On the convergence of a modified version of the SVMlight algorithm.

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

On the convergence of a modified version of SVM^{light} algorithm

L. Palagi [a]; M. Sciandrone [b]
[a] Dipartimento di Informatica e Sistemistica 'Antonio Ruberti', Università di Roma 'La Sapienza', Roma, Italy [b] Istituto di Analisi dei Sistemi ed Informatica 'Antonio Ruberti', Consiglio Nazionale delle Ricerche, Roma, Italy

Online Publication Date: 01 April 2005

To cite this Article Palagi, L. and Sciandrone, M.(2005)'On the convergence of a modified version of SVM^{light} algorithm',Optimization Methods and Software,20:2,317 — 334
To link to this Article: DOI: 10.1080/10556780512331318209
URL: http://dx.doi.org/10.1080/10556780512331318209

Optimization Methods and Software
Vol. 20, No. 2–3, April–June 2005, 317–334

Taylor & Francis
Taylor & Francis Group

On the convergence of a modified version of SVM^{light} algorithm

L. PALAGI† and M. SCIANDRONE‡*

†Dipartimento di Informatica e Sistemistica 'Antonio Ruberti',
Università di Roma 'La Sapienza', Via Buonarroti 12, 00185 Roma, Italy
‡Istituto di Analisi dei Sistemi ed Informatica 'Antonio Ruberti',
Consiglio Nazionale delle Ricerche, Viale Manzoni 30, 00185 Roma, Italy

(Received 29 November 2002; in final form 8 September 2003)

In this work, we consider the convex quadratic programming problem arising in support vector machine (SVM), which is a technique designed to solve a variety of learning and pattern recognition problems. Since the Hessian matrix is dense and real applications lead to large-scale problems, several decomposition methods have been proposed, which split the original problem into a sequence of smaller subproblems. SVM^{light} algorithm is a commonly used decomposition method for SVM, and its convergence has been proved only recently under a suitable block-wise convexity assumption on the objective function. In SVM^{light} algorithm, the size q of the working set, i.e. the dimension of the subproblem, can be any even number. In the present paper, we propose a decomposition method on the basis of a proximal point modification of the subproblem and the basis of a working set selection rule that includes, as a particular case, the one used by the SVM^{light} algorithm. We establish the asymptotic convergence of the method, for any size $q \geq 2$ of the working set, and without requiring any further block-wise convexity assumption on the objective function. Furthermore, we show that the algorithm satisfies in a finite number of iterations a stopping criterion based on the violation of the optimality conditions.

Keywords: Support vector machines; SVM^{light} algorithm; Decomposition methods; Proximal point

1. Introduction

The support vector machine (SVM) [1,2] is a promising technique for solving a variety of machine learning, classification and function estimation problems. Given a training set of input-target pairs (x^i, y^i), $i = 1, \ldots, l$, with $x^i \in R^n$, and $y^i \in \{-1, 1\}$, the SVM technique requires the solution of the following convex quadratic programming problem

$$\begin{aligned} \min \quad & f(\alpha) = \frac{1}{2}\alpha' Q\alpha - e'\alpha \\ \text{s.t.} \quad & y'\alpha = 0 \\ & 0 \leq \alpha \leq Ce, \end{aligned} \tag{1}$$

*Corresponding author. Email: sciandro@iasi.rm.cnr.it

Optimization Methods and Software
ISSN 1055-6788 print/ISSN 1029-4937 online © 2005 Taylor & Francis Group Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/10556780512331318209

where $\alpha \in R^l$, Q is a $l \times l$ positive semidefinite matrix, $e \in R^l$ is the vector of all ones, $y \in \{-1, 1\}^l$ and C is a positive scalar. The generic element q_{ij} of the matrix Q is given by $y^i y^j K(x^i, x^j)$, where $K(x, z) = \phi(x)'\phi(z)$ is the kernel function related to the nonlinear function ϕ that maps the data from the input space into the feature space.

Problem (1) is a convex problem with a very simple structure; however, since Q is a fully dense matrix, traditional optimization methods cannot be directly employed when the dimension l, i.e. the number of training data, is extremely large, as it happens in many real applications. This has motivated the study and design of block decomposition methods [3–5] which involve the solution of many subproblems of smaller dimension in place of the original problem.

