JOTA manuscript No.
(will be inserted by the editor)

Globalizing Stabilized Sequential Quadratic Programming Method by Smooth Primal-Dual Exact Penalty Function

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

Received: date / Accepted: date

Abstract An iteration of the stabilized sequential quadratic programming method consists in solving a certain quadratic program in the primal-dual space, regularized in the dual variables. The advantage with respect to the classical sequential quadratic programming is that no constraint qualifications are required for fast local convergence (i.e., the problem can be degenerate). In particular, for equality-constrained problems the superlinear rate of convergence is guaranteed under the only assumption that the primal-dual starting point is close enough to a stationary point and a noncritical Lagrange mul-

A. F. Izmailov
VMK Faculty, OR Department, Lomonosov Moscow State University (MSU)
Uchebniy Korpus 2, Leninskiye Gory, 119991 Moscow, Russia
izmaf@ccas.ru

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

E. I. Uskov
Tambov State University
Sovetskaya Str. 93, 392000 Tambov, Russia
ydoom@narod.ru

tiplier (the latter being weaker than the second-order sufficient optimality condition). However, unlike for the usual sequential quadratic programming method, designing natural globally convergent algorithms based on the stabilized version proved quite a challenge and, currently, there are very few proposals in this direction. For equality-constrained problems, we suggest to use for the task linesearch for the smooth two-parameter exact penalty function, which is the sum of the Lagrangian with squared penalizations of the violation of the constraints and of the violation of the Lagrangian stationarity with respect to primal variables. Reasonable global convergence properties are established. Moreover, we show that the globalized algorithm preserves the superlinear rate of the stabilized sequential quadratic programming method under the weak conditions mentioned above. We also present some numerical experiments on a set of degenerate test problems.

Keywords Stabilized sequential quadratic programming · Superlinear convergence · Global convergence · Exact penalty function · Second-order sufficiency · Noncritical Lagrange multiplier

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

1 Introduction

The stabilized sequential quadratic programming method (sSQP) was first introduced in [1] for inequality-constrained optimization problems (see also [2–4] and [5, Chapter 7]), with the motivation of achieving superlinear rate of convergence in the degenerate cases (in particular, when Lagrange multipliers associated to a given primal solution may not be unique). By contrast, the classical sequential quadratic programming method (SQP) requires uniqueness of the multipliers for proving local convergence, also known as the strict Mangasarian–Fromovitz constraint qualification (which in the equality-constrained case reduces to the usual linear independence of constraints' gradients); see [5, Chapter 4]. The sSQP method was extended in [6] to general variational problems with equality and inequality constraints, with no constraint qualifications of any kind. The method converges superlinearly when initialized close enough to a stationary point-multiplier pair satisfying the second-order sufficient optimality condition [6]. In the case of equality-constrained problems, even the weaker assumption that the multiplier in question is noncritical does the job [7]. Those assertions are stated formally in Theorem 5.1 below; see [5, Chapter 7] for details. In view of these very appealing local convergence properties, it is natural to think about possible approaches to globalization of sSQP. This issue proved quite a challenge, and is currently a matter of interest for several research groups (the relevant work is cited below). The essential difficulty is that merit functions commonly used for globalization of constrained optimization algorithms, such as the l_1-penalty function for example, do not seem to be suitable – solutions of sSQP subproblems do not provide a descent direction for such functions.

In this article, we propose a globalization of sSQP using linesearch to force descent for a certain smooth two-parameter primal-dual penalty function, originally introduced in [8] (for its properties, see also [9, Section 4.3] and [10]).

We now survey the few other approaches proposed to globalize sSQP so far. One possibility is to combine sSQP with the augmented Lagrangian method. This makes sense, as the two iterative schemes are related, and the augmented Lagrangian method has strong global convergence properties on degenerate problems [11,12], which are the primary target of sSQP. In fact, sSQP can be regarded in some sense as a "linearization" of the augmented Lagrangian method. Thus, one can try to use sSQP directions for solving the augmented Lagrangian subproblems, at least when the linearization approximates well the primal-dual iteration system of the augmented Lagrangian method. The proposal to combine stabilized Newton-type steps for optimality systems with augmented Lagrangian methods dates back at least to [13] (see also [9, p. 240]). This idea is also used in the very recent works [14–17], where the so-called primal-dual augmented Lagrangian is employed. Alternatively, [18] employs the usual augmented Lagrangian for similar purposes. In both approaches, roughly speaking, sSQP steps are used as inner iterations to approximately minimize the augmented Lagrangian, with the multipliers fixed, until it is decided that the multipliers estimate can be updated. Then the process is repeated. In particular, no merit/penalty function is being minimized, and global convergence is guided by the augmented Lagrangian algorithm.

Combining sSQP with augmented Lagrangian methods is not the only possibility, of course. For example, the so-called hybrid strategies can employ the standard globalized SQP as an outer-phase algorithm, trying to switch to the full-step sSQP locally, when convergence to a degenerate solution is detected [19]. Another attempt to globalize sSQP is described in [20], where sSQP is combined with the inexact restoration strategy [21].

The approach we propose here is quite different from any of the mentioned above: we try to use sSQP directions for minimizing a merit function – specifically, the primal-dual exact penalty function of [8]. For the resulting algorithm, we obtain satisfactory global convergence properties, and also prove that the method reduces indeed to pure sSQP locally, thus inheriting its superlinear rate of convergence under the same weak assumptions. We also report numerical experiments on a collection of degenerate problems.

The rest of the paper is organized as follows. In Section 2, we state the problem, the basic sSQP scheme, and introduce the necessary notation. In Section 3, we establish the descent properties of sSQP directions with respect to the penalty function in question. With those properties at hand, we propose the globalized sSQP algorithm. In Section 4, we establish its global convergence properties. Local rate of convergence analysis is given in Section 5. We note that the latter requires some novel results concerning acceptance of the unit stepsize in linesearch descent methods, related to the classical Dennis-Moré theorem but different. In particular, unlike the Dennis-Moré theorem, our results on this subject are applicable not only to Newtonian methods and allow for non-isolated solutions. The latter feature is particularly important in the context of sSQP, motivated by degenerate problems. Numerical results on the DEGEN test collection [22] are reported in Section 6.

2 Problem Setting and Preliminaries

In this work, we consider the equality-constrained optimization problem

$$\text{minimize } f(x) \quad \text{subject to} \quad h(x) = 0, \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ are at least twice differentiable. Let $L : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}$ be the usual Lagrangian of the problem (1), i.e.,

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

Given the current primal-dual iterate $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^l$ and the value $\sigma > 0$ of the stabilization parameter, the sSQP method [5, Chapter 7] for the problem (1) solves the following QP in the primal-dual space, to generate the direction of change $(\xi, \eta) \in \mathbb{R}^n \times \mathbb{R}^l$:

$$\begin{array}{ll} \text{minimize}_{(\xi,\, \eta)} & \left[\langle f'(x), \xi \rangle + \dfrac{1}{2} \left\langle \dfrac{\partial^2 L}{\partial x^2}(x, \lambda)\xi, \xi \right\rangle + \dfrac{\sigma}{2} \|\lambda + \eta\|^2 \right] \\ \text{subject to} & h(x) + h'(x)\xi - \sigma\eta = 0. \end{array} \tag{2}$$

The pure local sSQP scheme then sets $(x + \xi, \lambda + \eta)$ as the next iterate. (Note that an iteration of the usual SQP can be formally regarded as solving (2) for $\sigma := 0$.)

In what follows, we shall develop a globally convergent algorithm using linesearch in sSQP directions for the following two-parameter primal-dual merit function [8]. Define $\varphi_{c_1, c_2} : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}$,

$$\varphi_{c_1, c_2}(x, \lambda) := L(x, \lambda) + \frac{c_1}{2}\|h(x)\|^2 + \frac{c_2}{2}\left\|\frac{\partial L}{\partial x}(x, \lambda)\right\|^2, \tag{3}$$

where $c_1 > 0$ and $c_2 > 0$ are penalty parameters. Recall that stationary points and associated Lagrange multipliers of the problem (1) are characterized by the Lagrange optimality system

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

with respect to $x \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^l$. Thus, the function in (3) is the sum of the Lagrangian for problem (1) with weighted penalizations of violations of the optimality conditions in (4). According to the theory in [9, Section 4.3] and [10], the penalty function defined in (3) is exact in the following sense. If $c_2 > 0$ is small enough and $c_1 > 0$ is large enough, every stationary point of $\varphi_{c_1,\, c_2}(\cdot)$ satisfies the Lagrange optimality system (4); if a primal-dual solution of (4) satisfies the linear independence constraint qualification and the second-order sufficient optimality condition, then it is a strict local minimizer of $\varphi_{c_1,\, c_2}(\cdot)$.

Let $\mathcal{M}(\bar{x})$ be the set of Lagrange multipliers associated with a stationary point $\bar{x}$ of the problem (1), i.e., the set of $\lambda \in \mathbb{R}^l$ satisfying (4) for $x = \bar{x}$. The following second-order conditions on the stationary point – Lagrange multiplier pairs are the key to local convergence analysis of the sSQP method. A Lagrange multiplier $\bar{\lambda} \in \mathcal{M}(\bar{x})$ satisfies the second-order sufficient optimality condition (SOSC) if

$$\left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x},\, \bar{\lambda})\xi,\, \xi \right\rangle > 0 \quad \forall\, \xi \in \ker h'(\bar{x}) \setminus \{0\}. \tag{5}$$

A Lagrange multiplier $\bar{\lambda} \in \mathcal{M}(\bar{x})$ is called critical if

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

and noncritical otherwise. As is easy to see, the multiplier is always noncritical if it satisfies the SOSC (5). We refer to [7, 23] for details on critical and noncritical multipliers and the roles of these notions for convergence of Newton-type algorithms (see also [5, Chapter 7] and the recent survey in [24]).

We finish this section with a few words about our notation (which is mostly standard). Throughout the paper $\langle\cdot,\, \cdot\rangle$ is the Euclidean inner product, and $\|\cdot\|$ is the associated norm (the space is always clear from the context). By $\mathbb{R}_+$ we denote the set of nonnegative reals. The distance from a point $u \in \mathbb{R}^\nu$ to a set $U \subset \mathbb{R}^\nu$ is defined by $\operatorname{dist}(u,\, U) := \inf_{v \in U} \|u - v\|$. The closed ball centered at $u \in \mathbb{R}^\nu$ of radius $\varepsilon > 0$ is denoted by $B(u,\, \varepsilon)$. For a linear operator A, the notation $\operatorname{im} A$ stands for its range space, and $\ker A$ for its null space. When talking about superlinear convergence, we mean the Q-rate (without saying so explicitly).

3 Globalized Stabilized SQP Algorithm

According to the theory in [9, Section 4.3] and [10], under natural assumptions mentioned in Section 2 the penalty function (3) can be considered to be exact, at least for $c_2 > 0$ small enough and $c_1 > 0$ large enough. Accordingly, when using it as a merit function for globalization of convergence of some algorithm, the intention is to keep c_2 small, increasing it only when it is really necessary. With this in mind, we first show that under certain assumptions, the sSQP direction is a descent direction for this merit function at a point violating the Lagrange optimality system (4), provided c_1 and c_2 are chosen appropriately.

Note that writing the optimality conditions for the sSQP subproblem (2), we can consider the sSQP direction $(\xi,\, \eta) \in \mathbb{R}^n \times \mathbb{R}^l$ as computed by solving the linear system

$$\frac{\partial^2 L}{\partial x^2}(x,\, \lambda)\xi + (h'(x))^{\mathrm{T}}\eta = -\frac{\partial L}{\partial x}(x,\, \lambda), \quad h'(x)\xi - \sigma\eta = -h(x). \tag{6}$$

For each $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^l$ and $\sigma > 0$, define the matrix

$$H_\sigma(x,\, \lambda) = \frac{\partial^2 L}{\partial x^2}(x,\, \lambda) + \frac{1}{\sigma}(h'(x))^{\mathrm{T}} h'(x). \tag{7}$$

It is worth mentioning that for values of $\sigma > 0$ small enough this matrix is nonsingular for all $(x,\, \lambda)$ that violate the Lagrange optimality system (4), but are close enough to one of its solutions with a noncritical multiplier; see [7].

