J Optim Theory Appl (2013) 156:417–449
DOI 10.1007/s10957-012-0114-6

A Gauss–Newton Approach for Solving Constrained Optimization Problems Using Differentiable Exact Penalties

Roberto Andreani · Ellen H. Fukuda · Paulo J.S. Silva

Received: 1 June 2012 / Accepted: 16 June 2012 / Published online: 29 June 2012

Abstract We propose a Gauss–Newton-type method for nonlinear constrained optimization using the exact penalty introduced recently by André and Silva for variational inequalities. We extend their penalty function to both equality and inequality constraints using a weak regularity assumption, and as a result, we obtain a continuously differentiable exact penalty function and a new reformulation of the KKT conditions as a system of equations. Such reformulation allows the use of a semismooth Newton method, so that local superlinear convergence rate can be proved under an assumption weaker than the usual strong second-order sufficient condition and without requiring strict complementarity. Besides, we note that the exact penalty function can be used to globalize the method. We conclude with some numerical experiments using the collection of test problems CUTE.

Keywords Exact penalty · Multipliers estimate · Nonlinear programming · Semismooth Newton method

Communicated by Gianni Di Pillo.

R. Andreani · E.H. Fukuda
Department of Applied Mathematics, IMECC, State University of Campinas, Campinas, Brazil

R. Andreani
e-mail: andreani@ime.unicamp.br

E.H. Fukuda
e-mail: ellen@ime.usp.br

P.J.S. Silva (✉)
Department of Computer Science, IME, University of São Paulo, São Paulo, Brazil
e-mail: pjssilva@ime.usp.br

1 Introduction

A popular framework to solve constrained nonlinear programming problems is to use penalty-based methods, such as quadratic penalty functions, augmented Lagrangians and exact penalty functions. The last one consists in replacing the original constrained problem with a single unconstrained one. Recently, André and Silva [1] proposed an exact penalty function for variational inequalities. Their idea is based on Di Pillo and Grippo's work [2, 3], which consists in incorporating a multipliers estimate in an augmented Lagrangian function. In this work, we propose a modified multipliers estimate and extend the exact penalty function for variational inequalities to optimization problems with general equality and inequality constraints. We use a generalized Newton method to solve the associated reformulation of the KKT conditions and define a suitable merit function to globalize the method.

The paper is organized as follows. In Sect. 2, we give some notations, definitions and the background concerning exact penalty functions. In Sect. 3, we construct the exact penalty function and present some results associated to the modified multipliers estimate. Exactness results are presented in Sect. 4, and in Sect. 5 we show a way to dynamically update the penalty parameter. Local convergence results for the semismooth Newton method are presented in Sect. 6. We finish in Sect. 7, with a globalization idea and some numerical experiments.

2 Preliminaries

Consider the following nonlinear programming problem:

$$\min \quad f(x) \quad \text{s.t.} \quad x \in X, \tag{NLP}$$

where the feasible set is assumed to be nonempty and is defined by

$$X := \big\{x \in \mathbb{R}^n : h(x) = 0, g(x) \leq 0\big\},$$

and $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$ and $h : \mathbb{R}^n \to \mathbb{R}^p$ are $\mathcal{C}^2$ functions. Roughly speaking, a function $w_c : \mathbb{R}^n \to \mathbb{R}$, that depends on a positive parameter $c \in \mathbb{R}$, is an *exact penalty* function for the problem (NLP) if there is an appropriate choice of the penalty coefficient c such that a single minimization of w_c recovers a solution of (NLP).

A well-known exact penalty function is the one proposed by Zangwill [4]. It can be shown that the solutions of the constrained problem (NLP) are solutions of the following unconstrained one:

$$\min_x \big[f(x) + c \max\big\{0, g_1(x), \ldots, g_m(x), \big|h_1(x)\big|, \ldots, \big|h_p(x)\big|\big\}\big],$$

under reasonable conditions and when c is sufficiently large [5, Sect. 4.3.1]. However, the maximum function contained in such penalty function makes it nondifferentiable, which demands special methods to solve this unconstrained problem. Besides, it is not easy to find the value of the parameter c that ensures the recovering of solutions of (NLP).

To overcome such difficulty, many authors proposed continuously differentiable exact penalty functions. The first ones were introduced by Fletcher [6] and by Glad and Polak [7], respectively, for problems with equality constraints and for those with both equality and inequality constraints. Another important contribution was done by Mukai and Polak [8]. In problems associated to such penalties, the variables are in the same space as the variables of the original problem. An alternative approach consists in defining an unconstrained minimization problem on the product space of variables and multipliers. This last case, called *exact augmented Lagrangian* methods, was introduced later by Di Pillo and Grippo [9, 10].

Considering only problems with equality constraints, these authors formulated the unconstrained problem as follows:

$$\min_{x,\mu}\left[f(x)+\langle\mu,h(x)\rangle+\frac{c}{2}\|h(x)\|^2+\frac{1}{2}\|\mathcal{M}(x)(\nabla f(x)+Jh(x)^T\mu)\|^2\right],$$

where $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ denote the Euclidean inner product and norm, respectively, $Jh(x)$ is the Jacobian of h at x and $\mathcal{M}(x)\in\mathbb{R}^{\ell\times n}$ is a $\mathcal{C}^2$ matrix with $p\le\ell\le n$ and such that $\mathcal{M}(x)Jh(x)^T$ has full rank. Suitable choices of $\mathcal{M}(x)$ make the above objective function quadratic in the dual variable μ. In such a case, we can write the multiplier in terms of x and incorporate it again in the function, obtaining an unconstrained problem in the space of the original variables. Some of the choices of $\mathcal{M}(x)$ recover the exact penalties of Fletcher and Mukai and Polak.

For problems with inequality constraints, we can add slack variables and write it in terms of x using an appropriate choice of matrix $\mathcal{M}(x)$. Nevertheless, it is not known a formula for $\mathcal{M}(x)$ that also isolates the multiplier. After this, Di Pillo and Grippo proposed a new continuously differentiable exact penalty, this time taking Glad and Polak's multipliers estimate [7] as a base. The idea was presented in [11] and [3] and it consists in constructing an exact penalty function by incorporating such estimate in the augmented Lagrangian of Hestenes, Powell, and Rockafellar [12–14]. To overcome some theoretical limitations of such penalty, Lucidi proposed in [15] another exact penalty function for problems with inequality constraints.

The same idea was extended recently by André and Silva [1] to solve variational inequalities with the feasible set defined by functional inequality constraints. In their work, they incorporated Glad and Polak's multipliers estimate in the augmented Lagrangian for variational inequalities, proposed by Auslender and Teboulle [16]. If the variational problem comes from the first-order necessary condition of an optimization problem, then this penalty is equivalent to the gradient of Di Pillo and Grippo's penalty excluding second-order terms. This is important from the numerical point of view, because otherwise it would be necessary to deal with third-order terms when second-order methods, like the Newton method, are applied to solve the problem. In the optimization case, Newton-type methods based on exact merit functions (exact penalties and exact augmented Lagrangians), without the knowledge of third-order terms, have been also proposed [17–20]. In particular, an exact penalty function was used in [17] for problems with equality constraints.

In this work, we extend André and Silva's exact penalty to solve optimization problems like (NLP), that is, with both equality and inequality constraints. We also

adapt Glad and Polak's multipliers estimate in order to use a weaker regularity assumption and compare it with the one proposed by Lucidi in [15]. The obtained function is semismooth or, under some conditions, strongly semismooth, which allows the use of the semismooth Newton method to solve the associated system of equations [21]. Exactness results are established and we adapt some results of Facchinei, Kanzow, and Palagi [22] to prove that the convergence rate is superlinear (or, in some cases, quadratic) without requiring strict complementarity or strong second-order sufficient condition. Moreover, we indicate a way to globalize the method using a specific merit function that makes it works as a Gauss–Newton-type method.

3 Constructing the Exact Penalty

The construction of the exact penalty function is based on Di Pillo and Grippo's [3, 11] and André and Silva's [1] papers. In both cases, the authors incorporate a Lagrange multipliers estimate in an augmented Lagrangian function. In particular, they both use the estimate proposed by Glad and Polak [7], which requires that the gradients of active inequality constraints $\nabla g_i(x)$, $i \in \{i : g_i(x) = 0\}$ and *all* the gradients of equality constraints $\nabla h_i(x)$, $i = 1, \ldots, p$ are linearly independent for all $x \in \mathbb{R}^n$. This condition is called *linear independence constraint qualification* (LICQ) and, in the optimization literature, is usually referred only at feasible points of the problem.

However, methods based on exact penalties may need to compute multipliers estimates in infeasible points, and hence LICQ has to be assumed to hold in the whole space $\mathbb{R}^n$. Thus, it is interesting to search for estimates that depend on a weaker assumption than LICQ in $\mathbb{R}^n$. With this in mind, we introduce the following condition.

Definition 3.1 A point $x \in \mathbb{R}^n$ satisfies the *relaxed linear independence constraint qualification (relaxed LICQ)* or, equivalently, it is called *regular*, if and only if the gradients

$$\nabla g_i(x), \quad i \in I_=(x), \qquad \nabla h_i(x), \quad i \in E_=(x)$$

are linearly independent, where

$$I_=(x) := \big\{i \in \{1, \ldots, m\} : g_i(x) = 0\big\} \quad \text{and} \quad E_=(x) := \big\{i \in \{1, \ldots, p\} : h_i(x) = 0\big\}.$$

Also, define the set of regular points in $\mathbb{R}^n$ as[1]

$$\mathcal{R} := \big\{x \in \mathbb{R}^n : x \text{ satisfies relaxed LICQ}\big\}.$$

This condition is more reasonable than LICQ because it allows more infeasible points to satisfy it. To illustrate such advantage, consider $n = 2$, $m = 2$, $p = 1$, and functions defined by $g_1(x) := -x_1 + x_2$, $g_2(x) := x_1 + x_2 - 1$ and $h_1(x) := x_2$. Taking the infeasible point $\tilde{x} := (0.5, 0.5)^T$, we observe that $I_=(\tilde{x}) = \{1, 2\}$ and $E_=(\tilde{x}) = \emptyset$. Therefore, $\tilde{x}$ satisfies relaxed LICQ, but not LICQ.

[1] The notation $\mathcal{R}$ comes from the word *regularity*.

By replacing LICQ with relaxed LICQ, we can adapt Glad and Polak's multipliers estimate. In this case, we obtain the estimate by solving the following unconstrained minimization problem, with variables $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^p$ associated to inequality and equality constraints, respectively:

$$\min_{\lambda,\mu} \left\|\nabla_x L(x,\lambda,\mu)\right\|^2 + \zeta^2\left(\left\|G(x)\lambda\right\|^2 + \left\|H(x)\mu\right\|^2\right), \tag{1}$$

where $L(x,\lambda,\mu) := f(x) + \langle\lambda, g(x)\rangle + \langle\mu, h(x)\rangle$ is the Lagrangian function, $\zeta > 0$, $G(x) := \operatorname{diag}(g_1(x),\ldots,g_m(x))$ and $H(x) := \operatorname{diag}(h_1(x),\ldots,h_p(x))$ are diagonal matrices with diagonal entries $g_i(x)$, $i = 1,\ldots,m$ and $h_i(x)$, $i = 1,\ldots,p$, respectively.

Considering the KKT conditions of (NLP), we observe that the above minimization problem forces the zero condition $\nabla_x L(x,\lambda,\mu) = 0$ and the complementary slackness $\langle g(x),\lambda\rangle = 0$. This also happens in Glad and Polak's estimate. The difference is that the new one adds the term $\|H(x)\mu\|^2$. Thus, it also enforces the complementarity of equality constraints $\langle h(x),\mu\rangle = 0$, even if it is irrelevant in the KKT conditions. The interesting fact is that this new estimate allows the use of the weaker assumption, the relaxed LICQ, and it does not lose any properties of Glad and Polak's estimate. In fact, the associated problem (1) is equivalent to

$$\min_{\lambda,\mu} \left\| \begin{bmatrix} Jg(x)^T & Jh(x)^T \\ \zeta G(x) & 0 \\ 0 & \zeta H(x) \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} - \begin{bmatrix} -\nabla f(x) \\ 0 \\ 0 \end{bmatrix} \right\|^2, \tag{2}$$

that is, it is a linear least squares problem. The proposition below gives some other properties associated to the modified multipliers estimate.

