A convergent decomposition method for box-constrained optimization problems.

Optim Lett (2009) 3:397–409
DOI 10.1007/s11590-009-0119-8

ORIGINAL PAPER

A convergent decomposition method for box-constrained optimization problems

Andrea Cassioli · Marco Sciandrone

Received: 3 February 2009 / Accepted: 20 February 2009 / Published online: 18 March 2009

Abstract In this work we consider the problem of minimizing a continuously differentiable function over a feasible set defined by box constraints. We present a decomposition method based on the solution of a sequence of subproblems. In particular, we state conditions on the rule for selecting the subproblem variables sufficient to ensure the global convergence of the generated sequence without convexity assumptions. The conditions require to select suitable variables (related to the violation of the optimality conditions) to guarantee theoretical convergence properties, and leave the degree of freedom of selecting any other group of variables to accelerate the convergence.

Keywords Decomposition methods · Gauss–Southwell method · Global convergence

1 Introduction

Let us consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{s.t.} & l \leq x \leq u \end{array} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function, $l, u \in \mathbb{R}^n$. This is a well-studied problem and several methods have been proposed (see, e.g., [2,4,7,11,12]).

A. Cassioli · M. Sciandrone (✉)
Dipartimento di Sistemi e Informatica, Università degli Studi di Firenze, Firenze, Italy
e-mail: sciandro@dsi.unifi.it

A. Cassioli
e-mail: cassioli@dsi.unifi.it

In many real applications, due to the particular structure of the problem and/or to the large sizes, the adoption of a decomposition approach may be the practicable way to efficiently solve the optimization problem (see, e.g., [5]).

In a general decomposition framework, at any iteration, some variables are kept fixed to their current values, and the other variables are determined by solving the corresponding subproblem. Formally, a general decomposition scheme is described below.

1.1 General decomposition scheme

Data. A feasible point x^0.

For $k = 0, 1, \ldots$

choose a working set $W^k \subset \{1, \ldots, n\}$;

find

$$\begin{aligned} & x^{k+1} \in \arg\min f(x) \\ & \quad \text{s.t.} \\ & \qquad l_i \leq x_i \leq u_i \quad i \in W^k \\ & \qquad x_i = x_i^k \forall \quad i \notin W^k \end{aligned} \tag{2}$$

Global convergence properties of the generated sequence $\{x^k\}$ depend on the rule for selecting the working set W^k, provided that (2) admits solution. Most convergent decomposition methods require suitable convexity assumptions on the objective function to ensure that $\|x^{k+1} - x^k\| \to 0$ for $k \to \infty$ (see, e.g., [3,8,10]). Indeed, as one may expect, this plays an important theoretical role for the convergence of decomposition algorithm, where the optimization is sequentially performed with respect to different blocks of variables.

Here we are interested to the case of nonconvex problems, and we refer to working set selection rules based on the maximum violation of the optimality conditions. Note that the optimality conditions for box constrained problems can be expressed in two equivalent ways: the first one involves the projected gradient, the second one involves the so-called reduced gradient.

The well-known Gauss–Southwell decomposition method employs a working set selection rule based on the maximum violation of the optimality conditions expressed by means of the projected gradient. Recently the global convergence of the method has been proved in [14] without convexity assumptions, and with the possibility of including in the working set, besides the variable that mostly violates the optimality conditions, any other group of variables. The decomposition methods analyzed in [14] and including the Gauss–Southwell method are designed for a more general class of linearly constrained problems.

Decomposition algorithms for quadratic convex problems with one linear equality constraint and box constraints (arising in the training of support vector machines) have been proposed in [6,8]. These methods, adapted to the case of box constrained problems, select variables that mostly violate the optimality conditions expressed in terms of reduced gradient. In particular, the components of the reduced gradient are

ordered in a decreasing order in terms of absolute value, and the first $m \geq 1$ ordered components define the subproblem variables. These methods are widely used since are very efficient. Their convergence has been proved in [8] under convexity assumptions on the objective function and under the condition that the working set is made only by the m variables that mostly violate the optimality conditions.

