

Fruitful uses of smooth exact merit functions in constrained optimization¹

Gianni Di Pillo (dipillo@dis.uniroma1.it)

Giampaolo Liuzzi (liuzzi@dis.uniroma1.it)

Stefano Lucidi (lucidi@dis.uniroma1.it)

Laura Palagi (palagi@dis.uniroma1.it)

*Dipartimento di Informatica e Sistemistica "A. Ruberti"
Università di Roma "La Sapienza"
Via Buonarroti 12, 00185 Roma, Italy*

Abstract

In this paper we are concerned with continuously differentiable exact merit functions as a mean to solve constrained optimization problems even of considerable dimension. In order to give a complete understanding of the fundamental properties of exact merit functions, we first review the development of smooth and exact merit functions. A recently proposed shifted barrier augmented Lagrangian function is then presented as a potentially powerful tool to solve large scale constrained optimization problems. This latter merit function, rather than directly minimized, can be more fruitfully used to globalize efficient local algorithms, thus obtaining methods suitable for large scale problems. Moreover, by carefully choosing the search directions and the linesearch strategy, it is possible to define algorithms which are superlinearly convergent towards points satisfying first and second order necessary optimality conditions. We propose a general scheme for an algorithm employing such a merit function.

Keywords: constrained optimization, continuously differentiable merit functions, primal-dual algorithms.

¹This work was supported by MIUR, National Research Program *Algorithms for Complex Systems Optimization*

1 Problem statement and notation

We consider the following minimization problem:

$$\begin{aligned} \min \quad & f(x) \\ & g(x) \leq 0, \end{aligned} \tag{1}$$

where $f : R^n \rightarrow R$, and $g : R^n \rightarrow R^m$ are twice continuously differentiable functions. We denote the feasible set of Problem (1) by

$$\mathcal{F} = \{x \in R^n : g(x) \leq 0\},$$

and the Lagrangian function by

$$L(x, \lambda) = f(x) + \lambda'g(x),$$

where $\lambda \in R^m$ is the KKT multiplier. A KKT pair $(\bar{x}, \bar{\lambda}) \in R^n \times R^m$ satisfies:

$$\nabla_x L(\bar{x}, \bar{\lambda}) = 0, \tag{2a}$$

$$G(\bar{x})\bar{\lambda} = 0, \quad \bar{\lambda} \geq 0, \quad g(\bar{x}) \leq 0, \tag{2b}$$

where $G(x) = \text{diag}_{i=1, \dots, m}(g_i(x))$.

Assumption A1 (LICQ) For every $x \in \mathcal{F}$ the gradients of the active constraints are linearly independent.

Under LICQ conditions (2) are first order necessary optimality conditions. In the following we assume that Assumption A1 holds.

Given a feasible point \bar{x} we denote the index set of the active constraints by

$$A_0(\bar{x}) = \{j : g_j(\bar{x}) = 0\};$$

moreover given a KKT pair $(\bar{x}, \bar{\lambda})$ we denote the index set of the strictly active constraints by

$$A_+(\bar{x}, \bar{\lambda}) = \{j \in A_0(\bar{x}) : \bar{\lambda}_j > 0\}.$$

A KKT pair $(\bar{x}, \bar{\lambda})$ satisfies the *strict complementarity condition* if $\bar{\lambda}_j > 0$ for all $j \in A_0(\bar{x})$.

A KKT pair $(\bar{x}, \bar{\lambda})$ satisfies *the second order necessary optimality conditions (SONC)* for Problem (1), if $(\bar{x}, \bar{\lambda})$ is a KKT pair and

$$z' \nabla_x^2 L(\bar{x}, \bar{\lambda}) z \geq 0, \quad \forall z : \nabla g_j(\bar{x})' z = 0 \quad j \in A_0(\bar{x}). \tag{3}$$

A KKT pair $(\bar{x}, \bar{\lambda})$ which satisfies SONC will be called in the sequel a *second order stationary point* of Problem (1).

The paper is organized as follows. In Section 2 we review the development of smooth exact merit functions. The emphasis on smoothness stems from the fact that

it makes possible to devise methods for the solution of smooth constrained problems that behave like unconstrained minimization methods for smooth objective functions, as concerns the convergence and rate of convergence properties, and in particular avoid the Maratos' effect [25]. In Section 3 we describe a shifted barrier augmented Lagrangian function L_G recently proposed in [11], that can be used as a powerful mean to solve large scale constrained optimization problems. In Section 4 we describe a general primal-dual algorithmic scheme PDALA* based on the merit function L_G . By carefully choosing the search direction and the linesearch strategy, it is possible to define algorithms which are superlinearly convergent towards points satisfying not only first but also second order necessary optimality conditions, as described in Sections 4.2 and 4.3. Furthermore, in Section 5 an "ad hoc" local and superlinearly convergent algorithm is described that is based on the satisfaction of the KKT conditions and that uses the merit function L_G within a general stabilization scheme.

We introduce the following notation. Given two vectors u, v , $\max\{u, v\}$ denotes the vector with components $\max\{u_i, v_i\}$. $\|v\|_p$ denotes the ℓ_p norm of the vector; when p is not specified, we mean $p = 2$.

2 Brief review on smooth exact merit functions

Much research work in Nonlinear Programming has been devoted to the *transformation* of the constrained problem (1) into the unconstrained minimization of a *merit function* depending on a penalty parameter ε .

We say that a merit function enjoys *exactness properties* if it is possible to establish some correspondence between its unconstrained minimizers and the solutions of the constrained Problem (1) for strictly positive values of the penalty parameter. See [7] for different definitions of "exactness" and [6] for an extended review of exact merit functions.

The constrained problem (1) is determined by the interaction of two distinct subproblems:

- the feasibility subproblem;
- the subproblem of minimizing the objective function.

A merit function should properly balance these two subproblems.

The initial idea for constructing merit functions consists in adding a weighted term $p(x)$, penalizing the violation of feasibility, to the original objective function. Such a term is of the form $p(x) = \|\max\{g(x), 0\}\|_s^s$ for some finite $s \geq 1$. Depending on the choice of the norm we obtain different merit functions that enjoy different properties. We are interested in continuously differentiable functions. Among these, the most popular is the quadratic penalty function:

$$P_S(x; \varepsilon) = f(x) + \frac{1}{\varepsilon} \|\max\{g(x), 0\}\|_2^2.$$

However P_S is not “exact”, since the sequence $\{x^k\}$ of unconstrained minimizers of $P_S(x; \varepsilon^k)$ converges to a minimizer x^* of the constrained problem only in the limit, for $\varepsilon^k \rightarrow 0$ [18].

Another merit function penalizing only the violation of feasibility is:

$$P_N(x; \varepsilon) = f(x) + \frac{1}{\varepsilon} |\max\{g(x), 0\}|.$$

In this case we gain exactness, however we lose differentiability.

The subsequent step is that of introducing merit functions which *better* characterize the connections between the feasibility subproblem and the minimizing subproblem. This means that further characteristics of the constrained minimum points, rather than just the feasibility, have to be taken into account.

A viable approach is to define terms which perform a penalization of the KKT conditions. To this aim an estimate of the KKT multiplier $\bar{\lambda}$ is required. A key role in the definition of such terms is played by the *non differentiable* function

$$\gamma(x, c\lambda) = \max\{g(x), -c\lambda\}, \quad (4)$$

where c is a positive constant. Indeed, it is easy to verify that:

Proposition 2.1 *Let $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^m$. Then $\max\{g(x), -c\lambda\} = 0$ if and only if $g(x) \leq 0$, $\lambda \geq 0$, $G(x)\lambda = 0$.*

Unfortunately, $\max\{g(x), -c\lambda\} = 0$ cannot be used directly to penalize the KKT conditions because it is not continuously differentiable. A possibility is to use the function

$$\phi(x, \lambda; c) = c\lambda' \gamma(x, c\lambda) + \frac{1}{2} \|\gamma(x, c\lambda)\|^2$$

which is continuously differentiable, with gradient given by

$$\begin{aligned} \nabla_x \phi(x, \lambda) &= c \nabla g(x) \lambda + \frac{1}{c} \nabla g(x) \gamma(x, c\lambda), \\ \nabla_\lambda \phi(x, \lambda) &= c \gamma(x, c\lambda). \end{aligned}$$

