Globally convergent block-coordinate techniques for unconstrained optimization.

Optimization Methods and Software
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713645924

Globally convergent block-coordinate techniques for unconstrained optimization

Luigi Grippof [a]; Marco Sciandrone [a]
[a] Dipartimento di Informatica e Sistemistica, Universitd di Roma “La Sapienza”, Roma, Italy

Online Publication Date: 01 January 1999

Optimization Methods & Software, Vol. 10, pp. 587–637
Reprints available directly from the publisher
Photocopying permitted by license only

GLOBALLY CONVERGENT BLOCK-COORDINATE TECHNIQUES FOR UNCONSTRAINED OPTIMIZATION*

LUIGI GRIPPO[†] and MARCO SCIANDRONE

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

(Received 15 April 1996; Revised 22 June 1997; In final form 27 February 1998)

In this paper we define new classes of globally convergent block-coordinate techniques for the unconstrained minimization of a continuously differentiable function. More specifically, we first describe conceptual models of decomposition algorithms based on the interconnection of elementary operations performed on the block components of the variable vector. Then we characterize the elementary operations defined through a suitable line search or the global minimization in a component subspace. Using these models, we establish new results on the convergence of the nonlinear Gauss–Seidel method and we prove that this method with a two-block decomposition is globally convergent towards stationary points, even in the absence of convexity or uniqueness assumptions. In the general case of nonconvex objective function and arbitrary decomposition we define new globally convergent line-search-based schemes that may also include partial global minimizations with respect to some component. Computational aspects are discussed and, in particular, an application to a learning problem in a Radial Basis Function neural network is illustrated.

Keywords: Unconstrained optimization; decomposition; block-coordinate methods; nonlinear Gauss–Seidel method

1 INTRODUCTION

We consider the problem of minimizing a continuously differentiable function $f: R^n \to R$ by means of block decomposition techniques.

* This research was supported by Agenzia Spaziale Italiana, Rome, Italy.
† Corresponding author.

The main motivation for the use of a block decomposition method can be that, when some variables are fixed, we often obtain one or more subproblems of a special structure in the remaining variables. This can be useful in the solution of many optimization problems, especially when the structure of the subproblems can be conveniently exploited, by using, for instance, parallel optimization techniques.

One of the best known approaches to variable decomposition is the minimization version of the *block-nonlinear Gauss–Seidel method* [4,17], based on successive global minimizations with respect to each component vector. The convergence of this technique has been studied under suitable convexity assumptions on f (see, for instance, [4]). In the convex case, decomposition techniques have also been considered for constrained problems, with reference to the *alternating direction method of multipliers* (see, e.g. [4–6,9–11]) and to projection techniques [14,22]. In the special case of the coordinate method *with exact line searches*, convergence has been proved either under pseudoconvexity assumptions on f [25] or under a uniqueness assumption on the global one-dimensional minimizer along a line (see, e.g., [2,17,26]), which may require strict (generalized) convexity assumptions on f, as a function of each component. Moreover, it has been shown in [20], through a set of counterexamples, that, when these assumptions are not satisfied, the coordinate method with exact searches (and hence the Gauss–Seidel method) may not converge towards stationary points, in the sense that there are cases in which convergent subsequences are generated with gradients bounded away from zero.

An alternative (but related) approach to variable decomposition is that of performing successive searches along descent directions in the component subspaces. In this case we can regard the resulting algorithm (which is often called *block coordinate descent method*) as an ordinary descent method with search directions having zeros in certain positions. Now the difficulty is that these directions may not be *gradient related* [3], unless suitable rules are adopted for choosing the current direction. For instance, the so-called *Gauss–Southwell rule* [14] consists in choosing a direction which is related to the partial gradient of largest norm and actually yields a globally convergent (but expensive) technique. In the special case of the cyclic coordinate method *with inexact line searches*, the convergence results are based on the uniform linear independence of the search directions and on suitable assumptions on the line searches [17]. Inexact line searches that may avoid the need of evaluating the derivatives of the objective function and do not rely on convexity assumptions, have been considered in [1,12,19,21]. In the general case of nonconvex objective function and search directions that are related to the (block) partial gradients, convergence has been established for techniques employing constant stepsizes, under Lipschitz continuity conditions on the gradient [4,23]. Under similar assumptions, convergent algorithms have been also proposed in the context of parallel optimization (see, e.g., [4,24]).

In the present paper, with reference to the general nonconvex case, we present additional results for unconstrained decomposition algorithms and we propose new globally convergent schemes.

More specifically, we first state sufficient convergence criteria, expressed in terms of conditions on the elementary operations performed on each block component, and of suitable (sequential or parallel) connection rules. This allows us to simplify the analysis of various decomposition schemes, which can be viewed as the interconnection of different elementary mappings.

Then we characterize an elementary operation consisting of an inexact line search along a direction in the component subspace, which is based on the techniques proposed in [7,12]. The line search mapping yields both a constructive device for deriving convergence proofs and an effective computational tool. A similar analysis, based on known results [4,17] is also performed on the minimization mapping used in the Gauss–Seidel method, and the dependency of the stepsize on strict convexity assumptions is evidenced.

On the basis of these results we reconsider the convergence analysis of the nonlinear Gauss–Seidel method. Under pseudoconvexity assumptions on f we extend the results of [25] to the case of a block decomposition; moreover, we prove that in case of a two-block decomposition, the Gauss–Seidel method is globally convergent towards stationary points, even in the absence of any convexity or uniqueness assumption.

In the general case of nonconvex problems and arbitrary decomposition, we define a globally convergent scheme, where each step consists of a two-phase procedure. The first phase is an inexact line search along a search direction in the component subspace, which yields reference values for the objective function and the stepsize. In the second phase we can compute a further updating of the current component by means of any minimization method in the same subspace, provided that suitable acceptability conditions are satisfied. Under usual assumptions, we show that the proposed algorithm (which can be viewed as a line-search-based block-coordinate descent technique) is globally convergent, without any convexity assumption.

We also show that, when the objective function is a strictly quasi-convex function of some component, for fixed values of the other components, then convergence can be achieved by means of a hybrid scheme in which a partial Gauss–Seidel method is employed for a subset of components.

More particular schemes are defined in the case of a two-block decomposition, since, in this case, the convergence conditions are less demanding.

Finally, we discuss the computational aspects and the potential advantages of the techniques considered here and we describe the application of a decomposition approach to a learning problem in a Radial Basis Function neural network [13,18].

2 NOTATION

We consider the problem

$$\begin{array}{c} \text{minimize } f(x) \\ x \in R^n \end{array}$$

where, unless otherwise stated, the objective function $f: R^n \to R$ is assumed to be continuously differentiable on R^n.

We suppose that the vector $x \in R^n$ is partitioned into $m \leq n$ component vectors $x_i \in R^{n_i}$ with

$$\sum_{i=1}^{m} n_i = n,$$

that is we get: $x = (x_1, \dots, x_i, \dots, x_m)$.

The algorithms we will consider generate a sequence $\{x^k\}$ of points

$$x^k := (x_1^k, \dots, x_i^k, \dots, x_m^k) \in R^n,$$

in a way that the transition from x^k to x^{k+1} is performed through suitable inner steps that update the individual components of x^k.

A subsequence of $\{x^k\}$ corresponding to an infinite index set K will be denoted by $\{x^k\}_K$.

We indicate by $\|\cdot\|$ the Euclidean norm (on the appropriate space) and therefore, if $y \in R^n$ is partitioned in the form $y = (y_1, \ldots, y_i, \ldots, y_m)$ with $y_i \in R^{n_i}$, we can write:

$$\|x - y\| = \left[\sum_{i=1}^{m} \|x_i - y_i\|^2\right]^{1/2} \leq \sum_{i=1}^{m} \|x_i - y_i\|.$$

In correspondence to the given partition of x, the function value $f(x)$ is also indicated by $f(x_1, \ldots, x_i, \ldots, x_m)$.

The gradient of f with respect to x is denoted by $\nabla f \in R^n$ and $\nabla_i f \in R^{n_i}$ is the partial gradient of f with respect to x_i.

When f is assumed to be twice continuously differentiable, the Hessian matrix of f with respect to x is denoted by $\nabla^2 f \in R^{n \times n}$ and $\nabla_i^2 f \in R^{n_i \times n_i}$ is the partial Hessian matrix of f with respect to the component x_i.

We denote by $\mathcal{L}$ the level set of f corresponding to the given initial point $x^0 \in R^n$ that is:

$$\mathcal{L} := \{x \in R^n \colon f(x) \leq f(x^0)\}.$$

Finally we recall from [17] the notion of *forcing function*, which is a function $\sigma : R^+ \to R^+$, such that

$$\lim_{k \to \infty} \sigma(t_k) = 0 \text{ implies } \lim_{k \to \infty} t_k = 0.$$

3 CONVERGENCE CONDITIONS

In this section we introduce some sufficient convergence conditions that will be exploited in the sequel for the analysis and the construction of globally convergent decompostion algorithms. The models we will consider are essentially based on the interconnection of suitable elementary operations performed on each block component of the current vector x^k.

Borrowing the notation of [4], we represent an elementary operation performed on the ith block component by introducing a mapping $T_i : R^n \rightarrow R^{n_i}$ that associates to the vector $y^k \in R^n$ in a sequence $\{y^k\}$ a block-component $T_i(y^k)$. We note that, in general, the mapping T_i may be dependent on k; however, in order to simplify notation, we omit the explicit indication of the iteration index. As we do not use continuity assumption on T_i, this can be acceptable.

The mappings we will consider for given i are descent mappings such that the following condition holds.

Condition 1 If $\{y^k\}$ is a sequence of points in R^n we have, for all k:

$$f(y_1^k, \ldots, y_{i-1}^k, T_i(y^k), y_{i+1}^k, \ldots, y_m^k) \leq f(y^k).$$

We also need Lyapunov-type conditions on T_i that ensure convergence towards stationary points in the y_i-space, with respect to the partial gradient $\nabla_i f$. One of the weakest requirements we can impose is the following.

Condition 2 If $\{y^k\}$ is a sequence of points in R^n converging to some $\bar{y} \in R^n$ and such that

$$\lim_{k \to \infty} f(y^k) - f(y_1^k, \ldots, T_i(y^k), \ldots, y_m^k) = 0,$$

then we have:

$$\nabla_i f(\bar{y}) = 0.$$

In some instances we must require, in addition, that the stepsize $\|T_i(y^k) - y_i^k\|$ goes to zero. This is expressed formally in the following condition.

Condition 3 If $\{y^k\}$ is a sequence of points in R^n converging to some $\bar{y} \in R^n$ and such that

$$\lim_{k \to \infty} f(y^k) - f(y_1^k, \ldots, T_i(y^k), \ldots, y_m^k) = 0,$$

then we have:

$$\lim_{k \to \infty} \|T_i(y^k) - y_i^k\| = 0.$$

When the level sets of f are unbounded and the existence of a limit point is not postulated, Conditions 2 and 3 can be replaced, respectively, with the following stronger requirements.

Condition 4 If $\{y^k\}$ is a sequence of points in R^n such that

$$\lim_{k\to\infty} f(y^k) - f(y_1^k, \dots, T_i(y^k), \dots, y_m^k) = 0,$$

then we have:

$$\lim_{k\to\infty} \nabla_i f(y^k) = 0.$$

Condition 5 If $\{y^k\}$ is a sequence of points in R^n such that

$$\lim_{k\to\infty} f(y^k) - f(y_1^k, \dots, T_i(y^k), \dots, y_m^k) = 0,$$

then we have:

$$\lim_{k\to\infty} ||T_i(y^k) - y_i^k|| = 0.$$

Concrete examples of elementary mappings will be analyzed in the sequel. Here we state obvious conditions under which, given a mapping $\tilde{T}_i$ satisfying some of the properties stated above, we can generate a new mapping T_i in a way that the same properties are preserved. This possibility is specified in the next proposition.

PROPOSITION 3.1 *Let $\tilde{T}_i$ be a given mapping satisfying Condition* 1 *and suppose we define a new mapping T_i such that, if $\{y^k\}$ is a sequence in R^n we have, for all k*:

$$f(y_1^k, \dots, T_i(y^k), \dots, y_m^k) \le f(y_1^k, \dots, \tilde{T}_i(y^k), \dots, y_m^k), \qquad (1)$$

then T_i satisfies Condition 1 *and, moreover, we have*:

(i) *if $\tilde{T}_i$ satisfies Condition* 2 *or* 4, *then the same condition holds for T_i*;

(ii) *if $\tilde{T}_i$ satisfies Condition* 3 *or* 5 *and*

$$\lim_{k\to\infty} ||\tilde{T}_i(y^k) - y_i^k|| = 0 \text{ implies } \lim_{k\to\infty} ||T_i(y^k) - y_i^k|| = 0,$$

then the same condition holds for T_i.

