Preprint ANL/MCS-P760-0699

On the Rate of Convergence of Sequential Quadratic Programming with Nondifferentiable Exact Penalty Function in the Presence of Constraint Degeneracy

Mihai Anitescu*
Mathematics and Computer Science Division, Argonne National Laboratory
9700 S. Cass Avenue, Argonne, IL 60439, USA
e-mail anitescu@mcs.anl.gov

June 18, 1999

Abstract

We analyze the convergence of the sequential quadratic programming (SQP) method for nonlinear programming for the case in which the Jacobian of the active constraints is rank deficient at the solution and/or strict complementarity does not hold for some or any feasible Lagrange multipliers. We use a nondifferentiable exact penalty function, and we prove that the sequence generated by the SQP is locally R-linearly convergent if the matrices of the quadratic program are uniformly positive definite and bounded, provided that the Mangasarian-Fromowitz constraint qualification and some second-order sufficiency conditions hold.

Keywords: Linear Convergence, Nondifferentiable Exact Penalty, Degenerate Nonlinear Program.

1 Introduction

Recently, there has been renewed interest in analyzing and modifying sequential quadratic programming (SQP) algorithms for constrained nonlinear optimization for cases where the traditional regularity conditions do not hold [7, 11, 13]. This research has been motivated by the fact that large-scale nonlinear programming problems tend to be almost degenerate (large condition numbers for the Jacobian of the active constraints). It is therefore important to establish to what extent the convergence properties of the SQP methods are dependent on the ill-conditioning of the constraints.

To this end in this paper we relax both traditional regularity assumptions [5]: that the gradients of the active constraints are linearly independent and that there exists a choice of Lagrange multipliers that satisfies strict complementarity. To ensure good global convergence properties, we use a nondifferentiable exact penalty function as a

*Mathematics and Computer Science Division, Argonne National Laboratory. The work of this author was supported by the Mathematical, Information and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, U.S. Department of Energy, under Contract W-31-109-Eng-38.

measure of the progress of the SQP algorithm [2]. Previous work has shown that, if certain sufficient conditions are satisfied for one feasible choice of the Lagrange multipliers, an exact penalty function can be defined that has an unconstrained local minimum at a solution x^* of the nonlinear program [8]. However, in order to obtain good convergence properties, it is desirable to ensure that x^* is an isolated stationary point of the penalty function. Otherwise, the SQP could stop at a stationary point different from x^*, no matter how close to x^* it was initialized. In [4] it is proven that if, among other conditions, x^* is a solution and an isolated local minimum of the nonlinear program, it is an isolated local minimum of the penalty function. In [12] it is shown that, if the conditions from [8] hold at x^* for any feasible choice of Lagrange multipliers, x^* is an isolated local minimum of the nonlinear program under consideration. As a consequence of [4], x^* is an isolated local minimum of the penalty function, for some choice of the penalty parameter.

In this paper we provide an alternative proof of the fact that the second-order sufficient conditions from [12] result in the existence of an L^∞ penalty function for which x^* is a strict local minimum and an isolated stationary point. The advantage of our proof is that we can estimate the decrease of the penalty function at each step of the algorithm as a function of the distance to x^*. This guarantees that nondifferentiable exact penalty SQP algorithms with uniformly positive definite and bounded matrices of the quadratic programs achieve at least linear convergence even in the presence of degeneracy, provided that the weaker conditions from [12] hold. The result proven here is equivalent to the linear convergence of descent methods of unconstrained optimization that use directions making acute (bounded away from 90 degrees) angles with the gradient [9].

In our developments, an important part is played by a function that is an extension of the usual augmented Lagrangian (Theorem 2.1). The novelty of our approach consists in the fact that the augmentation is different for zero and nonzero multipliers. This new object is just once continuously differentiable and has, because of the second-order sufficient conditions from [12] being satisfied at x^*, a strictly monotone gradient near x^* for all feasible Lagrange multipliers.

1.1 Background and Assumptions

For details on the results of this section, see [1, 2, 12]. The nonlinear program (NLP) to be solved is

$$\begin{aligned} &\text{minimize} && f(x) && (1)\\ &\text{subject to} && h(x)=0, \quad g(x)\leq 0, && (2)\end{aligned}$$

where $h(x) = (h_1(x), ..., h_m(x))$ and $g(x) = (g_1(x), ..., g_r(x))$. In this paper, we assume that f, g_i, and h_j are twice continuously differentiable functions from R^n to R. We denote by $\boldsymbol{x^*}$ a solution of the nonlinear program (1–2).

The *Lagrangian function* of (1–2) is

$$\mathcal{L}(x,\lambda,\mu) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{i=1}^{r} \mu_j g_j(x) = f(x) + \lambda^T h(x) + \mu^T g(x), \quad (3)$$

where $\lambda \in R^m$ and $\mu \in R^r$ are the Lagrange multipliers.

When a *constraint qualification* condition holds (see discussion below), there exist λ^*, $\mu^* \geq 0$ at x^* such that the first-order conditions

$$\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0, \quad h(x^*) = 0, \quad g(z^*) \leq 0, \quad (\mu^*)^T g(z^*) = 0 \quad (4)$$

are satisfied. A point x satisfying (4) is called a stationary point of the NLP. These are the well-known Karush-Kuhn-Tucker (KKT) conditions. The *active set* at x^* is defined by

$$\mathcal{B} = \{i = 1, \ldots, r | g_i(x^*) = 0\}. \tag{5}$$

We assume that the Mangasarian-Fromowitz constraint qualification (MFCQ) holds at the local minimum x^* of (1–2): The gradients of the equality constraints, $\nabla_x h_j(x^*)$, $j = 1 \ldots, m$, are linearly independent and

$$\exists d \in R^n \text{ such that } \quad \nabla_x h_j(x^*)^T d = 0, \; j = 1, \ldots m \quad \nabla_x g_i(x^*)^T d < 0, \; i \in \mathcal{B}. \tag{6}$$

This ensures that the set of multipliers $\lambda^*, \mu^* \geq 0$, for which (4) is satisfied is nonempty and bounded [6]. We denote this set by $\boldsymbol{\mathcal{M}(x^*)}$.

In this paper we assume the following second-order sufficient condition: There exists a $\sigma > 0$ such that for all $(\lambda^*, \mu^*) \in \mathcal{M}(x^*)$

$$\begin{aligned}
& & w^T \nabla_{xx} \mathcal{L}(x^*, \lambda^*, \mu^*) w > \sigma \|w\|^2 & & (7)\\
\forall w \quad & \text{such that} & \nabla_x h_i(x^*)^T w = 0, \quad i = 1, \ldots, m & & (8)\\
& & \nabla_x g_i(x^*)^T w = 0, \quad i \in \mathcal{B}, \mu_i^* > 0 & & (9)\\
& & \nabla_x g_i(x^*)^T w \leq 0, \quad i \in \mathcal{B}, \mu_i^* = 0. & & (10)
\end{aligned}$$

We denote the set of w's satisfying (8–10) as $\boldsymbol{\mathcal{F}(x^*, \lambda^*, \mu^*)}$. This ensures that x^* is a strict local minimum and an isolated stationary point of (1–2) [5, 12]. The traditional second-order sufficiency condition assumes that the gradients of the equality constraints and active inequality constraints ($i \in \mathcal{B}$) are linearly independent and that at least one of the μ^*'s has positive $\mu_i^*, i \in \mathcal{B}$ (*strict complementarity*) [5]. In this paper these conditions are replaced by the more general MFCQ (6) and (7–10).

The *nondifferentiable penalty function* considered here is

$$P(x) = \max\{0, |h_1(x)|, \ldots, |h_m(x)|, g_1(x), \ldots, g_r(x)\}. \tag{11}$$

We are looking for unconstrained minima of the function

$$\phi(x) = f(x) + cP(x),$$

where c is a sufficiently large constant. Descent directions d of $\phi(x)$ at the point x can be obtained by solving the following quadratic program [2]:

$$\begin{aligned}
\text{minimize} \quad & \nabla f(x)^T d + \tfrac{1}{2} d^T H d + c\zeta & & (12)\\
\text{subject to} \quad & -\zeta \leq h_i(x) + \nabla h_i(x)^T d \leq \zeta \qquad i = 1, \ldots, m & & (13)\\
& g_j(x) + \nabla g_j(x)^T d \leq \zeta \qquad j = 0, 1, \ldots, r & & (14)
\end{aligned}$$

where H is some positive definite matrix and $g_0(x) = 0$. The matrices H considered here satisfy

$$\gamma_0 \|x\|^2 \leq x^T H x \leq \Gamma_0 \|x\|^2, \quad \forall x \in R^n. \tag{15}$$

We define $\boldsymbol{\mathcal{H}}$ to be the set of matrices satisfying (15).

At the current point x^k of an iterative procedure that attempts to determine x^*, (12–14) generate the descent direction d^k. H can change over the iterations; its current value is denoted by H^k at each iteration. The next iterate is $x^{(k+1)} = x^k + \alpha^k d^k$, where α^k is obtained by a line search procedure. Usual stepsize rules are the minimization rule, the limited minimization rule and the Armijo rule [2]. For these rules, any limit point of $\{x^k\}$ is a stationary point of $\phi(x)$, and the descent procedure is therefore globally convergent in this sense [2].