In a general decomposition framework, at each iteration k, the vector of variables α^k is partitioned into two subvectors $(\alpha^k_W, \alpha^k_{\overline{W}})$, where $W \subset \{1, \ldots, l\}$ identifies the variables of the subproblem to be solved and is called the *working set*, and $\overline{W} = \{1, \ldots, l\} \setminus W$ (for notational convenience the dependence of W and $\overline{W}$ on k is omitted). Then, starting from the current vector $\alpha^k = (\alpha^k_W, \alpha^k_{\overline{W}})$, which is a feasible point, the subvector α^{k+1}_W is computed as the solution of the following subproblem

$$\begin{aligned} \min_{\alpha_W} \quad & f(\alpha_W, \alpha^k_{\overline{W}}) \\ & y'_W \alpha_W = -y'_{\overline{W}} \alpha^k_{\overline{W}} \\ & 0 \le \alpha_W \le C e_W. \end{aligned} \tag{2}$$

The subvector $\alpha^{k+1}_{\overline{W}}$ is unchanged, i.e. $\alpha^{k+1}_{\overline{W}} = \alpha^k_{\overline{W}}$, and the new iterate is given by $\alpha^{k+1} = (\alpha^{k+1}_W, \alpha^{k+1}_{\overline{W}})$. In general, the cardinality q of the working set, i.e. the dimension of the subproblem, is prefixed according to, for instance, the available computational capability, and is kept constant for all iterates. The rule used for selecting the working set W at each iteration plays a crucial role, since it influences the convergence properties of the generated sequence $\{\alpha^k\}$. Note that the most popular convergent decomposition methods for nonlinear optimization, such as the successive over-relaxation algorithm and the Jacobi and Gauss–Seidel algorithms are applicable only when the feasible set is the Cartesian product of subsets defined in smaller subspaces [6]. Since problem (1) contains an equality constraint, such decomposition methods cannot be employed.

A very simple decomposition method for SVM is the sequential minimal optimization (SMO) algorithm [5], where only two variables are selected in the working set at each iteration, i.e. $q = 2$, so that an analytical solution of the subproblem (2) can be found, and this eliminates the need to use an optimization software. The choice of the two variables with respect to optimization is performed, is determined by some heuristic devoted to individuate which ones may provide a better contribution to the progress towards the solution.

A modified version of SMO has been proposed in ref. [7], where the two indices of the working set are those corresponding to the 'maximal violation' of the Karush–Kuhn–Tucker (KKT) conditions. This modification of SMO algorithm can in turn be viewed as a special case of the SVM^{light} algorithm [3], which is based on a specific procedure for choosing the q elements of the working set, where q is any even number.

SVM^{light} algorithm is a commonly used decomposition method for SVM, and its convergence properties have been established only recently. In particular, for any even size q of the working set, the asymptotic convergence of the algorithm has been proved in ref. [8] under a suitable strict block-wise convexity assumption on f. However, as remarked in ref. [9], this assumption may not hold if, for instance, some data points in the training set are the same.

In ref. [12], the convergence of the algorithm is proved, for the special case of $q = 2$, without requiring the strict block-wise convexity assumption on f.

In this work, we define a decomposition method which is similar to the SVM^{light} algorithm. The differences are in the selection rule and in the objective function of the subproblem to be solved at each iteration. In particular, we introduce a working set selection (WWS) rule that includes, as a particular case, the one used by the SVM^{light} algorithm, but does not restrict the size q of the working set to be an even number (the only constraint is $q \geq 2$). Moreover, alternatively to the standard subproblem (2), we define a modified subproblem of the form

$$\begin{aligned} \min_{\alpha_W} \quad & f(\alpha_W, \alpha_{\overline{W}}^k) + \tau \|\alpha_W - \alpha_W^k\|^2 \\ & y_W' \alpha_W = -y_{\overline{W}}' \alpha_{\overline{W}}^k \\ & 0 \leq \alpha_W \leq C e_W, \end{aligned}$$

where the objective function contains the additional quadratic *proximal point term* $\tau \|\alpha_W - \alpha_W^k\|^2$, where $\tau > 0$. Roughly speaking, the proximal point term plays the role of a 'convexifying' term of the objective function of the subproblem with respect to the subvector α_W. This allows us to remove the block-wise convexity assumption on f needed to prove convergence of the SVM^{light} algorithm. In particular, under the only assumption that f is convex, we prove that any limit point of the sequence $\{\alpha^k\}$ generated by our decomposition method is a solution of problem (1). The convergence analysis is based on some key ideas exploited in ref. [8], but follows a different guideline inspired from preceding papers [10,11] concerning decomposition methods for nonlinear optimization.

We emphasize that the focus of this paper is theoretical, namely the study of the convergence properties of the proposed SVM^{light}-type decomposition algorithm. However, we believe that the proximal point modification may be helpful also from a numerical point of view when using iterative methods to solve the subproblems (hence in the case $q > 2$). In our opinion, the study of methods for solving the subproblems and the definition of suitable truncated criteria deserve attention and need further work, but this is out of the scope of this paper.

The paper is organized as follows. In section 2, we state some definitions and technical results that we use to prove convergence of the method. In section 3, we introduce the WSS rule and the decomposition algorithm (called proximal point decomposition, PPD, algorithm). Section 4 is devoted to the convergence analysis of PPD algorithm, and we prove that every limit point of the sequence generated is a global minimum of Problem (1). In section 5, we show that a stopping criterion, derived in ref. [7], used in ref. [12] and analysed in ref. [13], which is based on the gap of the violation of the optimality conditions, can be used in PPD algorithm. Finally, section 6 contains some concluding remarks.