In this work we present a new algorithm based on a suitable modification of the maximum violation rule expressed in terms of reduced gradient. The rule we introduce requires to select, similarly to algorithms of [6] and [8], variables related to the violation of optimality conditions expressed in terms of reduced gradient, and (as in the Gauss–Southwell method) leaves the degree of freedom of selecting any other group of variables to accelerate the convergence. We observe that this latter point can be of great interest from a computational point of view, and that the rules computationally advantageous for selecting the additional group of variables are problem dependent. We prove the global convergence of the defined decomposition method without requiring convexity assumptions on the objective function, and hence we extend the class of convergent decomposition methods for nonconvex problems.

2 Notation and preliminary results

For convenience of the reader, we recall in this section some preliminary results needed in Sect. 3 and in the Appendix. We introduce also some basic notation to be used through the following sections.

The feasible set of (1) will be denoted as $\mathcal{F}$. In correspondence to any feasible point x, we define the set $L(x)$ of active lower bounds and the set $U(x)$ of active upper bounds:

$$L(x) = \{i \in 1 \dots n : x_i = l_i\}$$
$$U(x) = \{i \in 1 \dots n : x_i = u_i\}$$

We denote by $D(x)$ the set of feasible direction at x:

$$D(x) = \{d \in \mathbb{R}^n : d_i \geq 0 \; i \in L(x), \quad d_i \leq 0 \; i \in U(x)\}$$

On stationarity conditions

Given $\bar{x} \in \mathcal{F}$, we say $\bar{x}$ is a stationary point if

$$\nabla f(\bar{x})^T d \geq 0 \qquad \forall d \in D(\bar{x}), \tag{3}$$

where ∇f is the gradient of f.

In correspondence to any point $x \in \mathbb{R}^n$, we denote by $[x]^+$ the orthogonal projection of x onto the feasible set $\mathcal{F}$, so that we have

$$[x]_i^+ = \max\{l_i, \min\{u_i, x_i\}\} \qquad i = 1, \ldots, n.$$

For any point $x \in \mathcal{F}$, the reduced gradient $\nabla^{\text{red}} f(x)$ has components $\nabla_i^{\text{red}} f(x)$, with $i = 1, \ldots, n$, defined as follows

$$\nabla_i^{\text{red}} f(x) = \begin{cases} \min\{0, \nabla_i f(x)\} & \text{if } \ x_i = l_i \\ \nabla_i f(x) & \text{if } \ l_i < x_i < u_i \\ \max\{0, \nabla_i f(x)\} & \text{if } \ x_i = u_i \end{cases} \tag{4}$$

The stationarity condition (3) is equivalent to two conditions as stated in the following proposition.

Proposition 1 *Given a point $\bar{x} \in \mathcal{F}$, the following statements are equivalent:*

1. *$\bar{x}$ is a stationary point;*
2. *$\bar{x} = [\bar{x} - s\nabla f(\bar{x})]^+ \quad \forall s > 0$;*
3. *$\nabla^{\text{red}} f(\bar{x}) = 0$.*

For the convergence analysis presented in the next sections we need the following result proved in [9] in a more general setting (namely in the case of feasible set defined by linear inequalities).

Proposition 2 *Let $\{x^k\}$ be a sequence of points such that $x^k \in \mathcal{F}$ for all k. Assume that*

$$\lim_{k\to\infty} x^k = \bar{x}. \tag{5}$$

Then, given any direction $\bar{d} \in D(\bar{x})$, there exists a scalar $\hat{\beta} > 0$ such that, for sufficiently large values of k, we have

$$x^k + \beta\bar{d} \in \mathcal{F}, \quad \forall \beta \in [0, \hat{\beta}]. \tag{6}$$

On the Armijo-type line search

We recall the well-known Armijo-type line search along a feasible direction, and we state a theoretical result employed in our convergence analysis.