If, in addition,

$$c > \sum_{i=1}^{m} |\lambda_i^*| + \sum_{j=1}^{r} \mu_j^* \tag{16}$$

for some $(\lambda^*, \mu^*) \in \mathcal{M}(x^*)$, then x^* is a critical point of $\phi(x)$ [1].

A suitable value for c is not available in the early stages of the algorithm. Hence, it is common to choose [1]

$$c^{k+1} = \max\{c^k, \gamma + \sum_{i=1}^{m} |\lambda_{i1}^k - \lambda_{i2}^k| + \sum_{j=1}^{r} \mu_j^k\}, \tag{17}$$

where λ_{i1} and λ_{i2} correspond, respectively, to the left and right inequality constraints in (13). γ is a prescribed safety factor.

Consider the quadratic program

$$\begin{array}{lll} \text{minimize} & \nabla f(x)^T d + \frac{1}{2} d^T H d & \qquad (18) \\ \text{subject to} & h_i(x) + \nabla h_i(x)^T d = 0 \quad i = 1, \ldots, m & \qquad (19) \\ & g_j(x) + \nabla g_j(x)^T d \leq 0 \quad j = 0, 1, \ldots r. & \qquad (20) \end{array}$$

We denote the solution $\boldsymbol{d}$ of this program by $\boldsymbol{d(H, x)}$ and the set of its multipliers by $\boldsymbol{\mathcal{M}(H, x)}$. With these notations, $d(H, x^*) = 0$ and $\mathcal{M}(H, x^*) = \mathcal{M}(x^*)$, for any H satisfying (15). At $x = x^k$ and $H = H^k$ we denote the solution of (18–19) by $d^{\dagger k}, \lambda^{\dagger k}, \mu_m^{\dagger k}$. If

$$c^k = c > \gamma + \sum_{i=1}^{m} |\lambda_i^{\dagger k}| + \sum_{j=1}^{r} \mu_j^{\dagger k} \tag{21}$$

(for at least one set of multipliers $(\lambda^{\dagger k}, \mu^{\dagger k})$), then $(d^{\dagger k}, \zeta^k = 0)$ is a solution of (12–14), if $H = H^k$ [1, p. 195]. We therefore concentrate on the quadratic program (18–20), since, if c^k is large enough and we are sufficiently close to x^*, it generates the same descent direction as (12–14), thus sharing its global convergence property.

2 Properties of the Nondifferentiable Exact Penalty Function for the Degenerate Case

Theorem 2.1 constructs an object that is similar to the augmented Lagrangian. The augmentation is, however, different for the zero and nonzero multipliers corresponding to the active inequalities. This new function has one but not two continuous derivatives, and the Hessian is not defined. However, we prove that its gradient is a monotone mapping, and this suffices to ensure the right descent properties of the penalty function $\phi(x)$.

Theorem 2.1 *There exists a neighborhood $\mathcal{V}(x^*)$ and $\nu > 0$, $\beta > 0$, $\omega > 0$, such that $\forall(\lambda^*, \mu^*) \in \mathcal{M}(x^*)$ and $x \in \mathcal{V}(x^*)$*
a) $\mathcal{L}_\omega(x, \lambda^*, \mu^*) - \mathcal{L}_\omega(x^*, \lambda^*, \mu^*) \geq \beta\|x - x^*\|^2$,
b) $(x - x^*)^T \nabla_x \mathcal{L}_\omega(x, \lambda^*, \mu^*) \geq \beta\|x - x^*\|^2$,
where

$$\begin{aligned} \mathcal{L}_\omega(x, \lambda^*, \mu^*) &= f(x) + (\lambda^*)^T h(x) + (\mu^*)^T g(x) \\ &+ \omega(\|h(x)\|^2 + \sum_{i \in \mathcal{B},\, \mu_i^* > \nu} |g_i^-(x)|^2 + \sum_{i \in \mathcal{B}} |g_i^+(x)|^2) \end{aligned} \tag{22}$$

and $g_i^+(x) = \max\{0, g_i(x)\}$, $g_i^-(x) = \max\{0, -g_i(x)\}$ *are the nonnegative and non-positive part of the constraint* $g_i(x)$*, respectively.*

Lemma 2.2 *We have that*

$$\textbf{a)} \qquad \lim_{x\to x^*} \frac{\mathcal{L}(x,\xi^*) - \mathcal{L}(x^*,\xi^*) - \frac{1}{2}(x-x^*)\nabla_{xx}\mathcal{L}(x^*,\xi^*)(x-x^*)}{\|x-x^*\|^2} = 0 \qquad (23)$$

$$\textbf{b)} \qquad \lim_{x\to x^*} \frac{\mathcal{L}(x,\xi^*) - \mathcal{L}(x^*,\xi^*) - \frac{1}{2}\nabla_x\mathcal{L}(x,\xi^*)^T(x-x^*)}{\|x-x^*\|^2} = 0 \qquad (24)$$

uniformly as $\xi^* = (\lambda^*, \mu^*) \in \mathcal{M}(x^*)$*. Here* $\mathcal{L}(x,\xi^*) = \mathcal{L}(x,\lambda^*,\mu^*)$ *is the Lagrangian function defined in (3).*

Proof As a consequence of the fact that (6) is satisfied, $\mathcal{M}(x^*)$, the set of the optimal multipliers, is bounded, convex, closed, and polygonal. Let $\xi^{*j} = (\lambda^{*j}, \mu^{*j}), j = 1,\ldots,p$ be the set of its vertices. For each $\xi^* = (\lambda^*, \mu^*)$, there exists a set of numbers $0 \le u_j \le 1$, $j = 1,\ldots,p$, such that $\sum_{j=1}^p u_j = 1$ and $\xi^* = \sum_{j=1}^p u_j\xi^{*j}$. Since $f(x), g(x), h(x)$ are twice continuously differentiable, so is $\mathcal{L}(x,\xi^*)$. By the definition of differentiability, we have that, for all $\xi^* \in \mathcal{M}(x^*)$,

$$\lim_{x\to x^*} \frac{\mathcal{L}(x,\xi^*) - \mathcal{L}(x^*,\xi^*) - \frac{1}{2}(x-x^*)\nabla_{xx}\mathcal{L}(x^*,\xi^*)(x-x^*)}{\|x-x^*\|^2} = 0 \qquad (25)$$

because $\nabla_x\mathcal{L}(x^*,\xi^*) = 0, \forall\xi^* \in \mathcal{M}(x^*)$, by the KKT conditions (4). Since $\mathcal{L}(x,\xi^*)$ is an affine mapping with respect to the variable ξ^*, we have that

$$\frac{\mathcal{L}(x,\xi^*) - \mathcal{L}(x^*,\xi^*) - \frac{1}{2}(x-x^*)\nabla_{xx}\mathcal{L}(x^*,\xi^*)(x-x^*)}{\|x-x^*\|^2} = \qquad (26)$$

$$\sum_{j=1}^p u_j \frac{\mathcal{L}(x,\xi^{*j}) - \mathcal{L}(x^*,\xi^{*j}) - \frac{1}{2}(x-x^*)\nabla_{xx}\mathcal{L}(x^*,\xi^{*j})(x-x^*)}{\|x-x^*\|^2} \qquad (27)$$

for some $u_j > 0, j = 1,\ldots,p$, where $\sum_{j=1}^p u_j = 1$. Thus

$$\frac{|\mathcal{L}(x,\xi^*) - \mathcal{L}(x^*,\xi^*) - \frac{1}{2}(x-x^*)\nabla_{xx}\mathcal{L}(x^*,\xi^*)(x-x^*)|}{\|x-x^*\|^2} \le \qquad (28)$$

$$\sum_{j=1}^p \frac{|\mathcal{L}(x,\xi^{*j}) - \mathcal{L}(x^*,\xi^{*j}) - \frac{1}{2}(x-x^*)\nabla_{xx}\mathcal{L}(x^*,\xi^{*j})(x-x^*)|}{\|x-x^*\|^2}. \qquad (29)$$

Since the right-hand side is independent of ξ^* and its limit is zero as $x \to x^*$, the conclusion of part a follows. Part b is proved in an identical manner, after we note that

$$\lim_{x\to x^*} \frac{\mathcal{L}(x,\xi^*) - \mathcal{L}(x^*,\xi^*) - \frac{1}{2}\nabla_x\mathcal{L}(x,\xi^*)^T(x-x^*)}{\|x-x^*\|^2} = 0$$

for all $\xi^* \in \mathcal{M}(x^*)$, by part a and since

$$\lim_{x\to x^*} \frac{\|\nabla_x\mathcal{L}(x,\xi^*) - \nabla_{xx}\mathcal{L}(x^*,\xi^*)(x-x^*)\|}{\|x-x^*\|} = 0$$

by definition of differentiability. $\square$

As a result of Lemma 2.2, we can write

$$\mathcal{L}(x,\xi^*) - \mathcal{L}(x^*,\xi^*) - \frac{1}{2}(x-x^*)\nabla_{xx}\mathcal{L}(x^*,\xi^*)(x-x^*) = o(\|x-x^*\|^2),$$

where $o(\|x-x^*\|^2)$ denotes a quantity whose ratio to $\|x-x^*\|^2$ tends to zero as $x \to x^*$, uniformly for all $\xi^* \in \mathcal{M}(x^*)$.