Proof The assertions are immediate consequences of the assumptions. We must only note that the assumptions made imply

$$f(y_1^k, \ldots, T_i(y^k), \ldots, y_m^k) \le f(y_1^k, \ldots, \tilde{T}_i(y^k,), \ldots, y_m^k) \le f(y^k),$$

and hence the limit:

$$\lim_{k\to\infty} f(y^k) - f(y_1^k, \ldots, T_i(y^k), \ldots, y_m^k) = 0$$

implies

$$\lim_{k\to\infty} f(y^k) - f(y_1^k, \ldots, \tilde{T}_i(y^k), \ldots, y_m^k) = 0.$$

Now suppose that a set of mappings T_i for $i = 1, \ldots, m$, has been defined for every i, so that the construction of a decomposition algorithm for the minimization of f can be performed by choosing suitable connection rules.

We consider, in particular, two basic connection schemes:

- the sequential connection;
- the parallel connection.

In each of these schemes we assume that each major step (indexed by k) updates all components of x^k, through a set of elementary operations.

We admit also the possibility that some component is left unchanged during a major step, and therefore we introduce a nonempty index set $I^k \subseteq \{1, \ldots, m\}$ for specifying the components that are actually updated through a "serious" operation T_i.

3.1 Sequential Connection

In this scheme, starting from a given initial point x^0, a sequence $\{x^k\}$ of points

$$x^k := (x_1^k, \ldots, x_i^k, \ldots, x_m^k) \in R^n$$

is constructed by updating in sequence the components of x^k.

This produces the vectors $z(k,i) \in R^n$, such that $z(k,1) = x^k$ and

$$z(k,i) := (x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^k, x_{i+1}^k, \ldots, x_m^k) \quad \text{for } i = 2, \ldots, m.$$

Given $z(k,i)$ and the index set I^k, we compute the updated ith component by letting

$$x_i^{k+1} := \begin{cases} T_i(z(k,i)) & \text{if } i \in I^k, \\ x_i^k & \text{if } i \notin I^k, \end{cases}$$

so that, either $z(k,i+1)$ is the result of a "serious" operation T_i performed on $z(k,i)$ that updates the ith component, or we have $z(k,i+1) = z(k,i)$. For notational convenience, we set also

$$z(k,m+1) = z(k+1,1) = x^{k+1}.$$

This scheme is illustrated in Fig. 1, in the case $m = 3$ and $I^k = \{1,3\}$. We note that in the example considered the component x_2 is left unchanged during the major step.

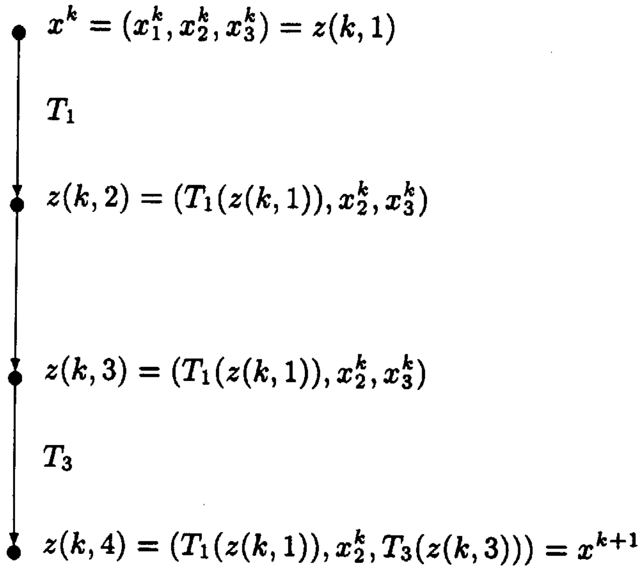

FIGURE 1 Sequential connection.

First of all, letting $I^k = \{1, \ldots, m\}$, we extend to our case the convergence results given in [25] in connection with the cyclic coordinate method with exact line searches. More specifically, we prove the following theorem.

THEOREM 3.2 *Suppose that f is a pseudoconvex function and that $\mathcal{L}$ is compact. Let $T_i: R^n \to R^{n_i}$ be given mappings that satisfy Conditions 1 and 2 for every $i = 1, \ldots, m$. Let $\{x^k\}$ be a sequence such that $z(k, i) = x^k$ and*

$$z(k, i+1) = (x_1^k + 1, \ldots, x_{i-1}^{k+1}, T_i(z(k, i)), x_{i+1}^k, \ldots, x_m^k).$$

Then every limit point of $\{x^k\}$ is a global minimizer of f.

Proof Recalling Condition 1, by definition of $z(k, i)$ (which implies $z(k, 1) = x^k$ and $z(k, m+1) = x^{k+1}$), we have:

$$f(x^{k+1}) \leq f(z(k, i+1)) \leq f(z(k, i)) \leq f(x^k) \quad \text{for } i = 1, \ldots, m. \tag{2}$$

Then, the sequences $\{z(k, i)\}$, for $i = 1, \ldots, m+1$, belong to the compact set $\mathcal{L}$. In particular, we have that the sequence $\{x^k\}$ admits at least one limit point. In order to prove the thesis, by contradiction, let us assume that there exists a subsequence $\{x^k\}_K$ such that

$$\lim_{k \to \infty,\, k \in K} x^k = \bar{x}, \tag{3}$$

$$\lim_{k \to \infty,\, k \in K} z(k, i) = \bar{z}^i, \quad i = 2, \ldots, m+1 \tag{4}$$

and

$$\|\nabla_j f(\bar{x})\| > 0 \quad \text{for some } j \in \{1, \ldots, m\}. \tag{5}$$

We observe that we can write

$$z(k, i) = z(k, i-1) + d(k, i-1) + d(k, i-1) \quad \text{for } i = 2, \ldots, m+1, \tag{6}$$

where the block components $d_h(k, i-1) \in R^{n_h}$ of the vector $d(k, i-1)$, with $h \in \{1, \ldots, m\}$, are such that $d_h(k, i-1) = 0$ if $h \neq i-1$. Therefore, for $i = 2, \ldots, m+1$, from (4) we get

$$\bar{z}^i = \bar{z}^{i-1} + \bar{d}^{i-1}, \tag{7}$$

where

$$\bar{d}^{i-1} = \lim_{k \to \infty,\, k \in K} d(k, i-1), \tag{8}$$

and

$$\bar{d}_h^{i-1} = 0, \quad h \neq i-1. \tag{9}$$

The continuity assumption on f and the compactness of $\mathcal{L}$ imply that f is bounded below. As $\{f(x^k)\}$ is nonincreasing, recalling (2), we have

$$\lim_{k \to \infty} (f(z(k,i)) - f(z(k, i+1))) = 0, \quad i = 1, \ldots, m. \tag{10}$$

By using the continuity assumption on f it follows that

$$f(\bar{x}) = f(\bar{z}^i), \quad i = 1, \ldots, m+1. \tag{11}$$

Recalling Condition 2 (where we identify, for each $i \in \{1, \ldots, m\}$, the sequence $\{y^k\}$ with the subsequence $\{z(k,i)\}_K$), by (10) we obtain

$$\nabla_i f(\bar{z}^i) = 0 \quad \text{for } i = 1, \ldots, m. \tag{12}$$

By (12), taking into account the pseudoconvexity assumption on f, we can write

$$\bar{z}_{i-1}^{i-1} = \arg \min_{\xi \in R^{n_{i-1}}} f(\bar{z}_1^{i-1}, \ldots, \xi, \ldots, \bar{z}_m^{i-1}) \quad \text{for } i = 2, \ldots, m+1. \tag{13}$$

Therefore, from (11), recalling (7) and (9), and using (13), we obtain

$$\nabla_{i-1} f(\bar{z}^i) = 0 \quad \text{for } i = 2, \ldots, m+1. \tag{14}$$

Now we prove that, if $\ell \in \{1, \ldots, m\}$ and we assume

$$\nabla_\ell f(\bar{z}^j) = 0, \tag{15}$$

then it follows

$$\nabla_\ell f(\bar{z}^{j-1}) = 0. \tag{16}$$

By (8) and (9) we have

$$\bar{z}^j = \bar{z}^{j-1} + \bar{d}^{j-1},$$

where $\bar{d}_h^{j-1} = 0$ for $h \neq j-1$.

For any given vector $\eta \in R^{n_\ell}$ define

$$w(\eta) = \bar{z}^{j-1} + d(\eta),$$

where $d_h(\eta) = 0$ for $h \neq \ell$ and $d_\ell(\eta) = \eta \in R^{n_\ell}$.

Then, from assumption (15) and (14), we obtain

$$\begin{aligned}\nabla f(\bar{z}^j)^{\mathrm{T}}(w - \bar{z}^j) &= \nabla f(\bar{z}^j)^{\mathrm{T}}(d(\eta) - \bar{d}^{j-1}) \\ &= \nabla_\ell f(\bar{z}^j)^{\mathrm{T}}\eta - \nabla_{j-1} f(\bar{z}^j)^{\mathrm{T}}\bar{d}_{j-1}^{j-1} = 0.\end{aligned}$$

It follows by the pseudoconvexity of f that

$$f(\bar{z}^{j-1} + d(\eta)) \geq f(\bar{z}^j).$$

On the other hand, $f(\bar{z}^j) = f(\bar{z}^{j-1})$, and therefore we have:

$$f(\bar{z}^{j-1} + d(\eta)) \geq f(\bar{z}^{j-1}) \quad \text{for all } \eta \in R^{n_\ell},$$

which, recalling the definition of $d(\eta)$, implies (16).

Finally, taking into account (12), and using the fact that (15) implies (16), by induction we obtain

$$\nabla_j f(\bar{z}^1) = \nabla_j f(\bar{x}) = 0,$$

which contradicts (5).

In the preceding theorem a crucial role is played by the pseudoconvexity assumption on f. If we remove this assumption, we must impose stronger requirements on the mappings T_i, which may enforce the stepsizes to go to zero in case of convergence. This allows us also to consider more general updating rules that take into account a finite number of previous iterations.

THEOREM 3.3 *Let $T_i: R^n \to R^{n_i}$ be given mappings that satisfy Conditions 1–3 for every $i = 1, \ldots, m$. Let $\{x^k\}$ be a sequence such that $z(k, i) = x^k$ and*

$$z(k, i+1) = (x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^{k+1}, x_{i+1}^k, \ldots, x_m^k),$$

where

$$x_i^{k+1} := \begin{cases} T_i(z(k, i)) & \text{if } i \in I^k, \\ x_i^k & \text{if } i \notin I^k, \end{cases}$$

and $I^k \subseteq \{1, \ldots, m\}$ is a nonempty index set. Let $\nu \geq 0$ be given integer; let

$$I_*^k := \bigcup_{j=0}^{\nu} I^{k-j},$$

and suppose that for every sufficiently large $k > \nu$ we have:

$$\sum_{i \in I_*^k} \|\nabla_i f(z, (h_i^k, i))\| \geq \sigma\left(\sum_{i=1}^{m} \|\nabla_i f(z(s_i^k, i))\|\right), \tag{17}$$

where σ is a forcing function and the integers h_i^k, for I_*^k, and s_i^k, for $i \in \{1, \ldots, m\}$, are indices satisfying:

$$0 \leq k - h_i^k \leq \nu, \;\; i \in I_*^k, \quad 0 \leq k - s_i^k \leq \nu, \;\; i \in \{1, \ldots, m\}.$$

Then, every limit point of $\{x^k\}$ is a stationary point of f. Moreover, if $\mathcal{L}$ is compact we have:

$$\lim_{k \to \infty} \nabla f(x^k) = 0$$

and there exists at least one limit point that is a stationary point of f.

Proof First we note that, whenever $i \notin I^k$, we have $f(z(k, i+1)) = f(z(k, i))$. Taking this into account and recalling Condition 1, by definition of $z(k, i)$, we have:

$$f(x^{k+1}) \leq f(z(k, i) + 1)) \leq f(z(k, i)) \leq f(x^k) \quad \text{for } i = 1, \ldots, m. \tag{18}$$

Now, let us consider an infinite subset $K \subseteq \{0, 1, \ldots\}$ such that

$$\lim_{k \to \infty,\, k \in K} x^k = \bar{x}. \tag{19}$$

Then, the continuity of f and the convergence of $\{x^k\}_K$ imply that the sequence $\{f(x^k)\}$ has a convergent subsequence. As $\{f(x^k)\}$ is non-increasing, this, in turn, implies that $\{f(x^k)\}$ is bounded from below and has a limit. Therefore, recalling (18) we have also

$$\lim_{k \to \infty} (f(z(k,i)) - f(z(k,i+1))) = 0, \quad i = 1, \ldots, m. \tag{20}$$

Now, using Condition 3 and the fact that

$$\|z(k,i+1) - z(k,i)\| = \|x_i^{k+1} - x_i^k\| = 0 \quad \text{for all } i \notin I^k, \tag{21}$$

we prove, by induction on i, that the subsequences $\{z(k,i)\}_K$ converge to $\bar{x}$ for all i.

