

On the convergence of a Jacobi-type algorithm for Singly Linearly-Constrained Problems Subject to simple Bounds

Giampaolo Liuzzi* Laura Palagi† Mauro Piacentini‡

Abstract

In this work we define a block decomposition Jacobi-type method for nonlinear optimization problems with one linear constraint and bound constraints on the variables. We prove convergence of the method to stationary points of the problem under quite general assumptions.

Keywords: Decomposition method Jacobi-type iteration

*Consiglio Nazionale delle Ricerche - Istituto di Analisi dei Sistemi ed Informatica "A. Ruberti" (CNR - IASI) - Viale Manzoni 30 - Rome, Italy. Tel.: +39-06-7716439, Fax: +39-06-7716461, e-mail: liuzzi@iasi.cnr.it

†Sapienza Università di Roma - Dipartimento di Informatica e Sistemistica A. Ruberti (DIS) - Via Ariosto 25 - Rome, Italy. Tel.: +39-06-77274081, e-mail: palagi@dis.uniroma1.it

‡Sapienza Università di Roma - Dipartimento di Informatica e Sistemistica A. Ruberti (DIS) - Via Ariosto 25 - Rome, Italy. Tel.: +39-06-77274085, e-mail: piacentini@dis.uniroma1.it

1 Introduction

Let us consider the problem

$$\min_{x \in \mathcal{F}} f(x) \tag{1}$$

where

$$\mathcal{F} = \{x \in \mathbb{R}^n : a^T x = b, l \leq x \leq u\}$$

and $a, l, u \in \mathbb{R}^n$, with $-\infty \leq l < u \leq +\infty$, $b \in \mathbb{R}$. We allow the possibility that some of the variables are unbounded by permitting both $l_i = -\infty$ and $u_i = \infty$ for some $i \in \{1, \dots, n\}$.

There are many problems that can be formulated as special cases of problem (1). In particular, training a Support Vector Machine (SVM) (see e.g. [19]) leads to a problem of type (1) where $f(x)$ is a convex quadratic function, $b = 0$ and $a_i \in \{-1, 1\}$, $0 = l_i < u_i = C$ with $C > 0$ for $i = 1, \dots, n$.

Another problem of type (1) is the so called Standard Quadratic Programming problem (StQP) where $f(x)$ is an indefinite quadratic form, $b = 1$ and $a_i = 1$, $l_i = 0$, $u_i = \infty$ for all $i = 1, \dots, n$. StQP problems arises, for example, as formulations of maximum clique problems (see e.g.[16, 2]).

In this paper, for the solution of Problem (1), we are interested in decomposition methods, which involve the solution of subproblems of smaller dimensions in place of the original one. In literature, decomposition methods for unconstrained problems can be roughly classified into two main classes: Gauss-Seidel methods (see e.g. [1, 6, 7, 3]) and Jacobi methods (see e.g. [6, 1, 5]). Gauss-Seidel methods define the new iterate x^{k+1} by sequentially updating the variables. On the other hand, Jacobi methods define the new iterate x^{k+1} by simultaneously updating all the variables.

Convergence results for these two classes of methods have been proved for unconstrained problems in e.g. [6, 1, 5]. For constrained problems, the block Gauss-Seidel method has been proved to be convergent when the feasible set is the Cartesian product of closed convex sets [7]. The presence of the linear constraint in problem (1) does not allow to use such results for defining decomposition methods. However, in the context of SVM training, much effort has been devoted to the definition of convergent decomposition schemes that fit in the class of block Gauss-Seidel-type methods. In this decomposition framework, starting from a feasible point, at each iteration k a subset $W^k \subset \{1, \dots, n\}$ of indices of variables is chosen and the new iterate x^{k+1} is defined by updating only the variables with indices belonging to W^k . The choice of set W^k at each iteration plays a crucial role in proving convergence of the sequence $\{x^k\}$. In particular, in most decomposition methods for SVM problem, the indices in W^k are selected on the basis of the violation of the optimality conditions at x^k (see e.g. [8, 9, 10, 11, 12, 13, 15, 17]). In [14] a decomposition algorithm for Problem (1) has been defined that differs from the other ones in that the selection rule does not require neither to apply any specific ordering procedure nor to exploit information about the current iterate x^k . Convergence of the scheme has been proved under either the convexity assumption of the objective function or by using a proximal point modification.

In this paper, we define a block Jacobi-type convergent decomposition method for problem (1) which makes use of a cyclic selection rule for W that falls in the rule proposed in [14]. Up to our knowledge, this is the first attempt to extend a block Jacobi-type iteration to the solution of constrained problems of type (1).

The paper is organized as follows. In Section 2, we introduce some notations and definitions. In Section 3 we define the block Jacobi-type decomposition method for Problem (1) which we call **CoJac**. The convergence properties of Algorithm **CoJac** are studied in Section 4 under some general conditions on the search directions used. Section 5 is devoted to the choice of directions that fulfill the conditions used in the convergence analysis.