The first two items in the following result lead to a way of choosing the penalty parameters to ensure that sSQP directions are of descent for the penalty function in consideration. Observe that as, by (6), it holds that $\langle h(x),\, h'(x)\xi\rangle = -\|h(x)\|^2 + \sigma\langle h(x),\, \eta\rangle$, the first item concerns violation of the constraints, while the second refers to violation of the Lagrangian stationarity.

Lemma 3.1 *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ be twice differentiable at a given $x \in \mathbb{R}^n$, and let $(\xi,\, \eta)$ be a solution of (6) for this x and some given $\lambda \in \mathbb{R}^l$ and $\sigma > 0$.*

Then the following assertions are valid:

(a) If $\langle h(x),\, h'(x)\xi\rangle < 0$*, then*

$$\langle \varphi'_{c_1,\, c_2}(x,\, \lambda),\, (\xi,\, \eta)\rangle \leq -\omega \tag{8}$$

holds for any $\omega > 0$*, any* $c_1 \geq \bar{c}_1(\omega;\, x,\, \lambda;\, \xi,\, \eta)$ *and any* c_2*, where*

$$\bar{c}_1(\omega;\, x,\, \lambda;\, \xi,\, \eta) := -\frac{\left\langle \frac{\partial L}{\partial x}(x,\, \lambda),\, \xi \right\rangle + \langle h(x),\, \eta\rangle + \omega}{\langle h(x),\, h'(x)\xi\rangle}. \tag{9}$$

(b) If $\frac{\partial L}{\partial x}(x,\, \lambda) \neq 0$*, then the relation* (8) *holds for any* $\omega > 0$*, any* c_1 *and any* $c_2 \geq \bar{c}_2(\omega;\, x,\, \lambda;\, \xi,\, \eta;\, c_1)$*, where*

$$\bar{c}_2(\omega;\, x,\, \lambda;\, \xi,\, \eta;\, c_1) := \frac{\left\langle \frac{\partial L}{\partial x}(x,\, \lambda),\, \xi \right\rangle + \langle h(x),\, \eta\rangle + c_1\langle h(x),\, h'(x)\xi\rangle + \omega}{\left\| \frac{\partial L}{\partial x}(x,\, \lambda) \right\|^2}. \tag{10}$$

(c) If the matrix defined in (7) *is nonsingular, then* $(\xi,\, \eta)$ *is uniquely defined by*

$$\xi = -(H_\sigma(x,\, \lambda))^{-1}\left(\frac{\partial L}{\partial x}(x,\, \lambda) + \frac{1}{\sigma}(h'(x))^{\mathrm{T}} h(x)\right), \tag{11}$$

$$\eta = \frac{1}{\sigma}\left(h(x) - h'(x)(H_\sigma(x,\, \lambda))^{-1}\left(\frac{\partial L}{\partial x}(x,\, \lambda) + \frac{1}{\sigma}(h'(x))^{\mathrm{T}} h(x)\right)\right). \tag{12}$$

Proof By direct differentiation of (3), for any c_1 and c_2 it holds that

$$\varphi'_{c_1,\, c_2}(x,\, \lambda) = \begin{pmatrix} I + c_2\frac{\partial^2 L}{\partial x^2}(x,\, \lambda) & c_1(h'(x))^{\mathrm{T}} \\ c_2 h'(x) & I \end{pmatrix}\begin{pmatrix} \frac{\partial L}{\partial x}(x,\, \lambda) \\ h(x) \end{pmatrix}. \tag{13}$$

Hence,

$$\begin{aligned} \langle \varphi'_{c_1,\, c_2}(x,\, \lambda),\, (\xi,\, \eta)\rangle &= \left\langle \frac{\partial L}{\partial x}(x,\, \lambda),\, \xi \right\rangle + \langle h(x),\, \eta\rangle + c_1\langle h(x),\, h'(x)\xi\rangle \\ &\quad + c_2\left\langle \frac{\partial L}{\partial x}(x,\, \lambda),\, \frac{\partial^2 L}{\partial x^2}(x,\, \lambda)\xi + (h'(x))^{\mathrm{T}}\eta \right\rangle \\ &= \left\langle \frac{\partial L}{\partial x}(x,\, \lambda),\, \xi \right\rangle + \langle h(x),\, \eta\rangle + c_1\langle h(x),\, h'(x)\xi\rangle \\ &\quad - c_2\left\| \frac{\partial L}{\partial x}(x,\, \lambda) \right\|^2, \end{aligned} \tag{14}$$

where the last equality is by the first equation in (6). Assertions (a) and (b) are now evident.

Note that the second equality in (6) can be written in the form

$$\eta = \frac{1}{\sigma}(h(x) + h'(x)\xi). \tag{15}$$

Substituting this into the first equality in (6) and using (7), yields

$$H_\sigma(x,\, \lambda)\xi = -\frac{\partial L}{\partial x}(x,\, \lambda) - \frac{1}{\sigma}(h'(x))^{\mathrm{T}} h(x).$$

If $H_\sigma(x,\, \lambda)$ is nonsingular, the latter equation uniquely defines ξ and then (15) uniquely defines η. The corresponding expressions give (11) and (12), respectively. This proves (c). □

The properties (a) and (b) of Lemma 3.1 readily lead to fully implementable rules for choosing c_1 and c_2 in order to ensure that the computed sSQP direction is of descent for the penalty function with the corresponding parameters. We next state the resulting algorithm.

Define $\Phi : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}^n \times \mathbb{R}^l$, the mapping of the Lagrange optimality system (4):

$$\Phi(x,\, \lambda) := \left(\frac{\partial L}{\partial x}(x,\, \lambda),\, h(x)\right). \tag{16}$$

Algorithm 3.1 Choose the parameters $\bar{\sigma} > 0$, $c_1 > 0$, $c_2 > 0$, $C_1 > 0$, $C_2 > 0$, $\delta > 0$, $\rho > 0$, $q > 1$, $\varepsilon,\, \theta \in (0,\, 1)$. Fix some continuous functions $\psi_1,\, \psi_2,\, \psi_3 : \mathbb{R}_+ \to \mathbb{R}_+$ which are positive everywhere except at 0. Choose $(x^0,\, \lambda^0) \in \mathbb{R}^n \times \mathbb{R}^l$ and set $k := 0$.

1. If $\Phi(x^k,\, \lambda^k) = 0$ where $\Phi(\cdot)$ is defined in (16), stop.
2. Set $\sigma_k := \min\{\bar{\sigma},\, \|\Phi(x^k,\, \lambda^k)\|\}$ and compute $d^k := (\xi^k,\, \eta^k)$ as a solution of the system (6) for $\sigma := \sigma_k$ and $(x,\, \lambda) := (x^k,\, \lambda^k)$. If the system (6) is not solvable, go to step 4.
3. If

$$\langle \varphi'_{c_1,\, c_2}(x^k,\, \lambda^k),\, d^k\rangle \leq -\rho\|d^k\|^q, \tag{17}$$

go to step 5.

If

$$\|h(x^k)\| \geq \psi_1(\sigma_k) \tag{18}$$

and

$$\langle h(x^k),\, h'(x^k)\xi^k\rangle \leq -\psi_2(\|h(x^k)\|), \tag{19}$$

set $c_1 := \bar{c}_{1,\,k} + \delta$, where

$$\bar{c}_{1,\,k} := \bar{c}_1(\rho\|d^k\|^q;\, x^k,\, \lambda^k;\, \xi^k,\, \eta^k).$$

If $c_1 \leq C_1$, go to step 5; otherwise, set $c_1 := C_1$ and go to step 4.

If

$$\left\|\frac{\partial L}{\partial x}(x^k,\, \lambda^k)\right\| \geq \psi_3(\sigma_k), \tag{20}$$

set $c_2 := \bar{c}_{2,\,k} + \delta$, where

$$\bar{c}_{2,\,k} := \bar{c}_2(\rho\|d^k\|^q;\, x^k,\, \lambda^k;\, \xi^k,\, \eta^k;\, c_1).$$

If $c_2 \leq C_2$, go to step 5; otherwise, set $c_2 := C_2$.

4. Choose a symmetric positive definite $(n + l) \times (n + l)$ matrix Q_k, and set

$$d^k := -Q_k\varphi'_{c_1,\,c_2}(x^k,\, \lambda^k).$$

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

$$\varphi_{c_1,\,c_2}((x^k,\, \lambda^k) + \theta^j d^k) \leq \varphi_{c_1,\,c_2}(x^k,\, \lambda^k) + \varepsilon\theta^j \left\langle \varphi'_{c_1,\,c_2}(x^k,\, \lambda^k),\, d^k\right\rangle. \tag{21}$$

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

The bounds C_1 and C_2 on the penalty parameters in Algorithm 3.1 play mostly the role of a precaution. In practice, they should be taken large, so that not to interfere with the rules of setting c_1 and c_2, except possibly in some extreme situations (as a matter of computation, one would want to avoid very large values of parameters anyway).

We emphasize that the algorithm employs the safeguarding direction on step 4 in two cases only: either when the sSQP iteration system (6) is not solvable, or when the sSQP direction does not pass the tests on step 3. According to the computational results in Section 6, this is actually quite rare, i.e., the sSQP direction exists, passes the tests, and is thus used "almost always", when we look at all iterations for all the problems.

On step 4 of Algorithm 3.1, various choices of matrices Q_k are possible. For instance, [9] suggests to use some fixed scaling matrix $Q_k := Q$ in methods considered there and also based on minimizing the same penalty function. In our numerical experiments in Section 6, we use the BFGS updates (if such iterates are consecutive), and with the Armijo linesearch rule on step 5 replaced by the Wolfe rule when the direction is computed at step 4. The Wolfe rule is used in order to ensure that Q_k obtained by the BFGS scheme is positive definite. We note that while our global convergence results in Section 4 for generic matrices Q_k refer to the Armijo rule, it is clear that they also hold if the Wolfe rule is used instead.

It is easily seen that the algorithm is well-defined: it either terminates at step 1 with a point satisfying the Lagrange optimality system (4), or a descent direction for the penalty function is obtained (possibly after updating its parameters), and then the linesearch condition (21) accepts a positive stepsize after a finite number of trial steps.

4 Global Convergence Properties

We now show that Algorithm 3.1 has reasonable global convergence properties. Observe that if the matrix in the right-hand side of (13) is nonsingular for $(x, \lambda) = (\bar{x}, \bar{\lambda})$, then $(\bar{x}, \bar{\lambda})$ satisfying the stationarity condition (22) below necessarily satisfies the Lagrange optimality system (4). It can be further seen that the matrix in question is nonsingular for generic values of c_1 and c_2 (i.e., "almost always").

Theorem 4.1 *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ be twice continuously differentiable on $\mathbb{R}^n$. Let $\{(x^k, \lambda^k)\}$ be an iterative sequence generated by Algorithm 3.1. Assume that matrices Q_k chosen on step 4 of the algorithm are uniformly positive definite and bounded.*

Then every accumulation point $(\bar{x}, \bar{\lambda})$ *of the sequence* $\{(x^k, \lambda^k)\}$ *satisfies*

$$\varphi'_{c_1, c_2}(\bar{x}, \bar{\lambda}) = 0. \tag{22}$$

Proof At each iteration of the algorithm, the values of c_1 and c_2 either remain unchanged, or one (and only one) of these parameters increases.

Observe that any of the penalty parameters can only be changed if the descent direction condition (17) is not satisfied with the previous values. By assertion (a) of Lemma 3.1, in the case of (19) this means that before the update we had $c_1 < \bar{c}_1(\rho\|d^k\|^q;\, x^k,\, \lambda^k;\, \xi^k,\, \eta^k)$, as otherwise (17) would have been satisfied. This shows that the update of c_1 in step 3 either increases its value by some quantity bigger than $\delta > 0$, or yields $c_1 = C_1$. In the latter case, no more changes of c_1 are allowed. Similarly, (17) not being satisfied in the case of (20) means that for the previous value it holds that $c_2 < \bar{c}_2(\rho\|d^k\|^q;\, x^k,\, \lambda^k;\, \xi^k,\, \eta^k;\, c_1)$, by assertion (b) of Lemma 3.1. This shows that the update of c_2 in step 3 also either increases its value by some quantity bigger than $\delta > 0$, or yields $c_2 = C_2$. Again, in the latter case, no more changes of c_2 are allowed.