Let d^k be a feasible direction at $x^k \in \mathcal{F}$. We denote by $\beta_{\mathcal{F}}^k$ the maximum feasible steplength along d^k, namely $\beta_{\mathcal{F}}^k$ satisfies

$$l \le x^k + \beta d^k \le u \qquad \text{for all } \ \beta \in [0, \beta_{\mathcal{F}}^k],$$

and (since $-\infty \le l < u \le \infty$) we have that either $\beta_{\mathcal{F}}^k = +\infty$ or at least an index $i \in \{1, \ldots, n\}$ exists such that

$$x_i^k + \beta_{\mathcal{F}}^k d_i^k = l_i \quad \text{or} \quad x_i^k + \beta_{\mathcal{F}}^k d_i^k = u_i.$$

Let β_u be a positive scalar and set

$$\beta^k = \min\{\beta^k_{\mathcal{F}}, \beta_u\}. \tag{7}$$

Assumption 1 Assume that $\{d^k\}$ is a sequence of feasible search directions such that

(a) for all k we have $\|d^k\| \leq M$ for a given number $M > 0$;
(b) for all k we have $\nabla f(x^k)^T d^k < 0$.

An Armijo-type line search algorithm is described below.

Armijo-type line search ALS(x^k, d^k, β^k)

Data: Given $\alpha > 0$, $\delta \in (0, 1)$, $\gamma \in (0, 1/2)$ and the initial stepsize $\alpha^k = \min\{\beta^k, \alpha\}$.
Step 1. Set $\lambda = \alpha^k$, $j = 0$.
Step 2. If

$$f(x^k + \lambda d^k) \leq f(x^k) + \gamma\lambda\nabla f(x^k)^T d^k \tag{8}$$

then set $\lambda^k = \lambda$ and stop.
Step 3. Set $\lambda = \delta\lambda$, $j = j + 1$ and go to Step 2.

The properties of algorithm ALS are reported in the next proposition (see, e.g., [1]).

Proposition 3 *Let $\{x^k\}$ be a sequence of points belonging to the feasible set $\mathcal{F}$, and let $\{d^k\}$ be a sequence of search directions satisfying Assumption 1. Then:*

(i) *Algorithm* ALS *determines, in a finite number of iterations, a scalar λ^k such that condition (8) holds, i.e.,*

$$f(x^k + \lambda^k d^k) \leq f(x^k) + \gamma\lambda^k\nabla f(x^k)^T d^k; \tag{9}$$

(ii) *if $\{x^k\}$ converges to $\bar{x}$ and*

$$\lim_{k\to\infty} \Big(f(x^k) - f(x^k + \lambda^k d^k)\Big) = 0, \tag{10}$$

then we have

$$\lim_{k\to\infty} \beta^k \nabla f(x^k)^T d^k = 0, \tag{11}$$

where β^k is given by (7).

On the Gauss–Southwell decomposition algorithm

A well-known decomposition algorithm is based on the Gauss–Southwell rule for selection of the working set. According to the Gauss–Southwell rule, at any iteration k, the working set W^k must contain the index i^k corresponding to the variable that mostly violates the optimality condition 2 of Proposition 1, that is

$$|x^k_{i^k} - [x^k_{i^k} - \nabla_{i^k} f(x^k)]^+_{i^k}| \geq |x^k_j - [x^k_j - \nabla_j f(x^k)]^+_j| \qquad j = 1, \ldots, n, \tag{12}$$

Formally the Gauss–Southwell algorithm is described below.

Gauss–Southwell decomposition (GSD) algorithm

Data. $x^0 \in \mathcal{F}$.
For $k = 0, 1, \ldots$
choose any working set $W^k \subset \{1, \ldots, n\}$ such that $i^k \in W^k$, where i^k is an index such that (12) holds;
find

$$\begin{aligned} x^{k+1} \in \arg\min\ & f(x) \\ \text{s.t.}\ & \\ & l_i \leq x_i \leq u_i \quad i \in W^k \\ & x_i = x^k_i \forall \quad i \notin W^k \end{aligned} \tag{13}$$

Note that, besides the variable corresponding to the most violating index i^k, any other group of variables can be inserted in the working set. The following convergence result follows from Theorem 4.1 of [14]. For convenience of the reader we report in the Appendix an alternative proof.