2 Notation and definitions

In this section we introduce some useful definitions and the basic notation that will be used throughout the paper.

Given an index set $W \subset \{1, \dots, n\}$, we denote by \overline{W} its complement with respect to $\{1, \dots, n\}$, that is, $\overline{W} = \{1, \dots, n\} \setminus W$. Given a vector $v \in \mathbb{R}^n$ and an index set W , we denote by $v_W \in \mathbb{R}^{|W|}$ the subvector of v made up of the component v_i with $i \in W$. Further, for the sake of simplicity we use the notation $\nabla_W f$ for $(\nabla f)_W$.

We denote by $\mathcal{W} = \{W^1, \dots, W^M\} \subseteq 2^{\{1, \dots, n\}}$ a family of index sets of cardinality q^1, \dots, q^M .

The set of *feasible directions* at a point $x \in \mathcal{F}$ is the cone

$$D(x) = \{d \in \mathbb{R}^n \mid a^T d = 0, d_i \geq 0, \forall i : x_i = l_i, \text{ and } d_i \leq 0, \forall i : x_i = u_i\}.$$

Next we define a stationary point for Problem (1).

Definition 2.1 (Stationary point) *A point $x^* \in \mathcal{F}$, is stationary for Problem (1) if*

$$\nabla f(x^*)^T d \geq 0 \quad \text{for all } d \in D(x^*).$$

Given a feasible point \tilde{x} , and a subset $W \subset \{1, \dots, n\}$, let us define the subproblem $P_W(\tilde{x})$ as

$$\min_{x_W \in \mathcal{F}_W(\tilde{x})} f(x_W, \tilde{x}_{\overline{W}}) \tag{2}$$

where

$$\mathcal{F}_W(\tilde{x}) = \{x_W \in \mathbb{R}^{|W|} : a_W^T x_W = b - a_{\overline{W}}^T \tilde{x}_{\overline{W}}, l_W \leq x_W \leq u_W\}$$

At any feasible point $x \in \mathcal{F}$, let us denote with $D_W(x_W)$ the set of feasible directions at x_W with respect to $\mathcal{F}_W(x)$, that is

$$D_W(x_W) = \{d \in \mathbb{R}^{|W|} : a_W^T d = 0, d_i \geq 0, \forall i \in W : x_i = l_i, \text{ and } d_i \leq 0, \forall i \in W : x_i = u_i\}.$$

We introduce the definition of stationary points for Problem (2).

Definition 2.2 (Stationary point of $P_W(x)$) *Given $x \in \mathcal{F}$ and the corresponding problem $P_W(x)$, a point $x_W^* \in \mathcal{F}_W(x)$ is stationary for $P_W(x)$ if*

$$\nabla_W f(x_W^*, x_{\overline{W}}^*)^T d \geq 0 \quad \text{for all } d \in D_W(x_W^*).$$

3 A constrained Jacobi-type algorithm

In this section we introduce a block Jacobi-type decomposition algorithm for Problem (1). The iterate x^{k+1} is generated by using information on the simultaneous (approximated) minimizations with respect to the components of the vector. To be more precise, let $\mathcal{W} = \{W^1, \dots, W^M\}$ be a family of index sets which does not define necessarily a partition of $\{1, \dots, n\}$, so that, differently from the block Jacobi-type method defined in [6], $W^i \cap W^j$, with $i \neq j$, may be nonempty.

We require that the sequence be generated in such a way to satisfy

$$f(x^{k+1}) \leq \min_{W^i} f_{ref_{W^i}}^k$$

where $f_{ref_{W^i}}^k$ are reference values for subproblems $P_{W^i}(x^k)$, $i = 1, \dots, M$. The reference values could be obtained by performing exact minimizations with respect to each x_{W^i} , namely setting

$$f_{ref_{W^i}}^k = \min_{x_{W^i} \in \mathcal{F}_{W^i}(x^k)} f(x_{W^i}, x_{\overline{W^i}}^k),$$

but this could be a strict requirement, particularly when $|W^i| > 2$ or the objective function is not convex. Hence, following the idea in [14, 13] we relax this requirement by asking only for a “sufficient reduction”, and the values $f_{ref_{W^i}}^k$ are obtained by means of an Armijo-type LineSearch (LS) procedure along a feasible direction d_{W^i} at $x_{W^i}^k$ for $i = 1, \dots, M$.

Now, we describe the Armijo-type linesearch scheme.

Given a feasible point $x \in \mathcal{F}$, a set $W \in \mathcal{W}$, and a direction $d_W \in D_W(x_W)$, let $\beta_{\mathcal{F}}$ be the maximum feasible steplength along direction d_W with respect to the bound constraints. Namely, $\beta_{\mathcal{F}}$ is such that

$$l_W \leq x_W + \beta d_W \leq u_W \quad \text{for all } \beta \in [0, \beta_{\mathcal{F}}],$$

and (since $-\infty \leq l < u \leq \infty$) we have that either $\beta_{\mathcal{F}} = +\infty$ or at least an index $i \in W$ exists such that

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

Further, let $\beta_u > 0$ be a positive scalar.