Note also that, as long as at least one of the new values of c_1 or c_2 remains smaller than C_1 and C_2, respectively, the descent direction condition (17) is satisfied for these new values. In particular, the number of times c_1 is increased cannot be greater than C_1/δ. Similarly, the number of times c_2 is increased cannot be greater than C_2/δ. Therefore, the values of c_1 and c_2 do not change for all sufficiently large k.

Since the values of c_1 and c_2 do not change for all sufficiently large k, the "tail" of the sequence $\{(x^k, \lambda^k)\}$ is generated by a descent method with Armijo linesearch for a fixed smooth function φ_{c_1, c_2}.

We next show that the sequence $\{d^k\}$ of search directions is uniformly gradient-related in the terminology of [9, p. 24], which means that if for some infinite set K of iteration indices the subsequence $\{(x^k, \lambda^k) : k \in K\}$ converges to some point $(\bar{x}, \bar{\lambda})$ such that $\varphi'_{c_1, c_2}(\bar{x}, \bar{\lambda}) \neq 0$, then $\{d^k : k \in K\}$ is bounded and $\limsup_{K \ni k \to \infty} \langle \varphi'_{c_1, c_2}(x^k, \lambda^k),\, d^k \rangle < 0$.

Let $K_1 \subset K$ be a sequence of all indices $k \in K$ such that d^k is generated as a solution of the system (6). Then (17) holds for all $k \in K_1$. Hence, $\|\varphi'_{c_1, c_2}(x^k, \lambda^k)\|\|d^k\| \geq \rho\|d^k\|^q$. Employing the assumption that $q > 1$ and the continuity of φ'_{c_1, c_2} at $(\bar{x}, \bar{\lambda})$, this evidently implies that $\{d^k : k \in K_1\}$ is bounded. Furthermore, since $\varphi'_{c_1, c_2}(\bar{x}, \bar{\lambda}) \neq 0$, from (13) we conclude that $\Phi(\bar{x}, \bar{\lambda}) \neq 0$. It then follows from (6) that $\{d^k : k \in K_1\}$ cannot have zero as an accumulation point. Therefore, (17) implies that the subsequence $\{\langle \varphi'_{c_1, c_2}(x^k, \lambda^k), d^k\rangle : k \in K_1\}$ is separated from zero by some negative constant. It remains to note that for $k \in K \setminus K_1$ the search direction d^k is obtained on step 4 of Algorithm 3.1. That such directions are uniformly gradient-related, is well known; e.g., [9, p. 24].

Therefore, we conclude that $\{d^k\}$ is a uniformly gradient-related sequence of search directions. Hence, by [9, Theorem 1.8], every accumulation point $(\bar{x}, \bar{\lambda})$ of the sequence $\{(x^k, \lambda^k)\}$ satisfies (22). □

5 Convergence Rate Analysis

We now turn our attention to convergence rate analysis of Algorithm 3.1. The objective is to show that close to a solution with certain properties (in particular, those that guarantee the superlinear rate for pure sSQP), the algorithm accepts the sSQP direction and then the unit stepsize in this direction.

We start with recalling sSQP local convergence results in Section 5.1, which we would like our global algorithm to inherit. In Section 5.2, we present an interesting analysis concerning acceptance of the unit stepsize in generic descent methods, in the spirit of the Dennis-Moré theorem for Newtonian methods, but different – the method need not be Newtonian, and in fact, second derivatives need not even exist; moreover, solutions need not be isolated (the latter is an important feature in the context of sSQP). Then, Section 5.3 presents the rate of convergence results for Algorithm 3.1.

5.1 Local Convergence of Pure Stabilized SQP

Let Φ be the mapping of the Lagrange optimality system (4), defined in (16). The following is [7, Theorem 1] with some quantitative refinements which can be derived from the proof of [3, Theorem 1] or [25, Theorem 4.1]; see also [5, Chapter 7].

Theorem 5.1 *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ be twice differentiable in a neighbourhood of $\bar{x} \in \mathbb{R}^n$, with their second derivatives being continuous at $\bar{x}$. Let $\bar{x}$ be a stationary point of problem* (1) *with an associated noncritical Lagrange multiplier $\bar{\lambda} \in \mathbb{R}^l$.*

Then the following assertions are valid:

(a) For some neighbourhood $\mathcal{V}$ of the point $(\bar{x}, \bar{\lambda})$ there exists the unique mapping $d(\cdot) := (\xi(\cdot), \eta(\cdot)) : \mathcal{V} \to \mathbb{R}^n \times \mathbb{R}^l$ with the following properties: $(\xi(x, \lambda), \eta(x, \lambda))$ satisfies (6) *with $\sigma := \|\Phi(x, \lambda)\|$ for every $(x, \lambda) \in \mathcal{V}$, and $d(\bar{x}, \lambda) = 0$ if $\lambda \in \mathcal{M}(\bar{x})$.*

(b) The neighbourhood $\mathcal{V}$ can be chosen small enough, so that there exist $\ell > 0$ and a function $\chi : \mathbb{R}_+ \to \mathbb{R}_+$ continuous at zero and such that $\chi(t) = o(t)$ as $t \to 0$, and

$$\|d(x, \lambda)\| \leq \ell \|\Phi(x, \lambda)\|, \tag{23}$$

$$\|x + \xi(x, \lambda) - \bar{x}\| + \operatorname{dist}(\lambda + \eta(x, \lambda), \mathcal{M}(\bar{x})) \leq \chi(\|x - \bar{x}\| + \operatorname{dist}(\lambda, \mathcal{M}(\bar{x}))) \tag{24}$$

for all $(x, \lambda) \in \mathcal{V}$.

(c) For any $\varepsilon > 0$ there exists $\varepsilon_0 > 0$ such that for any $(x^0, \lambda^0) \in B((\bar{x}, \bar{\lambda}), \varepsilon_0)$ the method's iterative sequence $\{(x^k, \lambda^k)\}$ is uniquely defined by the equality $(x^{k+1}, \lambda^{k+1}) = (x^k, \lambda^k) + d(x^k, \lambda^k)$ for all k, this sequence satisfies $\{(x^k, \lambda^k)\} \subset B((\bar{x}, \bar{\lambda}), \varepsilon)$; and it converges superlinearly to $(\bar{x}, \lambda^*)$ for some $\lambda^* := \lambda^*(x^0, \lambda^0) \in \mathcal{M}(\bar{x})$.

5.2 On the Acceptance of the Unit Stepsize in Generic Descent Methods

Consider the unconstrained optimization problem

$$\text{minimize } \varphi(u), \quad u \in \mathbb{R}^\nu, \tag{25}$$

where the objective function $\varphi : \mathbb{R}^\nu \to \mathbb{R}$ is smooth. Consider a generic descent method for the problem (25). Specifically, take any $u^0 \in \mathbb{R}^\nu$, and for $k = 0, 1, \ldots$, set

$$u^{k+1} := u^k + \alpha_k d^k,$$

where $d^k \in \mathbb{R}^\nu$ is some direction of descent for φ at u^k, and the stepsize α_k is computed by the Armijo linesearch rule: for some fixed parameters ε and θ such that $0 < \varepsilon < 1$ and $0 < \theta < 1$, set $\alpha_k := \theta^j$, where j is the smallest nonnegative integer satisfying the inequality

$$\varphi(u^k + \theta^j d^k) \leq \varphi(u^k) + \varepsilon\theta^j \langle \varphi'(u^k), d^k\rangle.$$

We start with a simple proposition, which can be considered as a first-order version of the part of the Dennis–Moré theorem (see, e.g, [9, Proposition 1.15]) concerned with acceptance of the unit stepsize by linesearch methods of the kind stated above. As the Dennis–Moré theorem would be invoked for comparisons several times, we state a part of it for the readers' convenience.

Proposition 5.1 *Let $\varphi : \mathbb{R}^\nu \to \mathbb{R}$ be twice differentiable in a neighbourhood of $\bar{u} \in \mathbb{R}^\nu$, with its second derivative being continuous at $\bar{u}$. Let $\bar{u}$ be a stationary point of problem (25), i.e., $\varphi'(\bar{u}) = 0$, and let the second-order sufficient optimality condition hold at $\bar{u}$, i.e., $\varphi''(\bar{u})$ is positive definite. Let $\{u^k\}$ be an iterative sequence of the descent method specified above, where $0 < \varepsilon < 1/2$. Assume that $\{u^k\}$ converges to $\bar{u}$, and*

$$\|d^k + (\varphi''(\bar{u}))^{-1}\varphi'(u^k)\| = o(\|\varphi'(u^k)\|) \tag{26}$$

as $k \to \infty$.

Then $\alpha_k = 1$ for all k large enough.

Proposition 5.2 *Let* $\varphi : \mathbb{R}^\nu \to \mathbb{R}$ *be differentiable in a neighbourhood of* $\bar{u} \in \mathbb{R}^\nu$. *Let* $\{u^k\} \subset \mathbb{R}^\nu$ *be any sequence convergent to* $\bar{u}$, *and let* $\{d^k\} \subset \mathbb{R}^\nu$ *be any sequence convergent to zero. Suppose that there exists* $\beta > 0$ *such that*

$$\varphi(u^k + d^k) - \varphi(u^k) \le \beta\langle \varphi'(u^k),\, d^k\rangle + o(\|d^k\|^2) \tag{27}$$

as $k \to \infty$, *and there exists* $\gamma > 0$ *such that*

$$\langle \varphi'(u^k),\, d^k\rangle \le -\gamma\|d^k\|^2 \tag{28}$$

for all k *large enough.*

Then for any ε *satisfying* $0 < \varepsilon < \beta$, *it holds that*

$$\varphi(u^k + d^k) \le \varphi(u^k) + \varepsilon\langle \varphi'(u^k),\, d^k\rangle \tag{29}$$

for all k *large enough.*

Proof By simply combining (27) with (28), we obtain

$$\begin{aligned} \varphi(u^k + d^k) - \varphi(u^k) - \varepsilon\langle \varphi'(u^k),\, d^k\rangle &\le (\beta - \varepsilon)\langle \varphi'(u^k),\, d^k\rangle + o(\|d^k\|^2) \\ &\le -\gamma(\beta - \varepsilon)\|d^k\|^2 + o(\|d^k\|^2) \\ &\le 0 \end{aligned}$$

for all k large enough, giving the needed conclusion. □

Assumptions (27) and (28) may look somewhat unexpected (especially (27)), but they are justified (at least) by Corollary 5.1 below, which extends the Dennis–Moré conditions to the case of possibly non-isolated solutions. The key features of the simple fact stated in Proposition 5.2, as compared to other somewhat related results, are the following. First, the direction d^k is not connected to the gradient $\varphi'(u^k)$ in any way other than (27) and (28). Second, twice differentiability of φ is not assumed, and therefore, neither the Dennis–Moré condition (26) nor the positive definiteness of $\varphi''(\bar{u})$ can be invoked in this setting. Moreover, the method in question may have nothing to do with the Newton or quasi-Newton methods, and thus the full steps need not result in the superlinear convergence rate, in principle. In fact, the second derivative of φ may not even exist near $\bar{u}$, and the (usual) Newton method may simply be not defined.

The following fact follows from Proposition 5.2. It is already somewhat closer to the setting of the Dennis–Moré theorem, but still covers a much wider territory.

Corollary 5.1 *Let $\varphi : \mathbb{R}^\nu \to \mathbb{R}$ be twice differentiable in a neighbourhood of a point $\bar{u} \in \bar{U} := \{u \in \mathbb{R}^\nu : \varphi'(u) = 0\}$, with its second derivative being continuous at $\bar{u}$. Let $\{u^k\} \subset \mathbb{R}^\nu$ be any sequence convergent to $\bar{u}$, and let $\{d^k\} \subset \mathbb{R}^\nu$ be any sequence convergent to zero. Assume that*

$$\mathrm{dist}(u^k + d^k, \bar{U}) = o(\mathrm{dist}(u^k, \bar{U})) \tag{30}$$

as $k \to \infty$, and there exists $\gamma > 0$ such that

$$\langle \varphi''(\bar{u})d^k, d^k\rangle \geq \gamma \|d^k\|^2 \tag{31}$$

for all k large enough.