For $i = 1$, recalling that $z(k,1) = x^k$, this follows from (19). Then, we suppose that for $i \geq 1$ we have

$$\lim_{k \to \infty,\, k \in K} z(k,i) = \bar{x} \tag{22}$$

and we show that $\{z(k,i+1)\}_K$ converges to $\bar{x}$.

If $i \notin I^k$ for a $k \in K$, we have $z(k,i+1) = z(k,i)$, and hence, if $i \notin I^k$ for all $k \in K$ sufficiently large, the assertion is obvious. Therefore, let us assume that there exists a subsequence $\{z(k,i)\}_{K_1}$, with $K_1 \subseteq K$ such that $i \in I^k$ for $k \in K_1$. Then, by identifying the sequence $\{y^k\}$ appearing in Condition 3 with the subsequence $\{z(k,i)\}_{K_1}$, by (20) and (22) we get from Condition 3:

$$\lim_{k \to \infty,\, k \in K_1} \|z(k,i+1) - z(k,i)\| = \lim_{k \to \infty,\, k \in K_1} \|T_i(z(k,i)) - x_i^k\| = 0,$$

so that, again by (22) we have:

$$\lim_{k \to \infty,\, k \in K_1} z(k,i+1) = \bar{x}.$$

Thus we can assert that:

$$\lim_{k\to\infty,\, k\in K} z(k,i) = \bar{x}, \quad i = 1,\dots,m. \tag{23}$$

Suppose now that $k > \nu$. As $z(k-1, m+1) = x^k$, we can repeat the same reasoning by backward induction on i to get the limits

$$\lim_{k\to\infty,\, k\in K} z(k-1,i) = \bar{x}, \quad i = 1,\dots,m.$$

Then, continuing in this way for increasing values of $j \le \nu$ we have:

$$\lim_{k\to\infty,\, k\in K} z(k-j,i) = \bar{x}, \quad i = 1,\dots,m, \; j = 0,1,\dots,\nu. \tag{24}$$

As the number of different index sets I_*^k is finite, we can find an index set $\hat{I}$ and subsequences $\{z(h_i^k, i)\}_{K_2}$, for $i \in \hat{I}$, and $\{z(s_i^k, i)\}_{K_2}$, for $i \in \{1,\dots,m\}$, with $K_2 \subseteq K$ such that, for $k \in K_2$, we can write, by assumption (17), the inequality:

$$\sum_{i\in\hat{I}} \|\nabla_i f(z(h_i^k, i))\| \ge \sigma\left(\sum_{i=1}^{m} \|\nabla_i f(z(s_i^k, i))\|\right). \tag{25}$$

Then, taking into account the assumptions on the indices h_i^k and s_i^k and recalling (24), it is easily seen that each sequence $\{z(h_i^k, i)\}_{K_2}$ and $\{z(s_i^k, i)\}_{K_2}$ will converge to $\bar{x}$ for every i, so that from (25), taking limits for $k \in K_2$, we obtain:

$$\sum_{i\in\hat{I}} \|\nabla_i f(\bar{x})\| \ge \sigma\left(\sum_{i=1}^{m} \|\nabla_i f(\bar{x})\|\right). \tag{26}$$

On the other hand, for $k \in K_2$, if $i \in \hat{I} = I_*^k$, this implies that there exists an integer in $[0, \nu]$ (depending on i and k), which we denote by $j(i,k)$, such that $i \in I^{k-j(k,i)}$, so that, by definition of I^k, we have:

$$z(k-j(i,k), i+1) = (x_1^{k-j(k,i)+1}, \dots, T_i(z(k-j(i,k), i)), \dots, x_m^{k-j(k,i)}).$$

By (24), it is easily seen that each subsequence $\{z(k-j(i,k), i)\}_{K_2}$ converges to $\bar{x}$. Therefore, recalling Condition 2 (where we identify, for each $i \in \hat{I}$, the sequence $\{y^k\}$ with the subsequence $\{z(k - j(i,k), i)\}_{K_2}$), by (20) we obtain

$$\nabla_i f(\bar{x}) = 0, \quad i \in \hat{I},$$

so that by (26), we get $\nabla f(\bar{x}) = 0$.

The last assertion of the theorem follows immediately from the compactness of $\mathcal{L}$ and the continuity assumptions on ∇f.

When $\mathcal{L}$ is unbounded, in order to show that the gradient goes to zero in the limit, we replace Conditions 2 and 3 with Conditions 4 and 5 and we impose stronger continuity requirements on ∇f.

THEOREM 3.4 *Suppose that the assumptions of Theorem* 3.3 *are satisfied and that the mappings* T_i *satisfy Conditions* 4 *and* 5 *for every* $i = 1, \dots, m$. *Assume also that* f *is bounded below and that*

$$\lim_{k\to\infty} ||\nabla f(u^k) - \nabla f(w^k)|| = 0, \tag{27}$$

whenever $\{u^k\}, \{w^k\} \in \mathcal{L}$ *and*

$$\lim_{k\to\infty} ||u^k - w^k|| = 0.$$

Then we have:

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

Proof Reasoning as in the proof of Theorem 3.3, we can write

$$f(x^{k+1}) \le f(z(k, i+1)) \le f(z(k,i)) \le f(x^k) \quad \text{for } i = 1, \dots, m. \tag{28}$$

This shows that the points x^k and $z(k,i)$ remain in $\mathcal{L}$ and that the objective function is nonincreasing. As f is bounded below, we get the limits

$$\lim_{k\to\infty} (f(z(k,i)) - f(z(k, i+1))) = 0, \quad i = 1, \dots, m. \tag{29}$$

Now, for each $i \in \{1, \dots, m\}$ if $i \notin I^k$, we have $x_i^{k+1} = x_i^k$; on the other hand, if $i \in I^k$ for a subsequence $\{z(k,i)\}_K$, by (29) and Condition 5 (where we identify $\{y^k\}$ with $\{z(k,i)\}_K$), we obtain

$$\lim_{k\to\infty,\, k\in K} \|T_i(z(k,i)) - x_i^k\| = 0. \tag{30}$$

It can be concluded that:

$$\lim_{k\to\infty} \|x_i^{k+1} - x_i^k\| = 0, \quad i = 1,\dots,m,$$

which also yields

$$\lim_{k\to\infty} \|z(k,i+1) - z(k,i)\| = 0, \quad i = 1,\dots,m-1. \tag{31}$$

As

$$\|z(k,i) - x^k\| = \|z(k,i) - z(k,1)\| \le \sum_{j=1}^{i-1} \|z(k,j+1) - z(k,j)\|, \tag{32}$$

from (31) we obtain

$$\lim_{k\to\infty} \|x^k - z(k,i)\| = 0, \quad i = 1,\dots,m. \tag{33}$$

For $k > \nu$ we can repeat the same reasoning by backward induction, starting from the point $z(k-1, m+1) = x^k$ and hence we get the limits

$$\lim_{k\to\infty} \|x^k - z(k-j,i)\| = 0, \quad i = 1,\dots,m, \; j = 0,1,\dots,\nu. \tag{34}$$

Now for every $i \in \{1,\dots,m\}$, and every $j \in \{0,1,\dots,\nu\}$ we can write:

$$\|\nabla_i f(x^k) - \nabla_i f(z(k-j,i))\| \le \|\nabla f(x^k) - \nabla f(z(k-j,i))\|, \tag{35}$$

and hence, by (34) and (27), we obtain

$$\lim_{k\to\infty} \|\nabla_i f(x^k) - \nabla_i f(z(k-j,i))\| = 0, \quad i = 1,\dots,m, \; j = 0,1,\dots,\nu. \tag{36}$$

If the assertion is false there must exist at least one index i^* and a subsequence $\{x^k\}_K$ such that, for $k \in K$, we have:

$$\|\nabla_{i^*} f(x^k)\| \geq \eta$$

for some $\eta > 0$. This, in turn implies, by (36) that there must exist subsequences $\{z(k-j, i^*)\}_{K_1}$, with $K_1 \subseteq K$, for $j = 0, 1, \ldots, \nu$ such that

$$\|\nabla_{i^*} f(z(k-j, i^*))\| \geq \eta_j \quad \text{for every } j = 0, 1, \ldots, \nu \tag{37}$$

for some η_j, with $0 < \eta_j \leq \eta$.

As the number of different index sets I_*^k is finite, we can find an index set $\hat{I}$ and subsequences $\{z(h_i^k, i)\}_{K_2}$, for $i \in \hat{I}$, and $\{z(s_i^k, i)\}_{K_2}$, for $i \in \{1, \ldots, m\}$, with $K_2 \subseteq K_1$ such that, for $k \in K_2$, we can write, by assumption (17), the inequality:

$$\sum_{i \in \hat{I}} \|\nabla_i f(z(h_i^k, i))\| \geq \sigma \left(\sum_{i=1}^{m} \|\nabla_i f(z(s_i^k, i))\| \right). \tag{38}$$

Now, for $k \in K_2$, if $i \in \hat{I}$, there exists an integer $j(i,k) \in [0, \nu]$, such that $i \in I^{k-j(k,i)}$ and:

$$z(k - j(i,k), i+1) = (x_1^{k-j(k,i)+1}, \ldots, T_i(z(k-j(i,k), i)), \ldots, x_m^{k-j(k,i)}).$$

Then, recalling Condition 4 (where we identify, for each $i \in \hat{I}$, the subsequence $\{y^k\}$ with the subsequence $\{z(k-j(i,k), i)\}_{K_2}$), by (29) we obtain:

$$\lim_{k \to \infty,\, k \in K_2} \nabla_i f(z(k - j(i,k), i)) = 0, \quad i \in \hat{I}. \tag{39}$$

Moreover, by (34) and the assumptions on h_i^k we have

$$\lim_{k \to \infty,\, k \in K_2} \|z(k - j(i,k), i) - z(h_i^k, i)\| = 0, \quad i \in \hat{I},$$

and hence, using (27), it is easily seen that (39) implies the limits:

$$\lim_{k \to \infty,\, k \in K_2} \nabla_i f(z(h_i^k, i)) = 0, \quad i \in \hat{I}. \tag{40}$$

Then (38) yields

$$\lim_{k\to\infty,\, k\in K_2} \nabla_i f(z(s_i^k, i)) = 0, \quad i = 1,\dots,m,$$

and hence we have, in particular:

$$\lim_{k\to\infty,\, k\in K_2} \nabla_{i^*} f(z(s_{i^*}^k, i^*)) = 0,$$

so that, recalling the assumptions on s_i^k, we get a contradiction to (37).

We note that condition (17) appearing in the preceding theorems is satisfied, in particular, if we choose $\sigma(t) = t$, $\nu = 0$ and $I^k = \{1,\dots,m\}$ for all k. In this case, every component x_i^k is updated at each step by means of the mapping T_i. However, many different schemes can be devised, in which we can take into account the results of the previous iterations in order to choose a subset of components to be updated at step k. In particular, we could define algorithms based on an approximate and less expensive implementation of the Gauss–Southwell rule, by evaluating the norm of the gradient components during a finite set of past iterations, and also various *almost cyclic* rules of the kind discussed in [14].

Note also that the continuity assumption on ∇f used in Theorem 3.4 is satisfied, in particular, if there exist numbers $p > 0$ and $c > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \le c\|x - y\|^p \quad \text{for all } x, y \in \mathcal{L}.$$

3.2 Parallel Connection

A parallel connection scheme, with initial point x^0, can be defined by computing at each k, for $i \in I^k \subseteq \{1,\dots,m\}$, the components $T_i(x^k)$ and the points

$$w(k,i) := (x_1^k,\dots,x_{i-1}^k, T_i(x^k), x_{i-1}^k,\dots,x_m^k),$$

and then constructing the updated point x^{k+1} by means of some rule.

In this case, we will assume that information on past iterations is not taken into account and this allows us to introduce less demanding conditions on the stepsizes.