Lemma 2.3 *There exists a neighborhood $\mathcal{V}_2(x^*)$ and $\omega > 0$ such that $\forall x \in \mathcal{V}_2(x^*)$ and $(\lambda^*,\mu^*) \in \mathcal{M}(x^*)$*

$$\mathcal{L}^*_\omega(x,\lambda^*,\mu^*) - \mathcal{L}^*_\omega(x^*,\lambda^*,\mu^*) \geq \frac{\sigma}{8}\|x-x^*\|^2,$$

where σ is the constant appearing in (7) and

$$\begin{aligned}\mathcal{L}^*_\omega(x,\lambda^*,\mu^*) &= f(x) + (\lambda^*)^T h(x) + (\mu^*)^T g(x) \\ +\omega(\|h(x)\|^2 &+ \sum_{i\in\mathcal{B},\,\mu_i^*>0}(g_i^-(x))^2 + \sum_{i\in\mathcal{B}}(g_i^+(x))^2).\end{aligned} \tag{30}$$

Note The difference between the definitions (30) and (22) resides in the terms $(g_i^-(x))$ that are considered for summation. Only terms corresponding to $\mu_i^* > \nu$ are summed in (22).

Proof We have, by (7–10), that

$$w^T\nabla_{xx}\mathcal{L}(x^*,\lambda^*,\mu^*)w \geq \sigma\|w\|^2$$

whenever $w \in \mathcal{F}(x^*,\lambda^*,\mu^*)$. It follows that there exists $\beta_1 > 0$ such that, for all $(\lambda^*,\mu^*) \in \mathcal{M}(x^*)$,

$$(w)^T\nabla_{xx}\mathcal{L}(x^*,\lambda^*,\mu^*)(w) \geq \tfrac{\sigma}{2}\|w\|^2 \tag{31}$$

$$\forall w \quad \text{such that} \quad -\beta_1\|w\| \leq \nabla_x h_i(x^*)^T w \leq \beta_1\|w\|, \quad i=1,\ldots,m \tag{32}$$

$$-\beta_1\|w\| \leq \nabla_x g_i(x^*)^T w \leq \beta_1\|w\| \quad i\in\mathcal{B}, \mu_i^* > 0 \tag{33}$$

$$\nabla_x g_i(x^*)^T w \leq \beta_1\|w\| \quad i\in\mathcal{B}, \mu_i^* = 0. \tag{34}$$

We denote the set of w's satisfying (32–34) as $\boldsymbol{\mathcal{F}_{\beta_1}(x^*,\lambda^*,\mu^*)}$. This can be proved by a continuity and compactness argument, using the fact that the set of Lagrange multipliers is bounded. Note that the bound in (7) is independent of the particular multiplier $(\lambda^*,\mu^*) \in \mathcal{M}(x^*)$. Let

$$\beta_2 = \max_{i=1,\ldots,m,j\in\mathcal{B}}\{\|\nabla_{xx}h_i(x^*)\|, \|\nabla_{xx}g_j(x^*)\|\}.$$

Let $\mathcal{V}_1(x^*)$ be a neighborhood of x such that

$$2\beta_2 \geq \max_{i=1,\ldots,m,j\in\mathcal{B}}\{\|\nabla_{xx}h_i(x)\|, \|\nabla_{xx}g_j(x)\|\}.$$

Also, let $\mathcal{V}_2(x^*)$ be the intersection between $\mathcal{V}_1(x^*)$ and the ball centered at x^* of radius $\frac{\beta_1}{4\beta_2}$. Now $x \in \mathcal{V}_2(x^*)$ such that $x - x^* \notin \mathcal{F}_{\beta_1}(x^*,\lambda^*,\mu^*)$. Then at least one of the constraints from (32–34) is violated, say, $\nabla g_i(x^*)^T(x-x^*) \geq \beta_1\|x-x^*\|$, where $i \in \mathcal{B}$ and $\mu_i^* = 0$. Then, by Taylor's remainder theorem,

$$g_i(x) = g_i(x^*) + \nabla g_i(x^*)^T(x-x^*) + \beta_3\|(x-x^*)\|^2,$$

where $|\beta_3| \le 2\beta_2$. Therefore

$$g_i(x) = \nabla g_i(x^*)^T (x - x^*) + \beta_3 \|x - x^*\|^2 \ge \frac{\beta_1}{2} \|x - x^*\|,$$

since $\beta_3 \|x - x^*\| \le 2\beta_2 \|x - x^*\| \le \frac{\beta_1}{2}$, and thus $g_i^+(x) = g_i(x) \ge \frac{\beta_1}{2}\|x - x^*\|$. Hence,

$$(\frac{\beta_1}{2})^2 \|x - x^*\|^2 \le \|h(x)\|^2 + \sum_{i \in \mathcal{B},\, \mu_i^* > 0} |g_i(x)|^2 + \sum_{i \in \mathcal{B},\, \mu_i^* = 0} |g_i^+(x)|^2. \tag{35}$$

With the same argument if another constraint from (32–34) is violated, it follows that (35) holds whenever $x \in \mathcal{V}_2(x^*)$ and $x - x^* \notin \mathcal{F}_{\beta_1}(x^*, \lambda^*, \mu^*)$. Now

$$\beta_4 = \max_{(\lambda^*, \mu^*) \in \mathcal{M}(x^*)} \|\nabla_{xx} \mathcal{L}(x^*, \lambda^*, \mu^*)\| \tag{36}$$

is finite because $\mathcal{M}(x^*)$ is bounded. Choose

$$\omega_1 = \frac{\frac{\sigma}{4} + \beta_4}{(\frac{\beta_1}{2})^2}. \tag{37}$$

It follows that, if $x \in \mathcal{V}_2(x^*)$ and $(\lambda^*, \mu^*) \in \mathcal{M}(x^*)$, we have that either $x - x^* \in \mathcal{F}_{\beta_1}(x^*, \lambda^*, \mu^*)$ and then by (31–34) it follows that

$$\frac{\sigma}{4} \|x - x^*\|^2 \le \frac{1}{2} (x - x^*)^T \nabla_{xx} \mathcal{L}(x^*, \lambda^*, \mu^*)(x - x^*), \tag{38}$$

or $x - x^* \notin \mathcal{F}_{\beta_1}(x^*, \lambda^*, \mu^*)$ and then by (36) and (35)

$$\begin{aligned}
&\omega_1(\|h(x)\|^2 + \textstyle\sum_{i \in \mathcal{B},\, \mu_i^* > 0} |g_i(x)|^2 + \sum_{i \in \mathcal{B},\, \mu_i^* = 0} |g_i^+(x)|^2) \ge \\
&\quad \omega_1 (\tfrac{\beta_1}{2})^2 \|x - x^*\|^2 \ge \tfrac{\sigma}{4} \|x - x^*\|^2 + \beta_4 \|x - x^*\|^2 \ge \\
&\quad \tfrac{\sigma}{4} \|x - x^*\|^2 - \tfrac{1}{2} (x - x^*)^T \nabla_{xx} \mathcal{L}(x^*, \lambda^*, \mu^*)(x - x^*).
\end{aligned} \tag{39}$$

As a result of (38) and (39), we have that for all $(\lambda^*, \mu^*) \in \mathcal{M}(x^*)$ and $x \in \mathcal{V}_2(x^*)$

$$\begin{aligned}
\frac{\sigma}{4} \|x - x^*\|^2 \quad \le \quad & \frac{1}{2} (x - x^*)^T \nabla_{xx} \mathcal{L}(x^*, \lambda^*, \mu^*)(x - x^*) \\
& + \omega_1 (\|h(x)\|^2 + \sum_{i \in \mathcal{B},\, \mu_i^* > 0} |g_i(x)|^2 + \sum_{i \in \mathcal{B},\, \mu_i^* = 0} |g_i^+(x)|^2).
\end{aligned} \tag{40}$$

By Lemma 2.2, part a, and (40) we have that

$$\mathcal{L}^*_{\omega_1}(x, \lambda^*, \mu^*) - \mathcal{L}^*_{\omega_1}(x^*, \lambda^*, \mu^*) = (x - x^*)^T \nabla_{xx} \mathcal{L}(x^*, \lambda^*, \mu^*)(x - x^*) \tag{41}$$

$$+ o(\|x - x^*\|^2) + \omega_1 (\|h(x)\|^2 + \textstyle\sum_{i \in \mathcal{B},\, \mu_i^* > 0} |g_i(x)|^2 + \sum_{i \in \mathcal{B},\, \mu_i^* = 0} |g_i^+(x)|^2) \tag{42}$$

$$\ge \tfrac{\sigma}{4} \|x - x^*\|^2 + o(\|x - x^*\|^2), \tag{43}$$

where $o(\|x - x^*\|^2)$ denotes a quantity whose ratio to $\|x - x^*\|^2$ tends to zero uniformly for all $(\lambda^*, \mu^*) \in \mathcal{M}(x^*)$. We can therefore choose a neighborhood $\mathcal{V}_3(x^*)$ of x^* in which the quantity in (43) is bounded below by $\frac{\sigma}{8} \|x - x^*\|^2$, for all $\lambda^*, \mu^* \in \mathcal{M}(x^*)$. For all $x \in \mathcal{V}_3(x^*)$ we have

$$\mathcal{L}^*_{\omega_1}(x, \lambda^*, \mu^*) - \mathcal{L}^*_{\omega_1}(x^*, \lambda^*, \mu^*) \ge \frac{\sigma}{8} \|x - x^*\|^2, \tag{44}$$

for all $(\lambda^*, \mu^*) \in \mathcal{M}(x^*)$. This completes the proof of the lemma. $\square$