We report below the Armijo-type procedure `Step_length` that returns the stepsize α_W along the direction d_W .

Procedure Step_length(x, W, d_W)

Parameter. $\gamma \in (0, 1/2)$, $\delta \in (0, 1)$.

Data. $x \in \mathcal{F}$, $W \in \mathcal{W}$, $d_W \in D_W(x_W)$ and $\beta = \min\{\beta_{\mathcal{F}}, \beta_u\}$.

Inizialization. If d_W satisfies $\nabla_W f(x)^T d_W \geq 0$, **Return** $\alpha_W = 0$;
otherwise set $\alpha = \beta$.

While $(f(x_W + \alpha d_W, x_{\overline{W}}) > f(x) + \gamma \alpha \nabla_W f(x)^T d_W)$

Set $\alpha = \delta \alpha$

End While

Return $\alpha_W = \alpha$.

The stepsize α_W is zero if and only if d_W does not satisfy the descent condition $\nabla_W f(x)^T d_W < 0$ at x .

The following proposition shows that Procedure **Step_length** is well-defined.

Proposition 3.1 *Let $W \in \mathcal{W}$. Assume $x \in \mathcal{F}$ and $d_W \in D_W(x_W)$. Then Procedure **Step_length** determines, in a finite number of iterations, a scalar α_W such that*

$$f(x + \alpha_W d) \leq f(x) + \gamma \alpha_W \nabla_W f(x)^T d_W. \quad (3)$$

Proof. The proof is quite standard and can be found e.g. in [1]. For the sake of completeness, we report it in the appendix. \square

Now we are ready to define the decomposition algorithm **CoJac**.

Constrained Jacobi-type (CoJac) Algorithm

Data. A point $x^0 \in \mathcal{F}$, and $\mathcal{W} = \{W^1, \dots, W^M\}$.

For $k = 0, 1, 2, \dots$

1. **For** $i = 1, \dots, M$

choose $d_{W^i}^k \in D_{W^i}(x^k)$;

compute $\alpha_{W^i}^k$ by means of procedure `Step_length`($x^k, W^i, d_{W^i}^k$);

set $y_{W^i}^k = x_{W^i}^k + \alpha_{W^i}^k d_{W^i}^k$.

2. **End For**

3. **Choose** x^{k+1} such that

$$f(x^{k+1}) \leq \min_{1 \leq i \leq M} f(y_{W^i}^k, x_{W^i}^k).$$

End For

We point out the degree of freedom on the choice for the next iterate x^{k+1} in Algorithm CoJac.

Convergence properties of the algorithm CoJac will be analyzed in the next section, under quite standard assumptions on the feasible directions used in the procedure `Step_length` and on the rule for the selection of the family of index sets \mathcal{W} .

4 Convergence results

In order to simplify the exposition, we first introduce an assumption on the behavior in the limit of the directions used by the algorithm. This is a basic assumption that can be satisfied by employing suitable methods for the computation of the search directions d_W^k .

Assumption 1 For a given working set $W \in \mathcal{W}$, let $\{x^k\}$ be a sequence of feasible points such that, for $K \subseteq \{0, 1, 2, \dots\}$,

$$\lim_{k \rightarrow \infty, k \in K} x^k = \hat{x}.$$

Given a sequence $\{d_W^k\}$ such that $d_W^k \in D_W(x_W^k)$ we have:

- (i) $x_W^k + d_W^k \in \mathcal{F}_W(x_W^k)$;
- (ii) a constant $M > 0$ exists such that $\|d_W^k\| \leq M$ for all $k \in K$;

(iii) if \hat{x} is not a stationary point for $P_W(\hat{x})$, then

$$\limsup_{k \rightarrow \infty, k \in K} \nabla_W f(x^k)^T d_W^k < 0.$$

Possible choices of directions satisfying Assumption 1 are reported and discussed in Section 5.

From the definition of procedure `Step_length`, when the set $W \in \mathcal{W}$ is held fixed, we can state the following preliminary convergence result.

Proposition 4.1 *Let $\{x^k\} \subset \mathcal{F}$ be a sequence of feasible points converging to a given point \hat{x} . For a fixed $W \in \mathcal{W}$, and every k , let $d_W^k \in D_W(x_W^k)$ satisfy Assumption (1) and*

$$y_W^k = x_W^k + \alpha_W^k d_W^k, \quad (4)$$

where α_W^k is computed by means of the `Step_length` Procedure. Then the sequence $\{y_W^k\}$, is such that:

(i) $f(y_W^k, x_W^k) \leq f(x^k)$, for all $k = 0, 1, 2, \dots$

(ii) if

$$\lim_{k \rightarrow \infty} \left(f(x^k) - f(y_W^k, x_W^k) \right) = 0 \quad (5)$$

then \hat{x} is a stationary point for $P_W(\hat{x})$.