Proposition 3.1 *Assume that $x \in \mathbb{R}^n$ satisfies the relaxed LICQ and define the matrix $N(x) \in \mathbb{R}^{(m+p)\times(m+p)}$ by*

$$N(x) := \begin{bmatrix} Jg(x)Jg(x)^T + \zeta^2 G(x)^2 & Jg(x)Jh(x)^T \\ Jh(x)Jg(x)^T & Jh(x)Jh(x)^T + \zeta^2 H(x)^2 \end{bmatrix}.$$

Then

(a) *The matrix $N(x)$ is positive definite.*
(b) *The solution of (1) (equivalently, (2)) is unique and it is given by*

$$\begin{bmatrix} \lambda(x) \\ \mu(x) \end{bmatrix} = -N^{-1}(x) \begin{bmatrix} Jg(x) \\ Jh(x) \end{bmatrix} \nabla f(x).$$

(c) *If $(x,\bar\lambda,\bar\mu) \in \mathbb{R}^{n+m+p}$ satisfies the KKT conditions, then $\bar\lambda = \lambda(x)$ and $\bar\mu = \mu(x)$.*
(d) *The Jacobian matrices of $\lambda(\cdot)$ and $\mu(\cdot)$ are given by*

$$\begin{bmatrix} J\lambda(x) \\ J\mu(x) \end{bmatrix} = -N^{-1}(x) \begin{bmatrix} R_1(x) \\ R_2(x) \end{bmatrix},$$

with

$$R_1(x) := Jg(x)\nabla^2_{xx}L\big(x,\lambda(x),\mu(x)\big) + 2\zeta^2\Lambda(x)G(x)Jg(x) + \sum_{i=1}^{m} e_i^m \nabla_x L\big(x,\lambda(x),\mu(x)\big)^T \nabla^2 g_i(x),$$

$$R_2(x) := Jh(x)\nabla^2_{xx}L\big(x,\lambda(x),\mu(x)\big) + 2\zeta^2 M(x)H(x)Jh(x) + \sum_{i=1}^{p} e_i^p \nabla_x L\big(x,\lambda(x),\mu(x)\big)^T \nabla^2 h_i(x),$$

where $\Lambda(x) := \operatorname{diag}(\lambda_1(x),\ldots,\lambda_m(x))$ *and* $M(x) := \operatorname{diag}(\mu_1(x),\ldots,\mu_p(x))$ *are diagonal matrices respectively with elements* $\lambda_i(x) := [\lambda(x)]_i$ *and* $\mu_i(x) := [\mu(x)]_i$, e_i^m, e_i^p *are the* i*-th elements of the canonical base of* $\mathbb{R}^m$ *and* $\mathbb{R}^p$, *respectively, and*

$$\nabla_x L\big(x,\lambda(x),\mu(x)\big) := \nabla_x L(x,\lambda,\mu)|_{\lambda=\lambda(x),\mu=\mu(x)},$$
$$\nabla^2_{xx} L\big(x,\lambda(x),\mu(x)\big) := \nabla^2_{xx} L(x,\lambda,\mu)|_{\lambda=\lambda(x),\mu=\mu(x)}.$$

Proof (a) Let $A(x) \in \mathbb{R}^{(n+m+p)\times(m+p)}$ be the matrix associated to the linear least squares problem (2), that is,

$$A(x) := \begin{bmatrix} Jg(x)^T & Jh(x)^T \\ \zeta G(x) & 0 \\ 0 & \zeta H(x) \end{bmatrix}. \tag{3}$$

Without loss of generality, we can write $Jg(x)^T = [Jg(x)_=^T \mid Jg(x)_\neq^T]$, where $Jg(x)_=$ and $Jg(x)_\neq$ correspond to the parts of $Jg(x)$ where $g_i(x) = 0$ and $g_i(x) \neq 0$, respectively. In the same way, we can define the matrices $Jh(x)_=$, $Jh(x)_\neq$, $G(x)_\neq$ and $H(x)_\neq$. Thus,

$$A(x) = \begin{bmatrix} Jg(x)_=^T & Jg(x)_\neq^T & Jh(x)_=^T & Jh(x)_\neq^T \\ 0 & 0 & 0 & 0 \\ 0 & \zeta G(x)_\neq & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \zeta H(x)_\neq \end{bmatrix},$$

and we can see that it has linearly independent columns, by relaxed LICQ (so the first and third block columns of $A(x)$ are linearly independent) and because of the nonzero block diagonal matrices $G(x)_\neq$ and $H(x)_\neq$. Furthermore, it is easy to see that $N(x) = A(x)^T A(x)$, so we can conclude that $N(x)$ is nonsingular and positive definite.

(b) Differentiating the objective function of problem (2) and setting the result to zero, we have

$$A(x)^T A(x) \begin{bmatrix} \lambda(x) \\ \mu(x) \end{bmatrix} = A(x)^T \begin{bmatrix} -\nabla f(x) \\ 0 \\ 0 \end{bmatrix},$$

where $A(x)$ is defined in (3). The result follows since $N(x) = A(x)^T A(x)$ is nonsingular from (a).

(c) From the KKT conditions, $\nabla_x L(x, \bar{\lambda}, \bar{\mu}) = 0$, $G(x)\bar{\lambda} = 0$ and $H(x)\bar{\mu} = 0$, so the objective function's value of (1) at $(\bar{\lambda}, \bar{\mu})$ is zero. The result follows since the solution of (1) is unique from (b), and because the objective function's value is always nonnegative.

(d) From (b), we have:

$$\begin{aligned} -Jg(x)\nabla f(x) &= \big(Jg(x)Jg(x)^T + \zeta^2 G(x)^2\big)\lambda(x) + Jg(x)Jh(x)^T \mu(x), \\ -Jh(x)\nabla f(x) &= Jh(x)Jg(x)^T \lambda(x) + \big(Jh(x)Jh(x)^T + \zeta^2 H(x)^2\big)\mu(x), \end{aligned}$$

which is equivalent to

$$Jg(x)\nabla_x L\big(x, \lambda(x), \mu(x)\big) + \zeta^2 G(x)^2 \lambda(x) = 0, \tag{4}$$

$$Jh(x)\nabla_x L\big(x, \lambda(x), \mu(x)\big) + \zeta^2 H(x)^2 \mu(x) = 0. \tag{5}$$

Note that Eq. (4) gives

$$\sum_{i=1}^{m} e_i^m \nabla g_i(x)^T \nabla_x L\big(x, \lambda(x), \mu(x)\big) + \zeta^2 G(x)^2 \lambda(x) = 0.$$

Thus, deriving it with respect to x, we obtain

$$\begin{aligned} 0 &= \sum_{i=1}^{m} e_i^m \nabla_x L\big(x, \lambda(x), \mu(x)\big)^T \nabla^2 g_i(x) + 2\zeta^2 \Lambda(x) G(x) Jg(x) + \zeta^2 G(x)^2 J\lambda(x) \\ &\quad + Jg(x)\big(\nabla^2_{xx} L\big(x, \lambda(x), \mu(x)\big) + Jg(x)^T J\lambda(x) + Jh(x)^T J\mu(x)\big) \\ &= R_1(x) + Jg(x)Jg(x)^T J\lambda(x) + Jg(x)Jh(x)^T J\mu(x) + \zeta^2 G(x)^2 J\lambda(x). \end{aligned}$$

Analogously, Eq. (5) yields

$$0 = R_2(x) + Jh(x)Jh(x)^T J\mu(x) + Jh(x)Jg(x)^T J\lambda(x) + \zeta^2 H(x)^2 J\mu(x).$$

These two equations give the desired result. □

Indeed, the idea of weakening the assumption LICQ in $\mathbb{R}^n$ was already investigated by Lucidi in [15]. His idea consists in adding another term in the objective function of Glad and Polak's estimate. Since Lucidi considered only problems with inequality constraints, we adapt his idea in order to solve (NLP). Then, for any

$x \in \mathbb{R}^n$, we can obtain an estimate by solving the following linear least squares problem:

$$\min_{\lambda,\mu} \left\| \nabla_x L(x,\lambda,\mu) \right\|^2 + \zeta_1^2 \left\| G(x)\lambda \right\|^2 + \zeta_2^2 \alpha(x) \left(\|\lambda\|^2 + \|\mu\|^2 \right), \tag{6}$$

where $\zeta_1, \zeta_2 > 0$ and $\alpha(x) := \sum_{i=1}^m \max\{g_i(x), 0\}^{q_1} + \sum_{i=1}^p [h_i(x)]^{q_2}$, with $q_1, q_2 \geq 2$. The assumption required by this estimate is weaker than Glad and Polak's and the new estimate (1). In fact, it only asks LICQ in the set of *feasible* points.

Also, Proposition 3.1 can be rewritten if we replace the new estimate (1) by this adaptation of Lucidi's. In particular, the matrix $A(x)$ in (3) is replaced by

$$\bar{A}(x) := \begin{bmatrix} Jg(x)^T & Jh(x)^T \\ \zeta_1 G(x) & 0 \\ \zeta_2 \alpha(x)^{1/2} I & 0 \\ 0 & \zeta_2 \alpha(x)^{1/2} I \end{bmatrix}.$$

Note that $\bar{A}(x)$ has linearly independent columns if x is infeasible since in this case $\alpha(x) \neq 0$. This also holds if x is feasible because of the LICQ assumption. Results analogous to Proposition 3.1 can be also proved. Even if the assumption required by (6) is the weakest one proposed, we observe that the term $\alpha(x)(\|\lambda\|^2 + \|\mu\|^2)$ can introduce a dependence among the multipliers that is absent in the new estimate. This fact will be clearly shown in Sect. 7 with some numerical experiments. Because of this, for now on we will focus only at the new estimate (1), although all the next results can be proved if we replace it by (6).

Now, let us show precisely the idea given by Di Pillo and Grippo [3, 11] for building an exact penalty function. Essentially, they considered the incorporation of Glad and Polak's multipliers estimate [7] in the classical augmented Lagrangian function, given by Hestenes, Powell, and Rockafellar [12–14], that is,

$$\begin{aligned} L_c(x,\lambda,\mu) := f(x) + \langle \lambda, g(x) \rangle + \frac{c}{2} \left\| g(x) \right\|^2 - \frac{1}{2c} \sum_{i=1}^m \max\left\{0, -\lambda_i - c g_i(x)\right\}^2 \\ + \langle \mu, h(x) \rangle + \frac{c}{2} \left\| h(x) \right\|^2. \end{aligned}$$

In the same way, we can use the new multipliers estimate $(\lambda(\cdot), \mu(\cdot))$ defined in Proposition 3.1b. Thus, a possible exact penalty function is given by

$$w_c(x) := L_c\big(x, \lambda(x), \mu(x)\big), \tag{7}$$

which has the gradient

$$\begin{aligned} \nabla w_c(x) = \nabla f(x) + Jg(x)^T \lambda(x) + \left(cJg(x)^T + J\lambda(x)^T \right) \left(g(x) + y_c(x) \right) \\ + Jh(x)^T \mu(x) + \left(cJh(x)^T + J\mu(x)^T \right) h(x), \end{aligned} \tag{8}$$

where

$$y_c(x) := \max\left\{0, -\frac{\lambda(x)}{c} - g(x)\right\}.$$

Recently, André and Silva [1] proposed an exact penalty for variational inequalities. Here, we are interested only in variational problems that come from the first-order necessary optimality conditions of (NLP). More precisely, we are interested in finding a feasible point $\bar{x} \in X$ such that $\langle \nabla f(\bar{x}), x - \bar{x} \rangle \geq 0$ for all $x \in X$. To solve this kind of problem, these authors incorporated the multipliers estimate in the augmented Lagrangian for variational inequalities proposed by Auslender and Teboulle [16]:

$$\mathcal{L}_c(x, \lambda, \mu) := \nabla f(x) + Jg(x)^T \lambda + cJg(x)^T \max\left\{ g(x), -\frac{\lambda}{c} \right\} + Jh(x)^T \mu + cJh(x)^T h(x).$$

Note that $\mathcal{L}_c(x, \lambda, \mu)$ is equal to $\nabla_x L_c(x, \lambda, \mu)$, the gradient of the classical augmented Lagrangian for optimization with respect to the first variable. Once again, using the new estimate $(\lambda(\cdot), \mu(\cdot))$, we define

$$\begin{aligned} W_c(x) &:= \mathcal{L}_c\big(x, \lambda(x), \mu(x)\big) \\ &= \nabla_x L\big(x, \lambda(x), \mu(x)\big) + cJg(x)^T \big(g(x) + y_c(x)\big) + cJh(x)^T h(x). \end{aligned} \tag{9}$$

Observe that $\nabla w_c(x)$ is equal to $W_c(x)$ plus some terms that depend on second-order information, that is, $W_c(x)$ does not contain $J\lambda(x)$ and $J\mu(x)$, which have $\nabla^2 f(x)$, $\nabla^2 g_i(x)$ and $\nabla^2 h_i(x)$ in their formulas. This is important because, as we will see in the next section, KKT points of (NLP) are related to the system of equations $W_c(x) = 0$. Therefore, because of the absence of second-order terms, we can use Newton-type methods to search for such KKT points. Besides the interpretation associated to the augmented Lagrangian for variational inequalities, we point out that a similar approach, using exact penalty functions, was proposed by Bertsekas in [17]. However, his formulation considers only equality constraints, which allows us to ignore discussions about assumptions like the strict complementarity and the (strong) second-order sufficient conditions.

4 Exactness Results

Let us now present the exactness properties for W_c defined in (9). We point out that we follow the structure presented by Di Pillo and Grippo [3, 11] and André and Silva [1]. The difference, besides the multipliers estimate that defines W_c, is the enunciation of the results using the infeasibility measure defined below.

Definition 4.1 Let $\mathcal{F}\colon \mathbb{R}^n \to \mathbb{R}$ be the *infeasibility measure* defined by

$$\mathcal{F}(x) := \frac{1}{2}\big(\big\|\max\{0, g(x)\}\big\|^2 + \big\|h(x)\big\|^2\big),$$

with gradient

$$\nabla \mathcal{F}(x) = Jg(x)^T \max\{0, g(x)\} + Jh(x)^T h(x).$$

Note that $\mathcal{F}(x) = 0$ if and only if x is feasible. Moreover, we say that x is a stationary point of the infeasibility measure $\mathcal{F}$ if and only if $\nabla\mathcal{F}(x) = 0$. Clearly, a feasible point is also a stationary point of $\mathcal{F}$ but the converse is not always true.

We claim that a point satisfies the KKT conditions if and only if it is a solution of the system of equations $W_c(x) = 0$, under some conditions. One implication of this statement is given below.

Proposition 4.1 *Let (x, λ, μ) be a KKT triple associated to the problem* (NLP) *with $x \in \mathcal{R} \subset \mathbb{R}^n$. Then, $W_c(x) = 0$ for all $c > 0$.*

Proof Proposition 3.1(c) ensures that $\lambda = \lambda(x)$ and $\mu = \mu(x)$. Then the statement follows directly from the definition of W_c and the KKT conditions. □

The other implication of the statement can be true if c is large enough and if, for example, the zeros of W_c are bounded. It is not true only if, instead of a KKT point, we find a stationary point of $\mathcal{F}$ that is infeasible for (NLP). Before the main theorem, we consider two additional results and we introduce the simple notation

$$\mathbb{R}_{++} := \{c \in \mathbb{R} \colon c > 0\}$$

for positive numbers.