Proof of Theorem 2.1 As a consequence of the fact that (6) is satisfied, $\mathcal{M}(x^*)$, the set of the optimal multipliers, is bounded, convex, closed, and polygonal. Let $\xi^{*j} = (\lambda^{*j}, \mu^{*j}), j = 1, \ldots, p$ be the set of its vertices. For each $\xi^* = (\lambda^*, \mu^*)$, there exists a set of numbers $0 \leq u_j \leq 1$, $j = 1, \ldots, p$, such that $\sum_{j=1}^p u_j = 1$ and $\xi^* = \sum_{j=1}^p u_j \xi^{*j}$. With this representation of $\mathcal{M}(x^*)$, $u_j = 0$ whenever $\mu_i^* = 0$ and $\mu_i^{*j} > 0$. For each ξ^*, we choose $u(\xi^*)$ to be the unique solution of the following quadratic programming problem.

$$\min \|u\|^2 \tag{45}$$
$$\text{such that} \quad \xi^* = \textstyle\sum_{i=1}^p u_j \xi^{*j} \tag{46}$$
$$\textstyle\sum_{i=1}^p u_i = 1 \qquad u_i \geq 0, 1 \leq i \leq p. \tag{47}$$

Since quadratic programs can be recast as linear complementarity problems [3], they share sensitivity properties, in particular that the solution set satisfies the upper Lipschitz property, with ξ^* regarded as the parameter. Since $u(\xi^*)$ is uniquely defined, that upper Lipschitz property reduces to the usual Lipschitz property, which implies that $u(\xi^*)$ is a continous function.

Therefore, from Lemma 2.3 applied to each vertex ξ^{*j}, and since the Lagrangians are linear with respect to the multipliers, we have that $\forall x \in \mathcal{V}_3(x^*)$,

$$\sum_{i=1}^p u_i(\xi^*)(\mathcal{L}^*_{\omega_1}(x, \xi^{*i}) - \mathcal{L}^*_{\omega_1}(x^*, \xi^{*i})) \geq \sum_{i=1}^p u_i(\xi^*) \frac{\sigma}{8} \|x - x^*\|^2 = \frac{\sigma}{8}\|x - x^*\|^2. \tag{48}$$

Since $\mathcal{L}^*_{\omega_1}(x^*, \xi^*) = f(x^*)$, $\forall \xi^* \in \mathcal{M}(x^*)$, and $\sum_{j=1}^p u_j(\xi^*) = 1$, we have that the last inequality can be rewritten as

$$\frac{\sigma}{8}\|x - x^*\|^2 \quad \leq \quad f(x) - f(x^*) + (\lambda^*)^T h(x) + (\mu^*)^T g(x) \tag{49}$$
$$+ \quad \omega_1(\|h(x)\|^2 + \sum_{i \in \mathcal{B}} |g_i^+(x)|^2 + \sum_{i \in \mathcal{B}} \psi_i(\xi^*)|g_i^-(x)|^2), \tag{50}$$

where
$$\psi_i(\xi^*) = \sum_{j=1,\ldots,p,\, \mu_i^{*j} > 0} u_j(\xi^*). \tag{51}$$

Since $u_i(\xi^*)$ are continuous functions, so is $\psi_i(\xi^*)$. From (46) it follows that, if $\mu_i^* = 0$, then $u_j(\xi^*) = 0$, whenever $\mu_i^{*j} > 0$. Therefore, if $\mu_i^* = 0$, then $\psi_i(\xi^*) = 0$. Also, from (47), $\psi_i(\xi^*) \leq 1, \forall \xi^* \in \mathcal{M}(x^*)$ and $i \in \mathcal{B}$. Now

$$\psi_i^*(\epsilon) = \max_{\mu_i^* < \epsilon,\, \xi^* \in \mathcal{M}(x^*)} \psi_i(\xi^*), \quad i \in \mathcal{B}.$$

Since $\mathcal{M}(x^*)$ is a compact set, it follows that $\psi_i^*(\epsilon)$ is a continuous function and $\psi_i^*(0) = 0$. Let C be such that $|g_i(x)|^2 \leq C\|x - x^*\|^2, \forall x \in \mathcal{V}_3(x^*)$, $i \in \mathcal{B}$ and $\epsilon > 0$ the largest one for which
$$\max_{i \in \mathcal{B}} \{\psi_i^*(\epsilon)\} \leq \frac{\sigma}{16 C r \omega_1},$$
where r is the total number of inequalities in (1–2). With this choice of ϵ it follows from $\psi(\xi^*) \leq 1$ that

$$\textstyle\sum_{i \in \mathcal{B}} \psi_i(\xi^*)|g_i^-(x^*)|^2 = \sum_{i \in \mathcal{B},\, \mu_i^* \geq \epsilon} \psi_i(\xi^*)|g_i^-(x^*)|^2 \tag{52}$$

$$+\sum_{i\in\mathcal{B},\,\mu_i^*<\epsilon}\psi_i(\xi^*)|g_i^-(x^*)|^2\le\sum_{i\in\mathcal{B},\mu_i^*\ge\epsilon}|g_i^-(x^*)|^2 \tag{53}$$

$$+\sum_{i\in\mathcal{B},\mu_i^*<\epsilon}\psi_i^*(\epsilon)|g_i^-(x)|^2\le\sum_{i\in\mathcal{B},\mu_i^*\ge\epsilon}|g_i^-(x^*)|^2 \tag{54}$$

$$+rC\|x-x^*\|^2\max_{i\in\mathcal{B}}\psi_i^*(\epsilon)\le\sum_{i\in\mathcal{B},\mu_i^*\ge\epsilon}|g_i^-(x^*)|^2+\tfrac{\sigma}{16\omega_1}\|x-x^*\|^2 \tag{55}$$

If we majorize with this inequality in (50), taking $\frac{\sigma}{16}\|x-x^*\|^2$ to the right-hand side and use the definition (22), the claim of part a follows by taking $\beta=\frac{\sigma}{16}$, $\mathcal{V}(x^*)=\mathcal{V}_3(x^*)$, $\nu=\epsilon$, and $\omega=\omega_1$.

b) Let

$$\begin{aligned}\mathcal{L}_{x^\dagger,\omega}(x,\lambda^*,\mu^*) &= f(x)+(\lambda^*)^Th(x)+(\mu^*)^Tg(x)+\omega(\|h(x)\|^2\\ &+\sum_{i\in\mathcal{B},\,\mu_i^*\ge\nu,g_i(x^\dagger)<0}|g_i(x)|^2+\sum_{i\in\mathcal{B},g_i(x^\dagger)>0}|g_i(x)|^2).\end{aligned} \tag{56}$$

Then both $\mathcal{L}_{x^\dagger,\omega}(x,\lambda^*,\mu^*)$ and its gradient coincide with $\mathcal{L}_\omega(x,\lambda^*,\mu^*)$ and its gradient, respectively, at $x^\dagger$ and x. Also, $\mathcal{L}_{x^\dagger,\omega}(x^*,\lambda^*,\mu^*)=\mathcal{L}(x^*,\lambda^*,\mu^*)=f(x^*)$ and $\mathcal{L}_{x^\dagger,\omega}$ has at least two continous derivatives. We then have, by an argument similar to part b of Lemma 2.2,

$$|\mathcal{L}_{x^\dagger,\omega}(x,\lambda^*,\mu^*)-\mathcal{L}_{x^\dagger,\omega}(x^*,\lambda^*,\mu^*)-\frac{1}{2}\nabla_x\mathcal{L}_{x^\dagger,\omega}(x,\lambda^*,\mu^*)^T(x-x^*)|=o(\|x-x^*\|^2)$$

for all $(\lambda^*,\mu^*)\in\mathcal{M}(x^*)$. Therefore, by taking the previous relation at $x=x^\dagger$ and replacing $x^\dagger$ with x and $\mathcal{L}_{x^\dagger,\omega}$ with $\mathcal{L}_\omega$, we get

$$|\mathcal{L}_\omega(x,\lambda^*,\mu^*)-\mathcal{L}_\omega(x^*,\lambda^*,\mu^*)-\frac{1}{2}\nabla_x\mathcal{L}_\omega^T(x,\lambda^*,\mu^*)(x-x^*)|=o(\|x-x^*\|^2),$$

where $o(\|x-x^*\|^2)$ again denotes a quantity that converges to zero as $x\to x^*$ uniformly with respect to $(\lambda^*,\mu^*)\in\mathcal{M}(x^*)$. From the last relation and part a of the theorem, there exists a neighborhood $\mathcal{V}_4(x^*)$ on which the b of the theorem holds, after eventually replacing β with $\frac{\beta}{2}$. □

The techniques from [2] can be used to ensure that if c is large enough, x^* is a stationary point of $\phi(x)$. However, to secure local convergence of the SQP to x^*, one would need to guarantee that x^* is a local minimum of the penalty function.