2. Notation and preliminary results

In this section, we state some results on problem (1) (whose proofs are reported in the Appendix) that will be used for the convergence analysis of the decomposition algorithm defined in the next section. Actually, these results, except for Proposition 3, have been proved in ref. [14], where a decomposition method for problem of type (1) is proposed that uses a different approach with respect to the SVM^{light} one for the WSS.

First, we introduce some basic notation and definitions. Throughout the paper, we denote by $\mathcal{F}$ the feasible set of problem (1), namely

$$\mathcal{F} = \{\alpha \in R^l : y'\alpha = 0,\ 0 \leq \alpha \leq Ce\},$$

and by $\nabla f = Q\alpha - e$ the gradient of f.

Given a vector $\alpha \in R^l$, and an index set $W \subseteq \{1, \ldots, l\}$, we have already introduced the notation $\alpha_W \in R^{|W|}$ to indicate the subvector of α made up of the component α_i with $i \in W$. Furthermore, given a matrix Q and two index sets $U, V \subseteq \{1, \ldots, l\}$, we denote by Q_{UV} the $|U| \times |V|$ submatrix made up of elements q_{ij} with $i \in U$ and $j \in V$.

For every feasible point α, we denote the sets of indices of active (lower and upper) bounds as follows:

$$L(\alpha) = \{i: \alpha_i = 0\}, \quad U(\alpha) = \{i: \alpha_i = C\}.$$

Since the feasible set $\mathcal{F}$ is compact, problem (1) admits solution. Moreover, as f is convex and the constraints are linear, a feasible point α^* is a solution of problem (1) if and only if the KKT conditions are satisfied, i.e. a scalar λ^* exists such that

$$(\nabla f(\alpha^*))_i + \lambda^* y_i \begin{cases} \geq 0 & \text{if } i \in L(\alpha^*) \\ \leq 0 & \text{if } i \in U(\alpha^*) \\ = 0 & \text{if } i \notin L(\alpha^*) \cup U(\alpha^*). \end{cases}$$

The KKT conditions can be written in a different form. To this aim, the sets L and U can be split into L^-, L^+ and U^-, U^+, respectively, where

$$L^-(\alpha) = \{i \in L(\alpha): y_i < 0\}, \quad L^+(\alpha) = \{i \in L(\alpha): y_i > 0\}$$
$$U^-(\alpha) = \{i \in U(\alpha): y_i < 0\}, \quad U^+(\alpha) = \{i \in U(\alpha): y_i > 0\}.$$

We report the KKT conditions in the following proposition.

PROPOSITION 1 (Optimality Conditions) *A point $\alpha^* \in \mathcal{F}$ is a solution of problem* (1) *if and only if there exists a scalar λ^* satisfying*

$$\begin{aligned} \lambda^* &\geq -\frac{(\nabla f(\alpha^*))_i}{y_i} \quad \forall\, i \in L^+(\alpha^*) \cup U^-(\alpha^*) \\ \lambda^* &\leq -\frac{(\nabla f(\alpha^*))_i}{y_i} \quad \forall\, i \in L^-(\alpha^*) \cup U^+(\alpha^*) \\ \lambda^* &= -\frac{(\nabla f(\alpha^*))_i}{y_i} \quad \forall\, i \notin L(\alpha^*) \cup U(\alpha^*). \end{aligned} \tag{3}$$

In correspondence to a feasible point α, the following index sets can be defined:

$$R(\alpha) = L^+(\alpha) \cup U^-(\alpha) \cup \{i: 0 < \alpha_i < C\},$$
$$S(\alpha) = L^-(\alpha) \cup U^+(\alpha) \cup \{i: 0 < \alpha_i < C\}.$$

These sets have been introduced in ref. [8] in the form

$$\begin{aligned} R(\alpha) &= \{i: (\alpha_i < C \text{ and } y_i > 0) \text{ or } (\alpha_i > 0 \text{ and } y_i < 0)\}, \\ S(\alpha) &= \{i: (\alpha_i < C \text{ and } y_i < 0) \text{ or } (\alpha_i > 0 \text{ and } y_i > 0)\}, \end{aligned} \tag{4}$$

where the indices in $R(\alpha)$ are called 'bottom' candidates, and the indices in $S(\alpha)$ are 'top' candidates.

We have the following results.