Proof. The proof is quite technical and is therefore reported in the appendix. \square

We stress the fact that $\{x^k\}$ is a given sequence that may not depend on Procedure `Step_length` in the sense that x^{k+1} is not necessarily defined on the basis of y_W^k .

Now we report the main result regarding convergence of the CoJac Algorithm. As standard in decomposition algorithms, we must require that the family of index sets \mathcal{W} satisfies a suitable rule. In particular we require that the family \mathcal{W} satisfies the following condition:

Pairwise Inclusion Property (PIP)¹

For each pair $\{i, j\} \subset \{1, \dots, n\}$, there exists at least an $\ell \in \{1, \dots, M\}$ such that, $\{i, j\} \subseteq W^\ell$.

We observe that the definition of a family of index sets \mathcal{W} satisfying Condition PIP does not require any information about the current iterate and can thus be defined a priori.

¹As an example, for $n = 6$ a possible family of index sets \mathcal{W} with $q^i = 4$ for all i satisfying condition (PIP) is $\mathcal{W} = \{W^1, \dots, W^3\}$ with $W^1 = \{1, 2, 3, 4\}$, $W^2 = \{1, 2, 5, 6\}$, $W^3 = \{3, 4, 5, 6\}$.

Proposition 4.2 *Let the family of index sets $\mathcal{W} = \{W^1, \dots, W^M\}$ satisfy condition PIP. Let $\{x^k\}$ and $\{d_{W^i}^k\}$ for each $i = 1, \dots, M$ be the sequences defined by Algorithm CoJac. Assume that, for each $i = 1, \dots, M$, sequences $\{d_{W^i}^k\}$ satisfy Assumption 1. Then, every limit point of $\{x^k\}$ is a stationary point for Problem (1).*

Proof. The proof of the proposition depends on a number of technical results and is therefore reported in the appendix. \square

5 Search directions

In this section, for any fixed W^h , $h = 1, \dots, M$, we report two well-known methods for calculating a feasible direction $d_{W^h}^k$ over a convex set satisfying Assumption 1. In particular, we consider the Frank-Wolfe direction and the Projected Gradient direction. First we report a useful theoretical result (whose proof can be found in, e.g., [14]) that will be used in the section.

Proposition 5.1 *Let $\{x^k\} \subset \mathcal{F}$ be a sequence of feasible points converging to a point $\tilde{x} \in \mathcal{F}$. Then, for sufficiently large values of k ,*

$$D(\tilde{x}) \subseteq D(x^k).$$

5.1 Frank-Wolfe direction.

For a given $W \in \mathcal{W}$ and a feasible point x^k , the Frank-Wolfe (F-W) direction is

$$d_W^k = \bar{x}_W^k - x_W^k,$$

where \bar{x}_W^k is the optimal solution of the following linear programming problem

$$\min_{x_W \in \mathcal{F}_W(x^k)} \nabla_W f(x^k)^T (x_W - x_W^k) \quad (6)$$

In the next proposition we show that the F-W direction is well defined and that the desired properties stated in Assumption 1 hold.

Proposition 5.2 *Assume that the feasible set \mathcal{F} is compact. Let $W \in \mathcal{W}$, $\{x^k\}$ be a sequence of feasible points and $\{d_W^k\}$ be the associated sequence of F-W directions. Then, sequence $\{d_W^k\}$ is well defined and the following conditions hold:*

- (i) for any k , $d_W^k \in D_W(x_W^k)$ and $x_W^k + d_W^k \in \mathcal{F}_W(x^k)$;
- (ii) for any k , $\nabla_W f(x^k)^T d_W^k < 0$ if and only if x^k is not a stationary point for $P_W(x^k)$;
- (iii) for any k , $\|d_W^k\| \leq U$ for a given constant $U > 0$;
- (iv) assume that

$$\lim_{k \rightarrow \infty, k \in K} x^k = \tilde{x},$$

where $K \subseteq \{0, 1, 2, \dots\}$ and \tilde{x} is not a stationary point of $P_W(\tilde{x})$, then

$$\limsup_{k \rightarrow \infty, k \in K} \nabla_W f(x^k)^T d_W^k < 0.$$

Proof. For any k , compactness of \mathcal{F} implies that problem (6) admits a solution \bar{x}_W^k so that d_W^k is well defined. Point (i) follows by definition of d_W^k . By definition of d_W^k , it holds that

$$\nabla_W f(x^k)^T d_W^k = \nabla_W f(x^k)^T (\bar{x}_W^k - x_W^k) \leq \nabla_W f(x^k)^T d, \quad \forall d \in D_W(x_W^k). \quad (7)$$

Hence, if x^k is not a stationary point of $P_W(x^k)$ we have that $\nabla_W f(x^k)^T d_W^k < 0$ and vice versa, so that point (ii) holds. Further, we observe that $\|d_W^k\| = \|\bar{x}_W^k - x_W^k\| \leq \|\bar{x}_W^k\| + \|x_W^k\|$ so that by compactness of \mathcal{F} we have point (iii).