The following theorem gives an alternative proof of this fact for the degenerate case considered in this paper, provided that the constraint qualification and second-order sufficiency condition hold as specified in Section 1.1. Note, however, that the theorem does not claim that x^* is an isolated stationary point, which is established later.

Corollary 2.4 *Let c be such that*

$$c>\sum_{i=1}^m|\lambda_i^*|+\sum_{j\in\mathcal{B}}\mu_i^*$$

for some $(\lambda^*,\mu^*)\in\mathcal{M}(x^*)$. Then there exists a neighborhood $\mathcal{W}(x^*)$ of x^* such that

$$\phi(x)-\phi(x^*)=f(x)+cP(x)-f(x^*)\ge\beta\|x-x^*\|^2,$$

where β is the constant from Theorem 2.1.

Proof By Theorem 2.1, there exists a neighborhood $\mathcal{V}(x^*)$ and $\omega > 0$ such that for all $x \in \mathcal{V}(x^*)$ we have

$$\beta\|x - x^*\|^2 \leq f(x) - f(x^*) + (\lambda^*)^T h(x) + (\mu^*)^T g(x) \tag{57}$$
$$+\omega(\|h(x)\|^2 + \sum_{i\in\mathcal{B},\, \mu_i^* > \nu} |g_i^-(x)|^2 + \sum_{i\in\mathcal{B}} |g_i^+(x)|^2). \tag{58}$$

By continuity, there exists a neighborhood $\mathcal{W}(x^*)$ such that for all $x \in \mathcal{W}(x^*)$ and $i \in \mathcal{B}$ we have $\nu \geq \ \omega|g_i(x)| \geq 0$ and

$$c > \sum_{i=1}^{m}(|\lambda_i^*| + \omega|h_i(x)|) + \sum_{j\in\mathcal{B}}(\mu_j^* + \omega g_j^+(x)).$$

With this choice of $\mathcal{W}(x^*)$ it follows that, for all $x \in \mathcal{W}(x^*)$, $(\mu_i^* - \omega g_i^-(x)) \geq 0$, whenever $\mu_i^* > \nu$ and, since $g_i(x) = g_i^+(x) - g_i^-(x)$,

$$\begin{aligned}
cP(x) \quad &\geq \quad \sum_{i=1}^{m}(|\lambda_i^*| + \omega|h_i(x)|)|h_i(x)| + \sum_{i\in\mathcal{B}}(\mu_i^* + \omega g_i^+(x))g_i^+(x) \\
&\geq \quad \sum_{i=1}^{m}(|\lambda_i^*| + \omega|h_i(x)|)|h_i(x)| + \sum_{i\in\mathcal{B},\, \mu_i^*>\nu}(-\mu_i^* + \omega g_i^-(x))g_i^-(x) \qquad (59) \\
&+ \quad \sum_{i\in\mathcal{B},\, \mu_i^*\leq\nu} -\mu_i^* g_i^-(x) + \sum_{i\in\mathcal{B}}(\mu_i^* + \omega g_i^+(x))g_i^+(x).
\end{aligned}$$

The conclusion follows by using inequality (59) in (58), since $\mu_i^* g(x) = \mu_i^*(g^+(x) - g^-(x))$. □

Note Using the same line of proof, we can prove Corollary 2.4 even when the second-order sufficient conditions (7–10) hold for just one Lagrange multiplier. In that case, Lemma 2.3, can be used instead of Theorem 2.1, because its proof does not rely on satisfying the second-order sufficient conditions for all Lagrange multipliers. Therefore, Corollary 2.4 could constitute an alternative proof of the fact that, if (6) and (7–10) are satisfied for one Lagrange multiplier, x^* is a local minimum for the original problem [5] and that a penalty function can be defined that has x^* as an unconstrained strict local minimizer [8].

An important consequence of the Mangasarian-Fromowitz constraint qualification (6) is that the solution of (18–20) is continuous with respect to x and H.

Lemma 2.5 *For any $\epsilon > 0$ there exists a neighborhood $\mathcal{V}_\epsilon(x^*)$ of x^* such that $\forall x \in \mathcal{V}_\epsilon(x^*)$, H satisfying (15), we have $\|d(H,x)\| \leq \epsilon$. In addition, there exists ϵ_0 such that, $\forall x \in \mathcal{V}_{\epsilon_0}(x^*)$ and $u \in \mathcal{M}(H,x)$,*

$$\min_{u^*\in\mathcal{M}(x^*)} \|u - u^*\| \leq \epsilon_u(\|d\| + \|x - x^*\|).$$

Proof Since Condition (6) holds at x^* for (1–2), it also holds in a neighborhood $\mathcal{V}(x^*)$ for all quadratic programs (18–20). Therefore, $d(H,x)$ is a continuous function on $\mathcal{V}(x^*) \times \mathcal{H}$, from [12, Corollary 4.3], since the second-order sufficient conditions (7–10) are satisfied at all points (the matrix H is positive definite).

Thus, the function

$$\psi(\delta) = \max_{H\in\mathcal{H}, \|x - x^*\|\leq\delta} \|d(H,x)\| \tag{60}$$

is a continuous function on some interval $[0,\Delta]$, $\Delta > 0$. $\psi(0) = 0$, since $d(H,x^*) = 0, \forall H \in \mathcal{H}$. By the continuity of ψ, it follows that, $\forall \epsilon_1$, there exists δ_1 such that

$$\forall H \in \mathcal{H} \quad \forall x, \text{ such that} \|x - x^*\| \leq \delta_1 \quad \Rightarrow d(H,x) \leq \epsilon_1. \tag{61}$$

We now define the following perturbed quadratic program:

$$\begin{aligned} &\text{minimize} \quad (\nabla f(x) + f_\delta)^T w + \tfrac{1}{2} w^T w && (62)\\ &\text{subject to} \quad h_i(x) + \nabla h_i(x)^T w = 0 \qquad i = 1,\ldots,m && (63)\\ &\qquad\qquad\quad g_j(x) + \nabla g_j(x)^T w \leq 0 \qquad j = 0,1,\ldots r && (64)\end{aligned}$$

At x^* and for $f_\delta = 0$, (62–64) satisfies the Mangasarian-Fromowitz constraint qualifications and the second-order sufficiency conditions. We regard f_δ and $x - x^*$ as perturbations. The sensitivity results from [12] therefore apply to this case: There exist ϵ_2, ϵ_3 such that whenever $\|f_\delta\| \leq \epsilon_2$, $\|x - x^*\| \leq \epsilon_2$, we have

$$\min_{u^* \in \mathcal{M}(x^*)} \|u - u^*\| \leq \epsilon_3(\|f_\delta\| + \|x - x^*\|)$$

for any u a Lagrange multiplier of (62–64). By inspection it follows that, if $f_\delta = (H - I)d(H,x)$, then $d(H,x)$ and $\mathcal{M}(H,x)$ are the solution andthe set of Lagrange multipliers of (62–64), respectively. If we choose δ_1 and ϵ_1 in (61) such that $(2\Gamma_0 + 1)\epsilon_1 \leq \epsilon_2$, it follows that, for all x such that $\|x - x^*\| \leq \delta_1$ and $u \in \mathcal{M}(H,x)$ we have

$$\min_{u^* \in \mathcal{M}(x^*)} \|u - u^*\| \leq \epsilon_3(\|(H - I)d(H,x)\| + \|x - x^*\|).$$

The conclusion now follows by taking $\epsilon_u = (\Gamma_0 + 1)\epsilon_3$ and $\epsilon_0 = \epsilon_1$. □

A stronger version of this result is proved in [12] where the right hand side in the conclusion of Lemma 2.5 does not contain a term involving $\|d\|$, for a given QP matrix H. The difference, however, is that the above bound is independent of $H \in \mathcal{H}$.

The following theorem is the main result of this section. It establishes a connection between the size of the direction generated by (18–20) and the distance from the current point x^*.

Theorem 2.6 *There exist a neighborhood $\mathcal{W}(x^*)$ and a constant σ_1 such that, $\forall x \in \mathcal{W}(x^*)$,*

$$\|d(H,x)\|^2 + P(x) + \mu^T g^-(x) \geq \sigma_1 \|x - x^*\|^2,$$

$\forall \mu$ such that $(\lambda,\mu) \in \mathcal{M}(H,d)$ for some λ.