Then for any ε such that $0 < \varepsilon < 1/2$, the inequality (29) holds for all k large enough.

Proof We shall show that the assumptions of this corollary imply the assumptions of Proposition 5.2 for $\beta := 1/2$.

For each k, let $\hat{u}^k$ be any projection of $u^k + d^k$ onto $\bar{U}$ (note that $\bar{U}$ need not be convex, so the projection need not be unique; any one can be taken as $\hat{u}^k$). Observe that since $\{u^k\} \to \bar{u}$ and $\{d^k\} \to 0$, it holds that $\{\hat{u}^k\} \to \bar{u}$. From (30) we then have that

$$\|u^k + d^k - \hat{u}^k\| = o(\mathrm{dist}(u^k, \bar{U})) = o(\|u^k - \hat{u}^k\|), \tag{32}$$

evidently implying that

$$u^k - \hat{u}^k = -d^k + o(\|d^k\|) \tag{33}$$

as $k \to \infty$.

By the mean-value theorem and by the continuity of φ'' at $\bar{u}$, it holds that

$$\varphi'(u^k+d^k) = \varphi'(u^k+d^k)-\varphi'(\hat{u}^k) = O(\|u^k+d^k-\hat{u}^k\|) = o(\|u^k-\hat{u}^k\|) = o(\|d^k\|)$$

as $k \to \infty$, where the last two estimates are by (32) and (33), respectively. On the other hand,

$$\varphi'(u^k + d^k) = \varphi'(u^k) + \varphi''(u^k)d^k + o(\|d^k\|),$$

and combining this with the previous estimate we get

$$\varphi'(u^k) + \varphi''(u^k)d^k = o(\|d^k\|) \tag{34}$$

as $k \to \infty$. Furthermore,

$$\begin{aligned} \varphi(u^k + d^k) - \varphi(u^k) &= \langle \varphi'(u^k),\, d^k\rangle + \frac{1}{2}\langle \varphi''(u^k)d^k,\, d^k\rangle + o(\|d^k\|^2) \\ &= \frac{1}{2}\langle \varphi'(u^k),\, d^k\rangle + o(\|d^k\|^2) \end{aligned} \tag{35}$$

as $k \to \infty$, where the last equality is by (34). This gives (27) with $\beta := 1/2$.

Finally, again employing (34), we obtain that

$$\langle \varphi'(u^k),\, d^k\rangle = -\langle \varphi''(\bar{u})d^k,\, d^k\rangle + o(\|d^k\|^2)$$

as $k \to \infty$, and therefore, (31) implies (28) with some $\gamma > 0$. □

The key differences between Corollary 5.1 and Proposition 5.1 (part of the Dennis–Moré theorem) are the following. Instead of the positive definiteness of $\varphi''(\bar{u})$, here it is directly assumed that the full-step method provides superlinear decrease of the distance to the solution set, and that the quadratic form given by the Hessian of the objective function is uniformly positive in the directions employed by the method. What is important is that this may be applicable when $\varphi''(\bar{u})$ is not positive definite, and when the stationary point in question is non-isolated. The latter is the key feature that would be needed for our developments below (recall that our objective is to establish superlinear convergence of a globalized sSQP algorithm, sSQP being specifically motivated by degenerate problems with non-isolated solutions).

We note that it can be verified that if φ is twice differentiable in a neighbourhood of $\bar{u}$, with its second derivative being continuous at $\bar{u}$, and if (31) holds (in particular, if $\varphi''(\bar{u})$ is positive definite), then the assumptions of Proposition 5.2 cannot hold with $\beta > 1$.

It is almost obvious that the assumptions of Proposition 5.1 imply the assumptions of Corollary 5.1, and hence, also of Proposition 5.2. At the same time, Proposition 5.2 can be applicable when Corollary 5.1 is not (and even less so is Proposition 5.1), even in the case when φ is twice differentiable in a neighbourhood of $\bar{u}$, with its second derivative being continuous at $\bar{u}$, and with positive definite $\varphi''(\bar{u})$. Indeed, under these assumptions, consider, for example, the scaled Newton directions $d^k := -\tau(\varphi''(u^k))^{-1}\varphi'(u^k)$ with the scaling factor τ satisfying $0 < \tau < 2$. Then

$$\langle \varphi'(u^k),\, d^k\rangle = -\frac{1}{\tau}\langle \varphi''(u^k)d^k,\, d^k\rangle = -\frac{1}{\tau}\langle \varphi''(\bar{u})d^k,\, d^k\rangle + o(\|d^k\|^2)$$

as $k \to \infty$, implying that (28) holds with some $\gamma > 0$. Furthermore, again employing the first equality in (35), we derive that

$$\varphi(u^k + d^k) - \varphi(u^k) = \left(1 - \frac{\tau}{2}\right)\langle \varphi'(u^k),\, d^k\rangle + o(\|d^k\|^2)$$

as $k \to \infty$, implying that (27) holds with any β satisfying $0 < \beta < 1 - \tau/2$. Therefore, according to Proposition 5.2, the method employing such scaled Newton directions will ultimately accept the unit stepsize. On the other hand, this scaled full-step method does not possess superlinear convergence rate unless $\tau = 1$. Indeed, if $\tau \neq 1$, then (34) does not hold, implying that (30) does not hold as well, and hence, Corollary 5.1 is not applicable.

5.3 Rate of Convergence of Globalized Stabilized SQP

Appropriate starting values of c_1 and c_2 in Algorithm 3.1 can be obtained using the techniques described in [10]. Allowing to increase these values (especially c_2, which is undesirable) is, of course, not needed for proving global convergence. However, the goal is also to ensure that the full sSQP steps are accepted near a qualified solution, i.e., near $(\bar{x}, \bar{\lambda})$ satisfying the Lagrange system (4), and such that $\bar{\lambda}$ is a noncritical Lagrange multiplier associated to $\bar{x}$. Accomplishing this without increasing c_2 is often possible, but may lead to a prohibitively rapid growth of c_1. The following example concerning the choices of penalty parameters is instructive.

Example 5.1 This is problem 20101 from the DEGEN test collection [22]. Let $n := l := 1$, $f(x) := h(x) := x^2$. Then $\bar{x} := 0$ is the unique solution of problem (1), the associated Lagrange multiplier set is the entire $\mathbb{R}$, and $\lambda := -1$ is the unique critical multiplier.

If $1+\lambda+2x^2/\sigma \neq 0$, the sSQP iteration system (6) has the unique solution

$$\xi = -S(\sigma;\, x,\, \lambda)x, \quad \eta = \frac{1}{\sigma}(1 - 2S(\sigma;\, x,\, \lambda))x^2, \tag{36}$$

where

$$S(\sigma;\, x,\, \lambda) := \frac{1+\lambda+x^2/\sigma}{1+\lambda+2x^2/\sigma}.$$

Assuming that $x \neq 0$ is close to 0, and λ is close to any multiplier $\bar{\lambda} \neq -1$, we have that $\sigma := \|\Phi(x,\, \lambda)\|$ is of order $|x|$, and $h(x) = O(\sigma^2)$. Therefore, the test (18) might be passed at $(x^k,\, \lambda^k) := (x,\, \lambda)$ if we take $\psi_1(t) = O(t^2)$. Furthermore, $S(\sigma;\, x,\, \lambda) = 1 + O(|x|)$, and hence,

$$h(x)h'(x)\xi = -2S(\sigma;\, x,\, \lambda)x^4 = -2x^4 + O(x^5).$$

Therefore, the test (19) might also be passed if we take $\psi_2(t) = O(t^2)$ (quite a natural choice), since in this case, $\psi_2(\|h(x)\|) = O(x^4)$. At the same time (by direct computation or employing (14)),

$$\begin{aligned}
\langle \varphi'_{c_1,\, c_2}(x,\, \lambda),\, (\xi,\, \eta)\rangle &= -2S(\sigma;\, x,\, \lambda)x^2(1+\lambda) + \frac{1}{\sigma}(1 - 2S(\sigma;\, x,\, \lambda))x^4 \\
&\quad -2c_1 S(\sigma;\, x,\, \lambda)x^4 - 4c_2 x^2(1+\lambda)^2 \\
&= -2(1+O(|x|))x^2(1+\lambda) - \frac{1}{\sigma}(1+O(|x|))x^4 \\
&\quad -2c_1(1+O(|x|))x^4 - 4c_2 x^2(1+\lambda)^2 \\
&= -2x^2(1+\lambda) + O(|x|^3) - 2c_1(x^4 + O(|x|^5)) \\
&\quad -4c_2 x^2(1+\lambda)^2.
\end{aligned}$$

If $\bar{\lambda} < -1$, which means that this multiplier violates SOSC (5), then in order to make the expression in the right-hand side negative by only selecting c_1 (i.e., when $c_2 \geq 0$ is some fixed small number), one has to take c_1 of order $1/x^2$, which also guarantees (17) provided $q \geq 2$ (if $q < 2$, then c_1 would have to grow even faster since $\|(\xi, \eta)\|$ is of order $|x|$). At the same time, in this case the needed descent property can be achieved by taking sufficiently large but *bounded* c_2. In order for this scenario to take place, we need to make the decrease of $\psi_1(t)$ slower than t^2 as $t \to 0+$, and $\psi_3(t) = O(t)$: with these choices, the test (18) will eventually fail, while the test (20) might be passed.

We start our local analysis with the following result concerning the acceptance of sSQP directions.

Lemma 5.1 *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *and* $h : \mathbb{R}^n \to \mathbb{R}^l$ *be twice differentiable in a neighbourhood of* $\bar{x} \in \mathbb{R}^n$*, with their second derivatives being continuous at* $\bar{x}$*. Let* $\bar{x}$ *be a stationary point of problem* (1) *with an associated noncritical Lagrange multiplier* $\bar{\lambda} \in \mathbb{R}^l$*. Let the functions* ψ_1*,* ψ_2*,* ψ_3 *in Algorithm 3.1 be taken as follows:* $\psi_1(t) := \rho_1 t$*,* $\psi_2(t) := \rho_2 t^2$*,* $\psi_3(t) := \rho_3 t$*, where* $\rho_1, \rho_3 > 0$*,* $\rho_1^2 + \rho_3^2 < 1$*,* $0 < \rho_2 < 1$*.*

Then for any $(x^k, \lambda^k) \in \mathbb{R}^n \times \mathbb{R}^l$ *violating the Lagrange system* (4) *and close enough to* $(\bar{x}, \bar{\lambda})$*, the system* (6) *with* $\sigma := \sigma_k$ *and* $(x, \lambda) := (x^k, \lambda^k)$ *has the unique solution* $d^k := (\xi^k, \eta^k)$*, and* (18) *and* (19) *are satisfied, or* (20) *is satisfied.*

Moreover, if $q \geq 2$*, then the quantity*

$$\bar{c}_1(\rho\|d^k\|^q;\, x^k,\, \lambda^k;\, \xi^k,\, \eta^k)$$

is uniformly bounded for all (x^k, λ^k) *close enough to* $(\bar{x}, \bar{\lambda})$ *and satisfying* (18) *and* (19)*, and the quantity*

$$\bar{c}_2(\rho\|d^k\|^q;\, x^k,\, \lambda^k;\, \xi^k,\, \eta^k;\, c_1)$$

is uniformly bounded for all (x^k, λ^k) *close enough to* $(\bar{x}, \bar{\lambda})$ *and satisfying* (20)*, and for bounded values of* c_1*. In particular, if* C_1 *and* C_2 *are chosen large* *enough, then for any* (x^k, λ^k) *close enough to* $(\bar{x}, \bar{\lambda})$*, Algorithm 3.1 accepts the sSQP direction* d^k *and sets* c_1 *and* c_2 *such that* (17) *holds for this direction.*

Proof The existence and uniqueness of d^k solving (6) is ensured by Theorem 5.1 (a).

With the specified choices of ψ_1 and ψ_3, if both tests (18) and (20) are violated for (x^k, λ^k) close enough to $(\bar{x}, \bar{\lambda})$, then

$$\sigma_k^2 = \|\Phi(x^k, \lambda^k)\|^2 \leq (\rho_1^2 + \rho_3^2)\sigma_k^2 < \sigma_k^2$$

(recall that $\rho_1^2 + \rho_3^2 < 1$), giving a contradiction. Therefore, at least one of these tests must be satisfied.