Proposition 4.2 *Let $\{x^k\} \subseteq \mathcal{R} \subset \mathbb{R}^n$ and $\{c_k\} \subset \mathbb{R}_{++}$ be sequences such that $c_k \to \infty$, $x^k \to \bar{x} \in \mathcal{R}$ and $W_{c_k}(x^k) = 0$ for all k. Then $\bar{x}$ is a stationary point of $\mathcal{F}$.*

Proof By definition of $W_{c_k}(x^k)$, we have

$$\begin{aligned}\nabla_x L\big(x^k, \lambda\big(x^k\big), \mu\big(x^k\big)\big) + c_k Jg\big(x^k\big)^T \max\big\{g\big(x^k\big), -\lambda\big(x^k\big)/c_k\big\} \\ + c_k Jh\big(x^k\big)^T h\big(x^k\big) = 0.\end{aligned}$$

Since $\lambda(\cdot)$ and $\mu(\cdot)$ are continuous by relaxed LICQ and recalling that f, g, and h are $\mathcal{C}^2$ functions, we can divide the above equality by c_k and take the limit to conclude that

$$Jg(\bar{x})^T \max\big\{g(\bar{x}), 0\big\} + Jh(\bar{x})^T h(\bar{x}) = 0,$$

that is, $\nabla\mathcal{F}(\bar{x}) = 0$. □

Proposition 4.3 *Let $\bar{x} \in \mathcal{R} \subset \mathbb{R}^n$ be a feasible point of the problem* (NLP). *Then there exist $\bar{c}$, $\bar{\delta} > 0$ (which depend on $\bar{x}$) such that if $\|x - \bar{x}\| \leq \bar{\delta}$ with $x \in \mathcal{R}$, $c \geq \bar{c}$ and $W_c(x) = 0$, then $(x, \lambda(x), \mu(x))$ is a KKT triple associated to* (NLP).

Proof First, it is easy to show that

$$Y_c(x)\lambda(x) = -cY_c(x)\big(g(x) + y_c(x)\big), \tag{10}$$

where $Y_c(x) := \operatorname{diag}((y_c)_1(x), \ldots, (y_c)_m(x))$ is a diagonal matrix with diagonal entries $(y_c)_i(x)$, $i = 1, \ldots, m$. Hence, from (4), we obtain

$$\begin{aligned} Jg(x)\nabla_x L\big(x, \lambda(x), \mu(x)\big) &= -\zeta^2 G(x)^2 \lambda(x) \\ &= -\zeta^2 G(x)\big(G(x) + Y_c(x)\big)\lambda(x) + \zeta^2 G(x) Y_c(x)\lambda(x) \\ &= -\zeta^2 G(x)\Lambda(x)\big(g(x) + y_c(x)\big) + \zeta^2 G(x) Y_c(x)\lambda(x). \end{aligned}$$

Combining the last result with (10), we have

$$\frac{1}{c} Jg(x)\nabla_x L\big(x, \lambda(x), \mu(x)\big) = -\zeta^2 G(x)\left(\frac{1}{c}\Lambda(x) + Y_c(x)\right)\big(g(x) + y_c(x)\big).$$

And the definition of W_c gives

$$\begin{aligned} \frac{1}{c} Jg(x) W_c(x) &= \frac{1}{c} Jg(x)\nabla_x L\big(x, \lambda(x), \mu(x)\big) + Jg(x) Jg(x)^T \big(g(x) + y_c(x)\big) \\ &\quad + Jg(x) Jh(x)^T h(x) \\ &= -\zeta^2 G(x)\left(\frac{1}{c}\Lambda(x) + Y_c(x)\right)\big(g(x) + y_c(x)\big) \\ &\quad + Jg(x) Jg(x)^T \big(g(x) + y_c(x)\big) + Jg(x) Jh(x)^T h(x). \end{aligned} \tag{11}$$

Moreover, Eq. (5) yields

$$Jh(x)\nabla_x L\big(x, \lambda(x), \mu(x)\big) = -\zeta^2 H(x)^2 \mu(x),$$

and thus

$$\begin{aligned} \frac{1}{c} Jh(x) W_c(x) &= -\frac{1}{c}\zeta^2 H(x)^2 \mu(x) + Jh(x) Jg(x)^T \big(g(x) + y_c(x)\big) \\ &\quad + Jh(x) Jh(x)^T h(x) \\ &= Jh(x) Jg(x)^T \big(g(x) + y_c(x)\big) \\ &\quad + \left(Jh(x) Jh(x)^T - \frac{1}{c}\zeta^2 H(x) M(x)\right) h(x). \end{aligned} \tag{12}$$

Combining the results from (11) and (12), we can write

$$\frac{1}{c}\begin{bmatrix} Jg(x) \\ Jh(x) \end{bmatrix} W_c(x) = K_c(x)\begin{bmatrix} g(x) + y_c(x) \\ h(x) \end{bmatrix}, \tag{13}$$

with

$$K_c(x) := \begin{bmatrix} (K_c(x))_1 & Jg(x) Jh(x)^T \\ Jh(x) Jg(x)^T & (K_c(x))_2 \end{bmatrix},$$

where

$$\big(K_c(x)\big)_1 := Jg(x) Jg(x)^T - \zeta^2 G(x)\big((1/c)\Lambda(x) + Y_c(x)\big),$$

$$\big(K_c(x)\big)_2 := Jh(x)Jh(x)^T - (1/c)\zeta^2 H(x)M(x).$$

Observing that $\bar{x}$ is feasible, if $c \to \infty$, then we have $y_c(\bar{x}) \to -g(\bar{x})$ and, therefore, $K_c(\bar{x}) \to N(\bar{x})$. Since relaxed LICQ implies that $N(\bar{x})$ is nonsingular, by continuity, there exist $\bar{c}$ and $\bar{\delta}$ such that if $\|x - \bar{x}\| \leq \bar{\delta}$, $c \geq \bar{c}$, then $K_c(x)$ is also nonsingular.

Now, consider any x and c such that $\|x - \bar{x}\| \leq \bar{\delta}$, $c \geq \bar{c}$ and $W_c(x) = 0$. Then Eq. (13) implies that $g(x) + y_c(x) = 0$ and $h(x) = 0$ because $K_c(x)$ is nonsingular. Plugging these equations into the definition of W_c gives $\nabla_x L(x, \lambda(x), \mu(x)) = 0$. Furthermore,

$$g(x) + y_c(x) = 0 \quad \Leftrightarrow \quad \max\big\{g(x), -\lambda(x)/c\big\} = 0 \quad \Rightarrow \quad 0 \geq g(x) \perp \lambda(x) \geq 0,$$

and we conclude that $(x, \lambda(x), \mu(x))$ is a KKT triple. □

Combining these results, we now obtain the following theorem.

Theorem 4.1 *Let $\{x^k\} \subseteq \mathcal{R} \subset \mathbb{R}^n$ and $\{c_k\} \subset \mathbb{R}_{++}$ be sequences such that $c_k \to \infty$ and $W_{c_k}(x^k) = 0$ for all k. Also, consider $\{x^{k_j}\}$ a subsequence of $\{x^k\}$ such that $x^{k_j} \to \bar{x} \subset \mathcal{R}$. Then, either there exists K such that $(x^{k_j}, \lambda(x^{k_j}), \mu(x^{k_j}))$ is a KKT triple associated to* (NLP) *for all $k_j > K$, or $\bar{x}$ is a stationary point of $\mathcal{F}$ that is infeasible for* (NLP).

Proof By Proposition 4.2, the point $\bar{x}$ is stationary of $\mathcal{F}$. If $\bar{x}$ is feasible, then we can conclude, using the Proposition 4.3, that there exists K such that $(x^{k_j}, \lambda(x^{k_j}), \mu(x^{k_j}))$ is KKT for all $k_j > K$. □

Observe that a subsequence $\{x^{k_j}\}$ of the above theorem exists if, for example, $\{x^k\}$ is bounded. The next result is an immediate consequence of this theorem. But in this case, we assume that all stationary points of the infeasibility measure $\mathcal{F}$ are feasible. This property holds if, for example, the functions g_i, $i = 1, \ldots, m$ are convex and h_i, $i = 1, \ldots, p$ are affine, which was assumed in André and Silva's work [1]. The property also holds under the extended Mangasarian–Fromovitz constraint qualification, used by Di Pillo and Grippo [3].

Corollary 4.1 *Assume that there exists $\bar{c} > 0$ such that the set*

$$Z := \big\{x \in \mathbb{R}^n : \ W_c(x) = 0, \ c > \bar{c}\big\}$$

is bounded with $Z \subset \mathcal{R}$. Assume that all stationary points of $\mathcal{F}$ are feasible for (NLP). *Then, there exists $\tilde{c} > 0$ such that if $W_c(x) = 0$ and $c > \tilde{c}$ then $(x, \lambda(x), \mu(x))$ is a KKT triple associated to* (NLP).

Proof Suppose that there is no such $\tilde{c}$. So, there exist sequences $\{x^k\} \subset \mathbb{R}^n$ and $\{c_k\} \subset \mathbb{R}_{++}$ with $W_{c_k}(x^k) = 0$ and $c_k \to \infty$ and such that $(x^k, \lambda(x^k), \mu(x^k))$ is not KKT. But for $c_k > \bar{c}$, we have $x^k \in Z$, which is bounded. So, there exists a convergent subsequence $\{x^{k_j}\}$ of $\{x^k\}$. This is not possible from Theorem 4.1 and because there is no stationary point of $\mathcal{F}$ that is infeasible. □

A drawback of the above result is the boundedness assumption, which is not easily verifiable. For nonlinear complementarity problems, that is, a particular case of variational inequalities, the boundedness was ensured by exploiting coercivity or monotonicity properties of the problem data [1]. For nonlinear programming, one approach is to use an extraneous compact set containing the problem solutions as in [2, 3, 11]. In this case, the sequence generated by an unconstrained algorithm could cross the boundary of this set, where the exactness property does not hold, or could not admit a limit point. To overcome such difficulty, the incorporation of barrier terms in the exact penalty function has been investigated [15, 18]. As a future work, we could extend the results given here to construct these functions, called *exact barrier functions* [18].

Let us show now that KKT points are not only equivalent, under some assumptions, to the system of equations $W_c(x) = 0$, but also to the system $\nabla w_c(x) = 0$, where w_c is defined in (7).

Corollary 4.2 *Let $\bar{x} \in \mathcal{R} \subset \mathbb{R}^n$ be a feasible point of* (NLP). *Then, there exist $\bar{c}$, $\bar{\delta} > 0$ (which depend on $\bar{x}$) such that if $\|x - \bar{x}\| \leq \bar{\delta}$ with $x \in \mathcal{R}$, $c \geq \bar{c}$ and $\nabla w_c(x) = 0$, then $(x, \lambda(x), \mu(x))$ is a KKT triple associated to* (NLP).

Proof The proof is analogous to the proof of Proposition 4.3. In this case, we have

$$K_c(x) := \begin{bmatrix} (K_c(x))_{11} & (K_c(x))_{12} \\ (K_c(x))_{21} & (K_c(x))_{22} \end{bmatrix},$$

where

$$\big(K_c(x)\big)_{11} := Jg(x)Jg(x)^T - \zeta^2 G(x)\left(\frac{1}{c}\Lambda(x) + Y_c(x)\right) + \frac{1}{c}Jg(x)J\lambda(x)^T,$$

$$\big(K_c(x)\big)_{12} := Jg(x)Jh(x)^T + \frac{1}{c}Jg(x)J\mu(x)^T,$$

$$\big(K_c(x)\big)_{21} := Jh(x)Jg(x)^T + \frac{1}{c}Jh(x)J\lambda(x)^T,$$

$$\big(K_c(x)\big)_{22} := Jh(x)Jh(x)^T - \frac{1}{c}\zeta^2 H(x)M(x) + \frac{1}{c}Jh(x)J\mu(x)^T.$$

Taking $c \to \infty$, we can also conclude that $K_c(\bar{x}) \to N(\bar{x})$ and the result follows. □

Corollary 4.3 *Let $\{x^k\} \subseteq \mathcal{R} \subset \mathbb{R}^n$ and $\{c_k\} \subset \mathbb{R}_{++}$ be sequences such that $c_k \to \infty$ and $\nabla w_{c_k}(x^k) = 0$ for all k. Consider $\{x^{k_j}\}$ a subsequence of $\{x^k\}$ with $x^{k_j} \to \bar{x} \subset \mathcal{R}$. Then, either there exists K such that $(x^{k_j}, \lambda(x^{k_j}), \mu(x^{k_j}))$ is a KKT triple associated to* (NLP) *for all $k_j > K$, or $\bar{x}$ is a stationary point of $\mathcal{F}$ that is infeasible for* (NLP).

Proof It is easy to show that Proposition 4.2 still holds if we replace W_c with ∇w_c. Then, the results follows directly from the proof of Theorem 4.1, replacing the Proposition 4.3 with Corollary 4.2. □

We proceed now to the equivalence of minimizers of (NLP) and the unconstrained penalized problem. Denote by $\mathcal{G}_f$ and $\mathcal{L}_f$ the sets of global and local solutions

of (NLP), respectively. Moreover, for each $c > 0$, denote the sets of global and local solutions of the penalized problem $\min_x w_c(x)$, respectively, by $\mathcal{G}_w(c)$ and $\mathcal{L}_w(c)$. First, let us prove that w_c is a *weakly exact penalty* function for the problem (NLP) in the sense of [3, 18], where we remove the extraneous compact set.