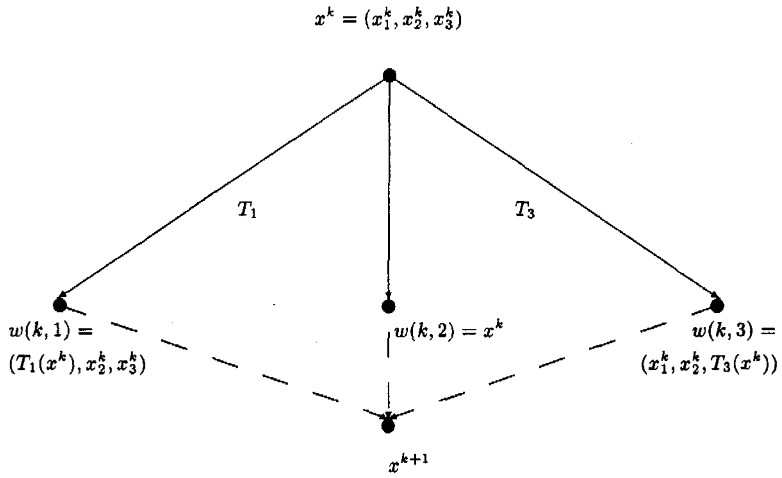

FIGURE 2 Parallel connection.

The connection scheme is illustrated in Fig. 2, in the case $m = 3$ and $I^k = \{1, 3\}$.

Here we will confine ourselves to consider the case in which the connection rule ensures that the function value does not increase.

Convergence results for parallel decomposition algorithms that use this criterion as a synchronization step were given in [16,8] both for unconstrained and constrained optimization problems.

When $I^k = \{1, \dots, m\}$, the following convergence conditions can be viewed as an abstraction of some of the results in [8,16].

THEOREM 3.5 *Let $T_i: R^n \to R^{n_i}$ be given mappings that satisfy Conditions 1 and 2 for every $i = 1, \dots, m$. For each k, define the points:*

$$w(k,i) := \begin{cases} (x_1^k, \dots, x_{i-1}^k, T_i(x^k), x_{i+1}^k, \dots, x_m^k) & \text{if } i \in I^k, \\ x^k & \text{if } i \notin I^k, \end{cases}$$

where $I^k \subseteq \{1, \dots, m\}$ is a nonempty index set such that

$$\sum_{i \in I^k} \|\nabla_i f(x^k)\| \geq \sigma\left(\sum_{i=1}^{m} \|\nabla_i f(x^k)\|\right), \tag{41}$$

for some forcing function σ.

Let $\{x^k\}$ *be a sequence such that*:

$$f(x^{k+1}) \leq f(w(k,i)) \quad \textit{for all } i = 1, \ldots, m. \tag{42}$$

Then, every limt point of $\{x^k\}$ *is a stationary point of* f. *Moreover, if* $\mathcal{L}$ *is compact we have*:

$$\lim_{k\to\infty} \nabla f(x^k) = 0$$

and there exists at least one limit point that is a stationary point of f.

Proof By Condition 1 and (42) we have

$$f(x^{k+1}) \leq f(w(k,i)) \leq f(x^k) \quad \text{for } i = 1, \ldots, m. \tag{43}$$

Now, let $\bar{x}$ be the limit point of some subsequence $\{x^k\}_K$. Then, reasoning as in the proof of Theorem 3.3 we get the limits

$$\lim_{k\to\infty} (f(x^k) - f(w(k,i))) = 0, \quad i = 1, \ldots, m. \tag{44}$$

Therefore, we can find an infinite subset $K_1 \subseteq K$ such that $I^k = \hat{I}$ for all $k \in K_1$. By Condition 2 (where we identify, for each $i \in \hat{I}$, the sequence $\{y^k\}$ with $\{x^k\}_{K_1}$) we obtain

$$\nabla_i f(\bar{x}) = 0, \quad i \in \hat{I},$$

which implies, together with (41), that $\nabla f(\bar{x}) = 0$.

The last assertion follows from the compactness of $\mathcal{L}$ and the continuity assumptions on ∇f.

In the next theorem we consider the case of unbounded level sets.

THEOREM 3.6 *Suppose that the assumptions of Theorem* 3.5 *are satisfied and that the mappings* T_i *satisfy Condition* 4 *for every* $i = 1, \ldots, m$. *Assume also that* f *is bounded below. Then we have*:

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

Proof The proof is similar to that of the preceding theorem. We must only note that (44) now follows from the assumptions made and the boundedness of f. Then the conclusion is established using Condition 4 in place of Condition 2.

As in the case of sequential connection, setting $I^k = \{1, \dots, m\}$ for all k, we have that condition (41) holds.

Condition (41) can also be satisfied, for instance, by choosing $I^k = \{i^*\}$, where:

$$i^* = \arg\max_i \{\|\nabla_i f(x^k)\| \; i = 1, \dots, m\}.$$

In fact, we have:

$$\|\nabla_{i^*} f(x^k)\| \geq \max_i\{\|\nabla_i f(x^k)\|\} \frac{1}{m} \sum_{i=1}^{m} \|\nabla_i f(x^k)\|,$$

so that (41) holds with $\sigma(t) = t/m$. This rule can be viewed as an extension of the Gauss–Southwell algorithm.

We observe also that in case of parallel decomposition we no more need Conditions 3 or 5 and that the continuity requirements on ∇f can be weakened.

4 LINE SEARCH MAPPING

In this section we define a mapping $T_i : R^n \to R^{n_i}$, by means of a line search technique, under the assumption that a gradient related search direction is employed, and we show that T_i satisfies the conditions of the preceding section.

Given a sequence $\{y^k\}$ in R^n we suppose that we can compute search directions $d_i^k \in R^{n_i}$ that satisfy the following assumption.

Assumption 1 Let $\{y^k\}$ be a given sequence in R^n. Then:

(i) $d_i^k = 0$ if and only if $\nabla_i f(y^k) = 0$;
(ii) there exists a forcing function $\sigma_i : R^+ \to R^+$ such that:

$$\frac{\nabla_i f(y^k)^{\mathrm{T}} d_i^k}{\|d_i^k\|} \leq -\sigma_i(\|\nabla_i f(y^k)\|) \tag{45}$$

for all k satisfying $\nabla_i f(y^k) \neq 0$.

We note that Assumption 1 is satisfied, in particular, if we set $\sigma_i(t) = t$ and:

$$d_i^k = -\nabla_i f(y^k).$$

Given a sequence $\{y^k\}$ in R^n, we define, for each k, a line search mapping by letting

$$T_i(y^k) = y_i^k + \alpha_i^k d_i^k,$$

where a_i^k is a number computed by means of some line search technique.

Line search algorithms with constant stepsizes have been used, in the context of decomposition methods, under Lipschitz continuity assumption on the gradient $\nabla_i f$ (see, for instance, [4,23]).

Here we define an Armijo-type algorithm, in which a sufficient decrease of f is enforced through line search rules of the kind studied in [7,12], in connection with no-derivative methods.

The algorithm does not depend on the knowledge of Lipschitz constants and allows us to establish that the convergence of the function values implies the limit

$$\lim_{k\to\infty} \alpha_i^k \|d_i^k\| = 0,$$

even in the absence of boundedness assumptions on $\{y^k\}$ and on $\{d_i^k\}$.

Line search algorithm (LS)

Data. $\rho_i > 0$, $\gamma_i > 0$, $\delta_i \in (0, 1)$.

Step 1. Choose $\Delta_i^k \geq \rho_i |d_i^{kT} \nabla_i f(y^k)| / \|d_i^k\|^2$,

Step 2. Compute $\alpha_i^k = \max_j \{\delta_i^j \Delta_i^k : j = 0, 1, \ldots\}$ such that

$$f(y_1^k, \ldots, y_i^k + \alpha_i^k d_i^k, \ldots, y_m^k) \leq f(y^k) - \gamma_i (\alpha_i^k)^2 \|d_i^k\|^2. \quad (46)$$

We note that conditions at Steps 1 and 2 can also be replaced with more practical line search rules. However, since we are not concerned here with computational implementations, we will refer to the simpler description given above.

In the next proposition we state some useful properties of Algorithm LS. Note that in what follows we assume that $\{y^k\}$ is a given sequence that may not depend on Algorithm LS, in the sense that y_i^{k+1} is not necessarily the result of a line search along d_i^k.

PROPOSITION 4.1 *Suppose that* $\nabla_i f(y^k) \neq 0$ *and that* d_i^k *satisfies Assumption* 1. *Then*:

(i) *there exists a finite integer* j *such that the number* $\alpha_i^k = \delta_i^j \Delta_i^k$ *satisfies the acceptability condition at Step* 2;
(ii) *there exist numbers* $\lambda_i^k \in [0, 1)$ *and* $\theta_i > 0$ *such that*

$$\alpha_i^k \|d_i^k\| \geq -\theta_i \frac{d_i^{k\mathrm{T}} \nabla_i f(\tilde{y}^k)}{\|d_i^k\|},$$

where:

$$\tilde{y}^k := \left(y_1^k, \ldots, y_i^k, \lambda_i^k \frac{\alpha_i^k}{\delta_i} d_i^k, \ldots, y_m^k\right).$$

Proof In order to prove assertion (i), let us assume, by contradiction, that $\nabla_i f(y^k) \neq 0$ and that condition (46) is violated for every $j \geq 0$, so that

$$\frac{f(y_1^k, \ldots, y_i^k + \delta_i^j \Delta_i^k d_i^k, \ldots, y_m^k) - f(y^k)}{\delta_i^j \Delta_i^k} - \gamma_i \delta_i^j \Delta_i^k \|d_i^k\|^2.$$

Then, taking limits for $j \to \infty$ we obtain:

$$\nabla_i f(y^k)^{\mathrm{T}} d_i^k \geq 0,$$

which contradicts the assumption $\nabla_i f(y^k)^{\mathrm{T}} d_i^k < 0$.

Now, to prove assertion (ii), let us distinguish the two cases $\alpha_i^k = \Delta_i^k$ and $\alpha_i^k < \Delta_i^k$, where Δ_i^k is the number defined at Step 1. In the first case, we have obviously:

$$\alpha_i^k \|d_i^k\| \geq \rho_i \frac{|\nabla_i f(y^k)^{\mathrm{T}} d_i^k|}{\|d_i^k\|} = -\rho_i \frac{\nabla_i f(\tilde{y}^k)^{\mathrm{T}} d_i^k}{\|d_i^k\|}, \tag{47}$$

where we set $\tilde{y}^k = y^k$ and $\lambda_i^k = 0$. Therefore, (ii) holds with $\lambda_i^k = 0$ and $\theta_i = \rho_i$.

Next suppose that $\alpha_i^k < \Delta_i^k$, so that α_i^k/δ_i violates the condition at Step 2 of Algorithm LS. In this case we can write, using the Mean Value Theorem:

$$f(y^k) + \frac{\alpha_i^k}{\delta_i} \nabla_i f(\tilde{y}^k)^{\mathrm{T}} d_i^k > f(y^k) - \gamma_i \left(\frac{\alpha_i^k}{\delta_i}\right)^2 \|d_i^k\|^2,$$

where $\lambda_i^k \in (0, 1)$ and

$$\tilde{y}^k := (y_i^k, \ldots, y_i^k + \lambda_i^k \frac{\alpha_i^k}{\delta_i} d_i^k, \ldots, y_m^k), \tag{48}$$

so that we obtain:

$$\alpha_i^k > -\left(\frac{\delta_i}{\gamma_i}\right) \frac{\nabla_i f(\tilde{y}^k)^{\mathrm{T}} d_i^k}{\|d_i^k\|^2} \tag{49}$$

and hence assertion (ii) holds with $\theta_i = \delta_i/\gamma_i$. This concludes the proof.

The next proposition allows us to show that the line search mapping T_i obtained by using Algorithm LS satisfies Conditions 1–3 and 5 of the preceding section.

PROPOSITION 4.2 *Let $\{y^k\}$ be a given sequence in R^n, let $\{d_i^k\}$ be a sequence of vectors, such that Assumption* 1 *is satisfied. Let α_i^k be computed by means of Algorithm* LS *when $\nabla_i f(y^k) \neq 0$ and set $\alpha_i^k = 0$ whenever $\nabla_i f(y^k) = 0$. Then:*

(i) *the limit*

$$\lim_{k\to\infty} f(y^k) - f(y_1^k, \ldots, y_i^k + \alpha_i^k d_i^k), \ldots, y_m^k) = 0 \tag{50}$$

implies

$$\lim_{k\to\infty} \alpha_i^k \|d_i^k\| = 0;$$

(ii) *the line search mapping T_i satisfies Conditions* 1–3 *and* 5 *of Section* 3.

Proof Let us define the point:

$$v^k := (y_1^k, \dots, y_i^k + \alpha_i^k d_i^k, \dots, y_m^k).$$

Then the acceptance rule of Algorithm LS ensures that:

$$f(y^k) - f(v^k) \geq \gamma_i (\alpha_i^k)^2 \|d_i^k\|^2,$$

so that the limit

$$\lim_{k\to\infty} f(y^k) - f(v^k) = 0$$

implies assertion (i) and also proves that Conditions 1, 3 and 5 of Section 3 holds.

