these tables is as follows (the information is reported separately for successful runs and for failures):

- E1S: cases when c_1 exceeded the threshold = 100 (% out of successful runs).
- E2S: cases when c_2 exceeded the threshold = 100 (% out of successful runs).
- D1S: cases when different values of c_1 were used on 2 last iterations, i.e., c_1 did not "stabilize" (% out of successful runs).

Table 2: Behavior of penalty parameters; $n = 10$

Triples			Successful runs				Failures			
n	l	r	E1S	E2S	D1S	D2S	E1F	E2F	D1F	D2F
10	2	1	0	2	0	2	11	6	15	12
10	3	1	2	4	1	3	14	4	19	2
10	3	2	0	1	0	1	7	2	13	2
10	4	1	3	7	0	4	27	3	15	4
10	4	2	0	5	0	2	9	1	11	9
10	4	3	0	1	0	0	2	2	10	10
10	5	1	2	12	0	4	15	2	11	7
10	5	2	1	4	0	2	9	1	10	8
10	5	3	2	1	0	0	6	1	9	10
10	5	4	1	0	0	0	5	1	6	10
10	6	1	2	11	0	4	17	3	12	7
10	6	2	0	3	0	1	13	1	8	6
10	6	3	1	1	0	1	10	1	8	9
10	6	4	1	1	0	0	6	1	6	7
10	6	5	1	0	0	0	4	1	4	8
10	7	1	2	20	0	5	13	4	7	4
10	7	2	1	3	0	4	13	1	9	5
10	7	3	1	2	2	1	11	2	8	7
10	7	4	1	2	1	1	9	1	6	7
10	7	5	2	1	0	0	5	1	4	6
10	7	6	0	0	1	0	4	1	3	8
10	8	1	4	15	0	4	18	6	9	4
10	8	2	2	3	0	4	18	3	11	4
10	8	3	1	4	0	3	14	1	6	3
10	8	4	1	2	1	2	12	3	7	5
10	8	5	1	1	0	1	7	2	4	7
10	8	6	2	1	1	0	6	2	3	7
10	8	7	1	0	0	1	5	1	4	6
10	9	1	4	21	0	4	21	6	8	4
10	9	2	1	14	0	4	19	5	11	5
10	9	3	2	5	0	2	15	4	9	4
10	9	4	3	4	1	1	12	2	7	5
10	9	5	3	3	1	1	11	2	7	4
10	9	6	2	1	0	0	7	2	4	6
10	9	7	1	1	1	1	7	2	4	6
10	9	8	2	0	1	0	7	2	4	6

– D2S: cases when different values of c_2 were used on 2 last iterations, i.e., c_2 did not "stabilize" (% out of successful runs).
– E1F: cases when c_1 exceeded the threshold = 100 (% out of failures).
– E2F: cases when c_2 exceeded the threshold = 100 (% out of failures).
– D1F: cases when different values of c_1 were used on 2 last iterations, i.e., c_1 did not "stabilize" (% out of failures).
– D2F: cases when different values of c_2 were used on 2 last iterations, i.e., c_2 did not "stabilize" (% out of failures).

Overall, we observe that usually c_1 and c_2 do not go higher than some rather moderate values, especially on successful runs. Naturally, this changes somewhat in the case of failures. However, even for failures, the behavior of penalty parameters looks "suspicious" in never more than 25% of cases (and usually much less). It thus appears that failures are mostly related to the inherent difficulties of degenerate problems.

Acknowledgments

The authors thank the two anonymous referees for their constructive comments.

The research of the first author was supported by the Russian Science Foundation Grant 17-11-01168 (Section 5). The second author is supported in part by CNPq Grant 303724/2015-3 and by FAPERJ Grant 203.052/2016. The third author is supported by the Volkswagen Foundation, and by the Russian Foundation for Basic Research Grant 17-01-00125.

References

1. D.P. Bertsekas. Constrained Optimization and Lagrange Multiplier Methods. Academic Press, New York, 1982.
2. D.P. Bertsekas. Enlarging the region of convergence of Newton's method for constrained optimization. J. Optim. Theory Appl. 36 (1982), 221–252.
3. G. Di Pillo and L. Grippo. A new class of augmented Lagrangians in nonlinear programming. SIAM J. Control Optim. 17 (1979), 618–628.
4. E.D. Dolan and J.J. More. Benchmarking optimization software with performance profiles. Math. Program. 91 (2002), 201–213.
5. F. Facchinei, A. Fischer, and M. Herrich. A family of Newton methods for nonsmooth constrained systems with nonisolated solutions. Math. Methods Oper. Res. 77 (2013), 433–443.
6. J.-Y. Fan and Y.-X. Yuan. On the quadratic convergence of the Levenberg–Marquardt method. Computing. 74 (2005), 23–39.
7. D. Fernández, E.A. Pilotta, and G.A. Torres. An inexact restoration strategy for the globalization of the sSQP method. Comput. Optim. Appl. 54 (2013), 595–617.
8. D. Fernández and M. Solodov. Stabilized sequential quadratic programming for optimization and a stabilized Newton-type method for variational problems. Math. Program. 125 (2010), 47–73.
9. A. Fischer, M. Herrich, A.F. Izmailov, and M.V. Solodov. Convergence conditions for Newton-type methods applied to complementarity systems with nonisolated solutions. Comput. Optim. Appl. 63 (2016), 425–459.
10. P.E. Gill, V. Kungurtsev, and D.P. Robinson. A stabilized SQP method: global convergence. IMA Journal of Numerical Analysis 37 (2017), 407–443.
11. W.W. Hager. Stabilized sequential quadratic programming. Comput. Optimizat. Appl. 12 (1999), 253–273.
12. A.F. Izmailov and M.V. Solodov. On attraction of Newton-type iterates to multipliers violating second-order sufficiency conditions. Math. Program. 117 (2009), 271–304.
13. A.F. Izmailov and M.V. Solodov. Stabilized SQP revisited. Math. Program. 133 (2012), 93–120.
14. A.F. Izmailov and M.V. Solodov. Critical Lagrange multipliers: what we currently know about them, how they spoil our lives, and what we can do about it. TOP 23 (2015), 1–26.

15. A.F. Izmailov and M.V. Solodov. Newton-Type Methods for Optimization and Variational Problems. Springer Series in Operations Research and Financial Engineering, Springer International Publishing, Switzerland, 2014.
16. A.F. Izmailov, M.V. Solodov, and E.I. Uskov. Combining stabilized SQP with the augmented Lagrangian algorithm. Comput. Optim. Appl. 62 (2015), 405–429.
17. A.F. Izmailov, M.V. Solodov, and E.I. Uskov. Globalizing stabilized sequential quadratic programming method by smooth primal-dual exact penalty function. J. Optim. Theory Appl. 169 (2016), 148–178.
18. A.F. Izmailov and E.I. Uskov. Subspace-stabilized sequential quadratic programming. Comput. Optim. Appl. 67 (2017), 129–154.
19. J. Nocedal and S.J. Wright. Numerical Optimization. Springer, New York, 2006. Second Edition.
20. SQPlab. `https://who.rocq.inria.fr/Jean-Charles.Gilbert/modulopt/optimization-routines/sqplab/sqplab.html`
21. S.J. Wright. Superlinear convergence of a stabilized SQP method to a degenerate solution. Comput. Optim. Appl. 11 (1998), 253–275.
22. S.J. Wright. Modifying SQP for degenerate problems SIAM J. Optim. 13 (2002), 470–497.
23. N. Yamashita and M. Fukushima. On the rate of convergence of the Levenberg–Marquardt method. Computing. Suppl. 15 (2001), 237–249.