Definition 4.2 The function w_c is a *weakly exact penalty* function for (NLP) if and only if there exists some $\bar{c} > 0$ such that $\mathcal{G}_w(c) = \mathcal{G}_f$ for all $c \geq \bar{c}$.

The following assumption is required to the proof.

Assumption 4.1 It holds that $\emptyset \neq \mathcal{G}_f \subset \mathcal{R}$ and $\mathcal{L}_w(c) \subset \mathcal{R}$ for c large enough. Note that the last one implies $\mathcal{G}_w(c) \subset \mathcal{R}$ for c sufficiently large.

Before presenting the main theorem concerning global minimizers, let us consider some additional results.

Lemma 4.1 *The function w_c defined in (7) at $x \in \mathbb{R}^n$ can be written as*

$$w_c(x) = f(x) + \langle \lambda(x), g(x) + y_c(x) \rangle + \frac{c}{2}\|g(x) + y_c(x)\|^2 + \langle \mu(x), h(x) \rangle + \frac{c}{2}\|h(x)\|^2.$$

Proof Observe that

$$\begin{aligned}
&\langle \lambda(x), g(x) + y_c(x) \rangle + \frac{c}{2}\|g(x) + y_c(x)\|^2 \\
&\quad = \langle \lambda(x), g(x) \rangle + \frac{c}{2}\|g(x)\|^2 + \langle \lambda(x), y_c(x) \rangle + \frac{c}{2}\|y_c(x)\|^2 + c\langle g(x), y_c(x) \rangle \\
&\quad = \langle \lambda(x), g(x) \rangle + \frac{c}{2}\|g(x)\|^2 + \left\langle y_c(x), \frac{c}{2} y_c(x) - c\left(-\frac{\lambda(x)}{c} - g(x)\right) \right\rangle. \qquad (14)
\end{aligned}$$

Let us consider two cases:

1. For i such that $[y_c(x)]_i = -\lambda_i(x)/c - g_i(x)$, we have

$$[y_c(x)]_i \left(\frac{c}{2}[y_c(x)]_i - c\left(-\frac{\lambda_i(x)}{c} - g_i(x)\right)\right) = -\frac{c}{2}[y_c(x)]_i^2.$$

2. Otherwise, for i such that $[y_c(x)]_i = 0$, we have

$$[y_c(x)]_i \left(\frac{c}{2}[y_c(x)]_i - c\left(-\frac{\lambda_i(x)}{c} - g_i(x)\right)\right) = 0 = -\frac{c}{2}[y_c(x)]_i^2.$$

Thus, (14) is equivalent to

$$\langle \lambda(x), g(x) \rangle + \frac{c}{2}\|g(x)\|^2 - \frac{c}{2}\|y_c(x)\|^2,$$

and the conclusion follows. □

Lemma 4.2 *Let (x, λ, μ) be a KKT triple associated to the problem (NLP) such that $x \in \mathcal{R}$. Then, $w_c(x) = f(x)$ for all $c > 0$.*

Proof Since x satisfies relaxed LICQ, $\lambda(x) = \lambda$. From the KKT conditions, $h(x) = 0$ and $0 \leq \lambda(x) \perp g(x) \leq 0$, which is equivalent to $g(x) + y_c(x) = 0$. Then the proof follows from the formula of w_c given in Lemma 4.1. □

Proposition 4.4 *Let $\{x^k\} \subseteq \mathcal{R} \subset \mathbb{R}^n$ and $\{c_k\} \subset \mathbb{R}_{++}$ be sequences such that $\{x^k\}$ is bounded, $c_k \to \infty$ and $x^k \in \mathcal{G}_w(c_k)$ for all k. If Assumption 4.1 holds, then there exists K such that $x^k \in \mathcal{G}_f$ for all $k > K$.*

Proof Assume that the assertion is false, that is, for all K, there exists $k > K$ such that $x^k \notin \mathcal{G}_f$. First, let $\hat{x} \in \mathcal{G}_f$, which exists by Assumption 4.1. Since $\hat{x}$ is a KKT point and satisfies relaxed LICQ (also from Assumption 4.1), we have, from Lemma 4.2,

$$w_{c_k}\left(x^k\right) \leq w_{c_k}(\hat{x}) = f(\hat{x}) \tag{15}$$

for all k. The boundedness assumption of $\{x^k\}$ guarantees that there exists a subsequence of $\{x^k\}$ converging to $\bar{x} \in \mathbb{R}^n$. Without loss of generality, we can write $\lim_{k\to\infty} x^k = \bar{x}$. So, taking the supremum limit in both sides of (15) gives

$$\limsup_{k\to\infty} w_{c_k}\left(x^k\right) \leq f(\hat{x}). \tag{16}$$

Now, from Lemma 4.1, w_{c_k} can be written as

$$w_{c_k}\left(x^k\right) = f\left(x^k\right) + \left\langle \lambda\left(x^k\right), g\left(x^k\right) + y_{c_k}\left(x^k\right)\right\rangle + \frac{c_k}{2}\left\| g\left(x^k\right) + y_{c_k}\left(x^k\right)\right\|^2 + \left\langle \mu\left(x^k\right), h\left(x^k\right)\right\rangle + \frac{c_k}{2}\left\| h\left(x^k\right)\right\|^2.$$

Thus, inequality (16) implies, by continuity of the involved functions, that $h(\bar{x}) = 0$ and $g(\bar{x}) + \max\{0, -g(\bar{x})\} = 0$, which implies $g(\bar{x}) \leq 0$. Moreover, it is easy to see that $f(\bar{x}) \leq \limsup_{k\to\infty} w_{c_k}(x^k)$. Therefore, $f(\bar{x}) \leq f(\hat{x})$ and $\bar{x} \in X$, that is, $\bar{x} \in \mathcal{G}_f$.

Since $\bar{x}$ is feasible and satisfies relaxed LICQ by Assumption 4.1, there exist $\bar{c}$ and $\bar{\delta}$ as in the Corollary 4.2. Let $\bar{K}$ be sufficiently large such that $\|x^k - \bar{x}\| \leq \bar{\delta}$, $c_k \geq \bar{c}$ and $x^k \in \mathcal{G}_w(c_k) \subset \mathcal{R}$ for all $k > \bar{K}$. Since $x^k \in \mathcal{G}_w(c_k)$ implies $\nabla w_{c_k}(x^k) = 0$, the same corollary ensures that x^k is KKT, and thus feasible, for all $k > \bar{K}$. Moreover, Lemma 4.2 and inequality (15) yield

$$f\left(x^k\right) = w_{c_k}\left(x^k\right) \leq f(\hat{x})$$

for all $k > \bar{K}$. We conclude that for such $\bar{K}$, $x^k \in \mathcal{G}_f$ for all $k > \bar{K}$, which gives a contradiction. □

Proposition 4.5 *Suppose that Assumption 4.1 holds. Then $\mathcal{G}_w(c) \subseteq \mathcal{G}_f$ implies that $\mathcal{G}_w(c) = \mathcal{G}_f$ for all $c > 0$.*

Proof Let $c > 0$ and $\tilde{x} \in \mathcal{G}_w(c)$. As $\mathcal{G}_w(c) \subseteq \mathcal{G}_f$, $\tilde{x}$ is a KKT point, and thus, by Lemma 4.2, $w_c(\tilde{x}) = f(\tilde{x})$. On the other hand, let $\hat{x} \in \mathcal{G}_f$ such that $\hat{x} \neq \tilde{x}$. Since $\hat{x}$ is also a KKT point and satisfies relaxed LICQ by Assumption 4.1, once again by Lemma 4.2, we have $w_c(\hat{x}) = f(\hat{x})$. Therefore, by the definition of global solutions,

$$w_c(\tilde{x}) = f(\tilde{x}) = f(\hat{x}) = w_c(\hat{x}).$$

We conclude that $\hat{x} \in \mathcal{G}_w(c)$ and the result follows. □

The above two propositions can be put together and the result is as follows.

Theorem 4.2 *If there exists $\bar{c} > 0$ such that $\bar{Z} := \bigcup_{c \geq \bar{c}} \mathcal{G}_w(c)$ is bounded and Assumption 4.1 holds, then w_c is a weakly exact penalty function for the problem.*

Proof From Proposition 4.5, it is sufficient to show that there exist $\tilde{c} > 0$ such that $\mathcal{G}_w(c) \subseteq \mathcal{G}_f$ for all $c \geq \tilde{c}$. Suppose that there are sequences $\{x^k\} \subset \bar{Z}$ and $\{c_k\} \subset \mathbb{R}_{++}$ with $c_k \geq \bar{c}$, $c_k \to \infty$ and $x^k \in \mathcal{G}_w(c_k)$ for all k. Since $\bar{Z}$ is bounded, Proposition 4.4 guarantees that there exists K such that $x^k \in \mathcal{G}_f$ for all $k > K$. The result follows taking $\tilde{c} = c_K$. □

The drawback of the definition of weakly exact penalty functions is that unconstrained minimization algorithms do not ensure to find global solutions in general. With this in mind, we prove that w_c is essentially an *exact penalty* function. Once again, following [3, 18] and removing the extraneous compact set, we have the following definition.

Definition 4.3 The function w_c is an *exact penalty* function for (NLP) if and only if there exists $\bar{c} > 0$ such that $\mathcal{G}_w(c) = \mathcal{G}_f$ and $\mathcal{L}_w(c) \subseteq \mathcal{L}_f$ for all $c \geq \bar{c}$.

In other words, w_c is an exact penalty if it is weakly exact and any local minimizer of the unconstrained problem is a local solution of the constrained one, when c is sufficiently large.

First, let us show that the equality from Lemma 4.2 becomes an inequality if the considered point is not KKT but only feasible.

Lemma 4.3 *Let $x \in \mathbb{R}^n$ be a feasible point for (NLP). Then, $w_c(x) \leq f(x)$ for all $c > 0$.*

Proof Fix some $c > 0$. From Lemma 4.1, $w_c(x)$ can be written as

$$w_c(x) = f(x) + \sum_{i=1}^{m} \big[\bar{y}_c(x)\big]_i + \big\langle \mu(x), h(x) \big\rangle + \frac{c}{2} \big\| h(x) \big\|^2,$$

where

$$\big[\bar{y}_c(x)\big]_i := \lambda_i(x)\big(g_i(x) + \big[y_c(x)\big]_i\big) + \frac{c}{2}\big(g_i(x) + \big[y_c(x)\big]_i\big)^2.$$

Since $h(x) = 0$, it is enough to see that $[\bar{y}_c(x)]_i \leq 0$ for all $i = 1, \ldots, m$. Thus, consider two cases:

1. For an index i such that $[y_c(x)]_i = -\lambda_i(x)/c - g_i(x)$, we have

$$\left[\bar{y}_c(x)\right]_i = \lambda_i(x)\left(-\frac{\lambda_i(x)}{c}\right) + \frac{c}{2}\left(-\frac{\lambda_i(x)}{c}\right)^2 = -\frac{\lambda_i^2(x)}{2c} \leq 0.$$

2. For an index i such that $[y_c(x)]_i = 0$, that is, $\lambda_i(x)/c + g_i(x) \geq 0$, we have

$$\left[\bar{y}_c(x)\right]_i = \lambda_i(x)g_i(x) + \frac{c}{2}g_i^2(x) = \frac{c}{2}g_i(x)\left[2\left(\frac{\lambda_i(x)}{c} + g_i(x)\right) - g_i(x)\right].$$

As $g_i(x) \leq 0$, the term between the box brackets is nonnegative. Thus, we have $[\bar{y}_c(x)]_i \leq 0$ and the conclusion follows. □

The results concerning local minimizers are similar to Theorem 4.1 and Corollary 4.3 in the sense that, if a local solution of the original problem is not recovered, then we end up in a stationary point of the infeasibility measure $\mathcal{F}$ that is infeasible for (NLP).

Theorem 4.3 *Let $\{x^k\} \subseteq \mathcal{R} \subset \mathbb{R}^n$ and $\{c_k\} \subset \mathbb{R}_{++}$ be sequences such that $c_k \to \infty$ and $x^k \in \mathcal{L}_w(c_k)$ for all k. Let $\{x^{k_j}\}$ be a subsequence of $\{x^k\}$ such that $x^{k_j} \to \bar{x} \subset \mathcal{R}$. If Assumption 4.1 holds, then either there exists K such that $x^{k_j} \in \mathcal{L}_f$ for all $k_j > K$, or $\bar{x}$ is a stationary point of $\mathcal{F}$ that is infeasible for (NLP).*

Proof Since $x^{k_j} \in \mathcal{L}_w(c_{k_j})$ implies $\nabla w_{c_{k_j}}(x^{k_j}) = 0$ for all k_j, from Corollary 4.3, there is K such that x^{k_j} is KKT for all $k_j > K$ or $\bar{x}$ is a stationary point of $\mathcal{F}$ that is infeasible. Considering the first case and fixing $k_j > K$, from Lemma 4.2, there exists a neighborhood $V(x^{k_j})$ of x^{k_j} such that

$$f\left(x^{k_j}\right) = w_{c_{k_j}}\left(x^{k_j}\right) \leq w_{c_{k_j}}(x) \quad \text{for all } x \in V\left(x^{k_j}\right).$$

The above statement is clearly true for all $x \in V(x^{k_j}) \cap X$. Thus, using Lemma 4.3, we conclude that $f(x^{k_j}) \leq w_{c_{k_j}}(x) \leq f(x)$ for all $x \in V(x^{k_j}) \cap X$. This means that $x^{k_j} \in \mathcal{L}_f$ for all $k_j > K$, which completes the proof. □

Corollary 4.4 *Assume that there exists $\bar{c} > 0$ such that $\bigcup_{c>\bar{c}} \mathcal{L}_w(c)$ is bounded. Consider also that Assumption 4.1 holds and that all stationary points of $\mathcal{F}$ are feasible for (NLP). Then, there exists $\tilde{c} > 0$ such that if $x \in \mathcal{L}_w(c)$ and $c > \tilde{c}$, then $x \in \mathcal{L}_f$.*

Proof Suppose that the statement is false. So, there exist sequences $\{x^k\} \subset \mathbb{R}^n$ and $\{c_k\} \subset \mathbb{R}_{++}$ with $x^k \in \mathcal{L}_w(c_k)$ and $c_k \to \infty$ and such that $x^k \notin \mathcal{L}_f$. But for $c_k > \bar{c}$, we have $x^k \in \bigcup_{c>\bar{c}} \mathcal{L}_w(c)$, which is bounded. So, there exists a convergent subsequence $\{x^{k_j}\}$ of $\{x^k\}$. The contradiction follows from Theorem 4.3 and because there is no stationary point of $\mathcal{F}$ that is infeasible. □

The previous results allow us to develop an algorithm to solve the problem (NLP). In particular, Theorem 4.1 and Corollary 4.3 show that we can find a KKT point if we solve the systems of equations $W_c(x) = 0$ or $\nabla w_c(x) = 0$ for a large enough c. Since W_c is semismooth and its formula does not contain second-order terms (see (9)), a semismooth Newton method can be used to solve $W_c(x) = 0$. It remains to show an easy way to choose the parameter c. Also, convergence results using the semismooth Newton method should be presented, as well as a globalization idea. The next three sections will be dedicated to these topics.