Now, suppose that y^k converges to $\bar{y}$. We can assume that $\nabla_i f(y^k) \neq 0$, for all sufficiently large k since, otherwise, the continuity of $\nabla_i f$ and the convergence of the sequence would imply $\nabla_i f(\bar{y}) = 0$.

By Assumption 1 it follows that $d_i^k \neq 0$ for all large k and hence, recalling assertion (ii) of Proposition 4.1, from (i) we have:

$$\lim_{k\to\infty} \frac{\nabla_i f(\tilde{y}^k)^{\mathrm{T}} d_i^k}{\|d_i^k\|} = 0, \tag{51}$$

where

$$\tilde{y}^k := (y_1^k, \dots, y_i^k + \lambda_i^k (\alpha_i^k / \delta_i) d_i^k, \dots, y_m^k), \tag{52}$$

for some $\delta_i > 0$ and $\lambda_i^k \in [0, 1)$.

Again by (i) we have:

$$\lim_{k\to\infty} \tilde{y}^k = \bar{y},$$

and therefore, as $\{d_i^k / \|d_i^k\|\}$ is bounded, we can find an infinite index set K such that

$$\lim_{k\to\infty,\, k\in K} \frac{d_i^k}{\|d_i^k\|} = \bar{d}_i,$$

for some $\bar{d}_i \in R^{n_i}$. Thus, by (51), taking limits for $k \to \infty$, $k \in K$ we obtain

$$\nabla_i f(\bar{y})^{\mathrm{T}} \bar{d}_i = 0. \tag{53}$$

On the other hand, we have also

$$\lim_{k\to\infty,\, k\in K} \frac{\nabla_i f(y^k)^{\mathrm{T}} d_i^k}{\|d_i^k\|} = \nabla_i f(\bar{y})^{\mathrm{T}} \bar{d}_i, \tag{54}$$

so that by (53) and (54) we obtain

$$\lim_{k\to\infty,\, k\in K} \frac{\nabla_i f(y^k)^{\mathrm{T}} d_i^k}{\|d_i^k\|} = 0.$$

Then Assumption 1 and the definition of forcing function implies

$$\lim_{k\to\infty} \nabla_i f(y^k) = \nabla_i f(\bar{y}) = 0,$$

which establishes Condition 2 of Section 3 and concludes the proof.

Remark 1 We note that, in general, a global one-dimensional minimizer α_i^{*k} along $\{d_i^k\}$ may be not acceptable for Algorithm LS.

However, if the objective function f is twice continuously differentiable and is a strongly convex function of x_i, when the other component vectors are held constant, we have that the assertions of Proposition 4.2 hold if we compute $\{\alpha_i^k\}$ through an exact line search.

More precisely, suppose that the following assumption is satisfied.

ASSUMPTION 2 The function f is twice continuously differentiable on an open bounded convex set $\mathcal{C} \supset \mathcal{L}$ and for every $x \in \mathcal{C}$ the Hessian matrix $\nabla_i^2 f(x)$ satisfies the condition

$$y^{\mathrm{T}} \nabla_i^2 f(x) y \geq \lambda \|y\|^2 \quad \text{for all } y \in R^{n_i} \tag{55}$$

for some $\lambda_i > 0$.

Let α_i^{*k} be the unique one-dimensional minimizer of the function

$$f(y_1^{k+1}, \ldots, y_i^k + \alpha_i d_i^k, \ldots, y_m^k)$$

with respect to α_i.

Then, can write:

$$\nabla_i f(y_1^{k+1}, \ldots, y_i^k + \alpha_i^{*k} d_i^k, \ldots, y_m^k)^{\mathrm{T}} d_i^k = 0. \tag{56}$$

Therefore, using Taylor's theorem we obtain:

$$f(y^k) = f(y_1^{k+1}, \ldots, y_i^k + \alpha_i^{*k} d_i^k, \ldots, y_m^k) + \tfrac{1}{2}(\alpha_i^{*k})^2 (d_i^k)^{\mathrm{T}} \nabla_i^2 f(v_i^k) d_i^k, \tag{57}$$

where $v_i^k = y_i^{k+1} + \zeta_i^k \alpha_i^{*k} d_i^k$ and $\zeta_i^k \in (0, 1)$.

Thus, from (57) we get, by the strong convexity assumption on f

$$f(y_1^{k+1}, \ldots, y_i^k + \alpha_i^{*k} d_i^k, \ldots, y_m^k) \le f(y^k) - \lambda_i (\alpha_i^{*k})^2 \|d_i^k\|^2$$

for some $\lambda_i > 0$.

Therefore, by (50) it follows that assertion (i) holds, which also allows us to prove, taking $\gamma_i = \lambda_i$ and reasoning as in the proof of Proposition 4.2, that assertion (ii) holds.

In order to prove that Condition 4 of Section 3 is valid, we must introduce stronger requirements on $\nabla_i f$.

This is the object of the next proposition.

PROPOSITION 4.3 *Let $\{y^k\}$ be a sequence in R^n, let d_i^k, α_i^k be defined as in Proposition 4.2 and define the points*:

$$v^k := (y_1^k, \ldots, y_i^k + \alpha_i^k d_i^k, \ldots, y_m^k).$$

Suppose that

$$\lim_{k\to\infty} \|\nabla f(u^k) - \nabla f(w^k)\| = 0, \tag{58}$$

whenever $\{u^k\}$, $\{w^k\}$ are sequences in R^n such that

$$\lim_{k\to\infty} \|u^k - w^k\| = 0.$$

Then the limit

$$\lim_{k\to\infty} f(y^k) - f(v^k) = 0$$

implies:

$$\lim_{k\to\infty} \nabla_i f(y^k) = 0,$$

so that the line search mapping T_i satisfies Condition 4 *of Section* 3.

Proof Suppose that the assertion is false. This implies that there exists a subsequence $\{y^k\}_K$ such that

$$\|\nabla_i f(y^k)\| \geq \varepsilon, \qquad (59)$$

for all $k \in K$ and some $\varepsilon > 0$.

By (i) of Proposition 4.2 we have:

$$\lim_{k\to\infty} \alpha_i^k \|d_i^k\| = 0 \qquad (60)$$

and therefore, by assertion (ii) of Proposition 4.1, we get:

$$\lim_{k\to\infty,\, k\in K} \frac{\nabla_i f(\tilde{y}^k)^{\mathrm{T}} d_i^k}{\|d_i^k\|} = 0, \qquad (61)$$

where $\tilde{y}^k$ is defined in (52). By (52) and (60) we get

$$\lim_{k\to\infty} \|\tilde{y}^k - y^k\| = 0. \qquad (62)$$

On the other hand, we can write

$$\frac{\nabla_i f(y^k)^{\mathrm{T}} d_i^k}{\|d_i^k\|} = \frac{\nabla_i f(\tilde{y}^k)^{\mathrm{T}} d_i^k}{\|d_i^k\|} + \frac{(\nabla_i f(\tilde{y}^k - \nabla_i f(y^k)^{\mathrm{T}} d_i^k)}{\|d_i^k\|},$$

so that by (61) and (62), recalling (58), we obtain

$$\lim_{k\to\infty,\, k\in K} \frac{\nabla_i f(y^k)^{\mathrm{T}} d_i^k}{\|d_i^k\|} = 0.$$

Then, from Assumption 1 we get a contradiction to the assumption (59) and this proves our thesis.

5 MINIMIZATION MAPPING

In this section we analyze the properties of the minimization mapping $T_i: R^n \to R^{n_i}$, defined as

$$T_i(y) = \arg\min_{\xi_i \in R^{n_i}} f(y_1, \ldots, \xi, \ldots, y_m). \tag{63}$$

First of all we prove that the minimization mapping satisfies Conditions 1 and 2 of Section 3, provided that the problem (63) has a solution and that a well defined solution is chosen.

PROPOSITION 5.1 *Suppose that the minimization mapping T_i is well defined. Then, it satisfies Conditions* 1 *and* 2 *of Section* 3. *Moreover, under the assumption of Proposition* 4.3, *the mapping T_i satisfies Condition* 4.

Proof Let $\{y^k\}$ be any sequence in R^n and define a line search mapping $T_i^{(\mathrm{LS})}$ by computing $d_i^k = -\nabla_i f(y^k)$ and letting:

$$T_i^{(\mathrm{LS})} = y_i^k + \alpha_i^k d_i^k,$$

where α_i^k is obtained by means of Algorithm LS. Then we have:

$$f(y_1^k, \ldots, T_i(y^k), \ldots, y_m^k) \le f(y_1^k, \ldots, T_i^{(\mathrm{LS})}(y^k), \ldots, y_m^k).$$

By Proposition 4.2, the mapping $T_i^{(\mathrm{LS})}$ satisfies Conditions 1 and 2. Therefore, from Proposition 3.1 we get that T_i satisfies Conditions 1 and 2.

Under the assumption of Proposition 4.3, we have that Condition 4 holds for $T_i^{(\mathrm{LS})}$, and then, invoking again Proposition 3.1, we have that T_i satisfies Condition 4.

Now, in order to prove that the minimization mapping satisfies also Condition 3, we must introduce strict (generalized) convexity hypotheses on f. More formally, we suppose that the following assumption holds.

ASSUMPTION 3 For every $x \in R^n$ and $y_i \in R^{n_i}$ such that $y_i \neq x_i$ we have for all $t \in (0, 1)$

$$f(x_1, \ldots, tx_i + (1-t)y_i, \ldots, x_m) < \max\{f(x), f(x_1, \ldots, y_i, \ldots, x_m)\}.$$

Then, we state the following proposition, whose proof requires only minor adaptations of the arguments used, for instance, in [4,17]; however the proof is reported in detail for completeness.

PROPOSITION 5.2 *Suppose that the minimization mapping T_i is well defined and that Assumption 3 holds. Then, it satisfies Conditions 1–3 of Section 3.*

Proof By Proposition 5.1, the minimization mapping satisfies Conditions 1 and 2. Now, in order to prove that it also satisfies Condition 3, let us consider a sequence $\{y^k\}$ converging to $\bar{y}$, and let us define the point:

$$v^k := (y_1^k, \dots, T_i(y^k), \dots, y_m^k).$$

Reasoning by contradiction, let us assume that Condition 3 does not hold, and hence that there exist an infinite subset K and a number $\beta > 0$ such that

$$\|v^k - y^k\| = \|T_i(y^k) - y_i^k\| \geq \beta \quad \text{for all } k \in K. \tag{64}$$

For $k \in K$, let $s^k = (v^k - y^k)/\|v^k - y^k\|$ and choose a point in the segment joining y^k and v^k, by assuming:

$$\tilde{v}^k := y^k + \lambda\beta s^k,$$

with $\lambda \in (0,1)$.

As $\|s^k\| = 1$, we can find an infinite subset $K_1 \subseteq K$ such that:

$$\lim_{k\to\infty,\, k\in K_1} y^k = \lim_{k\to\infty,\, k\in K_1} (y_1^k, y_2^k, \dots, y_m^k) = (\bar{y}_1, \dots, \bar{y}_i, \dots, \bar{y}_m) = \bar{y}$$

and

$$\begin{aligned} \lim_{k\to\infty,\, k\in K_1} \tilde{v}^k &= \lim_{k\to\infty,\, k\in K_1} (y_1^k, \dots, y_i^k + \lambda\beta s^k, \dots, y_m^k) \\ &= (\bar{y}_1, \dots, y_i^*, \dots, \bar{y}_m) = y^* \end{aligned}$$

with

$$\|\bar{y}_i - y_i^*\| = \lambda\beta > 0.$$

On the other hand, by Assumption 3, we have for all $t \in [0,1]$

$$f(y^k) \geq f(y_1^k, \ldots, (1-t)y_i^k + t(y_i^k + \lambda\beta s_i^k), \ldots, y_m^k) \geq f(v^k). \quad (65)$$

Suppose now, as stated in the assertion, that

$$\lim_{k\to\infty} f(y^k) - f(v^k) = 0. \quad (66)$$

Then, taking limits in (65), for $k \to \infty$, $k \in K_1$, by (66) and the continuity of f we get

$$f(\bar{y}) = f(\bar{y}_1, \ldots, (1-t)\bar{y}_i + ty_i^*, \ldots, \bar{y}_m) \quad \text{for all } t \in [0,1],$$

which contradicts Assumption 3. This concludes the proof.

6 CONVERGENCE OF THE GAUSS–SEIDEL METHOD

In this section we make use of the conditions established in the preceding sections for studying the convergence properties of the minimization version of the *block nonlinear Gauss–Seidel* (GS) method for the unconstrained minimization of f.

In particular, we reobtain some of the well known convergence results based on convexity assumptions and we show that the two-block version of this method is globally convergent towards stationary points of f, without any convexity hypothesis.

First we state the m-block GS method in the following form.

GS Method

Data. $x^0 \in R^n$.