PROPOSITION 2 *A feasible point α^* is a solution of problem* (1) *if and only if there exists no pair of indices i and j, with $i \in R(\alpha^*)$ and $j \in S(\alpha^*)$, such that*

$$-\frac{(\nabla f(\alpha^*))_i}{y_i} > -\frac{(\nabla f(\alpha^*))_j}{y_j}. \tag{5}$$

PROPOSITION 3 *Let $\{\alpha^k\}$ be a sequence of feasible points convergent to a point $\bar{\alpha}$. Then for sufficiently large values of k we have*

$$R(\bar{\alpha}) \subseteq R(\alpha^k) \quad \textit{and} \quad S(\bar{\alpha}) \subseteq S(\alpha^k).$$

The set of the feasible directions at α is the cone

$$D(\alpha) = \{d \in R^l : y'd = 0,\ d_i \geq 0,\ \forall\, i \in L(\alpha),\ \text{and } d_i \leq 0,\ \forall\, i \in U(\alpha)\}.$$

Then we can state the following result.

PROPOSITION 4 *Let $\hat{\alpha}$ be a feasible point. For each pair $i \in R(\hat{\alpha})$ and $j \in S(\hat{\alpha})$, the direction $d \in R^l$ such that*

$$d_i = \frac{1}{y_i}, \quad d_j = -\frac{1}{y_j}, \quad d_h = 0 \quad \text{for} \quad h \neq i, j$$

is a feasible direction at $\hat{\alpha}$, i.e. $d \in D(\hat{\alpha})$.

3. A proximal point modification of SVM^{light} algorithm

The basic strategy of a decomposition method is that of performing, at each iteration, the minimization of the objective function with respect to only a subset of variables, holding fixed the remaining variables. With reference to SVM problem (1), the subproblem (2) to be solved at any iteration k takes the form:

$$\begin{aligned} \min_{\alpha_W} f(\alpha_W, \alpha^k_{\overline{W}}) = &\frac{1}{2}\alpha'_W Q_{WW}\alpha_W - (e - Q_{W\overline{W}}\alpha^k_{\overline{W}})'\alpha_W \\ &y'_W\alpha_W = -y'_{\overline{W}}\alpha^k_{\overline{W}} \\ &0 \leq \alpha_W \leq Ce_W, \end{aligned} \tag{6}$$

where W is the working set at iteration k and $\overline{W} = \{1, \ldots, l\}\backslash W$ (for notational convenience we have omitted the dependence of W and $\overline{W}$ on the iteration counter k when this is not confusing). Note that, due to the presence of the linear equality constraint, the smallest number of variables that can be changed at each iteration to retain feasibility is two, so that the cardinality q of the working set W must be at least two.

As already observed in the introduction, a fundamental issue in the design of a decomposition method is the rule for selecting the working set W at each iteration. SVM^{light} algorithm is a commonly used decomposition method for SVM, and is based on a specific rule related to the violation of the optimality conditions.

In particular, the idea in ref. [3] is to find a steepest descent feasible direction with exactly q non-zero elements and to select in the working set the indices corresponding to these elements. This leads to solve the problem

$$\min_d \left\{ \nabla f(\alpha^k)'d : d \in D(\alpha^k),\ -e \leq d \leq e,\ \left|\{i : d_i \neq 0\}\right| = q \right\}.$$

A simple strategy to solve it, and hence to identify the indices in W, has been proposed in ref. [3]. In ref. [14] it has been point out that, in theory, a solution satisfying the constraint

$\left|\{i: d_i \neq 0\}\right| = q$ may not exist. Later in ref. [8], it has been proved that the procedure proposed in ref. [3] really solves the problem:

$$\min_d \left\{ \nabla f(\alpha^k)' d : d \in D(\alpha^k),\ -e \leq d \leq e,\ \left|\{i: d_i \neq 0\}\right| \leq q \right\}.$$

The procedure for the solution of this problem has been described in a compact form in refs. [8,13] using the sets $R(\alpha)$ and $S(\alpha)$ given in equation (4).

We introduce here a slightly more general rule than that of the SVM^{light}, which mimics one introduced in ref. [13]. To this aim, at any feasible point α, we define the index sets

$$I(\alpha) = \left\{ i : i = \arg \max_{h \in R(\alpha)} -\frac{(\nabla f(\alpha))_h}{y_h} \right\}, \quad J(\alpha) = \left\{ j : j = \arg \min_{h \in S(\alpha)} -\frac{(\nabla f(\alpha))_h}{y_h} \right\}. \quad (7)$$

At iteration k, the WSS rule can be described as follows.

Data: Integers $q_1, q_2 \geq 1$.

(i) Select q_1 indices in $R(\alpha^k)$ sequentially so that

$$-\frac{\nabla f(\alpha^k)_{i^1(k)}}{y_{i^1(k)}} \geq -\frac{\nabla f(\alpha^k)_{i^2(k)}}{y_{i^2(k)}} \geq \cdots \geq -\frac{\nabla f(\alpha^k)_{i^{q_1}(k)}}{y_{i^{q_1}(k)}}$$

with $i^1(k) \in I(\alpha^k)$.