5 Updating the Penalty Parameter

As we noted before, we need a way to choose the penalty parameter c. Following [1], we consider the dynamical update of parameter proposed by Glad and Polak [7]. The idea is to create a function, which is called *test function*, that measures the risk of computing a zero of W_c that is not KKT. First, define the following function:

$$a_c(x) := g(x) + y_c(x) = \max\left\{g(x), -\frac{\lambda(x)}{c}\right\}.$$

Note that for all $c > 0$, $a_c(x) = 0$ is equivalent to $\langle g(x), \lambda(x)\rangle = 0$, $g(x) \leq 0$ and $\lambda(x) \geq 0$. Hence, we can define a test function by

$$t_c(x) := -\left\|W_c(x)\right\|^2 + \frac{1}{c^\gamma}\left(\left\|a_c(x)\right\|^2 + \left\|h(x)\right\|^2\right),$$

with $\gamma > 0$. It is easy to show that t_c is continuous because of the continuity of the involved functions. Next proposition shows that t_c is in fact a test function.

Proposition 5.1 *The following statements are equivalent*:

(a) $(x, \lambda(x), \mu(x))$ *is a KKT triple associated to the problem.*
(b) $W_c(x) = 0$, $a_c(x) = 0$ *and* $h(x) = 0$.
(c) $W_c(x) = 0$ *and* $t_c(x) \leq 0$.

Proof It follows directly from the formulas of $W_c(x)$, $a_c(x)$, and $t_c(x)$ and the definition of KKT triple. □

Let us show now that for all $\bar{x} \in \mathcal{R}$, either $\bar{x}$ is a stationary point of $\mathcal{F}$ that is infeasible for (NLP), or there exists $\bar{c}$ large enough such that $t_c(x) \leq 0$ for all $c \geq \bar{c}$ and all x in a neighborhood of $\bar{x}$. From Proposition 5.1, this second case reveals us a way to update the parameter c. More precisely, for each computation of a zero of W_c, we increase the value of c if the test t_c at this point is greater that zero.

Lemma 5.1 *Let $S \subseteq \mathcal{R} \subset \mathbb{R}^n$ be a compact set with no KKT points. Then, either there exist $\bar{c}$, $\bar{\varepsilon}$ (which depend on S) such that $\|W_c(x)\| \geq \bar{\varepsilon}$ for all $x \in S$ and all $c \geq \bar{c}$; or there exist $\{x^k\} \subset S$, $\{c_k\} \subset \mathbb{R}_{++}$ such that $c_k \to \infty$, $\|W_{c_k}(x^k)\| \to 0$ and $\{x^k\}$ converges to a stationary point of $\mathcal{F}$ that is infeasible for* (NLP).

Proof If the first condition does not hold, then there exist two sequences $\{x^k\} \subset S$ and $\{c_k\} \subset \mathbb{R}_{++}$ such that $x^k \to \bar{x} \in S$, $c_k \to \infty$ and $\|W_{c_k}(x^k)\| \to 0$. Recalling the definition of W_{c_k} and using the continuity of the functions involved ($\lambda(\cdot)$ and $\mu(\cdot)$ are continuous by relaxed LICQ), we have

$$\begin{aligned} &\nabla f\big(x^k\big) + Jg\big(x^k\big)^T \lambda\big(x^k\big) + Jh\big(x^k\big)^T \mu\big(x^k\big) \\ &\quad + c_k \big(Jg\big(x^k\big)^T \max\big\{g\big(x^k\big), -\lambda\big(x^k\big)/c_k\big\} + Jh\big(x^k\big)^T h\big(x^k\big)\big) \to 0. \end{aligned} \tag{17}$$

As $c_k \to \infty$, $Jg(\bar{x})^T \max\{g(\bar{x}), 0\} + Jh(\bar{x})^T h(\bar{x}) = 0$, that is, $\bar{x}$ is a stationary point of $\mathcal{F}$. Suppose by contradiction that $\bar{x}$ is feasible and define

$$\begin{aligned} \bar{\lambda}^k &:= \lambda\big(x^k\big) + c_k \max\big\{g\big(x^k\big), -\lambda\big(x^k\big)/c_k\big\} = \max\big\{\lambda\big(x^k\big) + c_k g\big(x^k\big), 0\big\}, \\ \bar{\mu}^k &:= \mu\big(x^k\big) + c_k h\big(x^k\big). \end{aligned}$$

It follows from (17) that $\nabla f(x^k) + Jg(x^k)^T \bar{\lambda}^k + Jh(x^k)^T \bar{\mu}^k \to 0$. By continuity, we conclude that $\bar{\lambda}^k \to \bar{\lambda} \geq 0$, $\bar{\mu}^k \to \bar{\mu}$ and $\nabla f(\bar{x}) + Jg(\bar{x})^T \bar{\lambda} + Jh(\bar{x})^T \bar{\mu} = 0$. Also, the definition of $\bar{\lambda}^k$ shows that if $g(\bar{x}) < 0$ then $\bar{\lambda} = 0$. Therefore, $(\bar{x}, \bar{\lambda}, \bar{\mu})$ is a KKT triple, which is a contradiction because $\bar{x} \in S$ and S has no KKT points. □

Proposition 5.2 *For all $\bar{x} \in \mathcal{R} \subset \mathbb{R}^n$, either $\bar{x}$ is a stationary point of $\mathcal{F}$ that is infeasible for* (NLP)*, or there exist $\bar{c}$, $\bar{\delta} > 0$ such that if $c \geq \bar{c}$ and if $\|x - \bar{x}\| \leq \bar{\delta}$ with $x \in \mathcal{R}$, then $t_c(x) \leq 0$.*

Proof Suppose that the second assertion does not hold, that is, there are sequences $\{x^k\} \subset \mathbb{R}^n$ and $\{c_k\} \subset \mathbb{R}_{++}$ such that $x^k \to \bar{x}$, $c_k \to \infty$ and $t_{c_k}(x^k) > 0$ (which shows directly that, for all k, x^k is not a KKT point). Let us consider two cases.

1. Assume that $\bar{x}$ is not a KKT point. Consider the Lemma 5.1 applied to the set $S := \{x^k\} \cup \{\bar{x}\}$. Then, we have that $\bar{x}$ is either an infeasible stationary point of $\mathcal{F}$ or we have
$$t_{c_k}\big(x^k\big) \leq -\bar{\varepsilon}^2 + \frac{1}{c_k^\gamma}\big(\big\|a_{c_k}\big(x^k\big)\big\|^2 + \big\|h\big(x^k\big)\big\|^2\big)$$
for all k large enough. Since $c_k^\gamma \to \infty$, we have a contradiction in this last case and the claim follows.
2. Assume now that $\bar{x}$ is a KKT point. Equation (13) gives
$$K_{c_k}\big(x^k\big)\begin{bmatrix} a_{c_k}(x^k) \\ h(x^k) \end{bmatrix} = \frac{1}{c_k}\begin{bmatrix} Jg(x^k) \\ Jh(x^k) \end{bmatrix} W_{c_k}\big(x^k\big)$$
for all k. Let us recall that $K_{c_k}(x^k)$ converges to a nonsingular matrix $N(\bar{x})$ and $Jg(x^k)$, $Jh(x^k)$ converge respectively to $Jg(\bar{x})$ and $Jh(\bar{x})$. Hence, for sufficiently large k, we have
$$\left\|\begin{bmatrix} a_{c_k}(x^k) \\ h(x^k) \end{bmatrix}\right\| \leq \frac{1}{c_k}\big\|N^{-1}(\bar{x})\big\|\big\|W_{c_k}\big(x^k\big)\big\|\left\|\begin{bmatrix} Jg(\bar{x}) \\ Jh(\bar{x}) \end{bmatrix}\right\|.$$

Squaring both sides of the above inequality gives

$$\|a_{c_k}(x^k)\|^2 + \|h(x^k)\|^2 \leq \frac{1}{c_k^2}\|N^{-1}(\bar{x})\|^2 \|W_{c_k}(x^k)\|^2 \big(\|Jg(\bar{x})\|^2 + \|Jh(\bar{x})\|^2\big).$$

Thus,

$$\begin{aligned} t_{c_k}(x^k) &= -\|W_{c_k}(x^k)\|^2 + \frac{1}{c_k^\gamma}\big(\|a_{c_k}(x^k)\|^2 + \|h(x^k)\|^2\big) \\ &= \left(\frac{1}{c_k^{\gamma+2}}\|N^{-1}(\bar{x})\|^2\big(\|Jg(\bar{x})\|^2 + \|Jh(\bar{x})\|^2\big) - 1\right)\|W_{c_k}(x^k)\|^2, \end{aligned}$$

which is not positive as $c_k^{\gamma+2} \to \infty$, giving again a contradiction. □

We finish this section with a strategy to dynamically update the penalty parameter and a theorem associated to it.

Algorithm 5.1 Dynamical update of the penalty parameter.

1. Let $\mathcal{A}(x, c)$ be an algorithm that computes a zero of W_c.
   Initialize $x^0 \in \mathbb{R}^n$, $c_0 > 0$, $\xi > 1$ and $\gamma > 0$. Set $k = 0$.
2. If x^k is a KKT point of the problem, stop.
3. While $t_{c_k}(x^k) > 0$, do $c_k = \xi c_k$.
4. Compute $x^{k+1} = \mathcal{A}(x^k, c_k)$, set $k = k + 1$ and go to step 2.

Theorem 5.1 *Let $\{x^k\} \subseteq \mathcal{R} \subset \mathbb{R}^n$ be a sequence computed by Algorithm 5.1. If $\{x^k\}$ is bounded and infinite, then for each one of its accumulation points in $\mathcal{R}$, either it satisfies the KKT conditions or it is stationary point of $\mathcal{F}$ that is infeasible for* (NLP).

Proof Let $\bar{x}$ be an accumulation point of $\{x^k\}$. Then, by Proposition 5.2, if $\bar{x}$ is not a stationary point of $\mathcal{F}$ that is infeasible, then $t_{c_k}(x^k) \leq 0$ for all large enough k. Let $\bar{c}$ be the largest computed c_k value. Since $\bar{x}$ is a feasible accumulation point of an algorithm that computes a zero of W_c, we have $W_{\bar{c}}(\bar{x}) = 0$. Also, the continuity of $t_{\bar{c}}$ gives $t_{\bar{c}}(\bar{x}) \leq 0$ and we conclude that $\bar{x}$ is KKT. □

6 Local Convergence Results

In the same way as André and Silva's penalty function [1], W_c is not differentiable, but it is semismooth, which means that we can solve $W_c(x) = 0$ using an extension of the Newton method for these kinds of equations. The convergence theorem of semismooth Newton method [23, Chap. 7] shows that if x^* is a KKT point and all the elements of the B-subdifferential $\partial_B W_c(x^*)$ of W_c at x^* are nonsingular (with c large enough), then the method converges superlinearly. If, in addition, $\nabla^2 f$, $\nabla^2 g_i$, $i = 1, \ldots, m$ and $\nabla^2 h_i$, $i = 1, \ldots, p$ are locally Lipschitz continuous, then W_c is strongly semismooth and the convergence of the method is quadratic.

We present now the proof of the convergence rate following the structure of Facchinei, Kanzow and Palagi's manuscript [22]. First, we obtain a characterization of the elements of the B-subdifferential at a KKT point. To make it possible let us introduce some notations. Let x^* be a KKT point and λ^* the associated multiplier for inequality constraints. Then, denote by

$$I^* := I_=\big(x^*\big) \quad \text{and} \quad I_0^* := \big\{i \in I^* \colon \lambda_i^* = 0\big\}$$

the index set of active constraints at x^* (see Definition 3.1) and the indices of degenerate constraints, that is, those that are active with null multipliers.