Step 0. Set $k = 0$.

Step 1. Set $z(k,1) = x^k$. For $i = 2, \ldots, m$:

set

$$x_i^{k+1} = \arg\min_{\xi \in R^{n_i}} f(x_1^{k+1}, \ldots, \xi, \ldots, x_m), \quad (67)$$

and

$$z(k, i+1) = (x_1^{k+1}, \ldots, x_i^{k+1}, x_{i+1}^k, \ldots, x_m^k).$$

Step 2. Set

$$x^{k+1} = (x_1^{k+1}, \dots, x_m^{k+1}),$$

$k = k + 1$ and go to Step 1.

Under the hypotheses that the objective function is pseudoconvex and that the level set $\mathcal{L}$ is compact, convergence of the preceding algorithm is known in the particular case of $n_i = 1$, for $i = 1, \dots, m$ (see [25]).

The extension of this result can be obtained by using Proposition 5.1 and Theorem 3.2, which allow us to state the following theorem.

THEOREM 6.1 *Suppose that f is pseudoconvex and that $\mathcal{L}$ is compact. Then every limit point of the sequence generated by algorithm GS is a global minimizer of f.*

In [4,22], where the more general case of problems with convex constraints is considered, it has been proved that if f is convex and if, for each i, f is a strictly convex function of x_i, when the other components of x are held constant, then algorithm GS generates an infinite sequence such that every limit point of $\{x^k\}$ minimizes f on R^n. (Note that the compactness assumption on $\mathcal{L}$ is not required.)

From the proof of this result (see, for instance, [4], pp. 220–221), it is easy to see that, in order to prove convergence towards stationary points, the convexity assumption on f is not required and we only need to assume the strict convexity of f with respect to each component.

We can reobtain a slightly improved version of this result, noting that algorithm GS can be viewed as the sequential interconnection of the minimization mappings T_i defined by:

$$T_i(y) = \arg \min_{\xi_i \in R^{n_i}} f(y_1, \dots, \xi, \dots, y_m).$$

Using Theorem 3.3, where we set $I^k = \{1, \dots, m\}$ and $\nu = 0$, and recalling Proposition 5.2 we get immediately the following theorem.

THEOREM 6.2 *Suppose that, for each $i \in \{1, \dots, m\}$ the minimization mapping T_i is well defined and that the function f is strictly quasiconvex with respect to x_i, when the other components of x are held constant, that is, suppose that Assumption 3 holds for all i. Then, algorithm GS* *generates an infinite sequence* $\{x^k\}$ *such that:*

(i) *every limit point of* $\{x^k\}$ *is a stationary point of* f;
(ii) *if* $\mathcal{L}$ *is compact we have* $\lim_{k\to\infty} \nabla f(x^k) = 0$ *and there exists at least one limit point that is a stationary point of* f.

The rate of convergence of the Gauss–Seidel method, under the assumption that f is componentwise strongly convex (but not convex), has been studied in [15].

Now, let us consider the special case when $m=2$, that is let us describe our problem in the following form:

$$\underset{(x_1,x_2)\in R^{n_1}\times R^{n_2}}{\text{minimize}} \quad fx) = f(x_1, x_2) \tag{68}$$

Although this situation is quite particular in our setting, it includes many intersting applications of the block GS method. In fact, in many cases, a two-block decomposition yields subproblems of special structure, and often allows us to adopt parallel techniques for solving one subproblem.

We can prove the global convergence of the GS method, which we call 2Block GS algorithm, without imposing any convexity assumption on f. To see how this is possible, we may observe that the convexity assumption in Theorem 6.2 is needed for ensuring that the stepsizes go to zero, that is for ensuring satisfaction of Condition 3, through the result stated in Proposition 5.2. However, when $m=2$ we can equivalently represent the GS method in a way that Theorem 3.5 can be invoked, and this avoids the need of imposing conditions on the stepsizes. In fact, we reobtain the 2Block GS algorithm from the parallel connection scheme by assuming

$$w(k,1) = (T_1(x^k), x_2^k), \qquad w(k,2) = (x_1^k, x_2^k),$$

and by computing x^{k+1} through the minimization of f with respect to x_2, that is by letting

$$x^{k+1} = (T_1(x^k), T_2(w(k,1))),$$

where T_2 is the minimization mapping with respect to x_2. This allows us to state the following result.

THEOREM 6.3 *Suppose that the global minimization with respect to each component is well defined. Then, the 2Block GS method generates an infinite sequence* $\{x^k\}$ *such that*

(i) *every limit point of* $\{x^k\}$ *is a stationary point of f;*
(ii) *if* $\mathcal{L}$ *is compact we have* $\lim_{k\to\infty} \nabla f(x^k) = 0$ *and there exist at least one limit point that is a stationary point of f;*
(iii) *if f is pseudoconvex on* R^n, *every limit point is a global minimizer of f.*

Proof First we observe that, for all $k \geq 1$ we have:

$$x_2^k = \arg \min_{\xi \in R^{n_2}} f(x_1^k, \xi),$$

so that

$$\nabla_2 f(x^k) = 0.$$

Now, let us assume $I^k = \{1\}$ for all $k \geq 1$, so that we can consider the points

$$w(k,1) = (T_1(x^k), x_2^k), \qquad w(k,2) = (x_1^k, x_2^k),$$

where T_1 is the minimization mapping, which was assumed to be well defined. By Proposition 5.1, Conditions 1 and 2 of Section 3 are satisfied for $i = 1$.

On the other hand, we have

$$f(w(k,1)) \leq f(x^k), \qquad f(w(k,2)) = f(x^k)$$

and hence, as x^{k+1} is obtained by minimizing f with respect to x_2, starting from $w(k,1)$, we can write:

$$f(x^{k+1}) \leq f(w(k,1)) \leq f(x^k) = f(w(k,2)).$$

As $\nabla_2 f(x^k) = 0$, we have that condition (41) holds with $\sigma(t) = t$ so that the assumptions of Theorem 3.5 are satisfied and (i) follows from this theorem.

Then, assertions (ii) and (iii) are obvious consequences of (i).

By comparing Theorem 6.3 with Theorems 6.1 and 6.2 the important point to observe is that for $m = 2$ we require neither the pseudoconvexity hypothesis on f nor the strict convexity assumption on f as a function of each subvector.

When the level set α is unbounded we can state the following result.

THEOREM 6.4 *Suppose that the global minimization with respect to each component is well defined and suppose that f is bounded below and that*

$$\lim_{k\to\infty} \|\nabla f(u^k) - \nabla f(w^k)\| = 0,$$

whenever $\{u^k\}$, $\{w^k\}$ are sequences in R^n such that

$$\lim_{k\to\infty} \|u^k - w^k\| = 0.$$

Then

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

Proof Reasoning as in the proof of Theorem 6.3 it can be easily verified that Proposition 5.1 implies that the mapping T_1 satisfies Conditions 1 and 4. Thus the assertion follows from Theorem 3.6.

From the proof of Theorem 6.3, we can observe that the global minimization of f with respect to x_2 has the only motivation of providing a stationary point in the component subspace x_2, without increasing the objective function value. Therefore, we can obtain the same convergence results by replacing the global minimization with respect to x_2 with the computation of a point x_2^{k+1} such that

$$f(x_1^{k+1}, x_2^{k+1}) \le f(x_1^{k+1}, x_2^k) \quad \text{and} \quad \nabla_2 f(x_1^{k+1}, x_2^{k+1}) = 0.$$

Then, we can define the following scheme, which is a modified version of the 2Block GS method.

Modified 2Block GS Method

Data. $x^0 \in R^n$.
Step 0. Set $k = 0$.
Step 1. For $i = 1, 2$:
if $i = 1$ then set

$$x_1^{k+1} = \arg\min_{x_1} f(x_1, x_2^k); \tag{69}$$

if $i=2$ then *determine* x_2^{k+1} such that

$$f(x_1^{k+1}, x_2^{k+1}) \le f(x_1^{k+1}, x_2^k) \quad \text{and} \quad \nabla_2 f(x_1^{k+1}, x_2^{k+1}) = 0.$$

Step 2. Set

$$x^{k+1} = (x_1^{k+1}, x_2^{k+1}),$$

$k = k+1$ and go to Step 1.

Now we may ask whether similar convergence results can be obtained for the GS method when $m > 2$. In this case a negative result was established by Powell for $m = 3$. In fact, let us consider the following objective function [20]:

$$\begin{aligned} f(x) = &- x_1x_2 - x_2x_3 - x_1x_3 + (x_1 - 1)_+^2 + (-x_1 - 1)_+^2 \\ &+ (x_2 - 1)_+^2 + (-x_2 - 1)_+^2 + (x_3 - 1)_+^2 + (-x_3 - 1)_+^2, \end{aligned} \tag{70}$$

where

$$(t - c)_+^2 = \begin{cases} 0 & \text{if } t \le c, \\ (t - c)^2 & \text{if } t \ge c. \end{cases}$$

Powell showed that, if the starting point x^0 is the point $(-1-\epsilon, 1 + 1/2\epsilon, -1-1/4\epsilon)$ the steps of the GS method "tend to cycle round six edges of the cube whose vertices are $(\pm 1, \pm 1, \pm 1)$" and "on the limiting path the gradient vector of the objective function is bounded away from zero". This implies that the GS method generates a sequence which admits accumulation points that are not stationary points of f.

7 GLOBALLY CONVERGENT LINE-SEARCH-BASED ALGORITHMS

In the previous section we have seen that the Gauss–Seidel algorithm is guaranteed to converge either under suitable convexity assumptions on the objective function or in the case of a two-block decomposition.

Here we propose a new block-coordinate descent algorithm, whose convergence can be proved without convexity assumptions or restrictions on the number m of blocks. It is based on the line search technique described in Section 4, which does not require the knowledge of the Lipschitz constant to compute the stepsize aloug suitable search directions.

More specifically, we consider a sequential algorithm with $I^k = \{1,\dots,m\}$ and $\nu = 0$, where each elementary operation $x_i^{k+1} = T_i(z(k,i))$ is implicitly defined by means of a two-phases procedure. In the first phase we perform an inexact line search along a search direction d_i^k, which yields reference values for the objective function and the stepsize. In the second phase we compute a further updating of the current component by means of any minimization method in the component subspace, provided that suitable acceptability conditions are satisfied.

In the following conceptual model we assume that, for $i = 1,\dots,m$, the directions $d_i^k \in R^{n_i}$ satisfy Assumption 1, the numbers $\gamma_i > 0$ are parameters used in Algorithm LS, and $\{\xi_i^k\}$ are sequences of positive numbers converging to zero.

Algorithm 1

Data. $x^0 \in R^n$, numbers $\tau_i \geq 1/\gamma_i$. For $i = 1,\dots,m$.

Step 0. Set $k = 0$.

Step 1. Set $z(k,1) = x^k$. For $i = 1,\dots,m$:

(a) compute α_i^k by means of Algorithm LS (with $\alpha_i^k = 0$ if $\nabla_i f(z(k,i)) = 0$);

(b) choose x_i^{k+1} such that the following conditions are satisfied:

$$f(x_1^{k+1},\dots,x_i^{k+1},\dots,x_m^k) \leq f(x_1^{k+1},\dots,x_i^k + \alpha_i^k d_i^k,\dots,x_m^k); \tag{71}$$

$$\|x_i^{k+1} - x_i^k\|^2 \leq \tau_i \max\{\xi_i^k, \Delta f_i^k\}, \tag{72}$$

where

$$\Delta f_i^k = f(z(k,i)) - f(x_1^{k+1},\dots,x_i^{k+1},\dots,x_m^k);$$

(c) set

$$z(k, i+1) = (x_1^{k+1}, \ldots, x_i^{k+1}, x_{i+1}^k, \ldots, x_m^k).$$

Step 2. Set

$$x^{k+1} = (x_1^{k+1}, \ldots, x_m^{k+1}),$$

$k = k+1$ and go to Step 1.

In the conceptual model defined in Algorithm 1 we have not specified the method used for generating the point x_i^{k+1} starting from the knowledge of x_i^k. The conditions imposed at Step 1(b) have essentially the role of guaranteeing a sufficient decrease of f at each step and also that of ensuring that the limit $\|x_i^{k+1} - x_i^i\| \to 0$ is attained. Note that these conditions can be satisfied, for instance, by assuming

$$x_i^{k+1} = x_i^k + \alpha_i^k d_i^k.$$

In fact, it is easily verified that the acceptance rules of Algorithm LS eusure that the conditions of Step 1(b) hold. In this case Algorithm 1 reduces to a sequence of line searches along the directions d_i^k. We observe also that in the special case where $n_i = 1$ and $m = n$ we obtain a coordinate descent method with a suitable inexact line search (see, for instance, [12]).