Suppose that (18) holds. According to the estimate (23) in Theorem 5.1 (b), there exists $\ell > 0$ such that the bound

$$\|(\xi^k, \eta^k)\| \leq \ell \|\Phi(x^k, \lambda^k)\| \tag{37}$$

holds for (x^k, λ^k) close enough to $(\bar{x}, \bar{\lambda})$. Then, taking into account the second equation in (6), we obtain that

$$\begin{aligned}
\langle h(x^k),\, h'(x^k)\xi^k\rangle &= \langle h(x^k),\, -h(x^k) + \sigma_k \eta^k\rangle \\
&\leq -\|h(x^k)\|^2 + \sigma_k \|h(x^k)\| \|\eta^k\| \\
&\leq -\|h(x^k)\|^2 + \ell \sigma_k^2 \|h(x^k)\| \\
&\leq -\|h(x^k)\|^2 + \frac{\ell}{\rho_1^2} \|h(x^k)\|^3 \\
&= -\left(1 - \frac{\ell}{\rho_1^2}\|h(x^k)\|\right) \|h(x^k)\|^2,
\end{aligned}$$

implying that (19) with the specified ψ_2 is satisfied for x^k close enough to $\bar{x}$.

Furthermore, by the bound (37),

$$\left\langle \frac{\partial L}{\partial x}(x^k, \lambda^k),\, \xi^k \right\rangle + \langle h(x^k),\, \eta^k\rangle = O(\|\Phi(x^k, \lambda^k)\|^2)$$

as $(x^k, \lambda^k) \to (\bar{x}, \bar{\lambda})$, and the assertion concerning boundedness of $\bar{c}_1(\cdot)$ and $\bar{c}_2(\cdot)$ readily follows from (9), (10), and (37). The last assertion of the lemma is now evident. □

It remains to show that the full stepsize $\alpha_k = 1$ will be ultimately accepted by the Armijo rule, when the method converges to $(\bar{x}, \bar{\lambda})$ satisfying the Lagrange system (4) with a noncritical multiplier $\bar{\lambda}$.

Lemma 5.2 *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ be twice differentiable in a neighbourhood of $\bar{x} \in \mathbb{R}^n$, with their second derivatives being continuous at $\bar{x}$. Let $\bar{x}$ be a stationary point of problem* (1) *with an associated Lagrange multiplier $\bar{\lambda} \in \mathbb{R}^l$. Let $\sigma : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}_+$ be any function such that $\sigma(u) \to 0$ as $u \to \bar{u} := (\bar{x}, \bar{\lambda})$ and $\sigma(\bar{u}) = 0$, and assume that for each $u := (x, \lambda)$ close enough to $\bar{u}$ there exists $d(u) := (\xi, \eta)$ satisfying* (6) *with $\sigma := \sigma(u)$, and such that $d(u) \to 0$ as $u \to \bar{u}$ and $d(\bar{u}) = 0$.*

Then for any values of $c_1 > 0$ and $c_2 > 0$ there exists a neighbourhood $\mathcal{V}$ of $\bar{u}$ and a function $\chi : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}_+$, continuous at zero, and such that $\chi(u) = o(\|d(u)\|^2)$ as $u \to \bar{u}$, and

$$\varphi_{c_1,\, c_2}(u + d(u)) - \varphi_{c_1,\, c_2}(u) \leq \frac{1}{2} \left\langle \varphi'_{c_1,\, c_2}(u),\, d(u) \right\rangle + \chi(u) \tag{38}$$

for all $u \in \mathcal{V}$.

Proof Considering (6), we obtain the existence of a neighbourhood $\mathcal{V}_1$ of $\bar{u}$ and a function $\chi_1 : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}_+$ such that $d(u) := (\xi, \eta)$ with the required properties exists for each $u \in \mathcal{V}_1$, $\chi_1(u) = o(\|\xi\|)$ as $u \to \bar{u}$, and

$$\begin{aligned} \left\| \frac{\partial L}{\partial x}(x + \xi,\, \lambda + \eta) \right\| &\leq \left\| \frac{\partial L}{\partial x}(x,\, \lambda) + \frac{\partial^2 L}{\partial x^2}(x,\, \lambda)\xi + (h'(x))^{\mathrm{T}} \eta \right\| + \chi_1(u) \\ &= \chi_1(u), \end{aligned} \tag{39}$$

$$\|h(x + \xi)\| \leq \|h(x) + h'(x)\xi\| + \chi_1(u) = \sigma(u)\|\eta\| + \chi_1(u) \tag{40}$$

for all $u \in \mathcal{V}_1$.

Furthermore, from (40) we obtain that

$$L(x+\xi,\, \lambda+\eta) = L(x+\xi,\, \lambda) + \langle \eta,\, h(x+\xi) \rangle \leq L(x+\xi,\, \lambda) + \sigma(u)\|\eta\|^2 + \chi_1(u)\|\eta\|$$

for all $u \in \mathcal{V}_1$, and hence, there exists a neighbourhood $\mathcal{V} \subset \mathcal{V}_1$ of $\bar{u}$ and a function $\chi_2 : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}_+$ such that $\chi_2(u) = o(\|d(u)\|^2)$ as $u \to \bar{u}$, and

$$L(x+\xi,\, \lambda+\eta) - L(x,\, \lambda) \leq \left\langle \frac{\partial L}{\partial x}(x,\, \lambda),\, \xi \right\rangle + \frac{1}{2} \left\langle \frac{\partial^2 L}{\partial x^2}(x,\, \lambda)\xi,\, \xi \right\rangle + \chi_2(u) \tag{41}$$

for all $u \in \mathcal{V}$. On the other hand, again employing (6), we have that

$$\begin{aligned}\left\langle \frac{\partial L}{\partial x}(x,\,\lambda)+\frac{\partial^2 L}{\partial x^2}(x,\,\lambda)\xi,\,\xi\right\rangle &= -\langle (h'(x))^{\mathrm{T}}\eta,\,\xi\rangle \\ &= \langle \eta,\,h(x)-\sigma(u)\eta\rangle \\ &\leq \langle \eta,\,h(x)\rangle+\sigma(u)\|\eta\|^2,\end{aligned}$$

which further implies that

$$\begin{aligned}\langle L'(x,\,\lambda),\,d(u)\rangle &= \left\langle \frac{\partial L}{\partial x}(x,\,\lambda),\,\xi\right\rangle+\langle h(x),\,\eta\rangle \\ &\geq 2\left\langle \frac{\partial L}{\partial x}(x,\,\lambda),\,\xi\right\rangle+\left\langle \frac{\partial^2 L}{\partial x^2}(x,\,\lambda)\xi,\,\xi\right\rangle-\sigma(u)\|\eta\|^2\end{aligned}$$

for all $u \in \mathcal{V}_1$. Combining the latter with (41), we obtain the estimate

$$L(x+\xi,\,\lambda+\eta)-L(x,\,\lambda) \leq \frac{1}{2}\langle L'(x,\,\lambda),\,d(u)\rangle+\frac{1}{2}\sigma(u)\|\eta\|^2+\chi_2(u) \qquad (42)$$

for all $u \in \mathcal{V}$.

From (3), (39), (40), and (42), it follows that

$$\begin{aligned}\varphi_{c_1,\,c_2}(u+d(u))-\varphi_{c_1,\,c_2}(u) \leq\ &\frac{1}{2}\langle L'(x,\,\lambda),\,d(u)\rangle+\frac{1}{2}\sigma(u)\|\eta\|^2+\chi_2(u) \\ &+\frac{c_1}{2}(\sigma(u)\|\eta\|+\chi_1(u))^2+\frac{c_2}{2}(\chi_1(u))^2 \\ &-\frac{c_1}{2}\|h(x)\|^2-\frac{c_2}{2}\left\|\frac{\partial L}{\partial x}(x,\,\lambda)\right\|^2\end{aligned}$$

for all $u \in \mathcal{V}$.

Observe that for the function $\tilde{\chi}: \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}_+$ defined by

$$\tilde{\chi}(u) := \frac{c_1}{2}(\sigma(u)\|\eta\|+\chi_1(u))^2+\frac{c_2}{2}(\chi_1(u))^2+\frac{1}{2}\sigma(u)\|\eta\|^2+\chi_2(u),$$

it holds that $\tilde{\chi}(u) = o(\|d(u)\|^2)$ as $u \to \bar{u}$. Furthermore, recall that according to (14),

$$\langle \varphi'_{c_1,\,c_2}(x,\,\lambda),\,d(u)\rangle = \langle L'(x,\,\lambda),\,d(u)\rangle+c_1\langle h(x),\,h'(x)\xi\rangle-c_2\left\|\frac{\partial L}{\partial x}(x,\,\lambda)\right\|^2,$$

and hence, again employing (6),

$$\begin{aligned}
&\varphi_{c_1,\,c_2}(u+d(u)) - \varphi_{c_1,\,c_2}(u) - \frac{1}{2}\langle \varphi'_{c_1,\,c_2}(u),\, d(u)\rangle \\
&\le -\frac{c_1}{2}(\|h(x)\|^2 + \langle h(x),\, h'(x)\xi\rangle) + \tilde{\chi}(u) \\
&= -\frac{c_1}{2}\langle h(x),\, h(x) + h'(x)\xi\rangle + \tilde{\chi}(u) \\
&= -\frac{c_1}{2}\langle h(x),\, \sigma(u)\eta\rangle + \tilde{\chi}(u) \\
&= \sigma(u)\frac{c_1}{2}\langle h'(x)\xi - \sigma(u)\eta,\, \eta\rangle + \tilde{\chi}(u) \\
&\le \frac{c_1}{2}\sigma(u)(\|h'(x)\xi\| + \sigma(u)\|\eta\|)\|\eta\| + \tilde{\chi}(u)
\end{aligned}$$

for all $u \in \mathcal{V}$. This gives (38), once we define $\chi : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}_+$ as

$$\chi(u) := \frac{c_1}{2}\sigma(u)(\|h'(x)\xi\| + \sigma(u)\|\eta\|)\|\eta\| + \tilde{\chi}(u),$$

observing also that $\chi(u) = o(\|d(u)\|^2)$ as $u \to \bar{u}$. □

Assuming now that $q := 2$ instead of $q \ge 2$, and combining Lemmas 5.1 and 5.2 with Proposition 5.2, we obtain the following.

Theorem 5.2 *In addition to the assumptions of Lemma 5.1, suppose that in Algorithm 3.1 we take $q := 2$ and $0 < \varepsilon < 1/2$, and that C_1 and C_2 are taken large enough.*

Then for any sequence $\{(x^k,\, \lambda^k)\}$ generated by Algorithm 3.1 and convergent to $(\bar{x},\, \bar{\lambda})$, it holds that $\alpha_k = 1$ for all k large enough, and the rate of convergence is superlinear.

Proof Let a sequence $\{(x^k,\, \lambda^k)\}$ be generated by Algorithm 3.1 and converge to $(\bar{x},\, \bar{\lambda})$. Recalling the fact established in the proof of Theorem 4.1 and Lemma 5.1, for all k large enough penalty parameters c_1 and c_2 remain unchanged, the algorithm accepts the sSQP direction $d^k := d(u^k)$ where $u^k := (x^k,\, \lambda^k)$, and (17) holds with $q := 2$.