Moreover, if $\bar{x} \in \mathcal{F}$ then $\phi(\bar{x}, \lambda; c) = 0$ if and only if (\bar{x}, λ) satisfies conditions (2b) [10]. The first merit function based on the use of function ϕ was independently proposed by Hestenes-Powell-Rockafellar [23, 26, 28] and it is defined on the product space of the problem variables and of the KKT multipliers. Its expression is:

$$L_{HPR}(x, \lambda; \varepsilon) = f(x) + \frac{1}{\varepsilon} \phi(x, \lambda; \varepsilon).$$

L_{HPR} is continuously differentiable, in particular it is an SC^1 function, that is a continuous function with semismooth gradient [27]. Moreover it possesses some exactness properties. In particular, it is possible to prove that $\nabla L_{HPR}(\bar{x}, \bar{\lambda}; \varepsilon) = 0$ if and only if

$(\bar{x}, \bar{\lambda})$ is a KKT pair of Problem (1). Moreover if λ^* is the KKT multiplier associated to a global minimum point x^* of Problem (1), then $L_{HPR}(x, \lambda^*; \varepsilon)$ admits x^* as a global minimizer. However, the multiplier λ^* is not available and $L_{HPR}(x, \lambda; \varepsilon)$ is not bounded from below with respect to λ . This may cause unconstrained minimization methods to produce unbounded sequences (see [10] for a review on augmented Lagrangian functions).

To overcome this problem, it is necessary to define merit functions which convey more information about the KKT multipliers. Indeed, we have to penalize the distance $\|\lambda - \bar{\lambda}\|$, and this can be obtained by including a sort of penalty term on the condition $\nabla_x L(x, \lambda) = 0$. Different terms, which can be used to this aim, have been proposed (see [10, 11]). A possible example is the function

$$\eta(x, \lambda) = \|\nabla g(x)' \nabla_x L(x, \lambda) + G(x)^2 \lambda\|^2 = \|M(x)\lambda + \nabla g(x)' \nabla f\|^2, \quad (5)$$

where

$$M(x) = \nabla g(x)' \nabla g(x) + G^2(x). \quad (6)$$

Under Assumption A1 it is possible to prove that $\eta(\bar{x}, \lambda) = 0$ if and only if $\lambda = \bar{\lambda}$, where $(\bar{x}, \bar{\lambda})$ is a KKT pair of Problem (1). In particular, under Assumption A1, the matrix $M(x)$ is a positive definite matrix in a neighborhood of the feasible set [19].

However there are at least two ways to include this kind of information, that lead to two classes of continuously differentiable merit functions:

- *exact penalty functions* - in this case the KKT multipliers are obtained as a function of the variable x using a *multiplier function* $\lambda(x)$; exact penalty functions are defined on an open set \mathcal{P} , with $\mathcal{F} \subset \mathcal{P} \subseteq R^n$;
- *exact augmented Lagrangian functions* - in this case the KKT multipliers are variables on their own; exact augmented Lagrangian functions are defined on an open set $\mathcal{P} \times R^m$, again with $\mathcal{F} \subset \mathcal{P} \subseteq R^n$.

Continuously differentiable exact penalty functions. In this case the information on the KKT multiplier is conveyed in the merit function through the use of a *multiplier function* $\lambda(x)$ which yields an estimate of the KKT multiplier as function of the variable x , in the sense that $\bar{\lambda} = \lambda(\bar{x})$ when $(\bar{x}, \bar{\lambda})$ is a KKT pair. A typical form of exact penalty functions is:

$$P_E(x; \varepsilon) = L_{HPR}(x, \lambda(x); \varepsilon).$$

Expressions for $\lambda(x)$ can be obtained as the minimizer of a suitable term which penalizes the KKT conditions. A possible example of $\lambda(x)$ consists in obtaining it as the minimizer with respect to λ of the function $\eta(x, \lambda)$ defined by (5). In this case, whenever $M(x)$ is non singular, we get

$$\lambda(x) = -M(x)^{-1} \nabla g(x)' \nabla f(x).$$

The function $\lambda(x)$ is continuously differentiable and, if $(\bar{x}, \bar{\lambda})$ is a KKT pair, then $\lambda(\bar{x}) = \bar{\lambda}$.

Unfortunately the computation of $\lambda(x)$ is quite expensive when the number of constraints is large, because it requires the exact solution of a linear system at each merit function evaluation. This limits the applicability of continuously differentiable exact penalty functions when the number of constraints is large.

We do not enter into more details on exact penalty functions.

Exact augmented Lagrangian functions. In this case the multiplier λ is a variable on its own, and we must add terms penalizing the distance from the KKT multiplier $\bar{\lambda}$ to the augmented Lagrangian L_{HPR} . The typical form of functions in this class is

$$L_E(x, \lambda; \varepsilon) = L_{HPR}(x, \lambda; \varepsilon) + \eta(x, \lambda),$$

where the term $\eta(x, \lambda)$ can also be seen as a term that performs a “convexification” of the augmented Lagrangian function with respect to the variable λ .

For large problems exact augmented Lagrangian functions are preferable with respect to exact penalty functions, because they do not require the solution of a linear system at each merit function evaluation. However $L_E(x, \lambda; \varepsilon)$ may have unbounded level sets. Hence the last step is to define an augmented Lagrangian function with compact level sets. We describe a merit function with this property in the following section.

3 An exact augmented Lagrangian function with improved exactness properties

Following the ideas of barrier methods, in [11] a shifted barrier term has been introduced in the augmented Lagrangian $L_E(x, \lambda; \varepsilon)$. To be more precise, let \mathcal{P} be a suitable *open perturbation* of the feasible set \mathcal{F} (that is $\mathcal{P} \supset \mathcal{F}$) and consider a function $p(x, \lambda)$ such that:

$$\begin{aligned} p(x, \lambda) &> 0 \quad \forall (x, \lambda) \in \mathcal{P} \times \mathbb{R}^m, \\ \lim_{x \rightarrow \partial \mathcal{P}} p(x, \lambda) &= 0, \quad \lim_{\|\lambda\| \rightarrow \infty} p(x, \lambda) = 0. \end{aligned} \tag{7}$$

It appears that $1/p(x, \lambda)$ penalizes the fact that the variable x is too close to the boundary of \mathcal{P} and the fact that the norm of the vector λ is growing too much. Indeed $1/p(x, \lambda)$ tends to ∞ whenever $x \rightarrow \partial \mathcal{P}$ or $\|\lambda\| \rightarrow \infty$.

By including this barrier term in the augmented Lagrangian we get

$$L_G(x, \lambda; \varepsilon) = L_{HPR}(x, \lambda; \varepsilon p(x, \lambda)) + \eta(x, \lambda),$$

that can be written explicitly as follows:

$$L_G(x, \lambda; \varepsilon) = f(x) + \lambda' \max\{g(x), -\varepsilon p(x, \lambda)\lambda\} + \frac{1}{2\varepsilon p(x, \lambda)} \|\max\{g(x), -\varepsilon p(x, \lambda)\lambda\}\|^2 + \|\nabla g(x)' \nabla_x L(x, \lambda) + G(x)^2 \lambda\|^2.$$

The exactness properties of L_G can be established under suitable assumptions that depend on the particular choice of $p(x, \lambda)$.

A possible choice is to define

$$p(x, \lambda) = \frac{a(x)}{1 + \|\lambda\|^2},$$

where

$$a(x) = \alpha - \|\max\{g(x), 0\}\|_s^s, \quad (8)$$

and $\alpha > 0$, $s \geq 2$ are user-selected scalars. Correspondingly, the set \mathcal{P} is defined by:

$$\mathcal{P} = \{x \in \mathbb{R}^n : a(x) > 0\}.$$

With this choice of \mathcal{P} , we can state exactness properties of L_G under the following assumption.