As regards the convergence properties of Algorithm 1, we can state the following theorems.

THEOREM 7.1 *Let $\{x^k\}$ be the sequence generated by Algorithm* 1. *Then, every limit point of $\{x^k\}$ is a stationary point of f. Moreover, if $\mathcal{L}$ is compact, we have*

$$\lim_{k \to \infty} \nabla f(x^k) = 0,$$

and there exists at least one limit point that is a stationary point of f.

Proof It is easily seen that Algorithm 1 can be represented as the sequential interconnection of elementary mappings that update the components of x^k, according to the instructions specified at Step 1. Therefore, the assertions follow from Theorem 3.3 for $\nu = 0$ and $I^k = \{1, \ldots, m\}$, provided that we can prove that the mappings T_i

implicitly defined by the algorithm, for $i=1,\dots,m$, satisfy Conditions 1–3 of Section 3.

In order to show this, let $\{y^k\}$ be a given sequence in R^n and for each $i\in\{1,\dots,m\}$, let $\tilde{T}_i$ be the line search mapping defined by

$$\tilde{T}_i(y^k) = y_i^k + \alpha_i^k d_i^k,$$

where α_i^k is computed by means of Algorithm LS.

By Proposition 4.2 we have that $\tilde{T}_i$ satisfies Conditions 1–3 and 5 of Section 3.

Now consider the mapping T_i defined at Step 1(b) of Algorithm 1, which determines the point $T_i(y^k)$ in a way that

$$f(y_1^k,\dots,T_i(y^k),\dots,y_m^k) \le f(y_1^k,\dots,\tilde{T}_i(y^k),\dots,y_m^k),$$

and at least one of the following conditions is satisfied:

$$\|T_i(y^k) - y_i^k\| \le \tau_i(f(y^k) - f(y_1^k,\dots,T_i(y^k),\dots,y_m^k)), \tag{73}$$

$$\|T_i(y^k) - y_i^k\| \le \tau_i\xi_i^k \tag{74}$$

(where $\xi_i^k \to 0$ for $k\to\infty$). Then, by Proposition 3.1, it follows that the mapping T_i satisfies Conditions 1 and 2.

Now, if there exists a $\bar{k}$ such that condition (73) holds for all $k\ge\bar{k}$, then we have immediately that T_i satisfies also Condition 3. Therefore, let us assume that there exists an infinite subset $K_1 \subseteq \{0,1,\dots,\}$ such that, for all $k=K_1$, only condition (74) is satisfied. As $\xi_i^k\to 0$ for $k\to\infty$, from (74) it follows that

$$\lim_{k\to\infty,\, k\in K_1} \|T_i(y^k) - y_i^k\| = 0,$$

so that we can conclude that the mapping T_i satisfies also Condition 3.

Therefore the assertions follow from Theorem 3.3.

THEOREM 7.2 *Let $\{x^k\}$ be the sequence generated by Algorithm* 1. *Suppose that f is bounded below and that*

$$\lim_{k\to\infty} \|\nabla f(u^k) - \nabla f(w^k)\| = 0,$$

whenever $\{u^k\}$, $\{w^k\}$ *are sequences in* R^n *such that*

$$\lim_{k\to\infty} \|u^k - v^k\| = 0.$$

Then we have

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

Proof We can repeat the same reasonings used in the proof of Theorem 7.1, and then, using Theorem 3.4, we get the thesis.

We observe that the line search procedure prevents the occurrence of cycling that was evidenced by Powell [20] on problem (70), in connection with the coordinate descent method with exact line searches.

It can be easily verified that the level sets of the objective function in problem (70) are not compact; in fact, setting $x_2 = x_3 = x_1$ we can see that $f(x) \to -\infty$ as $\|x\| \to \infty$. Then, the convergence of Algorithm 1 is not guaranteed. However, if we use an inexact line search based on Algorithm LS it is easily seen from the preceding results that for every i the numbers $\alpha_i^k \|d_i^k\|$ would converge to zero if $\{f(x^k)\}$ is bounded and cycling cannot occur. In this problem, using a simple implementation of Algorithm LS, we obtained the results reported below where the unboundedness of f is eventually detected:

k	0	10	20	200
$f(x^k)$	1.0	−109.3	−228.2	−2309.7

Similar considerations can be repeated in connection with the other examples of [20].

As an alternative to the line search version of Algorithm 1, a different possibility is that of starting (either from x_i^k or from $x_i^k + \alpha_i^k d_i^k$) an unconstrained minimization of f with respect to x_i, holding fixed the remaining components.

We can adopt, in principle, any solution technique in the x_i-space, taking into account the fact that in case of failure to satisfy the criterion chosen, we can always set $x_i^{k+1} = x_i^k + \alpha_i^k d_i^k$.

Remark 2 Under the assumption that the objective function f is a twice continuously differentiable strongly convex function of x_i when the other component vectors are held constant, we can show that the acceptability conditions at Step 1 of Algorithm 1 are satisfied by the point

$$x_i^{k+1} = \arg\min_{x_i} f(x_1^{k+1}, \ldots, x_i, \ldots, x_m^k),$$

provided that a suitable value of the parameter τ_i is selected.

More precisely, given $i \in \{1, \ldots, m\}$ suppose that the Assumption 2 is satisfied.

Then, if $x_i^{k+1} = \arg\min_{x_i} f(x_1^{k+1}, \ldots, x_i, \ldots, x_m^k)$ is the global minimizer with respect to x_i, we can write:

$$f(x_1^{k+1}, \ldots, x_i^{k+1}, x_{i+1}^k \ldots, x_m^k) \le f(x_1^{k+1}, \ldots, x_i^k + \alpha_i^k d_i^k, x_{k+1}^k \ldots, x_m^k), \tag{75}$$

so that condition (71) is satisfied. On the other hand, by definition of x_i^{k+1} we can write

$$\nabla_i f(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^{k+1}, x_{i+1}^k, \ldots, x_m^k) = 0. \tag{76}$$

It follows that we have

$$\nabla_i f(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^{k+1}, x_{i+1}^k, \ldots, x_m^k)^{\mathrm{T}}(x_i^k - x_i^{k+1}) = 0. \tag{77}$$

Therefore, using Taylor's theorem we obtain:

$$\begin{aligned} f(x_1^{k+1}, \ldots, x_i^k, x_{i+1}^k \ldots, x_m^k) &= f(x_1^{k+1}, \ldots, x_i^{k+1}, x_{i+1}^k \ldots, x_m^k) \\ &\quad + \frac{1}{2}(x_i^k - x_i^{k+1})^{\mathrm{T}} \nabla_i^2 f(v_i^k)(x_i^k - x_i^{k+1}), \end{aligned} \tag{78}$$

where $v_i^k = x_i^{k+1} + \zeta_i^k (x_i^k - x_i^{k+1}) \in \mathcal{C}$ and $\zeta_i^k \in (0, 1)$.

Thus, from (78), recalling that

$$z(k, i) = (x_1^{k+1}, \ldots, x_i^k, x_{i+1}^k \ldots, x_m^k),$$

we get, by Assumption 2

$$f(x_1^{k+1},\dots,x_i^{k+1},x_{i+1}^k\dots,x_m^k) \leq f(z(k,i)) - \lambda_i\|x_i^{k+1} - x_i^k\|^2,$$

so that it can be concluded that also condition (72) at Step 1 is satisfied, provided that we take $\tau_i \geq 1/\lambda_i$.

Remark 3 Under the assumption that the objective function f is a pseudoconvex function and that $\mathcal{L}$ is compact, we can remove condition (72) at Step 1(b). In fact, we can repeat the same reasonings used in the proof of Theorem 7.1 to show that the mappings T_i, with $i = 1,\dots,m$, defined at Step 1(b) satisfy Conditions 1 and 2 of Section 3. Therefore, the global convergence of Algorithm 1 follows immediately from Theorem 3.2.

Now we extend our model to the case in which suitable (generalized) convexity conditions are imposed on f with reference to a subset of the components, in a way that the global minimization with respect to each of these components still yields a convergent algorithm. We will assume that the objective function of f is a *strictly quasiconvex* function of x_i, for $i \in I_c \subseteq \{1,\dots,m\}$, when the other component vectors are held constant. More formally, we suppose that the following assumption holds.

ASSUMPTION 4 There exists an index set $I_c \subseteq \{1,\dots,m\}$ such that, for every $i \in I_c$, $x \in R^n$ and $y_i \in R^{n_i}$ with $y_i \neq x_i$ we have for all $t \in (0,1)$

$$f(x_1,\dots,tx_i + (1-t)y_i,\dots,x_m) < \max\{f(x), f(x_1,\dots,y_i,\dots,x_m)\}.$$

Then we set $I_{nc} = \{1,\dots,m\} - I_c$, and we can define the following conceptual algorithm model, where we assume that the search directions $d_i^k \in R^{n_i}$, for $i \in I_{nc}$, satisfy Assumption 1.

Algorithm 2

Data. $x^0 \in R^n$, numbers $\theta_i \in (0,1)$ for $i = 1,\dots,m$.

Step 0. Set $k = 0$.

Step 1. Set $z(k,1) = x^k$. For $i = 1,\dots,m$:

If $i \in I_c$ then set

$$x_i^{k+1} = \arg\min_{x_i} f(x_1^{k+1},\dots,x_i,\dots,x_m^k); \tag{79}$$

If $i \in I_{nc}$ apply the instructions (a) and (b) of Step 1 in Algorithm 1; set

$$z(k, i+1) = (x_1^{k+1}, \dots, x_i^{k+1}, x_{i+1}^k, \dots, x_m^k).$$

Step 2. Set

$$x^{k+1} = (x_1^{k+1}, \dots, x_m^{k+1}),$$

$k = k+1$ and go to Step 1.

The convergence properties of Algorithm 2 are given in the next theorem.

THEOREM 7.3 *Let $\{x^k\}$ be the sequence generated by Algorithm* 2. *Then, every limit point of $\{x^k\}$ is a stationary point of f. Moreover, if $\mathcal{L}$ is compact, we have*

$$\lim_{k \to \infty} \nabla f(x^k) = 0,$$

and there exists at least one limit point that is a stationary point of f.

Proof For each $i \in \{1, \dots, m\}$, let T_i be the mappings that determines the point x_i^{k+1}. Then, whenever $i \in I_{nc}$ we can repeat the same reasonings used in the proof of Theorem 7.1, and we obtain that T_i satisfies Conditions 1–3 of Section 3. On the other hand, for each $i \in I_c$, by Proposition 5.2, it follows that the mappings T_i satisfy Conditions 1–3. Then, the assertion follows from Theorem 3.3.

In the special case of two-block decomposition, that is

$$\underset{(x_1, x_2) \in R^{n_1} \times R^{n_2}}{\text{minimize}} \; f(x) = f(x_1, x_2), \tag{80}$$

we can define a more particular scheme by replacing in the Modified 2Block GS method of Section 6 the global minimization with respect to x_1 with the same operations defined in Algorithm 1. This yields the following algorithm model, where the direction $d_1^k \in R^{n_1}$ satisfies Assumption 1.

Algorithm 3

Data. $x^0 \in R^n$.

Step 0. Set $k = 0$.

Step 1. For $i = 1, 2$:

if $i = 1$ then

(a) compute α_1^k by means of Algorithm LS (with $\alpha_1^k = 0$ if $\nabla_1 f(x^k) = 0$);
(b) choose x_1^{k+1} such that the following conditions are satisfied:

$$f(x_1^{k+1}, x_2^k) \leq f(x_1^k + \alpha_1^k d_1^k, x_2^k); \tag{81}$$

$$\|x_1^{k+1} - x_1^k\|^2 \leq \tau_i \max\{\xi_1^k, \Delta f_1^k\}, \tag{82}$$

where

$$\Delta f_1^k = f(x^k) - f(x_1^{k+1}, x_2^k);$$

if $i = 2$ then *determine* x_2^{k+1} such that

$$f(x_1^{k+1}, x_2^{k+1}) \leq f(x_1^{k+1}, x_2^k) \quad \text{and} \quad \nabla_2 f(x_1^{k+1}, x_2^{k+1}) = 0.$$

Step 2. Set

$$x^{k+1} = (x_1^{k+1}, x_2^{k+1}),$$

$k = k + 1$ and go to Step 1.

8 COMPUTATIONAL ASPECTS AND NUMERICAL EXAMPLES

The models described in this paper can be made the basis of various computational implementations.

In order to define a specific algorithm several choices have to be made and the following points deserve a special attention:

- the choice of the decomposition:
- the structure of the interconnection;
- the actual implementation of the elementary operations and, in particular, the realization of an effective line search procedure;

– the method used for generating the point x_i^{k+1} and the choice of a specific acceptability criterion.

Of course, the preceding choices have to be related to the concrete application in which a decomposition approach has to be employed.