Proposition 4 *Let* $\{x^k\}$ *be the sequence generated by* GSD *algorithm. Every limit point of* $\{x^k\}$ *is a stationary point.*

3 A new decomposition algorithm

A decomposition algorithm based on the violation of the optimality condition 3 of Proposition 1 has been proposed in [6] for quadratic convex problems with box constraints and one linear equality constraint. The decomposition algorithm, adapted to the case of feasible set defined by box constraints, is described below.

Maximum violation decomposition algorithm

Data. $x^0 \in \mathcal{F}$.
For $k = 0, 1, \ldots$
let $\{i_1, i_2, \ldots, i_n\}$ be such that

$$|\nabla_{i_1} f^{\text{red}}(x^k)| \geq |\nabla_{i_2} f^{\text{red}}(x^k)| \geq \cdots |\nabla_{i_n} f^{\text{red}}(x^k)|, \tag{14}$$

and choose a working set $W^k = \{i_1, \ldots, i_m\}$ with $1 \leq m < n$;
find

$$\begin{aligned} & x^{k+1} \in \arg\min f(x) \\ & \quad \text{s.t.} \\ & \qquad l_i \leq x_i \leq u_i \quad i \in W^k \\ & \qquad x_i = x_i^k \quad \forall i \notin W^k \end{aligned} \tag{15}$$

We observe that the working set selection rule requires to consider, as subproblem variables, those that mostly violate (in a decreasing order) the optimality condition expressed by means of the reduced gradient. Thus the degree of freedom for selecting the whole working set is limited. Furthermore, the global convergence of the algorithm holds under strict convexity assumptions on the objective function (see [8]).

We propose here a slightly different decomposition algorithm which overcomes the limitations above.

For any point $x^k \in \mathcal{F}$, we denote by $I_r(x^k)$ the index set such that, $i \in I_r(x^k)$ implies

$$|\nabla_i^{\text{red}} f(x)| \geq |\nabla_j^{\text{red}} f(x)| \qquad \text{for all} \quad j = 1, \ldots, n. \tag{16}$$

We also introduce, for a given scalar $\epsilon > 0$, the following indices (provided they exist)

$$j^k \in \arg\max_j \{\nabla_j f(x^k) : \ \nabla_j f(x^k) > 0, \qquad x_j^k \geq l_j + \epsilon\} \tag{17}$$

$$p^k \in \arg\min_p \{\nabla_p f(x^k) : \ \nabla_p f(x^k) < 0, \qquad x_p^k \leq u_p - \epsilon\} \tag{18}$$

and the index sets

$$I_L^k = \{h : x_h^k \leq l_h + \epsilon \qquad \text{and} \qquad \nabla_h^{\text{red}} f(x^k) > \nabla_{j^k}^{\text{red}} f(x^k)\} \tag{19}$$

$$I_U^k = \{h : x_h^k \geq u_h - \epsilon \qquad \text{and} \qquad \nabla_h^{\text{red}} f(x^k) < \nabla_{p^k}^{\text{red}} f(x^k)\} \tag{20}$$

Note that we set

- $I_L^k = \{h : x_h^k \leq l_h + \epsilon \quad \text{and} \quad \nabla_h^{\text{red}} f(x^k) > 0\}$ whenever j^k is not defined;
- $I_U^k = \{h : x_h^k \geq u_h - \epsilon \quad \text{and} \quad \nabla_h^{\text{red}} f(x^k) < 0\}$ whenever p^k is not defined.

ϵ-MVD algorithm

Data. $x^0 \in \mathcal{F}, \epsilon > 0$.
For $k = 0, 1, \ldots$
if there exists an index $i^k \in I_r(x^k)$ such that one of the following conditions holds

(a) $l_{i^k} + \epsilon \leq x^k_{i^k} \leq u_{i^k} - \epsilon$;
(b) $x^k_{i^k} \leq l_{i^k} + \epsilon$ and $\nabla_{i^k} f(x^k) < 0$;
(c) $x^k_{i^k} \geq u_{i^k} - \epsilon$ and $\nabla_{i^k} f(x^k) > 0$

then choose any working set W^k such that $i^k \in W^k$;
otherwise choose any working set W^k such that $\left(I^k_L \cup I^k_U \cup \{j_k\} \cup \{p_k\}\right) \subseteq W^k$;
find

$$\begin{aligned} x^{k+1} &\in \arg\min f(x) \\ &\quad \text{s.t.} \\ &\quad l_i \leq x_i \leq u_i \quad i \in W^k \\ &\quad x_i = x^k_i \quad \forall i \notin W^k \end{aligned} \tag{21}$$