Now, let us relabel as $\{x^k\}$ the subsequence converging to a feasible point \tilde{x} which is not a stationary point of $P_W(\tilde{x})$. Let $\tilde{d}_W \in D_W(\tilde{x}_W)$ be a feasible direction such that

$$\nabla_W f(\tilde{x})^T \tilde{d}_W < 0.$$

For sufficiently large values of k , by Proposition (5.1), we have

$$\tilde{d}_W \in D_W(x_W^k),$$

and by continuity of the gradient

$$\nabla_W f(x^k)^T \tilde{d}_W < 0.$$

So that, using (7), we also get that for sufficiently large values of k

$$\nabla_W f(x^k)^T d_W^k \leq \nabla_W f(x^k)^T \tilde{d}_W < 0.$$

By taking the limit, we obtain

$$\limsup_{k \rightarrow \infty} \nabla_W f(x^k)^T d_W^k \leq \nabla_W f(\tilde{x})^T \tilde{d}_W < 0, \quad (8)$$

which finally proves point (iv). \square

5.2 Projected Gradient direction.

We recall that the projection $P_S(x)$ of a point x over a non-empty closed convex set S is the solution of the following problem

$$\min_{y \in S} \|x - y\|.$$

The projection operator enjoys the following properties.

Proposition 5.3 *The projection operator is continuous and not expansive. Further, $P_S(x)$ is the projection of x over S if and only if*

$$(x - P_S(x))^T (y - P_S(x)) \leq 0, \quad \forall y \in S. \quad (9)$$

For a fixed $W \in \mathcal{W}$ and a feasible point x^k , let us define $P_{\mathcal{F}_W}^k$ as the projection operator over $\mathcal{F}_W(x^k)$. We consider the following Projected Gradient (PG) direction

$$d_W^k = \bar{x}_W^k - x_W^k$$

where

$$\bar{x}_W^k = P_{\mathcal{F}_W}^k(x_W^k - s\nabla_W f(x^k)) \quad (10)$$

and s is a positive scalar.

In the next proposition we show that the PG direction is well defined and that Assumption 1 holds.

Proposition 5.4 *Assume that \mathcal{F} is compact. Let $W \in \mathcal{W}$, $\{x^k\}$ be a sequence of feasible points and $\{d_W^k\}$ be the associated sequence of PG directions. Then, sequence $\{d_W^k\}$ is well defined and the following conditions hold.*

- (i) for any k , $d_W^k \in D_W(x_W^k)$ and $x_W^k + d_W^k \in \mathcal{F}_W(x^k)$;
- (ii) for any k , $\nabla_W f(x^k)^T d_W^k < 0$ if and only if x^k is not a stationary point for $P_W(x^k)$;
- (iii) for any k , $\|d_W^k\| \leq U$ for a given constant $U > 0$;
- (iv) assume that

$$\lim_{k \rightarrow \infty, k \in K} x^k = \tilde{x},$$

where $K \subseteq \{0, 1, 2, \dots\}$ and \tilde{x} is not a stationary point of $P_W(\tilde{x})$, then

$$\limsup_{k \rightarrow \infty, k \in K} \nabla_W f(x^k)^T d_W^k < 0.$$

Proof. For any k , the point \bar{x}_W^k as in (10) is always defined and hence also d_W^k . Furthermore, by definition $x_W^k + d_W^k = \bar{x}_W^k \in \mathcal{F}_W(x^k)$ so that we get point (i). By (9), \bar{x}_W^k satisfies

$$(x_W^k - s\nabla_W f(x^k) - \bar{x}_W^k)^T (y - \bar{x}_W^k) \leq 0, \quad \forall y \in \mathcal{F}_W(x^k), \quad (11)$$

so that, choosing $y = x_W^k$, we obtain by simple manipulations

$$\nabla_W f(x^k)^T d_W^k \leq -\frac{1}{s} \|d_W^k\|^2 \leq 0. \quad (12)$$

Recalling that x^k is a stationary point of Problem $\mathcal{P}_W(x^k)$ if and only if

$$x_W^k = P_{\mathcal{F}_W}^k(x_W^k - s\nabla_W f(x^k)),$$

for any scalar $s > 0$, we have that $d_W^k = 0$ if and only if x_W^k is a stationary point of Problem $\mathcal{P}_W(x^k)$. Hence point (ii) follows by (12).

Reasoning as in the proof of Proposition 5.2, we have also point(iii).