Assumption A2 *One of the two following conditions is satisfied:*

- a point $\hat{x} \in \mathcal{F}$ is known and $f(x)$ is coercive on $\bar{\mathcal{P}}$ (that is for any $\{x_k\} \subseteq \mathcal{P}$ such that $\|x_k\| \rightarrow \infty$ we have $f(x_k) \rightarrow \infty$);
- $\bar{\mathcal{P}}$ is a bounded set and at every point $x \in \mathcal{P}$ the following constraint qualification holds:

$$\sum_{i: g_i(x) > 0} c_i(x) \nabla g_i(x) = 0 \quad \implies \quad x \in \mathcal{F}, \quad (9)$$

where

$$c_i(x) = \left[1 + \frac{s \|\max\{g(x), 0\}\|_s^2 g_i(x)^{(s-2)}}{a(x)} \right] g_i(x). \quad (10)$$

Assumption A2 has been widely discuss in [11]. In the sequel we assume that Assumption A2 holds.

We can show that:

Proposition 3.1 (Exactness properties of L_G) [11]

- (i) $\forall \varepsilon > 0$, the level set

$$\Omega(x^0, \lambda^0; \varepsilon) = \{(x, \lambda) \in \mathcal{P} \times \mathbb{R}^m : L_G(x, \lambda; \varepsilon) \leq L_G(x^0, \lambda^0; \varepsilon)\}$$

is compact, and hence $L_G(x, \lambda; \varepsilon)$ admits a global minimizer;

- (ii) $\forall \varepsilon > 0$, if $(\bar{x}, \bar{\lambda})$ is a KKT pair of Problem (1) then $(\bar{x}, \bar{\lambda})$ is a stationary point of L_G ;
- (iii) $\forall \varepsilon > 0$, if $(\bar{x}, \bar{\lambda})$ is a stationary point of L_G such that $\gamma(\bar{x}, \varepsilon p(\bar{x}, \bar{\lambda})\bar{\lambda}) = 0$ then $(\bar{x}, \bar{\lambda})$ is a KKT pair of Problem (1);
- (iv) a threshold value $\bar{\varepsilon} > 0$ exists, such that for all $\varepsilon \in (0, \bar{\varepsilon}]$, if $(\bar{x}, \bar{\lambda})$ is a stationary point (local/global minimizer) of L_G then $(\bar{x}, \bar{\lambda})$ is a KKT pair (local/global minimizer) of Problem (1).

Proposition 3.1 states that suitable values of ε can be found such that

$$\min_{\substack{f(x) \\ g(x) \leq 0}} \quad \iff \quad \min_{(x, \lambda) \in \mathcal{P} \times \mathbb{R}^m} L_G(x, \lambda; \varepsilon)$$

Hence the merit function L_G can be used to define algorithms converging to a solution of the constrained problem. We discuss this topic in the next sections.

Under the assumption that f and g are three times continuously differentiable, and that $s > 2$ in (8), it is possible to perform a second order analysis of L_G . Since it is a SC^1 function in $\mathcal{P} \times \mathbb{R}^m$, the generalized Hessian $\partial^2 L_G(x, \lambda; \varepsilon)$, in Clarke's sense [3], can be defined. In particular we have $\partial^2 L_G(x, \lambda; \varepsilon) = \text{co}\{\partial_B^2 L_G(x, \lambda; \varepsilon)\}$ where the set $\partial_B^2 L_G(x, \lambda; \varepsilon)$ can be explicitly characterized in a neighborhood of a KKT pair $(\bar{x}, \bar{\lambda})$ of Problem (1).

Given a partition $\{A, N\}$ of the index set $\{1, \dots, m\}$, and by partitioning the vectors g and λ accordingly: $g = (g'_A \ g'_N)'$, $\lambda = (\lambda'_A \ \lambda'_N)'$, we introduce the $(n+m) \times (n+m)$ symmetric matrix $H(x, \lambda; \varepsilon, A)$ given block-wise by:

$$\begin{aligned} H_{xx}(x, \lambda; \varepsilon, A) &= \nabla_x^2 L(x, \lambda) + \frac{1}{\varepsilon p(x, \lambda)} \nabla g_A(x) \nabla g_A(x)' \\ &\quad + 2 \nabla_x^2 L(x, \lambda) \nabla g(x) \nabla g(x)' \nabla_x^2 L(x, \lambda), \\ H_{x\lambda}(x, \lambda; \varepsilon, A) &= \begin{bmatrix} \nabla g_A(x) & 0 \end{bmatrix} + 2 \nabla_x^2 L(x, \lambda) \nabla g(x) M_N(x), \\ H_{\lambda\lambda}(x, \lambda; \varepsilon, A) &= -\varepsilon p(x, \lambda) \begin{bmatrix} 0 & 0 \\ 0 & I_N \end{bmatrix} + 2 M_N(x) M_N(x), \end{aligned} \tag{11}$$

where

$$M_N(x) = \nabla g(x)' \nabla g(x) + \begin{pmatrix} 0 & 0 \\ 0 & G_N(x)^2 \end{pmatrix} \tag{12}$$

and I_N is the identity matrix of dimension $|N|$ and 0 is a zero matrix of proper dimensions.

Proposition 3.2 [11] *For every KKT pair $(\bar{x}, \bar{\lambda})$ of Problem (1) and every given ε , a neighbourhood \mathcal{B} of $(\bar{x}, \bar{\lambda})$ exists such that, for all (x, λ) in \mathcal{B} , we have:*

$$\partial_B^2 L_G(x, \lambda; \varepsilon) = \{H(x, \lambda; \varepsilon, A) + K(x, \lambda; \varepsilon, A) : A \in \mathcal{A}\},$$

where $\mathcal{A} = \{A : A_+(\bar{x}, \bar{\lambda}) \subseteq A \subseteq A_0(\bar{x})\}$, $H(x, \lambda; \varepsilon, A)$ is given by (11), and $K(x, \lambda; \varepsilon, A)$ is a matrix such that $\|K(x, \lambda; \varepsilon, A)\| \leq \zeta(x, \lambda)$, with $\zeta(x, \lambda)$ a non-negative continuous function such that $\zeta(\bar{x}, \bar{\lambda}) = 0$.

We note that at every KKT pair where the strict complementarity holds, $A_+(\bar{x}, \bar{\lambda}) = A_0(\bar{x})$, hence $\partial^2 L_G(\bar{x}, \bar{\lambda}; \varepsilon)$ reduces to a singleton. Therefore for every KKT pair $(\bar{x}, \bar{\lambda})$ of Problem (1) where the strict complementarity holds, a neighbourhood \mathcal{B} of $(\bar{x}, \bar{\lambda})$ exists such that L_G is twice continuously differentiable in \mathcal{B} and its Hessian matrix is given by:

$$\nabla^2 L_G(x, \lambda; \varepsilon) = H(x, \lambda; \varepsilon, A_0(\bar{x})) + K(x, \lambda; \varepsilon, A_0(\bar{x})).$$

We can state relationships among points satisfying the SONC for Problem (1) and some elements of the generalized Hessian of L_G .

Proposition 3.3 [15] *Let $(\bar{x}, \bar{\lambda})$ be a KKT pair of Problem (1) and let $\varepsilon > 0$ be given. If a positive semidefinite matrix $H \in \partial_B^2 L_G(\bar{x}, \bar{\lambda}; \varepsilon)$ exists, then the pair $(\bar{x}, \bar{\lambda})$ satisfies the second order necessary optimality conditions SONC.*

Under stronger second order assumptions, we can state additional second order results. In particular, we need the following condition.

Assumption A3 (SSOSC) *A KKT pair $(\bar{x}, \bar{\lambda})$ satisfies the Strong Second Order Sufficient Conditions (SSOSC) for \bar{x} to be an isolated local solution of Problem (1), that is*

$$z' \nabla_x^2 L(\bar{x}, \bar{\lambda}) z > 0, \quad z \neq 0, \quad z \in Z(\bar{x}),$$

where $Z(\bar{x}) = \{z : \nabla g_j(\bar{x})' z = 0, \quad j \in A_+(\bar{x})\}$.

Proposition 3.4 [11] *If a KKT pair $(\bar{x}, \bar{\lambda})$ satisfies the Assumptions A1 and A3 then a value $\bar{\varepsilon} > 0$ exists such that for all $\varepsilon \in (0, \bar{\varepsilon}]$ we have:*

- (i) $(\bar{x}, \bar{\lambda})$ is an isolated local minimum point for L_G ;
- (ii) all the matrices in $\partial^2 L_G(\bar{x}, \bar{\lambda}; \varepsilon)$ are positive definite.