Proof Let $(\lambda,\mu) \in \mathcal{M}(H,x)$, and let $(\lambda^*,\mu^*) \in \mathcal{M}(x^*)$ such that $\|(\lambda,\mu) - (\lambda^*,\mu^*)\| \leq \epsilon_u(\|x - x^*\| + \|d\|)$. By part b of Theorem 2.1 we have, for all x in a neighborhood $\mathcal{V}(x^*)$,

$$\begin{aligned}&(x - x^*)^T(\nabla f(x) + (\lambda^*)^T \nabla h(x) + (\mu^*)^T g(x) + 2\omega \nabla(h(x))^T h(x) \qquad (65)\\ &+2\omega \textstyle\sum_{i\in\mathcal{B},\mu_i>\nu} g_i^-(x)\nabla g_i(x) + 2\omega \sum_{i\in\mathcal{B}} g_i^+(x)\nabla g_i(x)) \geq \beta\|x - x^*\|^2.\end{aligned}$$

Since $d(H,x)$ is a solution of (18–20) it follows, by the first-order necessary conditions, that

$$\nabla f(x) = -Hd - \nabla h(x)^T \lambda - \nabla g(x)^T \mu.$$

Replacing this relation in (65), we get

$$\begin{aligned}&(x - x^*)^T(-Hd + \nabla h(x)^T(\lambda^* - \lambda) + \nabla g(x)^T(\mu^* - \mu) + 2\omega\nabla(h(x))^T h(x) \qquad (66)\\ &+2\omega \textstyle\sum_{i\in\mathcal{B},\mu_i>\nu} g_i^-(x)\nabla g_i(x) + 2\omega \sum_{i\in\mathcal{B}} g_i^+(x)\nabla g_i(x)) \geq \beta\|x - x^*\|^2. \qquad (67)\end{aligned}$$

Since $\|(\lambda, \mu) - (\lambda^*, \mu^*)\| \leq \epsilon_u(\|x - x^*\| + \|d\|)$, by Lemma 2.5, and $h(x) = \nabla h(x)^T(x - x^*) + O(\|x - x^*\|^2)$ as well as $g_j(x) = \nabla g_j(x)^T(x - x^*) + O(\|x - x^*\|^2)$, $\forall j \in \mathcal{B}$, we have that

$$\begin{aligned}(\mu_j - \mu_j^*)\nabla g_j(x)^T(x - x^*) &= (\mu_j - \mu_j^*)g_j + O(\|d\|\|x - x^*\|^2), \quad \forall j \in \mathcal{B} \\ (\lambda - \lambda^*)^T \nabla h(x)^T(x - x^*) &= (\lambda - \lambda^*)h(x) + O(\|x - x^*\|^3) + O(\|d\|\|x - x^*\|^2).\end{aligned}$$

Also, since $|g_j(x)| \leq C\|x - x^*\|, \forall j \in \mathcal{B}$, we have that

$$\begin{aligned}g_i^+(x)\nabla g_i(x)^T(x - x^*) &= (g_i^+(x))^2 + O(\|x - x^*\|^3), \\ g_i^-(x)\nabla g_i(x)^T(x - x^*) &= (g_i^-(x))^2 + O(\|x - x^*\|^3).\end{aligned}$$

By replacing the last four relations in (66–67) we get, after eventually restricting the neighborhood $\mathcal{V}(x^*)$ to get the terms $O(\|x - x^*\|^3)$ and $O(\|d\|\|x - x^*\|^2) = o(\|x - x^*\|^2)$ sufficiently small, that for all $x \in \mathcal{V}(x^*)$,

$$\begin{aligned}&-(x - x^*)^T Hd + h(x)^T(\lambda^* - \lambda) + g(x)^T(\mu^* - \mu) + 2\omega\|h(x)\|^2 \\ &+2\omega \textstyle\sum_{i\in\mathcal{B},\mu_i>\nu}(g_i^-(x))^2 + 2\omega \sum_{i\in\mathcal{B}}(g_i^+(x))^2 \geq \frac{\beta}{2}\|x - x^*\|^2.\end{aligned} \tag{68}$$

Since $g(x) = g^+(x) - g^-(x)$, we have that

$$g(x)^T(\mu^* - \mu) \leq \mu^T g(x)^- + (\mu^*)^T(g(x)^+ - g(x)^-).$$

There exists a neighborhood $\mathcal{W}(x^*)$ such that $\omega g_j^-(x) \leq \nu, \forall j \in \mathcal{B}$. On $\mathcal{W}(x^*)$ we have that

$$0 \geq -\mu^* g^-(x) + \omega \sum_{j\in\mathcal{B},\mu_j\geq\nu} g_j^-(x)^2.$$

Therefore, on $\mathcal{W}(x^*) \cap \mathcal{V}(x^*)$ we have, by using the previous relations and Lemma 2.5 in (68), as well as the boundedness of the multipliers as a result of (6), that

$$-(x - x^*)^T Hd + CP(x) + \mu^T g^-(x) \geq \frac{\beta}{2}\|x - x^*\|^2 \tag{69}$$

for some constant C. From our construction, it results that the bounds and neighborhood restrictions are uniform with respect to H. Using that $-(x - x^*)Hd \leq \|x - x^*\|\|Hd\|$ and denoting $A = \|x - x^*\|$, $B = \|Hd\|$, and $D = CP(x) + \mu^T g^-(x)$, we can write the previous inequality as

$$AB + D \geq \frac{\beta}{2}A^2.$$

From the quadratic formula, we have that

$$A^2 \leq \left(\frac{B + \sqrt{B^2 + 4D\frac{\beta}{2}}}{\frac{\beta}{2}}\right)^2 \leq \frac{4}{\beta^2}(B^2 + B^2 + 4D\frac{\beta}{2})$$

and, therefore, by an appropriate choice of a majorizing constant,

$$\|x - x^*\|^2 \leq C_1(\mu^T g^-(x) + P(x) + \|d(H, x)\|^2).$$

since $\|Hd\| \leq \Gamma_0 \|d\|$ by (15). The conclusion follows by choosing $\sigma_1 = \frac{1}{C_1}$. □

The following corollary establishes that a nondifferentiable penalty for the problem (1–2) can be defined for which x^* is a strict minimum and an isolated stationary point in an appropriate neighborhood. This fact can also be established based on [12] and [4], but our developments also give an upper bound of the constant c that makes the penalty function $\phi(x) = f(x) + cP(x)$ exact.

Corollary 2.7 *There exists a neighborhood $\mathcal{W}_1(x^*)$ such that x^* is the unique stationary point of $\phi(x) = f(x) + cP(x)$, where $c > \sum_{i=1}^m |\lambda_i^*| + \sum_{i\in\mathcal{B}} \mu_i^*$ for all $(\lambda^*, \mu^*) \in \mathcal{M}(x^*)$.*

Proof We take $\mathcal{W}_1(x^*)$ to be the intersection between $\mathcal{W}(x^*)$ (from the previous theorem) and the set where

$$c > \sum_{i=1}^{m} |\lambda_i| + \sum_{i\in\mathcal{B}} \mu$$

for all $(\lambda, \mu) \in \mathcal{M}(x)$. By Lemma 2.5, $\mathcal{W}_1(x^*)$ contains an open set centered at x^*.

Assume that $x \in \mathcal{W}_1(x^*)$ is another stationary point of $\phi(x)$. Then, $d(H, x) = 0$, since the solutions of (18–20) and (12–14) are identical under our assumptions concerning c. Hence, $P(x) = 0$, since $d(H, x)$ is a feasible point of (18–20). Then $\mu^T g^-(x) = \mu^T g(x) = -\mu^T \nabla g^T d$, by complementarity, and therefore $\mu^T g^-(x) = 0$. By Theorem 2.6, $x = x^*$, which proves our claim. □

3 Linear Convergence of the SQP with Nondifferentiable Exact Penalty $P(x)$

In this section we assume that $x^k \to x^*$ under the condition stated in Section 1.1. From the update rule (17) and Lemma 2.5, it follows that the update in (17) can be triggered only a *finite* number of times, or otherwise $\mathcal{M}(x^*)$ cannot be bounded. We can therefore assume, without loss of generality, that (21) is satisfied at all steps k and that d^k is obtained from the quadratic program (18–20). For this section, we introduce the following notation:

$$M(\mu, x) = \sum_{i\in\mathcal{B}} \mu_i g_i^-(x). \tag{70}$$

3.1 Outline of the Proof

The proof consists of two major steps (each statement is made for k sufficiently large).
Step 1 There exist $\overline{\alpha}, \overline{\alpha} \leq 1$, and a constant c_2 such that, for some $(\lambda^k, \mu^k) \in \mathcal{M}(H^k, x^k)$

$$\phi(x^k + \alpha d^k) - \phi(x^k) \leq -c_2\alpha(\|d^k\|^2 + P(x^k) + M(\mu^k, x^k)), \quad \forall \alpha \in [0, \overline{\alpha}]. \tag{71}$$

The major accomplishment of this step is that it connects the decrease of $\phi(x)$ with the value of the penalty function $P(x)$. Previous analyses concerning nondifferentiable penalty functions bound the decrease only by $(d^k)^T H^k d^k$ [2].
Step 2 There exists a constant c_3 such that

$$\phi(x^k) - \phi(x^*) \leq c_3(M(\mu^k, x^k) + P(x^k) + \|d^k\|^2 + \|x^k - x^*\|^2).$$

This step is proven by making use of the properties of the Lagrangian function defined in Theorem 2.1.

As a result of Theorem 2.6, there is a constant c_4 such that

$$\|x^k - x^*\|^2 \leq c_4(\|d^k\|^2 + P(x^k) + M(\mu^k, x^k)). \tag{72}$$

As a result of Step 2, there exists a constant c_6 such that

$$\phi(x^k) - \phi(x^*) \leq c_6(M(\mu^k, x^k) + P(x^k) + \|d^k\|^2).$$

Assuming that the length of the step is at least $\overline{\alpha}_1 \leq \overline{\alpha}$ at each iteration, we have from Steps 1 and 2 that

$$\phi(x^{(k+1)}) - \phi(x^k) \leq -c_2\overline{\alpha}_1(\|d^k\|^2 + P(x^k) + M(\mu^k, x^k)) \leq \tag{73}$$

$$-\tfrac{c_2\overline{\alpha}_1}{c_6}(\phi(x^k) - \phi(x^*)). \tag{74}$$

Adding $\phi(x^k) - \phi(x^*)$ to both sides, we obtain

$$\phi(x^{(k+1)} - \phi(x^*)) \leq (1 - \frac{c_2\overline{\alpha}_1}{c_6})(\phi(x^k) - \phi(x^*)),$$

which proves linear convergence with a rate of at most $(1 - \frac{c_2\overline{\alpha}_1}{c_6})$. Procedures that ensure that the stepsize is bounded below are described in Section 3.3.