Assume now that $\{x^k\}_K$ is a subsequence converging to a non-stationary point \tilde{x} of $\mathcal{P}_W(\tilde{x})$ and relabel it as $\{x^k\}$. Let $\tilde{d}_W \in D_W(\tilde{x}_W)$ be a feasible direction such that

$$\nabla_W f(\tilde{x})^T \tilde{d}_W < 0.$$

Using proposition (5.1) and the continuity of the gradient, we have, for sufficiently large values of k , that

$$\tilde{d}_W \in D_W(x_W^k), \quad (13)$$

and

$$\nabla_W f(x^k)^T \tilde{d}_W < 0. \quad (14)$$

Reasoning by contradiction, we suppose that point (iv) does not hold, namely

$$\limsup_{k \rightarrow \infty} \nabla_W f(x^k)^T d_W^k \geq 0,$$

so that as result of (12)

$$\limsup_{k \rightarrow \infty} \|d_W^k\|^2 = 0. \quad (15)$$

By (13), for k sufficiently large, a value $t^k > 0$ exists such that $x_W^k + t\tilde{d}_W \in \mathcal{F}_W(x^k)$ for all $t \in [0, t^k]$ and $\lim_{k \rightarrow \infty} t^k > 0$. Hence a value $\bar{t} > 0$ exists such that the point $y = x_W^k + \bar{t}\tilde{d}_W \in \mathcal{F}_W(x^k)$, for all k sufficiently large. Substituting in (11), we have that

$$0 \geq (x_W^k - s\nabla_W f(x^k) - \bar{x}_W^k)^T (x_W^k + t\tilde{d}_W - \bar{x}_W^k) = (-d_W^k - s\nabla_W f(x^k))^T (-d_W^k + \bar{t}\tilde{d}_W).$$

Rearranging the above inequality

$$\begin{aligned} \nabla_W f(x^k)^T d_W^k &\leq -\frac{1}{s}\|d_W^k\|^2 + \frac{\bar{t}}{s}\tilde{d}_W^T d_W^k + \bar{t}\nabla_W f(x^k)^T \tilde{d}_W \\ &\leq -\frac{1}{s}\|d_W^k\|^2 + \frac{\bar{t}}{s}\|\tilde{d}_W\|\|d_W^k\| + \bar{t}\nabla_W f(x^k)^T \tilde{d}_W \end{aligned}$$

Taking the limit of the above inequality, using (15) and (14), we get the contradiction

$$\limsup_{k \rightarrow \infty} \nabla_W f(x^k)^T d_W^k \leq \bar{t} \limsup_{k \rightarrow \infty} \nabla_W f(x^k)^T \tilde{d}_W < 0. \quad \square$$

In [4] and [18] methods to obtain the projected gradient direction have been proposed.

Appendix

A Proofs of the convergence results

Proof of Proposition 3.1. If d_W does not satisfy $\nabla f(x)^T d_W < 0$, the stepsize $\alpha_W = 0$ and the condition is obviously satisfied.

Assume now that d_W satisfies $\nabla f(x)^T d_W < 0$ and by contradiction that the algorithm does not terminate. Hence, let $d = (d_W, d_{\overline{W}})$ with $d_{\overline{W}} = 0$, we can write

$$f(x + \beta\delta^j d) > f(x) + \gamma\beta\delta^j \nabla f(x)^T d \quad \text{for all } j.$$

By applying the Mean Value theorem we have

$$\nabla f(x + \theta_j \beta \delta^j d)^T d > \gamma \nabla f(x)^T d \quad \text{for all } j, \quad (16)$$

with $\theta_j \in (0, 1)$. Taking limits in (16) for $j \rightarrow \infty$ we obtain

$$(1 - \gamma) \nabla f(x)^T d \geq 0,$$

which implies, together with the fact that $\gamma \in (0, 1)$, that $\nabla f(x)^T d \geq 0$, and this contradicts the descent assumption on d . \square

Proof of Proposition 4.1. Point (i) easily follows from the definition of the Armijo-type scheme and of y_W^k in (4).

Now, let us consider point (ii). As a result of closedness of the feasible set \mathcal{F} , the limit point \hat{x} is feasible. By definition of Armijo-type rule, we have

$$f(x^k) - f(y_W^k, x_{\overline{W}}^k) \geq |\alpha_W^k \nabla_W f(x^k)^T d_W^k|,$$

which, by (5), yields

$$\lim_{k \rightarrow \infty} \alpha_W^k \nabla_W f(x^k)^T d_W^k = 0. \quad (17)$$

Reasoning by contradiction, let us suppose that \hat{x} is not a stationary point. Then, by Assumption 1, we have that

$$\limsup_{k \rightarrow \infty} \nabla_W f(x^k)^T d_W^k < 0.$$

Let $K \subseteq \{0, 1, 2, \dots\}$ be a subset of the iteration set such that

$$\lim_{k \rightarrow \infty, k \in K} \nabla_W f(x^k)^T d_W^k < 0, \quad (18)$$

so that, for $k \in K$ and sufficiently large, d_W^k satisfies $\nabla_W f(x^k)^T d_W^k < 0$. Then, it follows from (17) that

$$\lim_{k \rightarrow \infty, k \in K} \alpha_W^k = 0. \quad (19)$$

By definition of the Step-length Procedure and by point (i) of Assumption 1 we have that, for all $k \in K$, the initial stepsize β is such that $\beta \geq \min\{1, \beta_u\} > 0$. Hence, by virtue of (19), the initial stepsize will be reduced at least once, so that, for $k \in K$ and sufficiently large,

$$f(x^k) - f(x_W^k + \frac{\alpha_W^k}{\delta} d_W^k, x_{\overline{W}}^k) < -\gamma \frac{\alpha_W^k}{\delta} \nabla_W f(x^k)^T d_W^k.$$