Finally, we want to mention that recently a new class of augmented Lagrangian function has been proposed for the case of two-sided constraints of the type $l \leq g(x) \leq u$ [8]. The definition of this function is based on a different way to write the KKT conditions that uses only m multipliers not constrained in sign (see [1]). In particular, it is possible to define a new function $\gamma(x, c\lambda)$:

$$\gamma_{\text{new}}(x, c\lambda) = \min\{g - l, -c\lambda\} + \max\{g - u, -c\lambda\} + c\lambda$$

with $c > 0$, that possess similar properties of (4). Moreover in [8] also a general framework that includes different type of barrier terms has been proposed.

4 A shifted barrier primal-dual algorithm

4.1 The basic algorithm model

Exact merit functions can be used in different ways to define globally convergent algorithms [16, 6]. However in order to obtain a true implementable algorithm, we must provide rules to determine the threshold value $\bar{\varepsilon}$ or a lower bound on it. Since we are interested in the use of the augmented Lagrangian function L_G , we consider primal-dual algorithm models, that is algorithms defined in the space of the variables (x, λ) .

We give the description of a general primal-dual algorithm model (PDALA_{*}) based on the augmented Lagrangian function L_G . PDALA_{*} incorporates an automatic adjustment of the penalty parameter and, under suitable assumptions, can be proved to be globally convergent to a KKT pair of the original constrained problem.

In the algorithm model we make use of an iteration map $LSA_* : \mathcal{P} \times \mathbb{R}^m \rightarrow \mathcal{P} \times \mathbb{R}^m$ and we denote by $LSA_*[(x^k, \lambda^k)]$, the new point produced by it. Here, we only assume that the iteration map LSA_* satisfies the following assumption:

Assumption A4 *For every fixed value ε and every starting point $(x^0, \lambda^0) \in \mathcal{P} \times \mathbb{R}^m$, the sequence $\{(x^k, \lambda^k)\}$ obtained as $(x^{k+1}, \lambda^{k+1}) = LSA_*[(x^k, \lambda^k)]$ belongs to the level set $\Omega(x^0, \lambda^0; \varepsilon)$, and all its limit points are stationary points of $L_G(x, \lambda; \varepsilon)$.*

These requirements on the map LSA_* can be easily satisfied by any globally convergent algorithm for the unconstrained minimization of $L_G(x, \lambda; \varepsilon)$ for fixed values of ε . In fact we can always ensure, by simple devices, that the trial points produced by the iteration map remain in $\Omega(x^0, \lambda^0; \varepsilon)$.

Different choices of the iteration map LSA_* determine different features of the overall algorithm, such as convergence towards points satisfying first or first and second order necessary optimality conditions, rate of convergence, heaviness of the computation, and so on.

Later in this section, we describe possible choices of the iteration map LSA_* in the class of linesearch methods.

Algorithm PDALA_{*} performs an outer iteration and an inner iteration. The outer iteration, indexed by j , monitors the decrease of the penalty parameter and provides a proper starting point (x^0, λ^0) for the inner iteration. In particular, the outer iteration produces the sequences $\{\varepsilon^j\} \subset \mathbb{R}^+$ and $\{(y^j, \mu^j)\} \subseteq \mathcal{P} \times \mathbb{R}^m$. The inner iteration, indexed by k , produces, for a fixed ε^j , a sequence $\{(x^k, \lambda^k)\}_j \subseteq \Omega(x^0, \lambda^0; \varepsilon^j)$ using LSA_* until some convergence criterion is satisfied.

Algorithm model PDALA*

Primal-Dual Augmented Lagrangian Algorithm (general scheme)

Data. $(y^0, \mu^0) \in \mathbb{R}^n \times \mathbb{R}^m$ and $\varepsilon^0 > 0$, $\theta \in (0, 1)$.

Choose the scalars $\alpha > 0$ and $s \geq 3$ appearing in (8) so that $y^0 \in \mathcal{P}$.

Step 0. Set $j = 0$ and $(x^0, \lambda^0) = (y^0, \mu^0)$ (outer iteration).

Step 1. Set $k = 0$ (inner iteration).

Repeat

If $\|\nabla L_G(x^k, \lambda^k; \varepsilon^j)\| \geq \|\max\{g(x^k), -\varepsilon^j p(x^k, \lambda^k)\lambda^k\}\|$ **then**

compute $(x^{k+1}, \lambda^{k+1}) = \text{LSA}_*[(x^k, \lambda^k)]$

set $k = k + 1$;

else (update ε and restart the inner iteration)

Set $\varepsilon^{j+1} = \theta \varepsilon^j$, $(y^{j+1}, \mu^{j+1}) = (x^k, \lambda^k)$, $j = j + 1$.

If $L_G(y^j, \mu^j; \varepsilon^j) \leq L_G(y^0, \mu^0; \varepsilon^j)$ set $(x^0, \lambda^0) = (y^j, \mu^j)$,

else set $(x^0, \lambda^0) = (y^0, \mu^0)$,

go to Step 1.

End If

Until (convergence)

The general scheme above can be refined by specifying both the stopping criterion for convergence and the iteration map LSA_* . In particular, we can enforce convergence to points satisfying only the KKT conditions or also the second order necessary optimality conditions for optimality.

4.2 Convergence to first order stationary points: PDALA₁

We consider a version of the algorithm model PDALA* converging to points satisfying the KKT conditions (2). In this case the convergence criterion used in the last line of PDALA* corresponds to the satisfaction of the KKT conditions (2) that, due to (iii) of Proposition 3.1, can be written as:

$$\|\nabla L_G(x^k, \lambda^k; \varepsilon^j)\| + \|\max\{g(x^k), -\varepsilon^j p(x^k, \lambda^k)\lambda^k\}\| = 0.$$

The iteration map can be an Armijo linesearch procedure for the unconstrained minimization of L_G , and we refer to it as LSA_1 . The resulting algorithm PDALA_1 was described in [11] with the name ALFA.

We assume that the direction $d^k = (d_x^k, d_\lambda^k)' \in \mathbb{R}^n \times \mathbb{R}^m$ used in LSA_1 satisfies:

Assumption A5 For all k , the direction $d^k \in \mathbb{R}^n \times \mathbb{R}^m$ is bounded and satisfies

- (a) $\nabla L_G(x^k, \lambda^k; \varepsilon)' d^k < 0$ if $\nabla L_G(x^k, \lambda^k; \varepsilon) \neq 0$;
- (b) $d^k = 0$ if $\nabla L_G(x^k, \lambda^k; \varepsilon) = 0$;
- (c) $\nabla L_G(x^k, \lambda^k; \varepsilon)' d^k \rightarrow 0$ implies $\begin{cases} \nabla L_G(x^k, \lambda^k; \varepsilon) \rightarrow 0 \\ d^k \rightarrow 0. \end{cases}$

Iteration map $\text{LSA}_1[(x^k, \lambda^k)]$

Data: $\gamma \in (0, \frac{1}{2})$, $\sigma \in (0, 1)$.

Step 1. Calculate a $d^k \in \mathbb{R}^{n+m}$ satisfying Assumption A5.

Step 2. Set $\eta \in (0, 1]$, $i = 0$.

Repeat

$$\begin{aligned} \eta &= \sigma^i \eta, \\ \begin{pmatrix} x^{k+1} \\ \lambda^{k+1} \end{pmatrix} &= \begin{pmatrix} x^k \\ \lambda^k \end{pmatrix} + \eta \begin{pmatrix} d_x^k \\ d_\lambda^k \end{pmatrix} \in \mathcal{P} \times \mathbb{R}^m, \\ i &= i + 1. \end{aligned}$$

Until $(L_G(x^{k+1}, \lambda^{k+1}; \varepsilon) \leq L_G(x^k, \lambda^k; \varepsilon) + \gamma \eta \nabla L_G(x^k, \lambda^k; \varepsilon)' d^k)$

Return (x^{k+1}, λ^{k+1}) .

Proposition 4.1 (Global convergence of PDALA_1) [11] *Assume that the direction d^k satisfies Assumption A5. Then, either the algorithm terminates at a KKT pair (x^p, λ^p) of Problem (1) or, after having updated the penalty parameter ε at most a finite number of times, it produces an infinite sequence $\{(x^k, \lambda^k)\}$ such that every limit point (x^*, λ^*) of $\{(x^k, \lambda^k)\}$ is a KKT pair of Problem (1).*

The convergence properties of Algorithm PDALA_1 depend on the satisfaction of Assumptions A1, A2. However, weaker convergence results can be established also in the case that the assumptions are not fulfilled (for details, see [11]).