Before stating the convergence result of the algorithm, we briefly explain the working set selection rule. In particular, the underlying rationale is to select the variables by means of an estimate of the active constraints to prevent possible "pathological" cases, due to the fact that the reduced gradient is not continuous. For instance, assume that

- i is the unique index that mostly violates the optimality condition at iteration k;
- $\nabla_i f(x^k) > 0$;
- x^k_i is sufficiently close to the lower bound l_i.

In this case, the value of $\nabla_i f(x^k) = \nabla^{\text{red}}_i f(x^k)$ does not contain enough information about the violation of the optimality conditions. Indeed, an high positive value of $\nabla_i f(x^k)$ says that the optimality conditions are strongly violated at the current point with respect to the i-th component, namely that $|\nabla^{\text{red}}_i f(x^k)| >> 0$, and hence the index i must be inserted in the working set. On the other hand, as the lower bound l_i is "quasi-active", we could roughly state that $|\nabla^{\text{red}}_i f(x^k)|$ is small, so that other suitable indices must be inserted in the working set to better capture the violation of the optimality conditions. Note that a "pathological" case can not occur whenever the component x^k_i is sufficiently far from the bounds [see condition (a)], or is close to the lower bound but we have $\nabla_i f(x^k) < 0$ [see condition (b)], or is close to the upper bound but we have $\nabla_i f(x^k) > 0$ [see condition (c)]. In the other cases, we must insert in the working set:

- the two indices j^k, p^k (provided they exist) corresponding to variables which "sufficiently" violate the optimality conditions and in the limit cannot have a "pathological" behaviour;

– all the indices corresponding to variables which locally "strongly" violate the optimality conditions but in the limit can have a "pathological" behaviour as above, namely the indices of I_L^k and of I_U^k.

Proposition 5 *Let* $\{x^k\}$ *be the sequence generated by* ϵ-MVD *algorithm. Every limit point of* $\{x^k\}$ *is a stationary point.*

Proof We proceed by contradiction. Let us assume that there exists an infinite subset $K \subseteq \{0, 1, \ldots\}$ such that

$$\lim_{k\to\infty, k\in K} x^k = \bar{x} \tag{22}$$

and $\bar{x}$ is not a stationary point, that is, there exists an index $\hat{i} \in \{1, \ldots, n\}$ such that

$$|\nabla_{\hat{i}}^{\text{red}} f(\bar{x})| > \eta > 0. \tag{23}$$

The instructions of the algorithm imply

$$f(x^{k+1}) \leq f(x^k),$$

and hence, using (22) and the continuity of f, it follows that the nonincreasing sequence $\{f(x^k)\}$ converges, i.e.,

$$\lim_{k\to\infty} f(x^k) = f(\bar{x}). \tag{24}$$

From (23) it follows that there exists a direction $d \in \{-e_{\hat{i}}, e_{\hat{i}}\}$ which is a feasible descent direction at $\bar{x}$. One of the following two cases can occur:

(I) there exists an infinite subset $\hat{K} \subseteq K$ such that $\hat{i} \in W^k$ for all $k \in \hat{K}$;
(II) for $k \in K$ and k sufficiently large $\hat{i} \notin W^k$.

In case (I), as $\hat{i} \in W^k$ for all $k \in \hat{K}$, we have that the instructions of the algorithm imply

$$f(x^{k+1}) \leq f(x^k + \lambda^k d) \leq f(x^k) + \gamma\lambda^k \nabla f(x^k)^T d \tag{25}$$

where λ^k is the stepsize determined by means of Armijo algorithm.