3.2 Proof of the Technical Results

All statements made in this section assume either that x is in a sufficiently small neighborhood of x^* or that k in x^k is sufficiently large.

Lemma 3.1

$$P(x^k + \alpha d^k) \leq (1-\alpha)P(x_k) + c_1\alpha^2\|d^k\|^2, \quad \forall \alpha \in [0,1].$$

Proof Since d^k is a feasible point of (18–20), we have that $\nabla g_i(x^k)^T d^k \leq -g_i(x^k), \forall i \in \mathcal{B}$. By Taylor's remainder theorem

$$g_i(x^k + \alpha d^k) \leq (1-\alpha)g_i(x^k) + d_i\alpha^2\|d^k\|^2, \forall \alpha \in [0,1]$$

for some nonnegative constants d_i, $i \in \mathcal{B}$. Similarly, $\nabla h_i(x^k)^T d^k = -h_i(x^k), \forall i = 1, \ldots, m$ and

$$h_i(x^k + \alpha d^k) = (1-\alpha)h_i(x^k) + \alpha^2 O(\|d^k\|^2).$$

Therefore,

$$|h_i(x^k + \alpha d^k)| \leq (1-\alpha)|h_i(x^k)| + e_i\alpha^2\|d^k\|^2, \forall \alpha \leq 1,$$

for some nonnegative constants $e_i, i = 1, \ldots, m$. Hence

$$\max\{|h_1(x)|, \ldots |h_m(x)|, g_1(x), \ldots g_r(x)\} \leq (1-\alpha)\max_{1\leq i\leq m, j\in\mathcal{B}}\{|h_i(x^k)|, g_i(x^k)\} + \alpha^2\|d^k\|^2 \max_{1\leq i\leq m, j\in\mathcal{B}}\{d_i, e_j\} \leq (1-\alpha)P(x^k) + c_1\alpha^2\|d^k\|^2, \quad \forall \alpha \in [0,1].$$

This completes the proof. $\square$

Lemma 3.2 *There exist $\overline{\alpha}$, $0 < \overline{\alpha} \leq 1$, and $c_2 > 0$ such that, for some $(\lambda^k, \mu^k) \in \mathcal{M}(H^k, x^k)$*

$$\phi(x^k + \alpha d^k) - \phi(x^k) \leq -\alpha\tfrac{1}{2}((d^k)^T H^k d^k + \tfrac{\gamma}{2}P(x^k) + M(\mu^k, x^k)) \leq -c_2\alpha(\|d^k\|^2 + P(x^k) + M(\mu^k, x^k)), \quad \forall \alpha \in [0, \overline{\alpha}].$$

Proof Writing the KKT conditions for (18–20), we obtain

$$H^k d^k + \nabla f(x^k) + \sum_{i \in \mathcal{B}} \mu_i^k \nabla g_i(x^k) + \sum_{j=1}^{m} \lambda_j^k \nabla h_j(x^k) = 0$$

and, hence,

$$\begin{array}{c} (d^k)^T H^k d^k + \nabla f(x^k)^T d^k + \sum_{i\in\mathcal{B}} \mu_i^k \nabla g_i(x^k)^T d^k + \sum_{j=1}^m \lambda_j^k \nabla h_j(x^k)^T d^k = 0 \\ (d^k)^T H^k d^k + \nabla f(x^k)^T d^k - \sum_{i\in\mathcal{B}} \mu_i^k g_i(x^k) - \sum_{j=1}^m \lambda_j^k h_j(x^k) = 0, \end{array}$$

since, by the complementarity conditions satisfied by the solution of (18–20), $\mu^k \nabla g(x^k)^T d^k = -\mu^k g(x^k), \forall i \in \mathcal{B}$ and $\nabla h(x^k)^T d^k = -g(x^k), \forall i = 1, \ldots, m$. Therefore, since $g_i(x^k) = g_i^+(x^k) - g_i^-(x^k)$,

$$\begin{array}{c} \nabla f(x^k)^T d^k = -(d^k)^T H^k d^k + \sum_{i\in\mathcal{B}} \mu_i^k (g_i^+(x^k) - g_i^-(x^k)) + \sum_{j=1}^m \lambda_j^k h_j(x^k)^T \leq \\ -(d^k)^T H^k d^k + P(x^k)(\sum_{i\in\mathcal{B}} \mu_i^k + \sum_{j=1}^m |\lambda_j^k|) - M(\mu^k, x^k) \leq \\ -(d^k)^T H^k d^k + (c - \frac{\gamma}{2}) P(x^k) - M(\mu^k, x^k) \end{array} \quad (75)$$

for sufficiently large k, by (11), (21), (70). By Taylor's remainder theorem,

$$f(x^k + \alpha d^k) \leq f(x^k) + \alpha \nabla f(x^k)^T d^k + \overline{c}_2 \alpha^2 \|d^k\|^2.$$

Hence, for $\alpha \in [0, 1]$,

$$\begin{array}{c} f(x^k + \alpha d^k) + cP(x^k + \alpha d^k) \leq f(x^k) + \nabla f(x^k)^T d^k + \overline{c}_2 \alpha^2 \|d^k\|^2 + \\ (1-\alpha) cP(x_k) + cc_1 \alpha^2 (d^k)^2 \leq f(x^k) + (1-\alpha) cP(x_k) + \\ \alpha(-(d^k)^T H^k d^k + (c - \frac{\gamma}{2}) P(x^k) - M(\mu^k, x^k)) + (cc_1 + \overline{c}_2) \alpha^2 \|d^k\|^2 = \\ f(x^k) + cP(x^k) - \alpha((d^k)^T H^k d^k + \frac{\gamma}{2} P(x^k) + M(\mu^k, x^k)) + (cc_1 + \overline{c}_2) \alpha^2 \|d^k\|^2 \end{array}$$

from (75) and Lemma 3.1. Therefore, for $\alpha \in [0, 1]$,

$$\phi(x^k + \alpha d^k) - \phi(x^k) \leq -\alpha((d^k)^T H^k d^k + \frac{\gamma}{2} P(x^k) + M(\mu^k, x^k)) + (cc_1 + \overline{c}_2) \alpha^2 \|d^k\|^2.$$

Since $(d^k)^T H^k d^k \geq \gamma_0 \|d^k\|^2$, the result of the statement follows by choosing $\overline{\alpha} = \min\{1, \frac{\gamma_0}{2(cc_1 + \overline{c}_2)}\}$ and $c_2 = \frac{1}{2} \min\{\gamma_0, \frac{\gamma}{2}, \frac{1}{2}\}$. □

Lemma 3.3 *There exists a constant c_5 such that, $\forall k \geq k_0$ and $\forall (\lambda^k, \mu^k) \in \mathcal{M}(H^k, x^k)$,*

$$\phi(x^k) - \phi(x^*) \leq c_5 (P(x^k) + \|x^k - x^*\|^2 + M(\mu^k, x^k) + \|d^k\|^2).$$

Proof Let p be the number of elements of $\mathcal{B}$, the active set. From (4) it follows, using Taylor's theorem, that, for a sufficiently small neighborhood of x,

$$\mathcal{L}(x, \lambda^*, \mu^*) - \mathcal{L}(x^*, \lambda^*, \mu^*) \leq \Sigma \|x - x^*\|^2 \quad \forall (\lambda^*, \mu^*) \in \mathcal{M}(x^*),$$

where $\Sigma = \max_{(\lambda^*, \mu^*) \in \mathcal{M}(x^*)} \{\|\nabla_{xx} \mathcal{L}(x, \lambda^*, \mu^*)\|\}$. Also, by Lemma 2.5, there exists a constant Σ_1 such that, $\forall (\lambda^k, \mu^k) \in \mathcal{M}(H^k, x^k)$, there exists a $(\lambda^*, \mu^*) \in \mathcal{M}(x^*)$ such that

$$|h(x^k)^T (\lambda^* - \lambda^k) + g(x^k)^T (\mu^* - \mu^k)| \leq \Sigma_1 (\|d^k\| + \|x^k - x^*\|) \|x^k - x^*\|.$$

Since $\mathcal{L}(x^*, \lambda^*, \mu^*) = f(x^*)$, we have that

$$f(x^k) - f(x^*) - \Sigma\|x^k - x^*\|^2 + (\lambda^k)^T h(x^k) + (\mu^k)^T g(x^k) \leq$$
$$(\lambda^k - \lambda^*)^T h(x^k) + (\mu^k - \mu^*)^T g(x) \leq \Sigma_1(\|x^k - x^*\|^2 + \|x^k - x^*\|\|d^k\|)$$

and, thus,

$$f(x^k) - f(x^*) \leq (\Sigma + \Sigma_1)\|x^k - x^*\|^2 + \Sigma_1\|x^k\|\|d^k\| - (\lambda^k)^T h(x^k) - (\mu^k)^T g(x^k). \quad (76)$$

Since $g(x) = g^+(x) - g^-(x)$, it follows that

$$-(\lambda^k)^T h(x^k) - (\mu^k)^T (g^+(x^k) - g^-(x^k)) \leq cP(x^k) + M(\mu^k, x^k).$$

We therefore have that

$$f(x^k) + cP(x^k) - f(x^*) \leq (\Sigma + \Sigma_1)\|x^k - x^*\|^2 + \Sigma_1\|x^k - x^*\|\|d^k\| + 2cP(x^k) + M(\mu^k, x^k).$$

The conclusion of the lemma follows by choosing $c_5 = \max\{\Sigma + 2\Sigma_1, 2c, 1\}$, since $2\|x^k - x^*\|\|d^k\| \leq \|x^k - x^*\|^2 + \|d^k\|^2$. □

3.3 Nondifferentiable Exact Penalty Algorithms and the Linear Convergence Theorem

The linearization algorithm [2, p.372] has the following form:

1. Choose $c^0 > 0$ and x^0.
2. Compute d^k from (12–14).
3. Choose α^k from a line search procedure, and set $x^{(k+1)} = x^k + \alpha^k d^k$.
4. Update c^k using (17), and restart with Step 2.