As regards the choice of d^k we indicate some possibilities in Section 5.

4.3 Convergence to second order stationary points: PDALA₂

We describe now a primal-dual augmented Lagrangian algorithm model, proposed in [15] with the name SOLA, converging to points satisfying the second order necessary optimality conditions (3). The possibility of defining such an algorithm is tied to the result of Proposition 3.3. Indeed, the “curvature” of L_G can be put in correspondence with the “curvature information” of the constrained Problem (1). This correspondence can be exploited to define algorithms converging towards points which satisfy the second order necessary optimality conditions SONC.

The algorithm model PDALA₂ is obtained by PDALA_{*} with the use of a *curvilinear line search* algorithm LSA₂. The curvilinear line search algorithm LSA₂ draw its inspirations from unconstrained minimization and makes use of an additional direction s^k and of a square matrix Q^k . We assume that s^k, Q^k satisfy:

Assumption A6

(i) For all k , the direction s^k and the matrix Q^k are bounded and satisfy:

- (a) $\nabla L_G(x^k, \lambda^k; \varepsilon)' s^k \leq 0$,
- (b) $(s^k)' Q^k s^k \leq 0$,
- (c) $(s^k)' Q^k s^k \rightarrow 0$ implies $s^k \rightarrow 0$;

(ii) let $\{x^k, \lambda^k\}$ and $\{z^k, \nu^k\}$ be sequences converging to a stationary point $(\bar{x}, \bar{\lambda})$ of the function L_G . Then, for every sequence of matrices $\{W^k\}$, with $W^k \in \partial^2 L_G(z^k, \nu^k; \varepsilon)$, we have:

$$(s^k)' (W^k - Q^k) s^k \leq \delta^k,$$

where $\{\delta^k\}$ is a sequence of numbers converging to 0;

(iii) let $\{x^k, \lambda^k\}$ be a sequence converging to a KKT pair $(\bar{x}, \bar{\lambda})$ of Problem (1). Then the directions s^k and the matrices Q^k are such that if $(s^k)' Q^k s^k \rightarrow 0$ then $(\bar{x}, \bar{\lambda})$ satisfies the second order necessary optimality conditions for Problem (1).

There are different possible choices for s^k, Q^k satisfying Assumption A6. One of them is described in Section 5.3.

In the case of PDALA₂, the convergence criterion in the last line of the algorithm model is given by the satisfaction of first and second order necessary optimality conditions. Due to Proposition 3.3 and Assumption A6 we can write

$$(\|\nabla L_G(x^k, \lambda^k; \varepsilon^j)\| + \|\max\{g(x^k), -\varepsilon^j p(x^k, \lambda^k)\lambda^k\}\| = 0) \text{ .and. } (Q^k \succeq 0),$$

where $Q^k \succeq 0$ stands for Q^k positive semidefinite.

Any gradient related direction d^k and any pair s^k, Q^k satisfying Assumption A6 can be used to define a curvilinear linesearch LSA_2 to be employed in the second order convergent algorithm model PDALA_2 .

Iteration map $\text{LSA}_2[(x^k, \lambda^k)]$

Data. $\gamma \in (0, \frac{1}{2}), \sigma \in (0, 1)$.

Step 1. Calculate $d^k \in \mathbb{R}^{n+m}$ satisfying Assumption A5.
Calculate $s^k \in \mathbb{R}^{n+m}$, and $Q^k \in \mathbb{R}^{(n+m) \times (n+m)}$ satisfying Assumption A6.

Step 2. Set $\eta \in (0, 1], i = 0$ and $t^k = 1 + \min\{0.5, \|\nabla L_G(x^k, \lambda^k; \varepsilon)\|\}$.

Repeat

$$\eta = \sigma^i \eta,$$

$$\begin{pmatrix} x^{k+1} \\ \lambda^{k+1} \end{pmatrix} = \begin{pmatrix} x^k \\ \lambda^k \end{pmatrix} + \eta^2 \begin{pmatrix} d_x^k \\ d_\lambda^k \end{pmatrix} + \eta^{t^k} \begin{pmatrix} s_x^k \\ s_\lambda^k \end{pmatrix},$$

$$i = i + 1.$$

Until $(L_G(x^{k+1}, \lambda^{k+1}; \varepsilon) \leq L_G(x^k, \lambda^k; \varepsilon) + \gamma \eta^2 \nabla L_G(x^k, \lambda^k; \varepsilon)' d^k + \frac{\gamma}{2} \eta^{2t^k} (s^k)' Q^k s^k)$

Return (x^{k+1}, λ^{k+1}) .

We note that the previous curvilinear search differs from similar unconstrained curvilinear search algorithms for the matrix Q^k and the term t^k . The matrix Q^k plays a role similar to the role played by the Hessian matrix in unconstrained algorithms and it must be chosen so as to provide, roughly speaking, some kind of second order information on the original constrained problem to the algorithm (see [15]). As regards the term t^k , in the unconstrained curvilinear search algorithms it is possible to set $t^k = 1$ since the Hessian matrix of the objective function is available; in our case $t^k = 1 + \min\{0.5, \|\nabla L_G(x^k, \lambda^k; \varepsilon)\|\}$ since L_G is only SC^1 .

We observe that, if we take $s^k = 0$, LSA_2 reduces to LSA_1 thus satisfying at least the properties of LSA_1 .

The following result holds.

Proposition 4.2 (Global convergence of PDALA_2) [15] *Assume that d^k satisfies Assumption A5 and s^k, Q^k satisfy Assumption A6. Then, either the algorithm terminates at a second order stationary pair (x^p, λ^p) of Problem (1) or, after having*

updated the penalty parameter ε at most a finite number of times, it produces an infinite sequence $\{(x^k, \lambda^k)\}$ such that every limit point (x^*, λ^*) of $\{(x^k, \lambda^k)\}$ is a second order stationary pair of Problem (1).

4.4 Superlinear rate of convergence

In the algorithm model PDALA_{*} we make use of iteration maps LSA₁ or LSA₂ that use a direction d^k to approach either a first or a second order stationary point of the function L_G .

Assumption A5 requires only some gradient related properties of the direction d^k . The simplest choice is to consider the steepest descent direction $d^k = -\nabla L_G$. Although this choice is quite simple and not expensive, it may result in a very slow algorithm, in the presence of narrow valley of the level sets of L_G .

It is well known that a proper choice for the search direction d^k guarantees a superlinear convergence rate in the unconstrained minimization of L_G . However, the use of the Newton direction is not suitable for the function L_G due to the fact that it is not twice continuously differentiable everywhere in $\mathcal{P} \times \mathbb{R}^m$, and to the fact that the evaluation of the Hessian matrix $\nabla^2 L_G$, where it exists, requires the evaluation of the third order derivatives of the problem functions f and g .

A possible way to overcome these drawbacks is to change the point of view. Rather than minimizing directly the augmented Lagrangian function L_G , we can use it as a merit function to measure the progress towards a solution of the iterates produced by an “ad hoc” local algorithm for Problem (1).

The basic idea is to define an efficient local algorithm, where “efficient” stands both for superlinearly convergent and for not computationally heavy. Indeed, it is possible to define directions d^k which can be used in LSA_{*} to produce sequences $\{(x^k, \lambda^k)\}$ which are locally convergent with a superlinear rate of convergence without requiring the evaluation of third order derivatives, so that the computation of d^k may be not very expensive.

However the sequence $\{(x^k, \lambda^k)\}$ produced by the local algorithm may be not globally convergent. Hence a stabilization scheme must be used to force global convergence. The merit function L_G can play different roles in the stabilization scheme. It can be used as a merit function to measure the progress of the iterates in a stabilization scheme for globalizing the local algorithm. In this way it ensures also the boundedness of $\{x^k, \lambda^k\}$, since the iterates remains in the level sets that are compact (see Proposition 3.1). Moreover the merit function provides low-cost alternative directions, such as $d^k = -\nabla L_G(x^k, \lambda^k)$, when required, and incorporates enough information on the “curvature” of Problem (1).

In order to define an efficient algorithm we make use, in the sequel, of the following assumption.