In this connection, we note that decomposition methods typically exhibit a much slower convergence rate in comparison with ordinary techniques. In spite of this, there are many practical contexts in which these methods can be very effective [4]. A first case is when the objective function can be put into the form:

$$f(x) = \psi_1(x_1) + \sum_{i=2}^{m} \psi_i(x_1)\phi_i(x_i).$$

We note, in fact, that once x_1 is fixed, the objective function can be minimized in parallel with respect to the components x_i for $i = 2, \dots, m$ and this can be advantageous, in some instances, with respect to ordinary methods.

A second interesting case is when the form of the objective function allows us to obtain subproblems of special structure in the component variables, so that the use of decomposition techniques may favor the application of specialized techniques for solving the subproblems.

The objective function of the minimization problem connected to the the "learning problem" for neural networks may exhibit these features. Indeed, the application of a decomposition approach to this class of problems has represented one of the original motivations of this work and it will be illustrated in the sequel on a particular example.

Let us consider a given set of data (input/output pairs)

$$T = \{(u^j, d^j),\ u^j \in R^m,\ d^j \in R,\ j = 1, \dots, P\}.$$

If we denote by $x \in R^n$ the vector of parameters of a neural network and by $\psi(x, u^j): R^n \to R$ the *output* of the network corresponding to an *input* u^j such that $(u^j, d^j) \in T$, we have that the learning problem can be formulated as the following least-squares problem

$$\min_{x \in R^n} f(x) = \sum_{j=1}^{P} (d^j - \psi(x; u^j))^2. \qquad (83)$$

For a radial basis function neural network (see [18]) with M *neurons*, the output ψ is given by

$$\psi(w, c^1, \ldots, c^M; u^j) = \sum_{i=1}^{M} w_i G(\|c^i - u^j\|),$$

where $G: R \to R$ is a fixed suitable function, and $w \in R^M$, $c^i \in R^m$, $i = 1, \ldots, M$, are the network parameters. Then, the vector $x \in R^{M(1+m)}$ is given by

$$x = (w, c^1, c^2, \ldots, c^M).$$

A possible choice for G is the *multiquadric function*, defined by:

$$G(r) = (r^2 + \sigma^2)^{1/2},$$

σ being a positive constant.

Starting from the least-squares problem (83), we have considered the following unconstrained optimization problem:

$$\min_{\substack{w \in R^M \\ c^1, \ldots, c^M \in R^m}} f(w, c^1, \ldots, c^M) = \sum_{j=1}^{P} \Big(d^j - \sum_{i=1}^{M} w_i G(\|c^i - u^j\|)\Big)^2 + \eta\Big(\|w\|^2 + \sum_{i=1}^{M} \|c^i\|^2\Big), \quad (84)$$

where $\eta > 0$.

We can observe that the objective function of (84) has compact level sets; moreover, it is a *strictly convex* function of w when the remaining variables are fixed, and for η "sufficiently small" tends to coincide with the objective function of (83).

Taking into account the special structure of the objective function, we have decomposed the vector x into two blocks as follows:

$$x_1 = w \qquad x_2 = (c^1, c^2, \ldots, c^M).$$

Then we have employed Algorithm 2 (ALG2) with $I_c = \{1\}$ and $I_{nc} = \{2\}$, which is equivalent, in this case, to Algorithm 3. It can be observed that when x_2 is fixed, from (84) we obtain a strictly convex

TABLE I Comparative results for problem (84)

n	n_i	n_f	n_g	cpu	$f(x^*)$	$\|\nabla f(x^*)\|$
EO4DGF						
17	526	539	539	1.34	0.196	0.7×10^{-3}
34	58393	58663	58663	287.89	0.087	0.9×10^{-3}
51	F	100545	100545	807.42	0.073	0.1×10^{-1}
85	F	100562	100562	1333.07	0.066	0.8×10^{-2}
119	F	100614	100614	1865.68	0.065	0.1×10^{-1}
ALG2						
17	30	96	30	0.2	0.196	0.9×10^{-3}
34	2642	8031	2642	36.11	0.087	0.9×10^{-3}
51	9266	38275	9266	217.27	0.069	0.9×10^{-3}
85	29355	89711	29355	1106.86	0.064	0.9×10^{-3}
119	35248	108293	35248	1875.21	0.063	0.9×10^{-3}

quadratic function in x_1, which was minimized exactly by employing the routine F04ASF of the NAG library, for solving the linear system

$$\nabla_1 f(x_1, x_2) = 0.$$

The component x_2 was updated by means of a line search procedure based on Algorithm LS described in Section 4.

The computational results were obtained, starting from random initial points, by assuming $m = 16$, $\sigma = 10$ and $\eta = 0.001$ in correspondence to an input/output data set taken from a letter recognition problem.† In order to keep the computing times within reasonable limits, we used only 50 input/output pairs. This may be not significant in the real application, but already constitutes a rather severe test for the optimization codes.

Some sample results are reported in Table I, by specifying the number $n = M(1 + m)$ of variables (depending on the number M of neurons), the number n_i of iterations required to attain convergence towards a point x^*, the number n_f of function evaluations, the number n_g of gradient evaluations, the cpu time in seconds, the objective function value $f(x^*)$, and the norm of the gradient $\|\nabla f(x^*)\|$. The termination criterion was

$$\|\nabla f(x^k)\| \leq 10^{-3} \quad \text{or} \quad k \geq 10^5.$$

†This problem was available via ftp from the UCI Repository of Machine Learning Databases and Domain Theories: ftp.ics.uci.edu: pub/machine-learning-databases.

In Table I we show also comparative figures obtained by employing a preconditioned limited memory quasi-Newton conjugate gradient method (E04DGF routine, NAG library).

Table I points out the advantages of Algorithm 2 with respect to algorithm E04DGF, both in terms of cpu time and of final accuracy. We note, in particular, that for $n > 51$ algorithm E04DGF fails to reach the prescribed accuracy within the prefixed number of iterations.

No significant progress can be obtained by increasing the number of iterations; in fact, for $n = 119$ and $n_i = 2 \times 10^5$ we obtained in the same problem the results shown in Table II.

Various other applications of block decomposition techniques can be envisaged in the field of learning problems for neural networks, such as the decomposition with respect to different layers or individual units. However, potential advantages still have to be assessed and additional work is needed.

The adoption of partial global minimizations with respect to some variables may also be advantageous in connection with global optimization. In fact, there are problems in which the global minimization with respect to some component, for fixed values of the remaining components, can be useful for escaping from a local minimizer. As an example, we can consider the problem of minimizing an objective function of the form:

$$f(x) = \sum_{i=1}^{n}(x_i - 1)^2 + 4\prod_{i=1}^{n} x_i + \prod_{i=1}^{n} x_i^2. \tag{85}$$

In this case Assumption 4 is satisfied for every i and the Gauss–Seidel method reduces to a coordinate descent method with exact line searches. The results obtained for various values of n, with random initial points, are compared in Table III with those obtained with algorithm E04DGF.

We can note that algorithm E04DGF is faster when both algorithms converge to the same local solution; however, the global minimization

TABLE II E04DGF performance in problem (84)

n	n_i	n_f	n_g	cpu	$f(x^*)$	$\|\nabla f(x^*)\|$
119	2×10^5	201233	201233	3731.50	0.064	0.3×10^{-2}

TABLE III Comparative results for problem (85)

n	E04DGF			GS		
	cpu	$f(x^*)$	$\|\nabla f(x^*)\|$	*cpu*	$f(x^*)$	$\|\nabla f(x^*)\|$
50	0.01	0.35	0.4×10^{-3}	0.01	0.35	0.5×10^{-3}
50	0.01	0.35	0.4×10^{-3}	1.19	−2.61	0.9×10^{-3}
50	0.01	0.35	0.6×10^{-3}	1.02	−2.61	0.9×10^{-3}
50	0.01	0.35	0.5×10^{-3}	0.01	0.35	0.7×10^{-3}
100	0.01	0.227	0.2×10^{-3}	12.98	−2.75	0.9×10^{-3}
100	0.01	0.227	0.4×10^{-3}	14.56	−2.75	0.9×10^{-3}
100	0.02	0.227	0.3×10^{-3}	12.97	−2.75	0.9×10^{-3}
100	0.01	0.227	0.5×10^{-3}	0.06	0.227	0.7×10^{-3}
200	0.03	0.143	0.5×10^{-3}	145.75	−2.85	0.9×10^{-3}
200	0.03	0.143	0.3×10^{-3}	134.35	−2.85	0.9×10^{-3}
200	0.03	0.143	0.8×10^{-3}	0.27	0.143	0.8×10^{-3}
200	0.71	−2.85	0.6×10^{-3}	144.21	−2.85	0.9×10^{-3}

with respect to the components, performed in the GS method, yields, in many cases, an improvement in the objective function.

Finally, we remark that the results obtained in the unconstrained case can also be extended to constrained problems. This extension will be the object of subsequent work.

Acknowledgments

The authors are indebted to the anonymous reviewers for their useful comments. In particular, the criticism and the constructive suggestions of one of them led to several improvements in the paper.

References

[1] N.V. Banitchouk, V.M. Petrov and R.L. Chernousko, Numerical solution of problems with variational limits by the method of local variations. *Zh. Vychisl. Mat. Mat. Fiz*, **6**: 947–961, 1966.

[2] M.S. Bazaraa, H.D. Sherali and C.M. Shetty, *Nonlinear Programming Theory and Algorithms*. John Wiley and Sons, New York, 1993.

[3] D.P. Bertsekas, *Nonlinear programming*. Athena Scientific, Belmond, MA, 1995.

[4] D.P. Bertsekas and J.N. Tsitsiklis, *Parallel and Distributed Computation*. Prentice-Hall International Editions, Englewood Cliffs, NJ, 1989.

[5] J. Cea, *Optimisation théorie et algorithmes*. Dunod, Paris, 1971.

[6] G. Chen and M. Teboulle, A proximal-based decomposition method for convex minimization problems. *Mathematical Programming*, **64**: 81–101, 1994.

[7] R. De Leone, M. Gaudioso and L. Grippo, Stopping criteria for linesearch methods without derivatives. *Mathematical Programming*, **30**: 285–300, 1984.

[8] M.C. Ferris and O.L. Mangasarian, Parallel variable distribution. *SIAM Journal on Optimization*, **4**(4): 815–832, 1994.
[9] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods: Applications to the Solution of Boundary-Valued Problems.* North-Holland, Amsterdam, 1983.
[10] M. Fukushima, Application of the alternating direction method of multipliers to separable convex programming problems. *Computational Optimization and Applications*, **1**: 93–111, 1992.
[11] D. Gabay and B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. *Computers and Mathematics with Applications*, **2**: 17–40, 1976.
[12] L. Grippo, F. Lampariello and S. Lucidi, Global convergence and stabilization of unconstrained minimization methods without derivatives. *Journal of Optimization Theory and Applications*, **56**(3): 385–406, 1988.
[13] S. Haykin, *Neural Networks.* Macmillan Publishing Company, Englewood Cliffs, 1994.
[14] Z.Q. Luo and P. Tseng, On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, **72**: 7–35, 1992.
[15] Z.Q. Luo and P. Tseng, Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, **46**: 157–178, 1993.
[16] O.L. Mangasarian, Parallel gradient distribution in unconstrained optimization. *SIAM Journal on Control and Optimization*, **33**(6): 1916–1925, 1995.
[17] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, New York, 1970.
[18] T. Poggio and F. Girosi, Networks for approximation and learning. *Proceedings of the IEEE*, **78**: 1481–1497, 1990.
[19] E. Polak, *Computational Methods in Optimization: A Unified Approach.* Academic Press, New York, 1985.
[20] M.J.D. Powell, On search directions for minimization algorithms. *Mathematical Programming*, **4**: 193–201, 1973.
[21] B.N. Pshenichny and Yu.M. Danilin, *Numerical Methods in Extremal Problems.* Mir Publishers Moscow, Mosca, 1978.
[22] P. Tseng, Decomposition algorithm for convex differentiable minimization. *Journal of Optimization Theory and Applications*, **70**: 109–135, 1991.
[23] P. Tseng, On the rate of convergence of a partially asynchronous gradient projection algorithm. *SIAM Journal on Optimization*, **1**(4): 603–619, 1991.
[24] P. Tseng, D.P. Bertsekas and J.N. Tsitsiklis, Partially asynchronous parallel algorithms for network flow and other problems. *SIAM Journal on Control and Optimization*, **28**(3): 678–710, 1990.
[25] N. Zadeh, A note on the cyclic coordinate ascent method. *Management Science*, **16**: 642–644, 1970.
[26] W.I. Zangwill, *Nonlinear Programming: A Unified Approach.* Prentice Hall, Englewood Cliffs, 1969.