The stepsize α^k is chosen by one of the following procedures [2, pp.372].

(a) *Minimization rule* Here α^k is chosen such that

$$\phi(x^k + \alpha^k d^k) = \min_{\alpha \geq 0}\{\phi(x^k + \alpha d^k).\}$$

(b) *Limited minimization rule* Here a fixed scalar $s > 0$ is selected, and α^k is chosen such that

$$\phi(x^k + \alpha^k d^k) = \min_{\alpha \in [0, s]}\{\phi(x^k + \alpha d^k)\}.$$

(c) *Armijo rule* Here fixed scalars s, τ, and σ with $s > 0$, $\tau \in (0, 1)$, and $\sigma \in (0, \frac{1}{2})$ are chosen and we set $\alpha^k = \tau^{m_k} s$, where m_k is the first nonnegative integer m for which

$$\phi(x^k) - \phi(x^k + \tau^m s d^k) \geq \sigma\tau^m s (d^k)^T H^k d^k.$$

It can be shown that the Armijo rule yields a stepsize after a finite number of iterations.

The following theorem establishes the convergence properties of the linearization algorithm. The global convergence properties, established in [1, Prop. 4.3.3], are also stated here for completeness.

Theorem 3.4 *Let x^k be a sequence generated by the linearization algorithm, where the stepsize α^k is chosen by the minimization rule, limited minimization rule or the Armijo rule. Assume that $c_k = c, \forall k \geq k_0$ and that the sequence H^k satisfies (15). Then any accumulation point of the sequence x^k is a stationary point of $\phi(x) = f(x) + cP(x)$. If $x^k \to x^*$, where x^* is a strict local minimum of the problem (1–2) satisfying the second-order sufficient properties (7–10) and the Mangasarian-Fromowitz constraint qualification 6, then $\phi(x^k) \to \phi(x^*)$ Q-linearly and $x^k \to x^*$ R-linearly.*

Proof The first part is an immediate consequence of [1, Prop. 4.3.3]. We prove the linear convergence statement only for the Armijo rule, the proof being similar for the other stepsize selection mechanisms. By Lemma 3.2

$$\begin{gathered}\phi(x^k) - \phi(x^k + \alpha d^k) \geq \alpha \tfrac{1}{2}((d^k)^T H^k d^k + \tfrac{\gamma}{2} P(x^k) + M(\mu^k, x^k)) \geq \\ \alpha \tfrac{1}{2}(d^k)^T H^k d^k > \sigma\alpha(d^k)^T H^k d^k\end{gathered}$$

for all $\alpha \in [0, \overline{\alpha}]$. Since m_k is the smallest integer m for which

$$\phi(x^k) - \phi(x^k + \tau^m s d^k) \geq \sigma\tau^m s(d^k)^T H^k d^k,$$

it follows that $\tau^m s \geq \tau\overline{\alpha}$. This therefore ensures that the stepsize is at least $\tau\overline{\alpha}$ for k sufficiently large. As a result of Lemma 3.2, we have that

$$\phi(x^k) - \phi(x^{(k+1)}) \geq c_2\tau\overline{\alpha}(\|d^k\|^2 + P(x^k) + M(x^k)). \tag{77}$$

On the other hand, by Lemma 3.3 we have that

$$\phi(x^k) - \phi(x^*) \leq c_5(P(x^k) + \|x^k - x^*\|^2 + \|d^k\|^2 + M(\mu^k, x^k)).$$

By (72) and (77) it follows that there exists c_6 such that

$$\phi(x^k) - \phi(x^*) \leq c_6(M(\mu^k, x^k) + P(x^k) + \|d^k\|^2) \leq \tag{78}$$

$$\tfrac{c_6}{\tau\overline{\alpha}c_2}(\phi(x^k) - \phi(x^{k+1})) = \delta(\phi(x^k) - \phi(x^{k+1})) = \tag{79}$$

$$\delta(\phi(x^k) - \phi(x^*)) - \delta(\phi(x^{(k+1)}) - \phi(x^*)), \tag{80}$$

where $\delta = \frac{c_6}{\tau\overline{\alpha}c_2}$. After some obvious manipulation, it follows that

$$\delta(\phi(x^{(k+1)}) - \phi(x^*)) \leq (\delta - 1)(\phi(x^k) - \phi(x^*)),$$

which proves the Q-linear convergence [10] of the sequence $\phi(x^k)$ to $\phi(x^*)$ with a linear rate of at most $\frac{\delta-1}{\delta}$. Therefore

$$\limsup_{k\to\infty} \sqrt[k]{\phi(x^k) - \phi(x^*)} \leq \frac{\delta - 1}{\delta}.$$

From Corollary 2.4

$$\phi(x^k) - \phi(x^*) \geq \beta\|x^k - x^*\|^2.$$

Therefore

$$\limsup_{k\to\infty} \sqrt[k]{\|x^k - x^*\|} \leq (\frac{\delta - 1}{\delta})^{\frac{1}{2}},$$

which proves the R-linear convergence [10] to 0 of the sequence $x^k - x^*$. The proof is complete. □

3.4 The Superlinear Convergence Issue

The algorithm described in the preceding section does not achieve superlinear convergence in general. Even when there is no degeneracy, a second-order correction may be necessary to ensure that a unit stepsize, which is necessary for superlinear convergence, results in the decrease of the penalty function [1]. When degeneracy is present, the quadratic program (18–20) needs to be modified to ensure Q-superlinear convergence, even assuming that the step is acceptable for the penalty function [13]. Whether the SQP can be modified such as to achieve both global convergence and local superlinear convergence is a question for future research.

4 Conclusions

In this paper we have analyzed the impact of constraint degeneracy on the behavior of sequential quadratic programming with nondifferentiable penalty function. We proved that if the Mangasarian-Fromowitz constraint qualification as well as some second-order sufficient conditions hold, then at least linear convergence of SQP algorithms with exact penalty function is maintained. These conditions do not require the existence of a Lagrange multiplier that satisfies strict complementarity.

In our analysis we have shown that it is possible to define an extension of the augmented Lagrangian that can accommodate lack of strict complementarity, by using different augmentations for zero and nonzero multipliers. The resulting object has only one continuous derivative, which is a strictly monotone map.

A conclusion of this work is that SQPs with exact penalties are fundamentaly robust, since global as well as linear local convergence can be secured under very mild assumptions.

Acknowledgment

Thanks to Jorge Moré and Stephen Wright for reading the manuscript and providing extremely valuable suggestions.

References

[1] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[2] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, Massachusets, 1995.

[3] R. W. Cottle, J.-S. Pang and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, Boston, 1992.

[4] G. Di Pillo and L. Grippo, "Exact Penalty Functions in Constrained Optimization", *SIAM Journal of Control and Optimization* 27 (1989),1333–1360.

[5] A. V. Fiacco, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Academic Press, New York, 1983.

[6] J. Gauvin, "A Necessary and Sufficient Regularity Condition to Have Bounded Multipliers in Nonconvex Programming", *Mathematical Programming* 12 (1977), 136–138.

[7] W. W Hager and M. S. Gowda "Stability in the Presence of Degeneracy and Error Estimation", Technical Report, Department of Mathematics, University of Florida, 1997.

[8] S.-P. Han, and O. L. Mangasarian, "Exact Penalty Functions in Nonlinear Programming", *Mathematical Programming* 17 (1979) , 251–269.

[9] D. G. Luenberger, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, Massachusetts, 1973.

[10] J. Ortega, and W. Rheinboldt, *Iterative Solutions of Nonlinear Equations in Several Variables*, Academic Press, New York, 1972.

[11] D. Ralph and S. J. Wright, "Superlinear Convergence of an Interior-Point Method Despite Dependent Constraints", Preprint ANL/MCS-P622-1196, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Ill., 1996.

[12] S. M. Robinson, " Generalized equations and their solutions, Part II: Applications to Nonlinear Programming" *Mathematical Programming Study* 19 (1980), 200–221.

[13] S. J. Wright, "Superlinear Convergence of a Stabilized SQP Method to a Degenerate Solution", Preprint ANL/MCS-P463-0297, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Ill., 1997.