Moreover, according to (37) (following from Theorem 5.1 (b)), $d(u^k) \to 0$ as $u^k \to \bar{u}$. Therefore, employing Proposition 5.2 (with $\varphi := \varphi_{c_1,\,c_2}$ and $\beta := 1/2$) and Lemma 5.2, we conclude that $\alpha_k = 1$ for all k large enough. Superlinear convergence rate now follows from Theorem 5.1 (c). □

In the rest of this section we present a somewhat different line of analysis demonstrating the ultimate acceptance of the full stepsize $\alpha_k = 1$. This analysis does not use any additional restrictions on the power q. Moreover, it actually applies to a more general framework than specifically Algorithm 3.1. In particular, our results hold for any descent method employing the sSQP search directions and the Armijo rule for $\varphi_{c_1,\,c_2}$, with any fixed $c_1 > 0$ and $c_2 > 0$ such that c_1 is large enough. The price paid for this generality is that we have to assume SOSC (5) rather than the weaker noncriticality of the limiting multiplier $\bar{\lambda}$, and the existence of third derivatives of f and h near $\bar{x}$. We note that the results cannot be obtained by applying the Dennis–Moré theorem, because the Hessian $\varphi''_{c_1,\,c_2}(\bar{x},\,\bar{\lambda})$ (which exists when f and h are three times differentiable) is always singular when $\operatorname{rank} h'(\bar{x}) < l$ (see Proposition 5.3 below). Our analysis employs Corollary 5.1 instead, which allows such singularity.

The following is a version of [9, Theorem 4.16 (a)], removing the regularity assumption on the constraints.

Proposition 5.3 *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* (1) *with an associated Lagrange multiplier $\bar{\lambda} \in \mathbb{R}^l$ satisfying SOSC* (5)*.*

Then for any value of $\bar{c}_2 > 0$ there exists $\bar{c}_1 \geq 0$ such that for all $c_1 \geq \bar{c}_1$ and $c_2 \geq \bar{c}_2$, the Hessian $\varphi''_{c_1,\,c_2}(\bar{x},\,\bar{\lambda})$ is positive semidefinite, and moreover, $\langle \varphi''_{c_1,\,c_2}(\bar{x},\,\bar{\lambda})(\xi,\,\eta),\,(\xi,\,\eta)\rangle = 0$ if and only if $\xi = 0$ and $\eta \in \ker(h'(\bar{x}))^{\mathrm{T}}$.

Proof Differentiating (13) and using the assumption that $(\bar{x},\,\bar{\lambda})$ satisfies the Lagrange system (4), for any c_1 and c_2 we obtain that

$$\varphi''_{c_1,\,c_2}(\bar{x},\,\bar{\lambda}) = \begin{pmatrix} \dfrac{\partial^2 L}{\partial x^2}(\bar{x},\,\bar{\lambda}) + c_1 (h'(\bar{x}))^{\mathrm{T}} h'(\bar{x}) & (h'(\bar{x}))^{\mathrm{T}} \\ h'(\bar{x}) & 0 \end{pmatrix}$$

$$+ c_2 \begin{pmatrix} \left(\dfrac{\partial^2 L}{\partial x^2}(\bar{x},\,\bar{\lambda})\right)^2 & \dfrac{\partial^2 L}{\partial x^2}(\bar{x},\,\bar{\lambda})(h'(\bar{x}))^{\mathrm{T}} \\ h'(\bar{x})\dfrac{\partial^2 L}{\partial x^2}(\bar{x},\,\bar{\lambda}) & h'(\bar{x})(h'(\bar{x}))^{\mathrm{T}} \end{pmatrix}.$$

Setting $x(\eta) := (h'(\bar{x}))^{\mathrm{T}}\eta$, we have that

$$\langle \varphi''_{c_1,\,c_2}(\bar{x},\,\bar{\lambda})(\xi,\,\eta),\,(\xi,\,\eta)\rangle = Q_{c_2}(\xi,\,x(\eta)) + c_1 R(\xi,\,x(\eta)), \tag{43}$$

where the quadratic form $Q_{c_2} : \mathbb{R}^n \times \mathrm{im}(h'(\bar{x}))^{\mathrm{T}} \to \mathbb{R}$ and the quadratic form $R : \mathbb{R}^n \times \mathrm{im}(h'(\bar{x}))^{\mathrm{T}} \to \mathbb{R}$ are defined by

$$Q_{c_2}(\xi,\,x) := \left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x},\,\bar{\lambda})\xi,\,\xi \right\rangle + 2\langle \xi,\,x\rangle + c_2 \left\| \frac{\partial^2 L}{\partial x^2}(\bar{x},\,\bar{\lambda})\xi + x \right\|^2,$$

$$R(\xi,\,x) := \|h'(\bar{x})\xi\|^2.$$

If $R(\xi,\,x) = 0$ for some $(\xi,\,x) \in (\mathbb{R}^n \times \mathrm{im}(h'(\bar{x}))^{\mathrm{T}}) \setminus \{(0,\,0)\}$, then it holds that $\xi \in \ker h'(\bar{x}) = (\mathrm{im}(h'(\bar{x}))^{\mathrm{T}})^{\perp}$. And hence, $\langle \xi,\,x\rangle = 0$. Therefore,

$$Q_{c_2}(\xi,\,x) = \left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x},\,\bar{\lambda})\xi,\,\xi \right\rangle + c_2 \left\| \frac{\partial^2 L}{\partial x^2}(\bar{x},\,\bar{\lambda})\xi + x \right\|^2.$$

If $\xi \neq 0$, then $Q_{c_2}(\xi,\,x) > 0$, by SOSC (5). And if $\xi = 0$ but $x \neq 0$, then $Q_{c_2}(\xi,\,x) = c_2\|x\|^2 > 0$ for any $c_2 > 0$. Taking into account that R is evidently positive semidefinite, by the Finsler–Debreu lemma (e.g., [9, Lemma 1.25]) we conclude that for any $c_2 > 0$ there exists $c_1 \geq 0$ such that the quadratic form $Q_{c_2} + c_1 R$ is positive definite.

Therefore, according to (43), $\langle \varphi''_{c_1,\,c_2}(\bar{x},\,\bar{\lambda})(\xi,\,\eta),\,(\xi,\,\eta)\rangle$ is always positive except for the case when $\xi = 0$ and $x(\eta) = 0$. Moreover, increasing c_1 and c_2 can only make this value larger. This completes the proof. □

Theorem 5.3 *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ be three times differentiable in a neighbourhood of $\bar{x} \in \mathbb{R}^n$, with their third derivatives being continuous at $\bar{x}$. Let $\bar{x}$ be a stationary point of problem* (1) *with an associated Lagrange multiplier $\bar{\lambda} \in \mathbb{R}^l$ satisfying SOSC* (5). *Let $0 < \varepsilon < 1/2$, $0 < \theta < 1$. Consider the iterative scheme $(x^{k+1},\,\lambda^{k+1}) := (x^k,\,\lambda^k) + \alpha_k d^k$, where for each sufficiently large k the search direction $d^k := (\xi^k,\,\eta^k)$ is a solution of the system* (6) *for $\sigma := \|\Phi(x^k,\,\lambda^k)\|$ and $(x,\,\lambda) := (x^k,\,\lambda^k)$, while $\alpha_k := \theta^j$, where j is the smallest nonnegative integer satisfying the Armijo inequality* (21) *with some fixed $c_1 \geq 0$ and $c_2 \geq 0$.*

Then for any $\bar{c}_2 > 0$ *there exists* $\bar{c}_1 \geq 0$ *such that for every* $c_1 \geq \bar{c}_1$ *and* $c_2 \geq \bar{c}_2$*, and for any sequence* $\{(x^k,\, \lambda^k)\} \subset (\mathbb{R}^n \times \mathbb{R}^l) \setminus (\{\bar{x}\} \times \mathcal{M}(\bar{x}))$ *generated by the iterative scheme specified above and convergent to* $(\bar{x},\, \bar{\lambda})$*, it holds that* $\alpha_k = 1$ *for all* k *large enough, and the convergence rate is superlinear.*

Proof Fix any $\bar{c}_2 \geq 0$ and consider any $\bar{c}_1 \geq 0$ such that the assertion of Proposition 5.3 holds. Fix any $c_1 \geq \bar{c}_1$ and $c_2 \geq \bar{c}_2$ and consider an arbitrary sequence $\{(x^k,\, \lambda^k)\} \subset (\mathbb{R}^n \times \mathbb{R}^l) \setminus (\{\bar{x}\} \times \mathcal{M}(\bar{x}))$ specified in the statement of the theorem.

According to Theorem 5.1 (b), close to the primal-dual solution satisfying SOSC, the full step of sSQP satisfies the estimate (24), and hence, (30) holds with $u^k := (x^k,\, \lambda^k)$ and $\bar{u} := (\bar{x},\, \bar{\lambda})$.

We next prove that under the stated assumptions there exists $\gamma > 0$ such that for all sufficiently large k it holds that

$$\langle \varphi''_{c_1,\, c_2}(\bar{x},\, \bar{\lambda}) d^k,\, d^k \rangle \geq \gamma \|d^k\|^2. \tag{44}$$

According to Proposition 5.3, we need to show that the normalized sequence $\{(\xi^k,\, \eta^k)/\|(\xi^k,\, \eta^k)\|\}$ does not have accumulation points of the form $(0,\, \eta)$ with $\eta \in \ker(h'(\bar{x}))^{\mathrm{T}}$. Suppose that, on the contrary, such an accumulation point exists (in which case $\|\eta\| = 1$) and assume, without loss of generality, that the entire sequence converges.

Let P be the orthogonal projector onto $(\operatorname{im} h'(\bar{x}))^{\perp}$. Then $Ph'(\bar{x}) = 0$, and employing (15) and the smoothness hypothesis, we obtain that

$$\begin{aligned}
\sigma_k P \eta^k &= P(h(x^k) + h'(x^k)\xi^k) \\
&= P(h'(\bar{x})(x^k - \bar{x}) + h'(\bar{x})\xi^k + (h'(x^k) - h'(\bar{x}))\xi^k)) + O(\|x^k - \bar{x}\|^2) \\
&= O(\|\xi^k\| \|x^k - \bar{x}\|) + O(\|x^k - \bar{x}\|^2)
\end{aligned} \tag{45}$$

as $k \to \infty$.

We next show that

$$x^k - \bar{x} = o(\|\eta^k\|) \tag{46}$$

as $k \to \infty$. Suppose that this is not the case. Then, passing onto a subsequence if necessary, $\eta^k = O(\|x^k - \bar{x}\|)$, and since $\xi^k = o(\|\eta^k\|)$, we conclude that $\xi^k = o(\|x^k - \bar{x}\|)$. Therefore, there exists $\mu > 0$ such that

$$\|x^k + \xi^k - \bar{x}\| \geq \mu \|x^k - \bar{x}\|$$

for all k large enough. Therefore, from (30) we obtain the estimate

$$\|x^k - \bar{x}\| + \operatorname{dist}(\lambda^k + \eta^k, \mathcal{M}(\bar{x})) = o(\|x^k - \bar{x}\| + \operatorname{dist}(\lambda^k, \mathcal{M}(\bar{x}))),$$

evidently implying that

$$\|x^k - \bar{x}\| + \operatorname{dist}(\lambda^k + \eta^k, \mathcal{M}(\bar{x})) = o(\operatorname{dist}(\lambda^k, \mathcal{M}(\bar{x}))) \tag{47}$$

as $k \to \infty$. Since

$$\operatorname{dist}(\lambda^k, \mathcal{M}(\bar{x})) \leq \operatorname{dist}(\lambda^k + \eta^k, \mathcal{M}(\bar{x})) + \|\eta^k\|,$$

the estimate (47) implies that $\operatorname{dist}(\lambda^k, \mathcal{M}(\bar{x})) = O(\|\eta^k\|)$, and hence, (46) holds, giving a contradiction.

Combining (45) with (46), we conclude that

$$\sigma_k P \eta^k = o(\|\eta^k\|^2) = o(\sigma_k \|\eta^k\|)$$

as $k \to \infty$, where the the last equality is by (37). This implies that $P\eta = 0$ or, equivalently, $\eta \in \ker P = \operatorname{im} h'(\bar{x})$.

At the same time, $\eta \in \ker(h'(\bar{x}))^{\mathrm{T}} = (\operatorname{im} h'(\bar{x}))^{\perp}$, which is only possible when $\eta = 0$. This gives the needed contradiction, thus establishing (44).

The result now follows by applying Corollary 5.1. □

Recall that the proof of Corollary 5.1 includes the demonstration of the fact that (28) holds with some $\gamma > 0$ for all k large enough. This implies that sSQP directions can indeed be employed by a descent method in Theorem 5.3, for k large enough.