Assumption 1 holds for the subsequence $\{d^k\}_K$ setting $d^k = d$ for all $k \in K$. Further, from (22), (24), and (25) we have that the assumptions of Proposition 3 hold for the subsequence $\{x^k\}_K$, and hence we have

$$\lim_{k\to\infty, k\in K} \beta^k \nabla f(x^k)^T d = 0, \tag{26}$$

where $\beta^k = \min\{\beta_{\mathcal{F}}^k, \beta_u\}$, $\beta_{\mathcal{F}}^k$ being the maximum feasible steplength along d, and β_u a prefixed positive number. Proposition 2 implies that $\beta^k \geq \bar{\beta} > 0$ for $k \in K$ and k sufficiently large, so that, recalling (26) and that $d \in \{e_{\hat{i}}, -e_{\hat{i}}\}$, we obtain

$$\lim_{k\to\infty, k\in K} |\nabla f(x^k)^T d| = |\nabla f(\bar{x})^T d| = |\nabla^{\text{red}}_{\hat{i}} f(\bar{x})| = 0, \tag{27}$$

and this contradicts (23).

In case (II), we have two possibilities:

(IIa) there exist an infinite subset $\bar{K} \subseteq K$ and an index $\bar{i} \in I_r(x^k)$ such that, for all $k \in \bar{K}$, one of the conditions (a), (b), (c) holds with $i^k = \bar{i}$;

(IIb) for $k \in K$ and k sufficiently large conditions (a), (b), (c) are never satisfied, and we necessarily have $\left(I^k_L, I^k_U, \{j^k\}, \{p^k\}\right) \subseteq W^k$.

Case (IIa)

First we observe that, as $\bar{i} \in I_r(x^k)$, we can write for $k \in \bar{K}$ and k sufficiently large

$$|\nabla^{\text{red}}_{\bar{i}} f(x^k)| \geq |\nabla^{\text{red}}_{\hat{i}} f(x^k)| \geq \eta > 0. \tag{28}$$

Assume that condition (a) holds for an infinite subsequence. Then we have $l_{\bar{i}} < \bar{x}_{\bar{i}} < u_{\hat{i}}$. In this case $\{e_{\bar{i}}, -e_{\bar{i}}\} \in D(\bar{x})$ are both feasible directions at $\bar{x}$, and one of them, denoted by d, is a descent direction, and as consequence, for $k \in K$ sufficiently large we can write

$$\nabla f(x^k)^T d < 0.$$

Now we can repeat the same reasonings used above to prove (27) and we obtain

$$\lim_{k\to\infty, k\in \bar{K}} |\nabla f(x^k)^T d| = |\nabla f(\bar{x})^T d| = |\nabla^{\text{red}}_{\bar{i}} f(\bar{x})| = 0,$$

and this contradicts (28).

Whenever either condition (b) or condition (c) holds for an infinite subsequence we would obtain, in a similar way, a contradiction with (28).

Case (IIb)

Assume that $\nabla_{\hat{i}} f(\bar{x}) > 0$. If $x^k_i \leq l_i + \epsilon$ then, as $\hat{i} \notin W^k$, we have

$$\hat{i} \notin I^k_L,$$

and hence, from the definition of I^k_L given in (19), we can write

$$0 < \nabla_{\hat{i}} f(x^k) \leq \nabla_{j^k} f(x^k) \tag{29}$$

where j^k is defined in (17).

If $x^k_i > l_i + \epsilon$ then again, as $\hat{i} \notin W^k$, it follows that $\hat{i} \neq j^k$ and (29) holds.

Then we can find an index $j \in W^k$ and an infinite subset $\tilde{K} \subseteq K$ such that for all $k \in \tilde{K}$

$$l_j + \epsilon \leq x^k_j, \tag{30}$$

$$\nabla_j f(x^k) \geq \nabla_{\hat{i}} f(x^k) \geq \eta > 0. \tag{31}$$

From (30) and (31) we have that the direction $d = -e_j$ is a feasible descent direction at $\bar{x}$. Using the fact that $j \in W^k$ and repeating the same reasonings employed above to prove (27) we can write

$$\lim_{k\to\infty, k\in K} |\nabla f(x^k)^T d| = |\nabla f(\bar{x})^T d| = |\nabla_j f(\bar{x})| = 0,$$

which contradicts (31).