(ii) Select q_2 indices in $S(\alpha^k)$ sequentially so that

$$-\frac{\nabla f(\alpha^k)_{j^1(k)}}{y_{j^1(k)}} \leq -\frac{\nabla f(\alpha^k)_{j^2(k)}}{y_{j^2(k)}} \leq \cdots \leq -\frac{\nabla f(\alpha^k)_{j^{q_2}(k)}}{y_{j^{q_2}(k)}}$$

with $j^1(k) \in J(\alpha^k)$

(iii) Set $W^k = \{i^1, \ldots, i^{q_1}, j^1, \ldots, j^{q_2}\}$.

We remark that the WSS rule employed in SVM^{light} algorithm is a particular case of WSS rule, with $q_1 = q_2 = q/2$, where q is an even number.

The asymptotic convergence of SVM^{light} algorithm has been established in ref. [8], under the assumption that

$$\min_{I:\ |I| \leq q} (\mathrm{eig}_{\min}(Q_{II})) > 0, \quad (8)$$

where I is any subset of $\{1, \ldots, l\}$ with $|I| \leq q$ and $\mathrm{eig}_{\min}(Q_{II})$ denotes the minimum eigenvalue of the matrix Q_{II}. Note that assumption (8) implies that the objective function is strictly convex with respect to block components of cardinality $\leq q$. However, it does not hold, for example, if some training data are the same. As showed in ref. [9], assumption (8) is not necessary for ensuring the convergence of SVM^{light} algorithm in the particular case of $q = 2$, which corresponds to the well-known SMO algorithm.

From the convergence analysis performed in ref. [8], we may deduce that the key role of hypothesis (8) stays in the fact that it permits to ensure that the distance between successive points of the sequence $\{\alpha^k\}$ generated by the decomposition methods tends to zero, i.e.

$$\lim_{k \to \infty} \|\alpha^{k+1} - \alpha^k\| = 0. \quad (9)$$

This is an important requirement to establish convergence properties in the context of a decomposition strategy. Indeed, in a decomposition method, at the end of each iteration k, only the satisfaction of the optimality conditions with respect to the variables associated to W^k is ensured. Therefore, to get convergence towards KKT points, it may be necessary to ensure that consecutive points, which are solutions of the corresponding subproblems, tend to the same limit point.

In order to ensure property (9) without requiring assumption (8), we employ a proximal point technique [11,15,16]. In particular, a proximal point term of the form $\tau \|\alpha_W - \alpha_W^k\|^2$, with $\tau > 0$, is added to the objective function of the subproblem (6), thus obtaining the following subproblem

$$\begin{aligned} \min_{\alpha_W} \quad & f(\alpha_W, \alpha_{\overline{W}}^k) + \tau \|\alpha_W - \alpha_W^k\|^2 \\ & y_W' \alpha_W = -y_{\overline{W}}' \alpha_{\overline{W}}^k \\ & 0 \le \alpha_W \le C e_W. \end{aligned} \tag{10}$$

Since f is quadratic, the objective function of problem (10) is still quadratic and can be written as follows

$$\frac{1}{2}\alpha_W' \left(Q_{WW} + 2\tau I_W\right) \alpha_W - \left(e_W - Q_{W\overline{W}}\alpha_{\overline{W}}^k - 2\tau\alpha_W^k\right)'\alpha_W,$$

where I_W denotes the identity matrix of dimension $|W|$. Note that problem (10) has the same structure of subproblem (6), but now, since the objective function is strictly convex, the solution is unique. Thus, the solution of problem (10) requires at most the same effort than the solution of subproblem (6).

We are ready to define formally the proximal point modification of the SVM^{light} decomposition method, which we call PPD algorithm, as follows.

PPD ALGORITHM

Data: a feasible point α^0, $\tau > 0$.
Inizialization: Set $k = 0$.

While (stopping criterion not satisfied)

1. Select the working set W^k according to the WSS rule.
2. Set $W = W^k$. Find the solution α_W^* of problem (10).
3. Set $\alpha_i^{k+1} = \begin{cases} \alpha_i^* & \text{if } i \in W \\ \alpha_i^k & \text{otherwise.} \end{cases}$
4. Set $k = k + 1$.

end while
Return $\alpha^* = \alpha^k$

In the next section, we prove the asymptotic convergence of PPD algorithm. In Section 5, we show that PPD algorithm satisfies the stopping criterion proposed in refs. [7,12].

4. Convergence analysis

We first prove some preliminary results that are independent of the WSS rule used in PPD algorithm for defining the working set W^k.