By using the mean value theorem, we obtain from the above relation

$$-\nabla_W f(x^k + \eta^k \frac{\alpha_W^k}{\delta} d_W^k, x^k) d_W^k < -\gamma \frac{\alpha_W^k}{\delta} \nabla_W f(x^k) d_W^k, \quad (20)$$

where $\eta^k \in (0, 1)$.

Since $\{d_W^k\}$ is bounded by Assumption 1, we can find a further set of indices, that we relabel again K , such that

$$\lim_{k \rightarrow \infty, k \in K} d_W^k = \hat{d}_W \quad \text{and} \quad \lim_{k \rightarrow \infty, k \in K} x^k = \hat{x}$$

Hence, taking the limit in (20) for $k \rightarrow \infty$ and $k \in K$, it follows that

$$0 \leq (1 - \gamma) \nabla_W f(\hat{x})^T \hat{d}_W,$$

which, recalling that $\gamma \in (0, 1/2)$, yields

$$0 \leq \nabla_W f(\hat{x})^T \hat{d}_W.$$

The proof follows by noting that the above inequality contradicts (18), that is,

$$\lim_{k \rightarrow \infty, k \in K} \nabla_W f(x^k)^T d_W^k = \nabla_W f(\hat{x})^T \hat{d}_W < 0.$$

□

To prove the main convergence results of the paper we need to introduce some more technical notation and preliminary results concerning problem (1).

Given a point $x \in \mathcal{F}$ let us define the following index sets

$$\begin{aligned} R(x) &= \{i : (x_i < u_i \text{ and } a_i > 0) \text{ or } (x_i > l_i \text{ and } a_i < 0)\}, \\ S(x) &= \{i : (x_i < u_i \text{ and } a_i < 0) \text{ or } (x_i > l_i \text{ and } a_i > 0)\}. \end{aligned}$$

Then, we denote by $D_{RS}(x)$ the set of directions $d^{i,j}$ with $i \in R(x)$ and $j \in S(x)$, namely

$$D_{RS}(x) = \bigcup_{\substack{i \in R(x) \\ j \in S(x) \\ i \neq j}} d^{i,j}.$$

Further, given indices $i, j \in \{1, \dots, n\}$, with $i \neq j$, we denote by $d^{i,j}$ a vector belonging to \mathbb{R}^n such that

$$d_h^{i,j} = \begin{cases} 1/a_i, & \text{if } h = i \\ -1/a_j, & \text{if } h = j \\ 0, & \text{otherwise} \end{cases}$$

In [13] it has been proved that for any feasible point x the set $D_{RS}(x)$ is a subset of the set of feasible directions at x and that it contains the generators of $D(x)$. The following proposition (that also has been proved in [13]) gives a characterization of the stationary points of Problem (1) using only the directions in $D_{RS}(x)$.

Proposition A.1 *A feasible point $x^* \in \mathcal{F}$ is stationary for Problem (1) if and only if*

$$\nabla f(x^*)^T d^{i,j} \geq 0 \quad \forall d^{i,j} \in D_{RS}(x^*).$$

Now, given a subset of indices W and a feasible point $x \in \mathcal{F}$, let us consider subproblem $P_W(x)$. We denote by $D_{W \cap RS}(x)$ the subset of feasible directions of $D_W(x)$ with exactly two nonzero components, namely

$$D_{W \cap RS}(x) = \bigcup_{\substack{i \in R(x) \cap W \\ j \in S(x) \cap W \\ i \neq j}} d_W^{i,j}. \quad (21)$$

Now we are ready to prove Proposition 4.2.

Proof of Proposition 4.2. Let \bar{x} be any limit point of a subsequence of $\{x^k\}$, i.e., there exists an infinite subset $K \subseteq \{0, 1, \dots\}$ such that

$$\lim_{k \rightarrow \infty, k \in K} x^k = \bar{x}.$$

By contradiction, let us assume that \bar{x} is not a stationary point for Problem (1). By Proposition A.1 there exists at least a pair $(i, j) \in R(\bar{x}) \times S(\bar{x})$, and a direction $d^{i,j} \in D_{RS}(\bar{x})$ such that:

$$\nabla f(\bar{x})^T d^{i,j} < 0. \quad (22)$$

By Condition PIP on the family of working sets, we know that $W^h \in \mathcal{W}$, with $h \in \{1, \dots, M\}$, exists such that $(i, j) \in W^h$. Let us consider the subvector $d_{W^h}^{i,j} \in D_{W^h \cap RS}(\bar{x})$, so that we have