Assumption A7 For all k , the direction $d^k \in \mathbb{R}^{n+m}$ satisfies a system of the kind:

$$\tilde{H}(x^k, \lambda^k; \varepsilon)d = -\nabla L_G(x^k, \lambda^k; \varepsilon), \quad (13)$$

where the matrix $\tilde{H}(x^k, \lambda^k; \varepsilon)$ has the property that, if the sequence $\{(x^k, \lambda^k)\}$ converges to a KKT pair $(\bar{x}, \bar{\lambda})$ for Problem (1), then

$$\lim_{k \rightarrow \infty} \text{dist}[\tilde{H}(x^k, \lambda^k; \varepsilon) | \partial_B^2 L_G(x^k, \lambda^k; \varepsilon)] = 0, \quad (14)$$

where $\text{dist}[\cdot | \cdot]$ denotes the distance function.

Search directions which satisfy Assumption A7 play a fundamental role in defining efficient local algorithm.

Indeed we can prove superlinear convergence in a neighborhood of a KKT pair satisfying some sufficient optimality conditions. Making use of the Strong Second order Sufficient Condition (Assumption A3) a strict connection between directions satisfying Assumption A7, and the merit function L_G can be proved. In particular, the following result holds.

Proposition 4.3 (Superlinear convergence rate) [11] *Let $(\bar{x}, \bar{\lambda})$ be a KKT pair where Assumption A3 holds. If $\{d^k\}$ satisfy Assumption A7, then an $\bar{\varepsilon}$ exists such that for all $\varepsilon \in (0, \bar{\varepsilon}]$:*

(a) *a neighborhood $\mathcal{B}(\bar{x}, \bar{\lambda})$ of $(\bar{x}, \bar{\lambda})$ exists such that, for all $(x^k, \lambda^k) \in \mathcal{B}(\bar{x}, \bar{\lambda})$:*

- the search direction d^k satisfies the conditions:

$$\begin{aligned} \nabla L_G(x^k, \lambda^k; \varepsilon)' d^k &\leq -c \|\nabla L_G(x^k, \lambda^k; \varepsilon)\|^2, \\ c \|d^k\| &\leq \|\nabla L_G(x^k, \lambda^k; \varepsilon)\|, \end{aligned}$$

where c is a positive constant;

- an Armijo-type linesearch accepts the unit stepsize;

(b) *if the sequence $\{(x^k, \lambda^k)\}$ obtained by LSA_* with $\eta = 1$ converges to $(\bar{x}, \bar{\lambda})$, then the rate of convergence is superlinear.*

In [11, 15] directions d^k satisfying Assumption A7 have been proposed. In section 5.2 we describe in some detail one of these choices.

4.5 A nonmonotone stabilization scheme

The use of nonmonotone stabilization techniques is *suitable* to minimize ill-conditioned functions [20, 21]. This is even more the case for the minimization of exact merit functions, that can easily have narrow curved valleys. The advantage of the use of a nonmonotone stabilization scheme to control the convergence of the local algorithm consists in the fact that it resorts to the merit function and its gradient only when the local algorithm does not satisfy some easily tested convergence criterion. In this way, the merit function and its gradient are not calculated at each iteration.

A very simple nonmonotone stabilization scheme can be obtained from PDALA* by using the nonmonotone linesearch algorithm LSA_{NM} described below.

Iteration map LSA_{NM}[(x^k, λ^k)]

Data: $\gamma \in (0, \frac{1}{2}), \sigma \in (0, 1), R^k \in \mathbb{R}$.

Step 1. Calculate a $d^k \in \mathbb{R}^{n+m}$ satisfying Assumption A5.

Step 2. Set $\eta = 1, i = 0$.

Repeat

$$\eta = \sigma^i \eta,$$

$$\begin{pmatrix} x^{k+1} \\ \lambda^{k+1} \end{pmatrix} = \begin{pmatrix} x^k \\ \lambda^k \end{pmatrix} + \eta \begin{pmatrix} d_x^k \\ d_\lambda^k \end{pmatrix} \in \mathcal{P} \times \mathbb{R}^m,$$

$$i = i + 1.$$

Until ($L_G(x^{k+1}, \lambda^{k+1}; \varepsilon) \leq R^k + \gamma \eta \nabla L_G(x^k, \lambda^k; \varepsilon)' d^k$)

Return (x^{k+1}, λ^{k+1}) .

The main difference with LSA₁ consists in the presence of a reference value R^k that in general is such that $R^k \geq L_G(x^k, \lambda^k; \varepsilon)$, so that the condition of sufficient reduction is not enforced at each iteration. The updating rule for the reference value must satisfy

$$L_G(x^k, \lambda^k; \varepsilon) \leq R^k \leq \max_{0 \leq i \leq \bar{m}} L_G(x^{k-i}, \lambda^{k-i}; \varepsilon),$$

where \bar{m} is a prefixed number of iterations to be taken into account.

Drawing inspiration from unconstrained minimization, it is also possible to define nonmonotone linesearch algorithm converging to second order stationary points [24]. More refined nonmonotone strategies, that do not even require the evaluation of the merit function at each step have been proposed [22, 24]. We do not enter into details here.

5 “Ad hoc” algorithm models

5.1 Preliminaries

In this section we propose a particular choice of the directions d^k and s^k and of the matrix Q^k that can be used in the definition of a globally and superlinearly convergent algorithm to second order stationary points of Problem (1). Of course different possibilities exist, and we refer to [11, 15] and references therein for more detailed discussions. In the preceding sections we have described a superlinearly convergent algorithm to second order stationary points without requiring the strict complementarity condition. Our aim in this section is to exploit the structure of the problem in order to reduce the computational effort in the calculation of d^k and s^k . In principle, no relationships among the two directions is needed. However, we will show that if the strict complementarity condition holds, we can define a direction s^k which can be computed as a by product of the computation of the direction d^k .

To be more precise, we define a direction d^k which is tied to the structure of Problem (1) rather than to the function L_G . To this aim, we make use of the following estimates of the index sets of active and non active constraints:

$$A_{\oplus}(x, \lambda) = \{i : g_i(x) \geq -\kappa\lambda_i\}, \quad N_{\oplus}(x, \lambda) = \{1, \dots, m\} \setminus A_{\oplus}(x, \lambda), \quad (15)$$

with $\kappa > 0$. These sets have interesting properties, as the following proposition states.

Proposition 5.1 [17] *Let $(\bar{x}, \bar{\lambda})$ be a KKT pair for Problem (1). Then there exists a neighborhood $\mathcal{B}(\bar{x}, \bar{\lambda})$ of $(\bar{x}, \bar{\lambda})$ such that for all $(x, \lambda) \in \mathcal{B}(\bar{x}, \bar{\lambda})$ we have $A_+(\bar{x}, \bar{\lambda}) \subseteq A_{\oplus}(x, \lambda) \subseteq A_0(\bar{x})$. Moreover if the strict complementarity holds at $(\bar{x}, \bar{\lambda})$, then, for all $(x, \lambda) \in \mathcal{B}(\bar{x}, \bar{\lambda})$ it results $A_{\oplus}(x, \lambda) = A_0(\bar{x})$.*

Therefore, under strict complementarity condition, the estimate is exact in a neighborhood of a KKT pair. Actually, exactness of the estimate is what we need for proving convergence to second order stationary points, so that any other estimate such that

$$A_{\oplus}(x, \lambda) = A_0(\bar{x}) \text{ for all } (x, \lambda) \in \mathcal{B}(\bar{x}, \bar{\lambda})$$

can be used to this aim. Whereas to prove the superlinear rate of convergence, we can get rid of the strict complementarity condition by using Assumption A3 (SSOSC).

For sake of simplicity we will assume in the following sections that the strict complementarity condition holds.

5.2 A particular choice of d^k

In this section we describe a particular direction that satisfies Assumptions A5 and A7 and is tied to the structure of Problem (1) rather than to the function L_G .