Theorem 6.1 *Let (x^*, λ^*, μ^*) be a KKT triple associated to the problem* (NLP) *with $x^* \in \mathcal{R}$. Then, for any $H \in \partial_B W_c(x^*)$, with $c > 0$ sufficiently large, there exists an index set $I \subseteq I_0^*$ (which depends on H) such that*

$$H = \nabla^2_{xx} L\big(x^*, \lambda^*, \mu^*\big) + \sum_{i \in I^* \setminus I} \nabla g_i\big(x^*\big)\big(\nabla \lambda_i\big(x^*\big)^T + c \nabla g_i\big(x^*\big)^T\big)$$
$$+ \sum_{i=1}^{p} \nabla h_i\big(x^*\big)\big(\nabla \mu_i\big(x^*\big)^T + c \nabla h_i\big(x^*\big)^T\big).$$

Proof Recalling the definition of W_c, we have

$$\begin{aligned} W_c(x) &= \nabla f(x) + Jg(x)^T\big(\lambda(x) + cg(x)\big) + cJg(x)^T y_c(x) \\ &\quad + Jh(x)^T\big(\mu(x) + ch(x)\big) \\ &= \nabla f(x) + \sum_{i=1}^{m}\big(\lambda_i(x) + cg_i(x)\big)\nabla g_i(x) + c\sum_{i=1}^{m}(y_c)_i(x)\nabla g_i(x) \\ &\quad + \sum_{i=1}^{p}\big(\mu_i(x) + ch_i(x)\big)\nabla h_i(x). \end{aligned}$$

Let $H \in \partial_B W_c(x^*)$ be arbitrarily given. The relaxed LICQ assumption ensures that $\lambda^* = \lambda(x^*)$ and $\mu^* = \mu(x^*)$. Then

$$\begin{aligned} H &= \nabla^2 f\big(x^*\big) + \sum_{i=1}^{m}\big(\lambda_i^* + cg_i\big(x^*\big)\big)\nabla^2 g_i\big(x^*\big) \\ &\quad + \sum_{i=1}^{p}\big(\mu_i^* + ch_i\big(x^*\big)\big)\nabla^2 h_i\big(x^*\big) + c\tilde{H} \\ &\quad + \sum_{i=1}^{m}\nabla g_i\big(x^*\big)\big(\nabla \lambda_i\big(x^*\big)^T + c\nabla g_i\big(x^*\big)^T\big) \\ &\quad + \sum_{i=1}^{p}\nabla h_i\big(x^*\big)\big(\nabla \mu_i\big(x^*\big)^T + c\nabla h_i\big(x^*\big)^T\big) \end{aligned}$$

$$= \nabla^2_{xx} L\big(x^*, \lambda^*, \mu^*\big) + c \sum_{i=1}^{m} g_i\big(x^*\big) \nabla^2 g_i\big(x^*\big)$$
$$+ \sum_{i=1}^{m} \nabla g_i\big(x^*\big)\big(\nabla \lambda_i\big(x^*\big)^T + c \nabla g_i\big(x^*\big)^T\big)$$
$$+ c\tilde{H} + \sum_{i=1}^{p} \nabla h_i\big(x^*\big)\big(\nabla \mu_i\big(x^*\big)^T + c \nabla h_i\big(x^*\big)^T\big), \tag{18}$$

for some $\tilde{H} \in \partial_B \tilde{W}_c(x^*)$, where

$$\tilde{W}_c(x) := \sum_{i=1}^{m} (y_c)_i(x) \nabla g_i(x).$$

By the definition of B-subdifferential, there is a sequence $\{x^k\} \subset \mathbb{R}^n$ converging to x^* such that x^k is a F-differentiable point of $\tilde{W}_c$ for all k and $\tilde{H} = \lim_{k\to\infty} J\tilde{W}_c(x^k)$. Now, let us analyze the differentiability of $(y_c)_i(x) := \max\{0, -\lambda_i(x)/c - g_i(x)\}$ for all $i = 1, \ldots, m$. To do this, we consider three cases:

1. If $i \in I^* \setminus I_0^*$, then $(y_c)_i$ is continuously differentiable around x^*. Furthermore, since $\lambda_i^* > 0$, we have for k big enough

$$\lambda_i\big(x^k\big) > 0 \quad \Rightarrow \quad -\frac{\lambda_i(x^k)}{c} < 0 \quad \Rightarrow \quad (y_c)_i\big(x^k\big) = 0.$$

   Hence,

$$\nabla (y_c)_i\big(x^*\big) = \lim_{k\to\infty} \nabla (y_c)_i\big(x^k\big) = 0.$$

2. If $i \notin I^*$, then $(y_c)_i$ is continuously differentiable around x^*. Since $g_i(x^*) < 0$, we have for k big enough,

$$g_i\big(x^k\big) < 0 \quad \Rightarrow \quad -g_i\big(x^k\big) > 0 \quad \Rightarrow \quad -\frac{\lambda_i(x^k)}{c} - g_i\big(x^k\big) > 0,$$

   where the last implication follows from the fact that $\lambda_i(x^k) \to 0$ by complementarity conditions. This implies $(y_c)_i(x^k) = -\lambda_i(x^k)/c - g_i(x^k)$ and, therefore,

$$\nabla (y_c)_i\big(x^*\big) = \lim_{k\to\infty} \nabla (y_c)_i\big(x^k\big) = -\frac{\nabla \lambda_i(x^*)}{c} - \nabla g_i\big(x^*\big).$$

3. If $i \in I_0^*$, then we have another three possibilities.
   (a) If $-\lambda_i(x^k)/c - g_i(x^k) > 0$ for all k (or for infinitely many k), then we have $(y_c)_i(x^k) = -\lambda_i(x^k)/c - g_i(x^k)$. Hence,

$$\lim_{k\to\infty} \nabla (y_c)_i\big(x^k\big) = -\frac{\nabla \lambda_i(x^*)}{c} - \nabla g_i\big(x^*\big). \tag{19}$$

(b) If $-\lambda_i(x^k)/c - g_i(x^k) < 0$ for all k, we have $(y_c)_i(x^k) = 0$. Then we obtain

$$\lim_{k\to\infty} \nabla(y_c)_i\big(x^k\big) = 0.$$

(c) If $-\lambda_i(x^k)/c - g_i(x^k) = 0$ for all k, then $(y_c)_i$ will be nondifferentiable unless $-\nabla\lambda_i(x^k)/c - \nabla g_i(x^k) = 0$. Since $\nabla g_i(x^k)$ is nonzero by relaxed LICQ and c is sufficiently large, we conclude that this equality does not hold.

Define now the following index set:

$$I := \big\{i \in I_0^* : \text{ equality (19) holds}\big\}.$$

Also note that $(y_c)_i(x^*) = -g_i(x^*)$. In fact, if $g_i(x^*) = 0$, then $(y_c)_i(x^*) = 0$, because $-\lambda_i^*/c \le 0$. On the other hand, if $g_i(x^*) < 0$, in view of complementarity conditions, we have $\lambda_i^* = 0$ and so $(y_c)_i(x^*) = -g_i(x^*)$. In this way, we have the following representation of $\tilde{H}$:

$$\begin{aligned}\tilde{H} &= \sum_{i=1}^{m}(y_c)_i\big(x^*\big)\nabla^2 g_i\big(x^*\big) + \sum_{i\notin I^*}\nabla g_i\big(x^*\big)\left(-\frac{\nabla\lambda_i(x^*)^T}{c} - \nabla g_i\big(x^*\big)^T\right)\\ &\quad + \sum_{i\in I}\nabla g_i\big(x^*\big)\left(-\frac{\nabla\lambda_i(x^*)^T}{c} - \nabla g_i\big(x^*\big)^T\right)\\ &= -\sum_{i=1}^{m} g_i\big(x^*\big)\nabla^2 g_i\big(x^*\big)\\ &\quad + \sum_{i\notin I^*\setminus I}\nabla g_i\big(x^*\big)\left(-\frac{\nabla\lambda_i(x^*)^T}{c} - \nabla g_i\big(x^*\big)^T\right). \end{aligned} \tag{20}$$

Finally, putting together equalities (18) and (20), we have

$$\begin{aligned} H &= \nabla_{xx}^2 L\big(x^*,\lambda^*,\mu^*\big) + c\sum_{i=1}^{m} g_i\big(x^*\big)\nabla^2 g_i\big(x^*\big)\\ &\quad + \sum_{i=1}^{m}\nabla g_i\big(x^*\big)\big(\nabla\lambda_i\big(x^*\big)^T + c\nabla g_i\big(x^*\big)^T\big)\\ &\quad - c\sum_{i=1}^{m} g_i\big(x^*\big)\nabla^2 g_i\big(x^*\big) - \sum_{i\notin I^*\setminus I}\nabla g_i\big(x^*\big)\big(\nabla\lambda_i\big(x^*\big)^T + c\nabla g_i\big(x^*\big)^T\big)\\ &\quad + \sum_{i=1}^{p}\nabla h_i\big(x^*\big)\big(\nabla\mu_i\big(x^*\big)^T + c\nabla h_i\big(x^*\big)^T\big) \end{aligned}$$

$$= \nabla^2_{xx} L\big(x^*, \lambda^*, \mu^*\big) + \sum_{i \in I^* \setminus I} \nabla g_i\big(x^*\big)\big(\nabla \lambda_i\big(x^*\big)^T + c \nabla g_i\big(x^*\big)^T\big)$$

$$+ \sum_{i=1}^{p} \nabla h_i\big(x^*\big)\big(\nabla \mu_i\big(x^*\big)^T + c \nabla h_i\big(x^*\big)^T\big),$$

which gives the desired result. □

In order to prove their convergence rate result, Facchinei, Kanzow, and Palagi [22] used an assumption called *weak regularity*, which is defined below.

Definition 6.1 A KKT triple (x^*, λ^*, μ^*) is said *weakly regular* if and only if, for all $I \subseteq I_0^*$, the Hessian $\nabla^2_{xx} L(x^*, \lambda^*, \mu^*)$ is not singular on the subspace

$$U_I := \big\{d \in \mathbb{R}^n : \big\langle \nabla g_i\big(x^*\big), d\big\rangle = 0,\ i \in I^* \setminus I,$$
$$\big\langle \nabla h_i\big(x^*\big), d\big\rangle = 0,\ i = 1, \ldots, p\big\}, \tag{21}$$

or, in other words, if

$$P_{U_I} \nabla^2_{xx} L\big(x^*, \lambda^*, \mu^*\big) d \neq 0 \quad \text{for all } d \in U_I, d \neq 0,$$

with P_{U_I} denoting the orthogonal projection onto U_I, and for all subset $I \subseteq I_0^*$.

Since they were interested in variational inequalities, we propose related conditions that guarantee fast convergence, which seems more natural in the optimization context. The first one is the well-known second-order sufficient condition.

Definition 6.2 A KKT triple (x^*, λ^*, μ^*) satisfies the *second-order sufficient condition* if and only if

$$\big\langle \nabla^2_{xx} L\big(x^*, \lambda^*, \mu^*\big) d, d\big\rangle > 0 \quad \text{for all } d \in U_{I_0^*} \cap U_0,\ d \neq 0,$$

where, $U_{I_0^*}$ is defined in (21) with $I = I_0^*$ and

$$U_0 := \big\{d \in \mathbb{R}^n : \big\langle \nabla g_i\big(x^*\big), d\big\rangle \leq 0 \text{ for all } i \in I_0^*\big\}. \tag{22}$$

Unfortunately, this condition is not sufficient to prove the convergence rate. An additional assumption that ensures fast convergence is the strict complementarity, that is, $I_0^* = \emptyset$. But this is a strong condition, in the sense that it does not hold in many cases. Classical results in sequential quadratic programming [24, Chap. 18], for example, use these two conditions, but more recent works in this area [25, 26] use instead the *strong second-order sufficient condition*, which we recall below.

Definition 6.3 A KKT triple (x^*, λ^*, μ^*) satisfies the *strong second-order sufficient condition* if and only if

$$\big\langle \nabla^2_{xx} L\big(x^*, \lambda^*, \mu^*\big) d, d\big\rangle > 0 \quad \text{for all } d \in U_{I_0^*},\ d \neq 0.$$

This condition is also required in recent works about exact merit functions [19, 20], and ensures the convergence rate. However, it can be proved that it is more restrictive than weak regularity.

Lemma 6.1 *Let (x^*, λ^*, μ^*) be a KKT triple. If it satisfies the strong second-order sufficient condition, then it is weakly regular.*

Proof Since $I \subseteq I_0^*$ implies $U_I \subseteq U_{I_0^*}$, the strong second-order sufficient condition can be written as

$$\left\langle \nabla_{xx}^2 L\left(x^*, \lambda^*, \mu^*\right) d, d \right\rangle > 0 \quad \text{for all } d \in U_I, d \neq 0 \text{ and all } I \subseteq I_0^*.$$

Fix $I \subseteq I_0^*$ and $d \in U_I$ with $d \neq 0$. Since P_{U_I} is symmetric and $d \in U_I$,

$$0 < \left\langle \nabla_{xx}^2 L\left(x^*, \lambda^*, \mu^*\right) d, d \right\rangle = \left\langle P_{U_I} \nabla_{xx}^2 L\left(x^*, \lambda^*, \mu^*\right) d, d \right\rangle.$$

This shows that $P_{U_I} \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*) d \neq 0$, and thus the weak regularity holds. □

Remark The converse implication is not necessarily true. Indeed, we have the following counterexample: $n = 2$, $m = 3$ and $p = 0$, with $f(x) := -x_2$, $g_1(x) := -x_1^2 + x_2$, $g_2(x) := -x_1$ and $g_3(x) := x_1$. In this case, we obtain $x^* = (0, 0)^T$, $I^* = \{1, 2, 3\}$ and $\nabla_x L(x^*, \lambda^*) = 0$ implies that $\lambda_1^* = 1$ and $\lambda_2^* = \lambda_3^*$. If $\lambda_2^* = \lambda_3^* = 0$, we have $I_0^* = \{2, 3\}$. Thus, we obtain $U_{\{2\}} = U_{\{3\}} = U_\emptyset = \{(0, 0)^T\}$, $U_0 = \{d \in \mathbb{R}^2 : d_1 = 0\}$, and $U_{I_0^*} = \{d \in \mathbb{R}^2 : d_2 = 0\}$. For the case $I = I_0^*$, we have

$$P_{U_{I_0^*}} \nabla_{xx}^2 L\left(x^*, \lambda^*\right) d = P_{U_{I_0^*}} \begin{bmatrix} -2d_1 \\ 0 \end{bmatrix} = \begin{bmatrix} -2d_1 \\ 0 \end{bmatrix} \neq 0$$

for all $d \in U_{I_0^*}$, $d \neq 0$. Therefore, the weak regularity is satisfied. However,

$$\left\langle \nabla_{xx}^2 L\left(x^*, \lambda^*\right) d, d \right\rangle = -2d_1^2 \leq 0,$$

which shows that the strong second-order sufficient condition is not satisfied.