PROPOSITION 5 *Assume that* PPD *algorithm does not terminate and let* $\{\alpha^k\}$ *be the sequence generated. Then we have*

$$\lim_{k\to\infty} \|\alpha^{k+1} - \alpha^k\| = 0.$$

Proof By the instructions of the algorithm, we have for all k

$$f(\alpha^{k+1}) + \tau\|\alpha^{k+1} - \alpha^k\|^2 = f(\alpha_W^{k+1}, \alpha_{\overline{W}}^k) + \tau\|\alpha_W^{k+1} - \alpha_W^k\|^2 \le f(\alpha_W^k, \alpha_{\overline{W}}^k) = f(\alpha^k), \tag{11}$$

so that the sequence $\{f(\alpha^k)\}$ is decreasing. Since $\{\alpha^k\}$ belongs to the feasible set, which is compact, then there exists a subsequence $\{\alpha^k\}_K$ such that $\lim_{k\to\infty, k\in K} \alpha^k = \bar{\alpha}$. As f is continuous, we have that $\{f(\alpha^k)\}_K$ converges to $f(\bar{\alpha})$, and this implies that the whole sequence $\{f(\alpha^k)\}$ converges to $f(\bar{\alpha})$. Then, the convergence of the sequence $\{f(\alpha^k)\}$ to a finite value and equation (11) imply that $\|\alpha^{k+1} - \alpha^k\| \to 0$. ■

As an immediate consequence of Proposition 5, we have the following result.

LEMMA 1 *Assume that* PPD *algorithm does not terminate and let* $\{\alpha^k\}$ *be the sequence generated. Let* $\{\alpha^k\}_K$ *be a subsequence convergent to a point* $\bar{\alpha}$*, i.e. there exists an infinite subset* $K \subseteq \{0, 1, \ldots\}$ *such that* $\alpha^k \to \bar{\alpha}$ *for* $k \to \infty$, $k \in K$*. Then, for any integer* p*, we have that*

$$\lim_{k\to\infty, k\in K} \alpha^{k+p} = \bar{\alpha}.$$

Proof Given any integer p, we can write

$$\|\alpha^{k+p} - \alpha^k\| \le \|\alpha^{k+p} - \alpha^{k+p-1}\| + \|\alpha^{k+p-1} - \alpha^{k+p-2}\| + \cdots + \|\alpha^{k+1} - \alpha^k\|. \tag{12}$$

By Proposition 5, we have that $\|\alpha^{k+j+1} - \alpha^{k+j}\| \to 0$ for all finite $j = 0, 1 \ldots$. From equation (12), we get that $\|\alpha^{k+p} - \alpha^k\| \to 0$ and hence, as $\alpha^k \to \bar{\alpha}$, we get also that $\alpha^{k+p} \to \bar{\alpha}$ for $k \to \infty$, $k \in K$. ■

In the proof of convergence of PPD algorithm we make use of the following result.

LEMMA 2 *Let* $\{\alpha^k\}$ *be the sequence generated by* PPD *algorithm. Assume that* (i, j) *is a pair such that:*

$$(i, j) \in W^k \quad \text{and} \quad (i, j) \in R(\alpha^{k+1}) \times S(\alpha^{k+1}).$$

Then

$$\nabla f(\alpha^{k+1})' d^{i,j} + 2\tau(\alpha^{k+1} - \alpha^k)' d^{i,j} \ge 0,$$

where $d^{i,j} \in R^l$ *is the direction defined as*

$$d_i^{i,j} = \frac{1}{y_i}, \quad d_j^{i,j} = -\frac{1}{y_j}, \quad d_h^{i,j} = 0 \text{ for } h \ne i, j.$$

Proof For simplicity, let $W = W^k$. By Proposition 4, we know that $d^{i,j}$ is a feasible direction at α^{k+1}. Let $d_W^{i,j}$ be the subvector of $d^{i,j}$ with elements in W; since $i, j \in W$ we have that $d_{\overline{W}}^{i,j} = 0$. Recalling that $\alpha_W^{k+1} = \alpha_W^*$ and $\alpha_{\overline{W}}^{k+1} = \alpha_{\overline{W}}^k$, it is immediate to verify that

the direction $d_W^{i,j}$ is a feasible direction for the subproblem (10) at α_W^*. Since equation (10) is a convex programming problem, the optimality conditions can be written as:

$$\nabla_W f(\alpha_W^*, \alpha_{\overline{W}}^k)' d_W^{i,j} + 2\tau(\alpha_W^* - \alpha_W^k)' d_W^{i,j} \geq 0,$$

where $\nabla_W f$ denotes the subvector of ∇f with components in W. Recalling again that $\alpha_W^{k+1} = \alpha_W^*$, $\alpha_{\overline{W}}^{k+1} = \alpha_{\overline{W}}^k$ and $d_{\overline{W}}^{i,j} = 0$, we get

$$\nabla f(\alpha^{k+1})' d^{i,j} + 2\tau(\alpha^{k+1} - \alpha^k)' d^{i,j} = \nabla_W f(\alpha_W^*, \alpha_{\overline{W}}^k)' d_W^{i,j} + 2\tau(\alpha_W^* - \alpha_W^k)' d_W^{i,j} \geq 0,$$

and hence the result. ■

Now we are ready to prove the asymptotic convergence of PPD algorithm.

PROPOSITION 6 *Assume that* PPD *algorithm does not terminate, and let* $\{\alpha^k\}$ *be the sequence generated by it. Then, every limit point of* $\{\alpha^k\}$ *is a solution of problem* (1).