The basic idea is to define a system of nonlinear equations $F(x, \lambda) = 0$ such that

$$F(x, \lambda) = 0 \iff (x, \lambda) \text{ is KKT pair,}$$

and to compute the direction $d = (d'_x, d'_\lambda)'$ as an *approximate Newton-type direction* for the system $F(x, \lambda) = 0$. Making use of the index sets (15), we define the system F as follows:

$$F(x, \lambda) = \begin{bmatrix} \nabla f(x) + \nabla g(x) \lambda \\ g(x)_{A_\oplus(x, \lambda)} \\ \lambda_{N_\oplus(x, \lambda)} \end{bmatrix} = 0. \quad (16)$$

The next proposition states that system (16) enjoys the required property.

Proposition 5.2 [14] *It results $F(\bar{x}, \bar{\lambda}) = 0$ if and only if $(\bar{x}, \bar{\lambda})$ is a KKT pair.*

The direction d^k is obtained as an approximated Newton direction of system (16) evaluated at (x^k, λ^k) . In particular, by partitioning the vectors g, λ according to the partition A_\oplus, N_\oplus , the direction $d \in R^{n+m}$ can be evaluated as follows.

$$\text{The direction } d^k = \begin{pmatrix} d_x^k \\ d_\lambda^k \end{pmatrix}$$

Solve the system:

$$\begin{bmatrix} \nabla_x^2 L(x^k, \lambda^k) & \nabla g_{A_\oplus}(x^k) \\ \nabla g_{A_\oplus}(x^k)' & 0 \end{bmatrix} \begin{bmatrix} d_x \\ d_{\lambda_{A_\oplus}} \end{bmatrix} = - \begin{bmatrix} \nabla f(x^k) + \nabla g_{A_\oplus}(x^k) \lambda_{A_\oplus}^k \\ g_{A_\oplus}(x^k) \end{bmatrix} \quad (17)$$

$$d_{\lambda_{N_\oplus}} = -\lambda_{N_\oplus}^k,$$

where A_\oplus and N_\oplus are given by (15) evaluated at (x^k, λ^k) .

If A_\oplus and N_\oplus were constant in a neighbourhood of (x^k, λ^k) , the direction d^k would be exactly the Newton direction for system (16). The following properties of the direction can be proved.

Proposition 5.3 [12] *Let $\{(x^k, \lambda^k)\}$ be a bounded sequence.*

(a) *Suppose that d^k exists for all k . If*

$$\lim_{k \rightarrow \infty} \|d^k\| = 0,$$

then every accumulation point $(\bar{x}, \bar{\lambda})$ of $\{(x^k, \lambda^k)\}$ is a KKT pair.

(b) *If every accumulation point $(\bar{x}, \bar{\lambda})$ of $\{(x^k, \lambda^k)\}$ is a KKT pair, then eventually the matrix $\nabla g_{A_\oplus}^k$ is full rank, the direction d^k is defined and it satisfies*

$$\lim_{k \rightarrow \infty} \|d^k\| = 0.$$

The previous direction can produce a sequence $\{(x^k, \lambda^k)\}$ which is locally convergent with a *quadratic convergence rate* towards a KKT pair where LICQ and SSOSC are satisfied [17].

Moreover, it is possible to prove that a “connection” between the merit function L_G and the direction d can be established. Indeed, in order to define an algorithm which reconciles the global convergence property with a local superlinear convergence rate, the merit function L_G must have some relationships with the local algorithm. In particular, the search direction of the local algorithm must be eventually a “good” descent direction for L_G . This connection can be established by further specifying the estimates A_{\oplus} and N_{\oplus} . In particular, in the expression (15) we set

$$\kappa = \varepsilon p(x, \lambda),$$

where ε and $p(x, \lambda)$ are respectively the parameter and the function which appear in the expression of L_G . Then we have the following main result.

Proposition 5.4 [9] *Let $(\bar{x}, \bar{\lambda})$ be a KKT pair where Assumption A3 holds. Then a neighbourhood $\mathcal{B}(\bar{x}, \bar{\lambda})$, a value $\hat{\varepsilon} > 0$ and a constant $c > 0$ exist such that for all $(x^k, \lambda^k) \in \mathcal{B}(\bar{x}, \bar{\lambda})$ and for all $\varepsilon \in (0, \hat{\varepsilon}]$, d^k satisfies the angle condition:*

$$\nabla L_G(x^k, \lambda^k; \varepsilon)' d^k \leq -c \|d^k\|^2.$$

Proposition 5.4 together with Proposition 5.3 and the exactness properties of L_G , ensure that the direction d^k satisfies (c) of Assumption A5, for small values of ε . However the threshold value $\hat{\varepsilon}$ is usually different from the value $\bar{\varepsilon}$ that ensures the exactness properties of L_G . So that the different schemes LSA_* must be slightly modified to check if the direction d^k calculated at Step 1 can be used as the search direction, or to decide whether an alternative gradient-related direction for L_G must be used, or the penalty parameter ε must be decreased. We do not enter into further details, and we refer the interested reader to [9].

5.3 A particular choice of s^k, Q^k

We introduce a possible choice for the direction s^k and the matrix Q^k used in LSA_2 . The interest in this choice lies in the fact that, as shown in the next section, s^k can be obtained as a by product of the calculation of the direction d^k described before.

Convergence to second order stationary points is enforced by investigating the curvature information on the Lagrangian L in the tangent space of the estimated active constraints.

The pair s^k, Q^k

The matrix Q^k is given by:

$$Q^k = \begin{pmatrix} \nabla_x^2 L(x^k, \lambda^k) & 0 \\ 0 & 0 \end{pmatrix}.$$

The vector $s^k = \begin{pmatrix} s_x^k \\ s_\lambda^k \end{pmatrix}$ is such that:

$$s_x^k \in \mathcal{N}^k \tag{18}$$

$$(s_x^k)' \nabla_x^2 L(x^k, \lambda^k) s_x^k \leq 0 \tag{19}$$

$$(s_x^k)' \nabla_x^2 L(x^k, \lambda^k) s_x^k \rightarrow 0 \implies \begin{cases} \min\{0, \lambda_{\min}(P^k \nabla_x^2 L(x^k, \lambda^k) P^k)\} \rightarrow 0 \\ s_x^k \rightarrow 0 \end{cases} \tag{20}$$

$$M_{N_\oplus}(x^k) s_\lambda^k = -(\nabla g(x^k))' \nabla_x^2 L(x^k, \lambda^k) s_x^k, \tag{21}$$

where:

\mathcal{N}^k denotes the null space of $\nabla g'_{A_\oplus}(x^k)$

P^k denotes the projection matrix on \mathcal{N}^k

$M_N(x)$ is defined in (12)

A_\oplus and N_\oplus are given by (15) evaluated at (x^k, λ^k) .

We point out that Assumption A1 guarantees that $M_{N_\oplus^k}$ is nonsingular in a neighborhood of the feasible set and hence that s_λ^k can be evaluated by (21).

The underlying idea of this choice of s^k, Q^k is that the second order necessary optimality conditions require that $\nabla_x^2 L$ is positive semidefinite on the tangent space of the active constraints $g_{A_0}(\bar{x})$. By Proposition 5.1, under the strict complementarity condition, a neighborhood of $(\bar{x}, \bar{\lambda})$ exists where $A_\oplus^k = A_0(\bar{x})$, so that by continuity we have that condition (20) ensures in the limit the satisfaction of the second order necessary optimality conditions (3).

Under the strict complementarity condition, the pair s^k, Q^k satisfies Assumption A6 [15].

5.4 Combined calculation of d^k, s^k

The interest in using the direction d^k and s^k , introduced in the preceding sections, is due to the fact that their calculation can be obtained simultaneously.

Indeed, algorithms superlinearly convergent to KKT pairs have been recently proposed that are based on the decomposition of the search direction d_x^k into a horizontal step $d_o^k \in \mathcal{N}^k$ and a vertical step $d_v^k \in \mathcal{R}^k$ (\mathcal{R}^k denotes the range space of $\nabla g_{A_\oplus}(x^k)$) so that $d_x^k = d_o^k + d_v^k$, with $d_o^k{}' d_v^k = 0$ (see [2, 13]).

By substituting in system (17) we get that the vertical step d_v^k satisfies the linearized systems of the (estimated) active constraints, namely

$$\nabla g_{A_\oplus}(x^k)' d_v^k = -g_{A_\oplus}(x^k),$$

where A_\oplus and N_\oplus are given by (15) evaluated at (x^k, λ^k) .