6 Numerical Results

In this section, we present computational experiments with the proposed globalized sSQP algorithm on degenerate problems, and compare it with some alternatives. These include linesearch quasi-Newton SQP, the augmented Lagrangian algorithm, and a combination of sSQP with the augmented Lagrangian. As shown in [26], the effect of attraction to critical Lagrange multipliers (which slows down local convergence) is much less persistent for sSQP than for SQP, though this undesirable behavior is not avoided by sSQP completely. One expects that the superlinear rate of sSQP in the cases of convergence to noncritical multipliers should make its performance on degenerate problems superior to that of the usual SQP, the latter typically converging (slowly) to critical multipliers. But it is also important not to lose too much to SQP in those cases when sSQP still converges to critical multipliers (and thus also does not achieve the superlinear rate), even if this scenario is much less frequent for sSQP.

In our numerical experiments, performed in Matlab environment, we compare Algorithm 3.1 with some well-established implementations of SQP and augmented Lagrangian algorithms, namely, with SNOPT [27] and ALGENCAN [28], respectively. We used ALGENCAN 2.3.7 with AMPL interface, and SNOPT 7.2-8 coming with AMPL, both with their default values of the parameters. The third algorithm used for comparisons is the combination of sSQP with the augmented Lagrangian algorithm, developed in [18], with all the parameters values as stated in that reference, with the option of sequential updating of the Hessian, and with the upper bound for the stabilization parameter (see [18, Section 4]). In what follows, the latter algorithm will be referred to as sSQP-AugL.

In Algorithm 3.1, parameters are as follows: $\delta := 10$, $\varepsilon := 0.1$, $\theta := 0.5$. Initial values of the penalty parameters were taken as suggested in [10]: $c_1 := 100$, $c_2 := 0.01$, with upper bounds $C_1 := C_2 := 10^{20}$ (as commented earlier, these are just safeguards, and in fact they were never activated in our computational experience reported below). For the tests (17)–(20) in Algorithm 3.1 we used the parameters $\rho := 0.1$, $q := 2$, and the functions $\psi_1(t) := \psi_3(t) := 0.5t$, and $\psi_2(t) := 0.5t^2$. The choices of $\bar{\sigma}$ are discussed further below. We define the matrices Q_k in Algorithm 3.1 (when they are requested) using BFGS updates, if the iterations of this type are consecutive. On the very first iteration where Q_k is needed after an iteration that used the sSQP direction, $Q_k := I$ is taken. Furthermore, when the BFGS direction is used, we employ the Wolfe linesearch rule instead of the Armijo one, i.e., the stepsize α_k must satisfy

$$\varphi_{c_1,\, c_2}((x^k,\, \lambda^k) + \alpha_k d^k) \leq \varphi_{c_1,\, c_2}(x^k,\, \lambda^k) + \varepsilon_1 \alpha_k \left\langle \varphi'_{c_1,\, c_2}(x^k,\, \lambda^k),\, d^k \right\rangle$$

and

$$\left\langle \varphi'_{c_1,\, c_2}((x^k,\, \lambda^k) + \alpha_k d^k),\, d^k \right\rangle \geq \varepsilon_2 \left\langle \varphi'_{c_1,\, c_2}(x^k,\, \lambda^k),\, d^k \right\rangle.$$

This rule guarantees that the matrices obtained by BFGS updates remain positive definite; see, e.g., [29, Theorem 4.5]. We use the parameters $\varepsilon_1 := 0.1$, $\varepsilon_2 := 0.9$.

Our first test set includes all the 34 equality-constrained problems from the DEGEN collection [22]. We comment that these are small but in some ways difficult (at least for achieving fast convergence) problems, ranging from 1 to 10 variables, and from 1 to 7 constraints, with various kinds of degeneracy (to be discussed later; in particular, we shall consider specifically the subset of problems where degeneracy is induced by the non-uniqueness of Lagrange multipliers associated to the primal solution). For each of these problems we performed 20 runs from random starting points satisfying $\|(x^0,\, \lambda^0)\|_\infty \leq 100$. The run is considered successful if the stopping criterion $\|\Phi(x^k,\, \lambda^k)\| \leq 10^{-8}$ was satisfied before the iteration count k passed 500.

In the first series of the experiments, a large value for the upper bound of the stabilization $\bar{\sigma}$ is taken. Specifically, $\bar{\sigma} := 10^4$. As a measure of efficiency of the algorithms we use the number of evaluations of the objective function (which is always the same as the number of evaluations of the constraints).

The results are presented in Figure 1 in the form of a performance profile, which is a version of the original proposal in [30], adapted for the case of

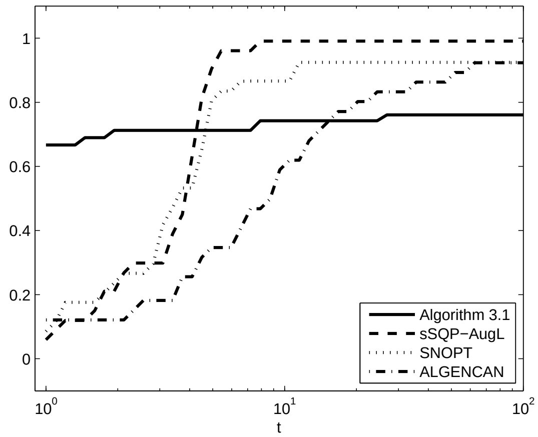

Fig. 1 Results for all equality-constrained problems from DEGEN.

multiple runs from different starting points for each test problem. For each algorithm the plotted function $\pi : [1, \infty) \to [0, 1]$ is defined as follows. Let k_p stand for the average iteration count of a given algorithm per one successful run for problem p. Let s_p denote the portion of successful runs on this problem. Let r_p be equal to the best (minimal) value of k_p over all algorithms. Then

$$\pi(\tau) := \frac{1}{P} \sum_{p \in R(\tau)} s_p,$$

where P is the number of problems in the test set (34 in our case) and $R(\tau)$ is the set of problems for which k_p is no more than τ times worse (larger) than the best result r_p:

$$R(\tau) := \{p = 1, \ldots, P : \ k_p \leq \tau r_p\}, \quad \tau \in [1, \infty).$$

In particular, the value $\pi(1)$ corresponds to the portion of runs for which the given algorithm demonstrated the best result. The values of $\pi(\tau)$ for large τ characterize robustness, that is, the portion of successful runs.

One can see from Figure 1 that Algorithm 3.1 is more efficient on most problems. On the other hand, it has more failures, and is still worse than the alternatives on some problems. We next comment on the observed behavior in more detail, which also leads us to consider the subset of problems with degeneracy corresponding specifically to non-uniqueness of Lagrange multipliers (which is, arguably, the most common type of degenerate problems).

We first note that the safeguarding BFGS steps in Algorithm 3.1 were activated rather rarely, and thus, the algorithm usually works indeed as (globalized) sSQP. BFGS steps showed up somehow systematically only for the problems 20303, 20308, 20310, 2DD01. Sometimes they save the run, but all successful runs where a BFGS step was encountered, actually ended up with sSQP steps (BFGS steps helped the algorithm to recover and get back to successful sSQP steps). However, most cases where BFGS steps were needed at all, eventually ended up with a failure. This seems to suggest that the degenerate problems for which this happened are somehow difficult for the given penalty function, regardless of which directions one uses to minimize it.

A closer look at the numerical results reveals some special features of problem instances for which Algorithm 3.1 did not perform well. First, 9 problems in DEGEN have non-isolated *primal* solutions. Somewhat surprisingly, quasi-Newton SQP behaves well on such problems, often converging superlinearly to some solution, while globalized sSQP converges rather slowly on some of these problems. In 4 other problems there exist no Lagrange multipliers associated to the primal solution. These two types of degenerate problems (non-isolated primal solutions and empty Lagrange multiplier sets) are, in fact, rather special. Moreover, sSQP was certainly not intended for tackling *these* kinds of degeneracy, and therefore, there are no reasons to expect its good performance on this type of problems, in general (that said, Algorithm 3.1 still behaves quite well on some of these problems). We recall that the purpose of sSQP is dual stabilization, in the degenerate cases when Lagrange multipliers exist but are not unique. It is thus natural to consider this class of problems, which gives 21 instances in DEGEN. For these problems, we observe the behavior demonstrated in Figure 2, and now Algorithm 3.1 definitely outperforms the alternatives by efficiency, and has almost the same robustness. With due caution, we may conclude that the presented globalized sSQP method does work quite well on degenerate problems with degeneracy induced by nonunique Lagrange multipliers. However, taking yet a further look at those problems where relatively slow convergence was observed, we notice that this usually happens in the non-fully degenerate cases, i.e., when the method converges to a degenerate solution $\bar{x}$ such that $h'(\bar{x}) \neq 0$. Such cases deserve further investigation. We discuss this next.

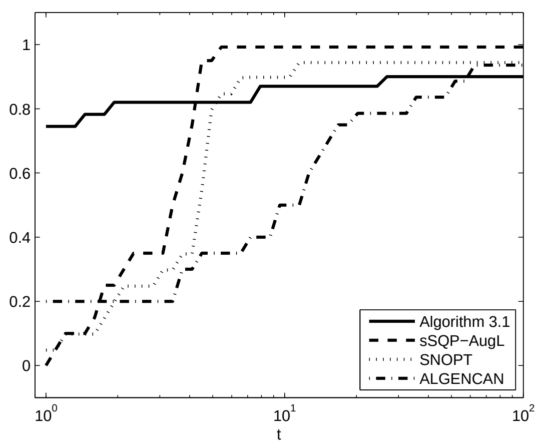

Fig. 2 Results for equality-constrained problems from DEGEN with non-unique multipliers associated to isolated primal solutions.

We observed that when convergence is slow, it does not seem to be caused by any deficiencies in the proposed globalization technique – often the method in fact takes full sSQP steps (i.e., $\alpha_k = 1$ is accepted). At issue seem to be some intrinsic properties of sSQP directions themselves, when far from a solution. To understand this better, consider the following *non-degenerate* example.

Example 6.1 Let $n := l := 1$, $f(x) := x^2/2$, $h(x) := x$. Then $\bar{x} := 0$ is the unique solution of the problem (1). This solution is non-degenerate, and the unique associated Lagrange multiplier is $\bar{\lambda} := 0$.

The sSQP iteration system (6) takes the form

$$\begin{pmatrix} 1 & 1 \\ 1 & -\sigma \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix} = - \begin{pmatrix} x + \lambda \\ x \end{pmatrix},$$

and the method generates

$$\begin{pmatrix} \xi^k \\ \eta^k \end{pmatrix} = \frac{1}{\sigma_k + 1} \begin{pmatrix} -\sigma_k & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} x^k + \lambda^k \\ x^k \end{pmatrix} = \frac{1}{\sigma_k + 1} \begin{pmatrix} -(\sigma_k + 1)x^k - \sigma_k \lambda^k \\ -\lambda^k \end{pmatrix}.$$

Therefore, $x^{k+1} = -S_k \lambda^k$, $\lambda^{k+1} = S_k \lambda^k$, where $S_k = \sigma_k/(\sigma_k + 1)$. If σ_k is large (e.g., $\sigma_k = \|\Phi(x^k, \lambda^k)\|$ for (x^k, λ^k) far from solutions), then S_k is close to 1, and thus the iterates move slowly. On the other hand, if σ_k is small, then S_k is close to 0, and thus convergence is fast.

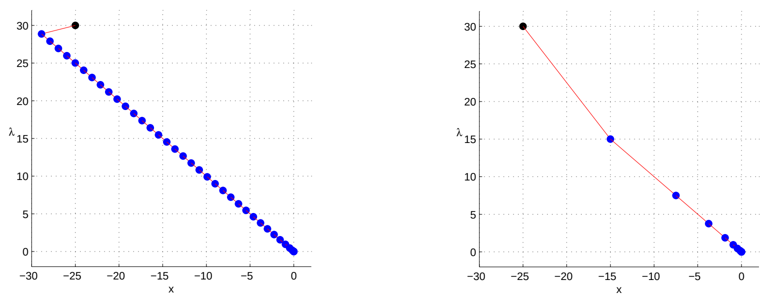

Fig. 3 Primal-dual sequences for Example 6.1; $(x^0, \lambda^0) := (-25, 30)$.

The sequence generated by pure (i.e., without globalization) sSQP iterates for this example is shown in the left graph of Figure 3. The stopping criterion is satisfied after 38 iterations only, even though the asymptotic convergence rate is superlinear.