Proof Let $\bar{\alpha}$ be any limit point of a subsequence of $\{\alpha^k\}$, i.e. there exists an infinite subset $K \subseteq \{0, 1, \ldots\}$ such that $\alpha^k \to \bar{\alpha}$ for $k \in K$, $k \to \infty$.

By contradiction, let us assume that $\bar{\alpha}$ is not a KKT point for problem (1). By Proposition 2, there exists at least a pair $(i, j) \in R(\bar{\alpha}) \times S(\bar{\alpha})$such that:

$$-\frac{(\nabla f(\bar{\alpha}))_i}{y_i} > -\frac{(\nabla f(\bar{\alpha}))_j}{y_j}. \tag{13}$$

According to the WSS rule, at iteration k, the indices $i^1(k) \in I(\alpha^k)$ and $j^1(k) \in J(\alpha^k)$ are inserted in the working set W^k [where $I(\alpha^k)$ and $J(\alpha^k)$ are defined in equation (7)].

The proof is divided in two parts.

a. Suppose first that there exists an integer $s \geq 0$ such that:

$$i^1(k + m(k)) \in R(\alpha^{k+m(k)+1}) \quad \text{and} \quad j^1(k + m(k)) \in S(\alpha^{k+m(k)+1})$$
$$\text{for some } m(k) \in [0, s]. \tag{14}$$

Since $i^1(k)$ and $j^1(k)$ belong to the finite set $\{1, \ldots, l\}$, we can extract a further subset of K, which we relabel again with K, such that

$$i^1(k + m(k)) = \widehat{i}, \quad j^1(k + m(k)) = \widehat{j} \quad \text{for all } k \in K.$$

Lemma 1 implies that $\alpha^{k+m(k)} \to \bar{\alpha}$ for $k \to \infty$, $k \in K$. Then, recalling that, by definition, $\widehat{i}, \widehat{j} \in W^{k+m(k)}$ for all $k \in K$, we can define a subsequence $\{\alpha^k\}_{K_1}$ such that for all $k \in K_1$

- $(\widehat{i}, \widehat{j}) \in W^k$
- $(\widehat{i}, \widehat{j}) \in R(\alpha^{k+1}) \times S(\alpha^{k+1})$
- $\alpha^k \to \bar{\alpha}$ for $k \to \infty, k \in K_1$.

Hence, we can apply Lemma 2 and write:

$$\nabla f(\alpha^{k+1})' d^{\widehat{i},\widehat{j}} + 2\tau(\alpha^{k+1} - \alpha^k)' d^{\widehat{i},\widehat{j}} \geq 0 \quad \text{for all } k \in K_1.$$

By Proposition 5, we have $\|\alpha^{k+1} - \alpha^k\| \to 0$, so that, recalling the continuity of ∇f and the definition of $d^{\hat{i},\hat{j}}$ in Lemma 2, taking limits for $k \to \infty$, $k \in K_1$, we obtain

$$\nabla f(\bar{\alpha})' d^{\widehat{i},\widehat{j}} = \frac{(\nabla f(\bar{\alpha}))_{\widehat{i}}}{y_{\widehat{i}}} - \frac{(\nabla f(\bar{\alpha}))_{\widehat{j}}}{y_{\widehat{j}}} \geq 0. \tag{15}$$

On the other hand, the indices i, j satisfying equation (13) are such that, by Proposition 3, $i \in R(\alpha^k)$ and $j \in S(\alpha^k)$ for $k \in K_1$ and k sufficiently large. Hence, taking into account the definition of $\widehat{i}$, $\widehat{j}$ and the WSS rule, we can write for $k \in K_1$ and k sufficiently large,

$$-\frac{(\nabla f(\alpha^k))_{\widehat{i}}}{y_{\widehat{i}}} \geq -\frac{(\nabla f(\alpha^k))_i}{y_i} \quad \text{and} \quad -\frac{(\nabla f(\alpha^k))_{\widehat{j}}}{y_{\widehat{j}}} \leq -\frac{(\nabla f(\alpha^k))_j}{y_j}.$$

Taking limits for $k \to \infty$, $k \in K_1$, we obtain

$$-\frac{(\nabla f(\bar{\alpha}))_{\widehat{i}}}{y_{\widehat{i}}} \geq -\frac{(\nabla f(\bar{\alpha}))_i}{y_i} \quad \text{and} \quad -\frac{(\nabla f(\bar{\alpha}))_{\widehat{j}}}{y_{\widehat{j}}} \leq -\frac{(\nabla f(\bar{\alpha}))_j}{y_j}.$$