The horizontal step d_o^k lies in the tangent space of the estimated active constraints; by using the projection matrix P^k on the null space of $\nabla g'_{A_\oplus}(x^k)$, we can write from the first equation of system (17)

$$P^k \nabla_x^2 L(x^k, \lambda^k) (d_o^k + d_v^k) = -P^k \nabla f(x^k).$$

Since any $d_o \in \mathcal{N}^k$ can be written as $d_o = P^k y$, with $y \in \mathbb{R}^n$, the horizontal step d_o^k can be obtained by computing a solution y^k of the following system:

$$P^k \nabla_x^2 L(x^k, \lambda^k) P^k y = -P^k (\nabla_x^2 L(x^k, \lambda^k) d_v^k + \nabla f(x^k)). \quad (22)$$

The direction $d_{\lambda_{A_\oplus}^k}$ is then obtained by solving the system

$$\nabla g_{A_\oplus}(x^k) d_{\lambda_{A_\oplus}^k} = -(\nabla_x^2 L(x^k, \lambda^k) d_x^k + \nabla f(x^k) + \nabla g_{A_\oplus}(x^k) \lambda_{A_\oplus}^k).$$

A solution of system (22) can be obtained by applying iterative schemes of conjugate gradient type [2, 13]. The use of iterative methods is particularly efficient in large scale setting where a truncated solution of the linear system (22) can be employed. By using standard results in inexact Newton methods [4, 5] it is possible to prove that the superlinear convergence rate is retained.

As a by product of the iterative solution of system (22) it is possible to obtain a vector r^k such that the direction $s_x^k = P^k r^k$ satisfies (18)-(20) (see [24] for similar approach in unconstrained methods).

6 Conclusion and final remarks

We have described a possible use of a smooth exact augmented Lagrangian function L_G in defining algorithms for constrained optimization which are globally convergent to KKT pairs, with superlinear convergence rate. Moreover, also convergence to second order stationary pairs can be attained.

We focused on one of the main advantages in the use of the augmented Lagrangian function, namely the possibility of defining locally convergent algorithms that exploit the structure of the constrained problem, and of using L_G as merit function in a

stabilization scheme. The local algorithm requires only the solution of linear systems that can be also performed inexactly, with increasing accuracy when approaching the solution, so that it is well-suited for large scale problems.

Moreover, thanks to the “indirect” use of L_G only to measure the progress of the iterate to the solution, the conditioning of the linear systems to be solved is independent on the value of the penalty parameter ε . Indeed, although ε is bounded away from zero, its threshold value can be small enough to cause numerical instability when minimizing “directly” L_G .

Since L_G is defined on an open set strictly containing the feasible set, we can also manage a certain degree of infeasibility, that can be prefixed by the user by a proper selection of the parameter α in (8). This can be helpful when a feasible point is not known in advance and/or the problem is highly nonlinear in the constraint functions, so that to retain feasibility at each iteration may be an heavy task and may slow down the rate of convergence.

References

- [1] D. P. Bertsekas. *Constrained Optimization and Lagrange Multipliers Methods*. Academic Press, New York, 1982.
- [2] R. H. Byrd, M. E. Hribar, and J. Nocedal. An interior point algorithm for large-scale nonlinear programming. *SIAM J. Optimization*, 9(4):877–900, 1999.
- [3] F. H. Clarke. *Optimization and Nonsmooth Analysis*. John Wiley and Sons, New York, 1983.
- [4] R. S. Dembo, S. Eisenstat, and T. Steihaug. Inexact Newton methods. *SIAM J. on Numerical Analysis*, 19:400–408, 1982.
- [5] R. S. Dembo and T. Steihaug. Truncated Newton algorithms for large-scale unconstrained optimization. *Math. Programming*, 26:190–212, 1983.
- [6] G. Di Pillo. Exact penalty methods. In E. Spedicato, editor, *Algorithms for Continuous Optimization: the State of the Art*, pages 1–45. Kluwer Academic Press, Boston, 1994.
- [7] G. Di Pillo and L. Grippo. Exact penalty functions in constrained optimization. *SIAM J. Control and Optimization*, 27:1333–1360, 1989.
- [8] G. Di Pillo, G. Liuzzi, S. Lucidi, and L. Palagi. An exact augmented Lagrangian function for two-sided constraints. Tech. Rep. 26-01, Department of Computer and Systems Science, University of Rome “La Sapienza”, Rome, Italy, 2001. Accepted for publication in *Computational Optimization and Applications*.

- [9] G. Di Pillo, G. Liuzzi, S. Lucidi, and L. Palagi. Use of a truncated Newton direction in an augmented Lagrangian framework. Tech. Rep. 18-02, Department of Computer and System Sciences, University of Rome “La Sapienza”, Rome, Italy, 2002. Submitted.
- [10] G. Di Pillo and S. Lucidi. On exact augmented Lagrangian functions in nonlinear programming. In G. Di Pillo and F. Giannessi, editors, *Nonlinear Optimization and Applications*, pages 85–100. Plenum Press, New York, 1996.
- [11] G. Di Pillo and S. Lucidi. An augmented Lagrangian function with improved exactness properties. *SIAM J. Optimization*, 12:376–406, 2001.
- [12] G. Di Pillo, S. Lucidi, and L. Palagi. A truncated Newton method for constrained optimization. In G. Di Pillo and F. Giannessi, editors, *Nonlinear Optimization and Related Topics*. Kluwer Academic, Dordrecht, 1999.
- [13] G. Di Pillo, S. Lucidi, and L. Palagi. Use of a truncated Newton direction in an augmented Lagrangian framework. Tech. rep., Department of Computer and Systems Science, University of Rome “La Sapienza”, Rome, Italy, 1999.
- [14] G. Di Pillo, S. Lucidi, and L. Palagi. A superlinearly convergent primal-dual algorithm for constrained optimization problems with bounded variables. *Optimization Methods and Software*, 14:49–73, 2000.
- [15] G. Di Pillo, S. Lucidi, and L. Palagi. Convergence to 2nd order stationary points of a primal-dual algorithm model for nonlinear programming. Tech. Rep. 10-01, Department of Computer and System Sciences, University of Rome “La Sapienza”, Rome, Italy, 2001. Submitted.
- [16] G. Di Pillo and L. Palagi. Nonlinear programming: introduction, unconstrained optimization, constrained optimization. In P. Pardalos and M. Resende, editors, *Handbook of Applied Optimization*, pages 263–298. Oxford University Press, New York, 2002.
- [17] F. Facchinei and S. Lucidi. Quadratically and superlinearly convergent algorithms for the solution of inequality constrained minimization problems. *J. Optim. Theory and Appl.*, 85:265–289, 1995.
- [18] A. V. Fiacco and G. P. McCormick. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. John Wiley and Sons, New York, 1969.
- [19] T. Glad and E. Polak. A multiplier method with automatic limitation of penalty growth. *Math Programming*, 17:140–155, 1979.
- [20] L. Grippo, F. Lampariello, and S. Lucidi. A nonmonotone line search technique for Newton’s method. *SIAM J. on Numerical Anal.*, 23(4):707–716, 1986.

- [21] L. Grippo, F. Lampariello, and S. Lucidi. A truncated Newton method with nonmonotone line search for unconstrained optimization. *J. Optim. Theory and Appl.*, 60(3):401–419, 1989.
- [22] L. Grippo, F. Lampariello, and S. Lucidi. A class of nonmonotone stabilization methods in unconstrained optimization. *Numer. Math.*, 59:779–805, 1991.
- [23] M. Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory and Application*, 4:303–320, 1969.
- [24] S. Lucidi, F. Rochetich, and M. Roma. Curvilinear stabilization techniques for truncated Newton methods in large scale unconstrained optimization. *SIAM J. Optimization*, 8:916–939, 1998.
- [25] N. Maratos. *Exact penalty function algorithm for finite dimensional and control optimization problems*. PhD thesis, University of London, London, England, 1978.
- [26] M. J. D. Powell. A method for nonlinear constraints in minimization problem. In R. Fletcher, editor, *Optimization*. Academic Press, New York, 1969.
- [27] L. Qi and Sun J. A nonsmooth version of Newton’s method. *Mathematical Programming*, 58:353–367, 1993.
- [28] R. Rockafellar. The multiplier method of Hestenes and Powell applied to convex programming. *J. Optim. Theory and Appl.*, 12:555–562, 1973.