By assuming that $\nabla_{\hat{i}} f(\bar{x}) < 0$ and by repeating similar reasonings we would obtain again a contradiction. □

The decomposition algorithm proposed in this work has the same theoretical convergence properties of the known Gauss–Southwell algorithm. Both the approaches are based on the strategy of selecting the variables that mostly violate the optimality conditions. The Gauss–Southwell algorithm refers to optimality conditions for a general class of problems with convex feasible set. These conditions are expressed in terms of distance between the current point and the one obtained by the projected gradient (involving $2n$ comparisons, $2n$ subtractions), so that they do not take into account the specificity of box constrained problems. The proposed algorithm exploits information of the reduced gradient (involving n comparisons) whose opposite identifies the normalized feasible steepest descent direction (see [5]). Then we may expect that the working set selection rule of the proposed algorithm may identify promising subproblem variables and could lead to faster convergence, preserving theoretical convergence properties and the possibility of including any other group of variables as in the Gauss–Southwell algorithm. However, in order to draw significant computational conclusions, extensive numerical experiments on different types of problems should be conducted, and this may be the object of a future work.

Appendix

Proof of Proposition 4 We proceed by contradiction. Let us assume that there exists an infinite subset $K \subseteq \{0, 1, \ldots\}$ such that

$$\lim_{k\to\infty, k\in K} x^k = \bar{x} \tag{32}$$

and $\bar{x}$ is not a stationary point, that is, there exists an index $\hat{i} \in \{1, \ldots, n\}$ such that

$$|\bar{x}_{\hat{i}} - [\bar{x} - \nabla f(\bar{x})]^+_{\hat{i}}| \geq \eta > 0. \tag{33}$$

The instructions of the algorithm imply

$$f(x^{k+1}) \leq f(x^k),$$

and hence, using (32) and the continuity of f, it follows that the nonincreasing sequence $\{f(x^k)\}$ converges, i.e.,

$$\lim_{k\to\infty} f(x^k) = f(\bar{x}). \tag{34}$$

Since $i^k \in \{1, \ldots, n\}$, we can extract a further subset, relabelled by K, such that $i^k = \bar{i}, \forall k \in K$.

By definition we have

$$|x^k_{\bar{i}} - [x^k - \nabla f(x^k)]^+_{\bar{i}}| \geq |x^k_{\hat{i}} - [x^k - \nabla f(x^k)]^+_{\hat{i}}|,$$

and thus, using (33) and the continuity of the projection mapping, we can write

$$|\bar{x}_{\bar{i}} - [\bar{x} - \nabla f(\bar{x})]^+_{\bar{i}}| \geq \eta > 0.$$

Let $\bar{y}$ be the point defined as follows

$$\bar{y}_h = \begin{cases} \bar{x}_h & h \neq \bar{i} \\ (\bar{x} - \nabla f(\bar{x}))_h & h = \bar{i} \end{cases}$$

As $[\bar{y}]^+$ belongs to the convex set $\mathcal{F}$, the direction

$$d = [\bar{y}]^+ - \bar{x}$$

is a feasible direction at $\bar{x}$, and is such that $d_{\bar{i}} \neq 0$ and $d_h = 0$ for $h \neq \bar{i}$.

Using the properties of the projection mapping, we have

$$\left(\bar{y} - [\bar{y}]^+\right)^T (\bar{x} - [\bar{y}]^+) = \left(\bar{x}_{\bar{i}} - \nabla_{\bar{i}} f(\bar{x}) - [\bar{y}]^+_{\bar{i}}\right) (\bar{x}_{\bar{i}} - [\bar{y}]^+_{\bar{i}}) \leq 0,$$

from which we get

$$\nabla_{\hat{i}} f(\bar{x})([\bar{y}]^+_{\bar{i}} - \bar{x}_{\bar{i}}) \leq -(\bar{x}_{\bar{i}} - [\bar{y}]^+_{\bar{i}})^2 \leq -\eta^2,$$

so that we can write

$$\nabla f(\bar{x})^T d = \nabla_{\bar{i}} f(\bar{x})([\bar{y}]^+_{\bar{i}} - \bar{x}_{\bar{i}}) < 0. \tag{35}$$