The behavior similar to that in Example 6.1 appears common if the problem is not fully degenerate: sSQP often generates long sequences of short steps far from solutions, before reaching the region where the superlinear convergence shows up. In fact, this has also been observed on some examples in [31] (which contains some experiments for sSQP without any globalization).

Somewhat informally, the situation can be described as follows. There are two areas around any qualified solution: the "small" one, where the superlinear convergence is guaranteed, and the "large" one, *outside* of which nothing reasonable can be expected from sSQP directions. Short steps happen inside the "large" area outside of the "small" area. In DEGEN problems with fully degenerate solutions, the "small" area often actually appears to be quite large, while for problems with non-fully degenerate solutions this is not the case. From a different viewpoint, nonzero singular values of the constraints' Jacobian appear to give rise to repulsion of sSQP iterates from solutions, which drastically slows down convergence as the current iterate is moved away from the solutions and the stabilization parameter grows accordingly. On the other hand, in the region where the repulsion and the attraction compensate for each other, the algorithm gets essentially stuck. This not only degrades its efficiency, but often results in failures.

One potential strategy to help avoid the negative effect described above, is to decrease the value of $\bar{\sigma}$ in Algorithm 3.1. The smaller is this value, the closer is the sSQP iteration system to the one of the usual SQP. In particular, limiting σ_k in Example 6.1 to some small value results in small S_k giving fast convergence. The sequence generated by pure sSQP iterations (without globalization) with σ_k bounded by 1 is shown in the right graph of Figure 3. It takes only 11 iterations for this version of the method to achieve the required stopping tolerance (cf. the left graph of Figure 3, where there are 38 iterations).

On the other hand, making sSQP "closer" to SQP facilitates attraction to critical multipliers, and therefore, decreasing $\bar{\sigma}$ may also degrade the convergence rate. To illustrate this effect, we consider again the problem from Example 5.1. It can be easily seen that (36) implies

$$x^{k+1} = s_k x^k, \quad \lambda^{k+1} = \lambda^k - s_k(\lambda^k + 1),$$

where

$$s_k := 1 - S(\sigma_k;\, x^k,\, \lambda^k) = \left(2 + \frac{\sigma_k(1+\lambda^k)}{(x^k)^2}\right)^{-1}.$$

Therefore, for small values of σ_k we have

$$x^{k+1} \approx \frac{1}{2}x^k, \quad \lambda^{k+1} + 1 \approx \frac{1}{2}(\lambda^k + 1).$$

In particular, $\{(x^k,\, \lambda^k)\}$ converges linearly to $(\bar{x},\, \lambda)$, where $\lambda := -1$ is the unique critical multiplier.

In our experience, the overall behavior of Algorithm 3.1 with $\bar{\sigma} \leq 1$ is rather poor, but it can be significantly improved by using some semi-heuristic modifications, such as second-order corrections (see, e.g., [29, Section 17.3]) and nonmonotone linesearch [32]. The idea of the latter is to replace $\varphi_{c_1,\, c_2}(x^k,\, \lambda^k)$ in the right-hand side of (21) by

$$\max_{j=0,\, \ldots,\, R-1} \varphi_{c_1,\, c_2}(x^{k-j},\, \lambda^{k-j}).$$

The value $R := 8$ appears optimal in our experiments.

Finally, our impression is that in cases of convergence to critical multipliers, the control of c_2 becomes crucially important. In particular, the following heuristic rule added to step 5 of Algorithm 3.1 considerably improves the performance: if $\alpha_k \leq 0.1$, replace c_2 by $\min\{10c_2,\, 10^{10}\}$.

Nonmonotone linesearch increases the efficiency if $\bar{\sigma}$ is taken large as well. The other heuristics mentioned above have almost no impact on the overall performance for large $\bar{\sigma}$.

Performance profile in Figure 4 compares the behavior of Algorithm 3.1 with all the mentioned heuristics implemented, and with different values of $\bar{\sigma}$. Decreasing $\bar{\sigma}$ improves robustness, but significantly degrades efficiency. In particular, this version of Algorithm 3.1 with small values $\bar{\sigma}$ is still outperformed by the other algorithms on many problems. One of the reasons is that Algorithm 3.1 with small $\bar{\sigma}$ is "close" to SQP globalized by the smooth penalty function, while it is known that globalization using nonsmooth penalty is better.

Perhaps not surprisingly, being only the first step in this direction, our numerical experience is rather mixed. We detected that far from solutions sSQP directions are often not very efficient, even when the unit stepsize is

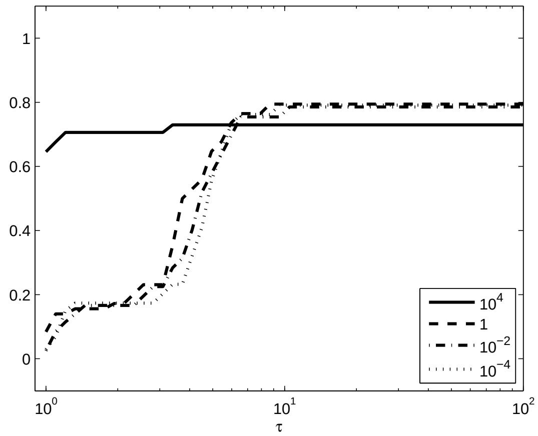

Fig. 4 Algorithm 3.1 with different values of $\bar{\sigma}$.

accepted. This should probably affect *any other* globalization, not just the one considered here. However, we believe this mixed news is nevertheless one of the *useful* results of the paper, as it makes it clear that some modifications to pure sSQP directions would probably be needed in the global phase. This would be the subject of future research. And/or, perhaps, some innovative ways of managing the dual stabilization parameter σ in (2) might be the key.

7 Conclusions

We presented a globalization of the stabilized SQP method, based on linesearch for a two-parameter smooth primal-dual penalty function. Global convergence properties of the algorithm were established. Moreover, we extended the classical Dennis-Moré analysis on acceptance of the unit stepsize to problems with non-isolated solutions. Based on these results, it was shown that the proposed globalized sSQP method retains local superlinear convergence under the same weak assumptions as those of the pure sSQP. Computational performance on a set of degenerate test problems was analyzed as well.

Acknowledgements This research is supported by the Russian Foundation for Basic Research Grant 14-01-00113, by the Russian Science Foundation Grant 15-11-10021, by CNPq Grants PVE 401119/2014-9 and 302637/2011-7 (Brazil), and by FAPERJ.

The authors also thank the three anonymous referees for their evaluation and helpful comments.

References

1. Wright, S.J.: Superlinear convergence of a stabilized SQP method to a degenerate solution. Comput. Optim. Appl. 11, 253–275 (1998)
2. Hager, W.W.: Stabilized sequential quadratic programming. Comput. Optim. Appl. 12, 253–273 (1999)
3. Fischer, A.: Local behavior of an iterative framework for generalized equations with nonisolated solutions. Math. Program. 94, 91–124 (2002)
4. Wright, S.J.: Modifying SQP for degenerate problems. SIAM J. Optim. 13, 470–497 (2002)
5. Izmailov, A.F., Solodov, M.V.: Newton-Type Methods for Optimization and Variational Problems. Springer Series in Operations Research and Financial Engineering. Springer International Publishing, Switzerland (2014)
6. Fernández, D., Solodov, M.: Stabilized sequential quadratic programming for optimization and a stabilized Newton-type method for variational problems. Math. Program. 125, 47–73 (2010)
7. Izmailov, A.F., Solodov, M.V.: Stabilized SQP revisited. Math. Program. 122, 93–120 (2012)
8. Di Pillo, G., Grippo, L.. A new class of augmented Lagrangians in nonlinear programming. SIAM J. Control Optim. 17, 618–628 (1979)
9. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods. Academic Press, New York (1982)
10. Bertsekas, D.P.: Enlarging the region of convergence of Newton's method for constrained optimization. J. Optim. Theory Appl. 36, 221–252 (1982)
11. Andreani, R., Haeser, G., Schuverdt, M.L., Silva, P.J.S.: A relaxed constant positive linear dependence constraint qualification and applications. Math. Program. 135, 255–273 (2012)
12. Izmailov, A.F., Solodov, M.V., Uskov E.I.: Global convergence of augmented Lagrangian methods applied to optimization problems with degenerate constraints, including problems with complementarity constraints. SIAM J. Optim. 22, 1579–1606 (2012)
13. Glad, S.T.: Properties of updating methods for the multipliers in augmented Lagrangian. J. Optim. Theory Appl. 28, 135–156 (1979)

14. Gill, P.E., Robinson, D.P.: A primal-dual augmented Lagrangian. Comput. Optim. Appl. 51, 1–25 (2012)
15. Gill, P.E., Robinson, D.P.: A globally convergent stabilized SQP method. SIAM J. Optim. 23, 1983–2010 (2013)
16. Gill, P.E., Kungurtsev, V., Robinson, D.P.: A regularized SQP method with convergence to second-order optimal points. UCSD Center for Computational Mathematics Technical Report CCoM-13-4, (2013)
17. Gill, P.E., Kungurtsev, V., Robinson, D.P.: A globally convergent stabilized SQP method: superlinear convergence. UCSD Center for Computational Mathematics Technical Report CCoM-13-4, (2014)
18. Izmailov, A.F., Solodov, M.V., Uskov E.I.: Combining stabilized SQP with the augmented Lagrangian algorithm. Comput. Optim. Appl. 62, 405–429 (2015)
19. Izmailov, A.F., Krylova, A.M., Uskov E.I.: Hybrid globalization of stabilized sequential quadratic programming method. In Russian. In: Bereznyov, V.A. (ed.): Theoretical and Applied Problems of Nonlinear Analysis, pp. 47–66. Computing Center RAS, Moscow, (2011)
20. Fernández, D., Pilotta, E.A., Torres, G.A.: An inexact restoration strategy for the globalization of the sSQP method. Comput. Optim. Appl. 54, 595–617 (2013)
21. Martínez, J.M., Pilotta, E.A.: Inexact restoration methods for nonlinear programming: advances and perspectives. In: Optimization and Control with Applications, pp. 271–292. Springer, New York (2005)
22. DEGEN. **http://w3.impa.br/~optim/DEGEN_collection.zip**.
23. Izmailov, A.F., Solodov, M.V.: On attraction of Newton-type iterates to multipliers violating second-order sufficiency conditions. Math. Program. 117, 271–304 (2009)
24. Izmailov, A.F., Solodov, M.V.: Critical Lagrange multipliers: what we currently know about them, how they spoil our lives, and what we can do about it. TOP 23, 1–26 (2015)
25. A.F. Izmailov and A.S. Kurennoy. Abstract Newtonian frameworks and their applications. SIAM J. Optim. 23 (2013), 2369–2396.
26. Izmailov, A.F., Solodov, M.V.: On attraction of linearly constrained Lagrangian methods and of stabilized and quasi-Newton SQP methods to critical multipliers. Math. Program. 126, 231–257 (2011)
27. P.E. Gill, W. Murray, and M.A. Saunders. SNOPT: An SQP algorithm for large-scale constrained optimization. SIAM J. Optim. **12**, 979–1006 (2002)
28. ALGENCAN: **http://www.ime.usp.br/~egbirgin/tango/**
29. Bonnans, J.F., Gilbert, J.Ch., Lemaréchal, C., Sagastizábal, C.. Numerical Optimization: Theoretical and Practical Aspects. Second Edition. Springer–Verlag, Berlin, Germany (2006)
30. Dolan, E.D., Moré, J.J.: Benchmarking optimization software with performance profiles. Math. Program. 91, 201–213 (2002)

31. Mostafa, E.M.E., Vicente, L.N., Wright, S.J.: Numerical behavior of a stabilized SQP method for degenerate NLP problems. In: Bliek, C., Jermann, C., Neumaier, A. (eds.): Global Optimization and Constraint Satisfaction. Lecture Notes in Computer Science 2861, pp. 123–141. Springer-Verlag, Berlin (2003)
32. Grippo, L., Lampariello, F., Lucidi, S.: A nonmonotone line search technique for Newton's method. SIAM J. Numer. Anal. 23, 707–716 (1986)