$$\nabla_{W^h} f(\bar{x})^T d_{W^h}^{i,j} < 0, \quad (23)$$

that is to say that \bar{x} is not a stationary point of problem $P_{W^h}(\bar{x})$ as well. By construction, we have

$$f(x^{k+1}) \leq f(y_{W^h}^k, x_{W^h}^k) \leq f(x^k)$$

so that the sequence $\{f(x^k)\}_K$ is not increasing and thus converges to $f(\bar{x})$. Hence, we get that

$$\lim_{k \rightarrow \infty, k \in K} \left(f(x^k) - f(y_{W^h}^k, x_{W^h}^k) \right) = 0. \quad (24)$$

Then, using Proposition 4.1, we have that

$$\nabla_{W^h} f(\bar{x})^T d^{r,s} \geq 0, \quad \forall d^{r,s} \in D_{W^h \cap RS}(\bar{x})$$

which, for $d_{W^h}^{i,j}$, contradicts (23) and hence also (22), thus concluding the proof. \square

References

- [1] D.P. BERTSEKAS AND J.N. TSITSIKLIS, *Parallel and Distributed Computation*. Prentice-Hall International Editions, Englewood Cliffs, NJ, 1989.
- [2] I.M. BOMZE. Evolution towards the Maximum clique. *Journal of Global Optimization* 10: 143-164, 1997.
- [3] A. CASSIOLI, M. SCIANDRONE, *A convergent decomposition method for box-constrained optimization problems*, *Optimization Letters*, 3, pp. 397–409, 2009.
- [4] Y.-H. DAI AND R. FLETCHER, *New algorithms for singly linearly constrained quadratic programs subject to lower and upper bounds*, *Mathematical Programming*, 106, Number 3, Pages 403-421, 2006.
- [5] M.C. FERRIS AND O. L. MANGASARIAN, *Parallel variable distribution*. *SIAM Journal on Optimization* 4: 815-832, 1994.
- [6] L. GRIPPO, M. SCIANDRONE, *Globally convergent block-coordinate techniques for unconstrained optimization*. *Optimization Methods and Software*, Vol. 10 (4), pp.587-637, 1999.
- [7] ———, *On the convergence of the block nonlinear Gauss-Seidel method under convex constraints*. *Operations Research Letters*, Vol. 26 (3), pp.127-136, 2000.
- [8] T. GLASMACHERS, C. IGEL, *Maximum-Gain Working Set Selection for SVMs*, *Journal of Machine Learning Research*, 7, pp. 1437–1466, 2006.
- [9] T. JOACHIMS, *Making large scale SVM learning practical*, in *Advances in Kernel Methods - Support Vector Learning*, C. B. B Schölkopf, C.J.C. Burges and A. Smola, eds., MA: MIT Press, Cambridge, 1998.
- [10] S.S. KEERTHI AND E.G. GILBERT, *Convergence of a Generalized SMO algorithm for SVM classifier design*. *Machine Learning*, Vol. 46, pp. 351–360, 2002.
- [11] C.-J. LIN, *On the convergence of the decomposition method for Support Vector Machines*, *IEEE Transactions on Neural Networks*, 12 (2001), pp. 1288–1298.
- [12] ———, *Asymptotic convergence of an SMO algorithm without any assumptions*, *IEEE Transactions on Neural Networks*, 13 (2002), pp. 248–250.
- [13] C.-J. LIN, S. LUCIDI, L. PALAGI, A. RISI, M. SCIANDRONE, *A decomposition algorithm model for singly linearly constrained problems subject to lower and upper bounds*. *Journal of Optimization Theory and Applications*, vol. 141, No. 1, pp. 107-126, 2009.
- [14] S. LUCIDI, L. PALAGI, A. RISI, M. SCIANDRONE, *A convergent decomposition algorithm for Support Vector Machines*. *Computational Optimization and Applications*, 38:217–234, 2007.
- [15] ———, *A convergent hybrid decomposition algorithm model for SVM training*. *IEEE on neural Networks*, Volume 20, Issue 6, June 2009, Page(s):1055 - 1060 DOI 10.1109/TNN.2009.2020908.
- [16] T.S. MOTZKIN AND STRAUSS E.G. (1965). *Maxima for graphs and a new proof of a theorem of Turan* *Canadian J. Math.* **17**, 533-540.

- [17] L. PALAGI, M. SCIANDRONE (2005). On the convergence of a modified version of SVM^{light} Algorithm. *Optimization methods and Software* vol.20, No.2-3, 2005, pp. 311-328.
- [18] T. SERAFINI, G. ZANGHIRATI, L. ZANNI. *Gradient Projection Methods for Large Quadratic Programs and Applications in Training Support Vector Machines*. *Optim. Meth. Soft.* 20 (2005), 353-378.
- [19] V.N. VAPNIK, *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.