From (35) and the continuity assumption on the gradient ∇f it follows that for $k \in K$ and k sufficiently large

$$\nabla f(x^k)^T d < 0, \tag{36}$$

that is d is a descent direction for f at x^k. Note that, by the working set selection rule of the algorithm, $\bar{i} \in W^k$ for $k \in K$ and k sufficiently large. Therefore, the instructions of the algorithm imply

$$f(x^{k+1}) \leq f(x^k + \lambda^k d) \leq f(x^k) + \gamma \lambda^k \nabla f(x^k)^T d \tag{37}$$

where λ^k is the stepsize determined by means of Armijo algorithm.

From (36) it follows that Assumption 1 holds for the subsequence $\{d^k\}_K$ setting $d^k = d$ for all $k \in K$. Further, from (32), (34), and (37) we have that the assumptions

of Proposition 3 hold for the subsequence $\{x^k\}_K$, and hence we have

$$\lim_{k\to\infty, k\in K} \beta^k \nabla f(x^k)^T d = 0, \tag{38}$$

where $\beta^k = \min\{\beta^k_{\mathcal{F}}, \beta_u\}$, being $\beta^k_{\mathcal{F}}$ the maximum feasible steplength along d, and β_u a prefixed positive number. Proposition 2 implies that $\beta^k \geq \bar{\beta} > 0$ for $k \in K$ and k sufficiently large, so that, recalling (38), we obtain

$$\lim_{k\to\infty, k\in K} \nabla f(x^k)^T d = \nabla f(\bar{x})^T d = 0,$$

which contradicts (35). □

References

1. Bertsekas, D.P.: Nonlinear Programming. 2nd edn. Athena Scientific, New York (1999)
2. Facchinei, F., Lucidi, S., Palagi, L.: A truncated Newton algorithm for large scale box constrained optimization. SIAM J. Optim. **12**, 1100–1125 (2002)
3. Grippo, L., Sciandrone, M.: On the convergence of the block nonlinear Gauss–Seidel method under convex constraints. Oper. Res. Lett. **26**, 127–136 (2000)
4. Han, C.G., Pardalos, P.M., Ye, Y.: Computational aspects of an interior point algorithm for quadratic programming problems with box constraints. Large Scale Numer. Optim. 92–112 (1990)
5. Hsu, C.-W., Lin, C.-J. A simple decomposition method for support vector machines. Mach. Learn. **46**, 291–314 (2002)
6. Joachims, T.: Making large scale SVM learning practical. In: Schölkopf, C.B.B., Smola, A. (eds.) Advances in Kernel Methods—Support Vector Learning. MIT Press, Cambridge (1998)
7. Lin, C.-J., Moré, J.J.: Newton's method for large bound-constrained optimization problems. SIAM J. Optim. **9**, 1100–1127 (1999)
8. Lin, C.-J.: On the convergence of the decomposition method for support vector machines. IEEE Trans. Neural Netw. **12**, 1288–1298 (2001)
9. Lin, C.-J., Lucidi, S., Palagi, L., Risi, A., Sciandrone, M.: A decomposition algorithm model for singly linearly constrained problems subject to lower and upper bounds. J. Optim. Theory Appl. (2009) (to appear)
10. Luo, Z.Q., Tseng, P.: On the convergence of the coordinate descent method for convex differentiable minimization. J. Optim. Theory Appl. **72**, 7–35 (1992)
11. Moré, J.J., Toraldo, G.: Algorithms for bound constrained quadratic programming problems. Numerische Mathematik **55**, 377–400 (1988)
12. Pardalos, P.M., Resende, M.G.C.: Handbook of Applied Optimization. Oxford University Press, New York (2002)
13. Pardalos, P.M., Kovoor, N.: An algorithm for a singly constrained class of quadratic programs subject to upper and lower bounds. In: Mathematical Programming. Springer, Heidelberg, vol. 46-1, pp. 321–328 (1990)
14. Tseng, P., Yun, S.: A coordinate descent method for nonsmooth separable minimization. Math. Program. B **117**, 387–423 (2009)