Recalling that we desire a natural condition in the context of optimization, and as an alternative, we propose to use the second-order sufficient condition with another condition associated to the nonsingularity of the Hessian $\nabla_{xx}^2 L(x^*, \lambda^*, \mu^*)$, which is similar to weak regularity.

Assumption 6.1 Let (x^*, λ^*, μ^*) be a KKT triple. Then the second-order sufficient condition (see Definition 6.2) holds and

$$P_{U_{I_0^*}} \nabla_{xx}^2 L\left(x^*, \lambda^*, \mu^*\right) d \neq 0 \quad \text{for all } d \in U_{I_0^*} \setminus U_0,\ d \neq 0,$$

recalling that $U_{I_0^*}$ is defined in (21) for $I = I_0^*$ and U_0 is defined in (22).

Theorem 6.2 *Let (x^*, λ^*, μ^*) be a KKT triple associated to the problem* (NLP) *with $x^* \in \mathcal{R}$. If Assumption* 6.1 *holds, then all matrices of the B-subdifferential $\partial_B W_c(x^*)$ are nonsingular for all sufficiently large $c > 0$.*

Proof Assume for the purpose of contradiction that there are sequences $\{c_k\} \subset \mathbb{R}_{++}$ and $\{H_k\} \subset \mathbb{R}^{n\times n}$ such that $c_k \to +\infty$ and $H_k \in \partial_B W_{c_k}(x^*)$ is singular for all k. Then there exists $\{d^k\} \subset \mathbb{R}^n$ with $\|d^k\| = 1$, such that $H_k^T d^k = 0$ for all k and, consequently, $\langle H_k^T d^k, d^k\rangle = 0$ for all k. Consider, without loss of generality, that $\{d^k\}$ converges to some vector $d^* \in \mathbb{R}^n$. From Theorem 6.1, for each k, there is a subset $I_k \subseteq I_0^*$ such that

$$H_k^T d^k = \nabla^2_{xx} L\big(x^*, \lambda^* \mu^*\big) d^k + \sum_{i \in I^* \setminus I_k} \big\langle \nabla g_i\big(x^*\big), d^k\big\rangle \big(\nabla \lambda_i\big(x^*\big) + c_k \nabla g_i\big(x^*\big)\big)$$
$$+ \sum_{i=1}^{p} \big\langle \nabla h_i\big(x^*\big), d^k\big\rangle \big(\nabla \mu_i\big(x^*\big) + c_k \nabla h_i\big(x^*\big)\big). \tag{23}$$

Since there are finitely many subsets I_k, we may assume that $I_k = I$ for all k. Thus,

$$\big\langle H_k^T d^k, d^k\big\rangle = \big\langle \nabla^2_{xx} L\big(x^*, \lambda^*, \mu^*\big) d^k, d^k\big\rangle + \sum_{i \in I^* \setminus I} \big\langle \nabla g_i\big(x^*\big), d^k\big\rangle \big\langle \nabla \lambda_i\big(x^*\big), d^k\big\rangle$$
$$+ c_k \sum_{i \in I^* \setminus I} \big\langle \nabla g_i\big(x^*\big), d^k\big\rangle^2 + \sum_{i=1}^{p} \big\langle \nabla h_i\big(x^*\big), d^k\big\rangle \big\langle \nabla \mu_i\big(x^*\big), d^k\big\rangle$$
$$+ c_k \sum_{i=1}^{p} \big\langle \nabla h_i\big(x^*\big), d^k\big\rangle^2.$$

Dividing the above expression by c_k, taking $c_k \to +\infty$ and recalling that, for all k, $\langle H_k^T d^k, d^k\rangle = 0$, it follows that

$$\sum_{i \in I^* \setminus I} \big\langle \nabla g_i\big(x^*\big), d^k\big\rangle^2 + \sum_{i=1}^{p} \big\langle \nabla h_i\big(x^*\big), d^k\big\rangle^2 \to 0,$$

which implies that $\langle \nabla g_i(x^*), d^*\rangle = 0$ for all $i \in I^* \setminus I$ and $\langle \nabla h_i(x^*), d^*\rangle = 0$ for all $i = 1, \ldots, p$, that is, $d^* \in U_I$. Now, observe that for all k,

$$\sum_{i \in I^* \setminus I} \big\langle \nabla g_i\big(x^*\big), d^k\big\rangle \nabla g_i\big(x^*\big) \in U_I^{\perp} \quad \text{and} \quad \sum_{i=1}^{p} \big\langle \nabla h_i\big(x^*\big), d^k\big\rangle \nabla h_i\big(x^*\big) \in U_I^{\perp}.$$

In fact, for all $d \in U_I$,

$$\sum_{i \in I^* \setminus I} \big\langle \nabla g_i\big(x^*\big), d^k\big\rangle \big\langle \nabla g_i\big(x^*\big), d\big\rangle = 0 \quad \text{and} \quad \sum_{i=1}^{p} \big\langle \nabla h_i\big(x^*\big), d^k\big\rangle \big\langle \nabla h_i\big(x^*\big), d\big\rangle = 0,$$

because of the definition of U_I. This fact, the equality (23) with $I_k = I$, and the fact that $H_k^T d^k = 0$ for all k, show that

$$0 = P_{U_I} \nabla^2_{xx} L\big(x^*, \lambda^* \mu^*\big) d^k + P_{U_I} \left[\sum_{i \in I^* \setminus I} \big\langle \nabla g_i\big(x^*\big), d^k \big\rangle \nabla \lambda_i\big(x^*\big) \right]$$

$$+ P_{U_I} \left[\sum_{i=1}^{p} \big\langle \nabla h_i\big(x^*\big), d^k \big\rangle \nabla \mu_i\big(x^*\big) \right]$$

for all k. Taking $k \to \infty$ in the above equality, we obtain

$$P_{U_I} \nabla^2_{xx} L\big(x^*, \lambda^* \mu^*\big) d^* = 0, \tag{24}$$

once again because $d^* \in U_I$.

Note now that if $d^* \notin U_0$, then we have a contradiction because of Assumption 6.1 and since $I \subseteq I_0^*$ implies $U_I \subseteq U_{I_0^*}$. On the other hand, if $d^* \in U_0$, then equality (24), the fact that $d^* \in U_I$ and the symmetry of P_{U_I} imply that

$$0 = \big\langle P_{U_I} \nabla^2_{xx} L\big(x^*, \lambda^* \mu^*\big) d^*, d^* \big\rangle = \big\langle \nabla^2_{xx} L\big(x^*, \lambda^* \mu^*\big) d^*, d^* \big\rangle,$$

which contradicts the second-order sufficient condition. □

From Lemma 6.1, it is easy to see that the convergence result above can be proved if we replace the Assumption 6.1 with the strong second-order sufficient condition or the weak regularity.

7 Globalization and Numerical Experiments

To globalize the method, we can just use the merit function $\|W_c(\cdot)\|^2/2$. But this happens to be nondifferentiable. Besides, recalling the formulas of ∇w_c and W_c in (8) and (9), we observe that a minimizer x of w_c should satisfy

$$\nabla w_c(x) = W_c(x) + \text{second-order terms} = 0.$$

Thus, a stationary point of w_c is a zero of W_c, if we ignore the second-order terms. In Theorem 4.1 and Corollary 4.3, we saw that these terms are ignored when x is a KKT point because, in this case, $\nabla w_c(x) = W_c(x) = 0$. This means that we can use the exact penalty w_c as the merit function. Also, the method can be considered as a *Gauss–Newton-type* method, in the sense that we can ignore second-order terms at the points that we are interested in.

The final algorithm is given below.

Algorithm 7.1 Gauss–Newton-type method to minimize w_c, using W_c to compute the search direction and with dynamical update of the penalty parameter.

1. Choose $x^0 \in \mathbb{R}^n$, $c_0 > 0$, $\xi > 1$, $\varepsilon_1 \geq 0$, $\varepsilon_2, \varepsilon_3 \in (0, 1)$ and $\sigma \in (0, 1/2)$. Set $k = 0$.

2. If $\|\nabla w_{c_k}(x^k)\| \leq \varepsilon_1$, stop.
3. While $t_{c_k}(x^k) > 0$, do $c_k = \xi c_k$.
4. Compute $W_{c_k}(x^k)$ and take $H_k \in \partial_B W_{c_k}(x^k)$.
5. Find d^k such that $H_k d^k = -W_{c_k}(x^k)$.
6. If $\langle \nabla w_{c_k}(x^k), d^k \rangle > -\varepsilon_2 \|d^k\| \|\nabla w_{c_k}(x^k)\|$ or $\|d^k\| < \varepsilon_3 \|\nabla w_{c_k}(x^k)\|$, set $d^k = -\nabla w_{c_k}(x^k)$.
7. Find $t_k \in (0, 1]$ such that $w_{c_k}(x^k + t_k d^k) \leq w_{c_k}(x^k) + \sigma t_k \langle \nabla w_{c_k}(x^k), d^k \rangle$ with a backtracking strategy.
8. Set $x^{k+1} = x^k + t_k d^k$, $k = k + 1$ and go to step 2.

Observe that in step 6 we verify if d^k is a sufficient descent direction and if the norm condition is satisfied. If not, we replace it with the steepest descent direction. Concerning the stepsize t_k, we do an Armijo-type line search at step 7. It is important to see that for each Armijo-type iteration, we have to compute the functional value $w_{c_k}(x^k + t_k d^k)$ that requires the multipliers estimates $\lambda(x^k + t_k d^k)$ and $\mu(x^k + t_k d^k)$, which means to solve a linear least squares problem. Since the matrix associated to this system of equations changes for each point, this strategy may be computationally expensive if many Armijo iterations are required.

To increase the robustness of the implementation, we also use nonmonotone line search [27] and quadratic interpolation for the backtracking strategy. When the steepest descent direction is taken instead of the Newton direction, we use the spectral step [28] to compute the initial stepsize of the iteration. One fundamental aspect in a globalization scheme is to verify if, eventually, the Newton direction is chosen and the unit stepsize is taken. When one of these conditions fails, then the globalization procedure deteriorates the superlinear convergence and the so-called "Maratos effect" occurs. Unfortunately, we have no formal proof that ensures that the "Maratos effect" does not show up. However, this seems not to affect the numerical behavior, as we will discuss later.

Let us now begin to describe the implementation of the Algorithm 7.1. We implemented it in Python and we consider the following values of parameters: $\gamma = 2$, $\xi = 10$, $\zeta = 2.0$, $\sigma = 10^{-4}$ (following [1]) and $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = 10^{-8}$ (following [29, 30]). To compute an element of the B-subdifferential $\partial_B W_{c_k}(x^k)$, we also have to choose a convexity parameter because of the maximum function contained in the formula of $W_{c_k}(x_k)$. We choose the one that vanishes λ_i and not g_i. Besides, we use the routine `lstsq` from SciPy library to solve the linear least squares problem associated to the estimates. Moreover, if $H_k \in \partial_B W_{c_k}(x^k)$ is computationally singular, then we add on it a multiple of the identity matrix in order to solve the system of equations (step 5 of Algorithm 7.1). At step 2, we also consider as an alternative stopping criterion that the point satisfies the KKT conditions within a prefixed tolerance of 10^{-8}. Furthermore, we limit the maximum number of iterations in 100000 and the time in 10 minutes. With all these implementation details, numerical experiments was done using the CUTE collection [31] modeled in AMPL [32].

We considered all constrained problems of such collection with at most 300 variables and constraints, which returns 328 problems. In order to have a comparison parameter, we also tested these problems with ALGENCAN [29, 30], an augmented Lagrangian method. The method proposed here does not make distinctions between box constraints and the other inequality ones as ALGENCAN do. Thus, to make

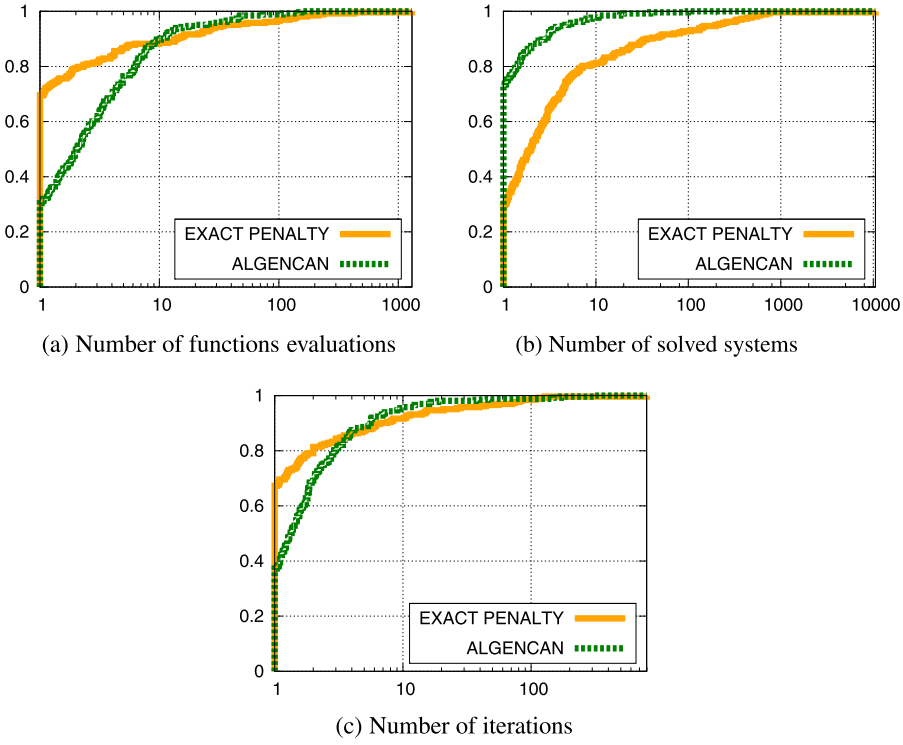
a more significant comparison, we modify ALGENCAN's interface with AMPL in order to not differentiate these box constraints. We also cut off extrapolations, Newtonian steps and acceleration of this solver, in order to make ALGENCAN to behave as a pure augmented Lagrangian algorithm. Moreover, in the same way of ALGENCAN, we make the initial penalty parameter of the exact penalty method being dependent of the problem. Since updates of the penalty parameter with the test function are not so frequent [1, 7], we also add a dynamical update of parameter using an analysis of the infeasibility measure used by ALGENCAN.

First, let us point out that the computational time is not considered here, because the implementation of the exact penalty method is not mature as ALGENCAN. We verified that ALGENCAN solved 297 problems, while the exact penalty method solved 282. If we ignore the (few) cases where the implementations fail because of functions evaluations, we have an effective robustness of 92.52 % and 87.31 %, respectively for ALGENCAN and for the exact penalty. Most of the failures of the exact penalty are due to the high condition numbers of the matrices associated to system of equations for the computation of the multipliers estimate, or the big values of penalty parameters, which cause numerical instabilities. Furthermore, in 7 problems, the exact penalty returned a stationary point of the infeasibility measure $\mathcal{F}$ that is infeasible. Concerning the total number of functions evaluations for each problem, the exact penalty method performs better. To clarify this fact, a performance profile [33], drawn on the subset of problems solved by both solvers, is presented in Fig. 1a. This result is interesting particularly for problems such that these evaluations are computationally expensive, which clearly appears in the literature [34, 35].

On the other hand, the total number of systems of equations is smaller with ALGENCAN, as shown in Fig. 1b. We note, however, that this comparison is not totally appropriate because it makes sense only in the case that the number of variables n do not differ a lot from the total number of constraints $m + p$. Let us recall that Algorithm 7.1 solves basically one $n \times n$ system for iteration during the computation of Newton's step, and at least one $(n + m + p) \times (m + p)$ system during the Armijo-type line search. In ALGENCAN's case, we only have $n \times n$ systems to solve for each inner iteration. Despite this, the comparison clearly shows that the Armijo-type line search is expensive, and thus should be avoided.

Another result refers to the number of iterations of the exact penalty against the number of inner iterations of ALGENCAN. In such a case, the exact penalty has the advantage, as it could be seen in Fig. 1c. Observe that the number of iterations in the exact penalty method is equal to the number of $n \times n$ systems, except for rare cases when the B-subdifferential matrix is computationally singular. This result suggests that if we replace the Armijo-type line search for some strategy that requires few $(n + m + p) \times (m + p)$ systems, then the total number of systems of equations in both algorithms could not differ so much. For further research, trust region methods could be a replacement for the Armijo-type line search.

As we noted before, we could not ensure that the "Maratos effect" does not occur, so we also search for a counterexample in the CUTE problem set that we had considered. We observe that there are 10 instances where the Newton directions with unit stepsizes are not eventually taken. However, in these cases the matrices associated to the systems of equations for the computation of Newton directions or the



(a) Number of functions evaluations

(b) Number of solved systems

(c) Number of iterations

Fig. 1 Performance profiles for the exact penalty algorithm and the ALGENCAN

multipliers estimates have a high condition number (at least 10^8). Numerically, this means that the Newton directions or the multipliers estimates are not computed accurately, and hence we can not really use them to decide that the "Maratos effect" occurs. Most likely, the numerical instabilities are the main reason for the numerical behavior in these problems. In all the other cases, if the assumption of relaxed LICQ at the KKT point is satisfied (which we ask in Theorem 6.2), the solver based on the exact penalty does not show the "Maratos effect." Although the theoretical question is not answered, the lack of a clear example of the "Maratos effect" in the 282 problems solved give high expectations about the global convergence of the proposed method.

Let us close this section with a brief comparison between the multipliers estimates (1) and (6), which we will refer just as the new estimate and Lucidi's estimate respectively. First, recall that the last one can be used to define the penalty function (7) in the same way as the new one. We also recall that Lucidi's estimate requires LICQ only in the set of feasible points, a weaker assumption than the relaxed LICQ, required by the new estimate, which is a clear theoretical advantage.

However, our numerical experiments with Lucidi's estimate showed that it may impact the robustness of the method unless a careful tuning of the parameters ζ_1 and ζ_2 is performed. We believe that the reason for this behavior is related to the term $\alpha(x)(\|\lambda\|^2 + \|\mu\|^2)$ appearing only in (6), where $\alpha(x)$ is a measure of the infeasi-

bility of x. This term introduces a tendency to select small multipliers whenever x is infeasible. In particular, infeasibility with respect to a single constraint impacts the multipliers associated to the other constraints. This behavior may lead to a larger increase of the penalty parameter, to achieve feasibility, when compared to an exact penalty based on the new estimate.

Depending on the values of the parameters ζ_1 and ζ_2, the necessity of a larger c can also be observed using the 328 problems from the CUTE test set. A large c can increase the possibility of converging to a stationary point of the infeasibility measure $\mathcal{F}$ that is infeasible for (NLP). In fact, while working on a Lucidi's version of the code we could see that it would fail returning such points in 7 up to 15 test problems, depending on the choice of parameters. This contrasts with a shorter range from 7 up to 10 problems observed with the penalty using the new multipliers estimate. Note, however, that with a careful choice of the parameters, both penalties solved 282 problems. The number of failures due to convergence to stationary points of $\mathcal{F}$ is also the same (in the case, 7 problems). Such result was established using the parameters $\zeta_1 = \zeta_2 = 2$ and it was the best robustness result that we obtained using Lucidi's estimate. One might conjecture that using small values for ζ_2 yields better robustness, but we could not confirm this behavior in our framework. Actually, for $\zeta_2 = 10^{-6}$ the number of problems solved decreases slightly to 279. Using $\zeta_2 = 10^{-8}$ the number of problems solved decreases more significantly to 266.

Finally, a natural question is if the multipliers estimate proposed by Glad and Polak in [7] performs well or not in practice. We recall that such estimate is obtained by solving a problem like (1) without the term $\|H(x)\mu\|^2$. We end up solving 271 problems from the CUTE test set that we are considering, and all unsolved problems by the new estimate were also not solved when using Glad and Polak's estimate.

8 Conclusion

As far as we know, numerical experiments with medium-sized test problems and exact merit functions for optimization (like Di Pillo and Grippo's [3]) had not been considered yet in the literature, except for exact augmented Lagrangians in the recent paper of Di Pillo et al. [20]. Here, we extend the penalty function for variational inequalities [1] in order to solve optimization problems, and we observe that in this case Di Pillo and Grippo's exact penalty could be used as a merit function. Further investigations into the implementation should be done, in particular, concerning the expensive Armijo-type line search.

Acknowledgements This work was supported by PRONEX-Optimization (PRONEX-CNPq/FAPERJ E-26/171.510/2006-APQ1), FAPESP (Grants 2010/20572-0, 2007/53471-0, 2006/53768-0, and 2005/02163-8) and CNPq (Grants 303030/2007-0 and 474138/2008-9). We are also grateful to the referees for their helpful comments.

References

1. André, T.A., Silva, P.J.S.: Exact penalties for variational inequalities with applications to nonlinear complementarity problems. Comput. Optim. Appl. **47**(3), 401–429 (2010)

2. Di Pillo, G., Grippo, L.: An exact penalty method with global convergence properties. Math. Program. **36**, 1–18 (1986)
3. Di Pillo, G., Grippo, L.: Exact penalty functions in constrained optimization. SIAM J. Control Optim. **27**(6), 1333–1360 (1989)
4. Zangwill, W.I.: Nonlinear programming via penalty functions. Manag. Sci. **13**, 344–358 (1967)
5. Bertsekas, D.P.: Nonlinear Programming, 2nd edn. Athena Scientific, Belmont (1999)
6. Fletcher, R.: A class of methods for nonlinear programming with termination and convergence properties. In: Abadie, J. (ed.) Integer and Nonlinear Programming, pp. 157–173. North-Holland, Amsterdam (1970)
7. Glad, T., Polak, E.: A multiplier method with automatic limitation of penalty growth. Math. Program. **17**(2), 140–155 (1979)
8. Mukai, H., Polak, E.: A quadratically convergent primal-dual algorithm with global convergence properties for solving optimization problems with equality constraints. Math. Program. **9**(3), 336–349 (1975)
9. Di Pillo, G., Grippo, L.: A new class of augmented Lagrangians in nonlinear programming. SIAM J. Control Optim. **17**(5), 618–628 (1979)
10. Di Pillo, G., Grippo, L.: A new augmented Lagrangian function for inequality constraints in nonlinear programming. J. Optim. Theory Appl. **36**(4), 495–519 (1982)
11. Di Pillo, G., Grippo, L.: A continuously differentiable exact penalty function for nonlinear programming problems with inequality constraints. SIAM J. Control Optim. **23**(1), 72–84 (1985)
12. Hestenes, M.R.: Multiplier and gradient methods. J. Optim. Theory Appl. **4**, 303–320 (1969)
13. Powell, M.J.D.: A method for nonlinear constraints in minimization problems. In: Fletcher, R. (ed.) Optimization, pp. 283–298. Academic Press, New York (1969)
14. Rockafellar, R.T.: Augmented Lagrange multiplier functions and duality in nonconvex programming. SIAM J. Control Optim. **12**(2), 268–285 (1974)
15. Lucidi, S.: New results on a continuously differentiable exact penalty function. SIAM J. Optim. **2**(4), 558–574 (1992)
16. Auslender, A., Teboulle, M.: Lagrangian duality and related multiplier methods for variational inequality problems. SIAM J. Optim. **10**(4), 1097–1115 (2000)
17. Bertsekas, D.P.: Constrained Optimization and Lagrange Multipliers Methods. Academic Press, New York (1982)
18. Di Pillo, G.: Exact penalty methods. In: Algorithms for Continuous Optimization: The State of the Art. NATO ASI Series, Mathematical and Physical Sciences, vol. 434, pp. 209–254. Kluwer Academic, Dordrecht (1993)
19. Di Pillo, G., Lucidi, S.: An augmented Lagrangian function with improved exactness properties. SIAM J. Optim. **12**(2), 376–406 (2001)
20. Di Pillo, G., Liuzzi, G., Lucidi, S., Palagi, L.: A truncated Newton method in an augmented Lagrangian framework for nonlinear programming. Comput. Optim. Appl. **45**(2), 311–352 (2010)
21. Qi, L., Sun, J.: A nonsmooth version of Newton's method. Math. Program. **58**(1–3), 353–367 (1993)
22. Facchinei, F., Kanzow, C., Palagi, L.: Personal communication (2007)
23. Facchinei, F., Pang, J.S.: Finite-Dimensional Variational Inequalities and Complementarity Problems, vol. II. Springer Series in Operations Research. Springer, New York (2003)
24. Nocedal, J., Wright, S.J.: Numerical Optimization, 1st edn. Springer, New York (1999)
25. Jian, J.B., Tang, C.M., Hu, Q.J., Zheng, H.Y.: A feasible descent SQP algorithm for general constrained optimization without strict complementarity. J. Comput. Appl. Math. **180**, 391–412 (2005)
26. Wright, S.J.: Modifying SQP for degenerate problems. SIAM J. Optim. **13**(2), 470–497 (2002)
27. Grippo, L., Lampariello, F., Lucidi, S.: A nonmonotone line search technique for Newton's method. SIAM J. Numer. Anal. **23**(4), 707–716 (1986)
28. Birgin, E.G., Martínez, J.M., Raydan, M.: Nonmonotone spectral projected gradient methods for convex sets. SIAM J. Optim. **10**(4), 1196–1211 (2000)
29. Andreani, R., Birgin, E.G., Martínez, J.M., Schuverdt, M.: On augmented Lagrangian methods with general lower-level constraints. SIAM J. Optim. **18**(4), 1286–1309 (2007)
30. Andreani, R., Birgin, E.G., Martínez, J.M., Schuverdt, M.: Augmented Lagrangian methods under the constant positive linear dependence constraint qualification. Math. Program. **111**(1–2), 5–32 (2008)
31. Bongartz, I., Conn, A.R., Gould, N.I.M., Toint, P.L.: CUTE: constrained and unconstrained testing environment. ACM Trans. Math. Softw. **21**, 123–160 (1995)
32. Vanderbei, R.J.: AMPL models. http://www.orfe.princeton.edu/~rvdb/ampl/nlmodels, Princeton University (2010)

33. Dolan, E.D., Moré, J.J.: Benchmarking optimization software with performance profiles. Math. Program. **91**(2), 201–213 (2002)
34. Bates, D., Hagström, M.: Nonlinear Analysis and Synthesis Techniques for Aircraft Control. Lecture Notes in Control and Information Sciences, vol. 365. Springer, Berlin (2007)
35. Biegler, L.T., Ghattas, O., Heinkenschloss, M., van Bloemen Waanders, B.: Large-Scale PDE-Constrained Optimization. Lecture Notes in Computational Science and Engineering, vol. 30. Springer, Berlin (2003)