Hence, using equation (15) we can write:

$$-\frac{(\nabla f(\bar{\alpha}))_i}{y_i} \leq -\frac{(\nabla f(\bar{\alpha}))_{\widehat{i}}}{y_{\widehat{i}}} \leq -\frac{(\nabla f(\bar{\alpha}))_{\widehat{j}}}{y_{\widehat{j}}} \leq -\frac{(\nabla f(\bar{\alpha}))_j}{y_j}$$

and this contradicts equation (13).

b. Thus, we can assume that condition (14) does not hold, so that, we must have for all $k \in K$ and for all $m \geq 0$

$$i^1(k+m) \in R(\alpha^{k+m}) \quad \text{and} \quad j^1(k+m) \in S(\alpha^{k+m})$$

and

$$i^1(k+m) \notin R(\alpha^{k+m+1}) \quad \text{or} \quad j^1(k+m) \notin S(\alpha^{k+m+1}).$$

For simplicity and without loss of generality, we consider only the case that $i^1(k+m) \notin R(\alpha^{k+m+1})$. Then we have

$$\begin{aligned} &i^1(k) \in R(\alpha^k) && \text{and} \quad i^1(k) \notin R(\alpha^{k+1}) \\ &i^1(k+1) \in R(\alpha^{k+1}) && \text{and} \quad i^1(k+1) \notin R(\alpha^{k+2}) \\ &\vdots && \vdots \end{aligned}$$

As $i^1(k)$ belongs to $\{1, \ldots, l\}$, we can extract a subset of K (that we relabel again K) such that for all $k \in K$ we can write

$$i^1(k + h(k)) = i^1(k + n(k)) = \widehat{i}, \quad \text{with} \quad 0 \leq h(k) < n(k) \leq l.$$

Then, we can define a subset K_1 such that for all $k_i \in K_1$,

$$i^1(k_i) = i^1(k_{i+1}) = \widehat{i}, \quad \text{with} \quad k_i < k_{i+1} \leq k_i + l,$$

and $\alpha^{k_i} \to \bar{\alpha}$ for $k_i \to \infty$ and $k_i \in K_1$. Hence, we can write

$$\widehat{i} \in R(\alpha^{k_i}), \quad \text{and} \quad \widehat{i} \notin R(\alpha^{k_i+1}) \quad \text{and} \quad \widehat{i} \in R(\alpha^{k_{i+1}}), \tag{16}$$

which means that index $\widehat{i}$ must have been inserted in the working set and modified by the optimization process between the iterates $k_i + 1$ and $k_{i+1} \leq k_i + l$.

Thus, for all $k_i \in K_1$, an index $p(k_i)$, with $k_i < p(k_i) \leq k_{i+1} \leq k_i + l$, exists such that

$$\widehat{i} \in S(\alpha^{p(k_i)}) \quad \text{and} \quad \widehat{i} \in W^{p(k_i)} \quad \text{and} \quad \widehat{i} \in R(\alpha^{p(k_i)+1}).$$

As $p(k_i) - k_i \leq l$, recalling Lemma 1, we can write

$$\lim_{k_i \to \infty,\, k_i \in K_1} \alpha^{p(k_i)} = \lim_{k_i \to \infty,\, k_i \in K_1} \alpha^{p(k_i)+1} = \bar{\alpha}. \tag{17}$$

We prove now that also the index j, defined in equation (13), must belong to the working set at iteration $p(k_i)$.

To this aim, we first show that

$$-\frac{(\nabla f(\bar{\alpha}))_{\widehat{i}}}{y_{\widehat{i}}} \geq -\frac{(\nabla f(\bar{\alpha}))_i}{y_i}. \tag{18}$$

Indeed if this were not true, namely if

$$-\frac{(\nabla f(\bar{\alpha}))_i}{y_i} > -\frac{(\nabla f(\bar{\alpha}))_{\widehat{i}}}{y_{\widehat{i}}},$$

by the continuity of the gradient we would have for $k_i \in K_1$ and k_i sufficiently large:

$$-\frac{(\nabla f(\alpha^{k_i}))_i}{y_i} > -\frac{(\nabla f(\alpha^{k_i}))_{\widehat{i}}}{y_{\widehat{i}}},$$

which in turn implies that $\widehat{i} \notin I(\alpha^{k_i})$ and hence $i^1(k_i) \neq \widehat{i}$ for $k_i \in K_1$ and sufficiently large. Since equation (18) holds, using equation (13) we get

$$-\frac{(\nabla f(\bar{\alpha}))_{\widehat{i}}}{y_{\widehat{i}}} > -\frac{(\nabla f(\bar{\alpha}))_j}{y_j}.$$

By the continuity of the gradient, we can write for all $k_i \in K_1$ sufficiently large and for all $m \geq 0$:

$$-\frac{(\nabla f(\alpha^{k_i-m}))_{\widehat{i}}}{y_{\widehat{i}}} > -\frac{(\nabla f(\alpha^{k_i-m}))_j}{y_j}. \tag{19}$$

On the other hand, by equation (17) and Proposition 3, as $j \in S(\bar{\alpha})$, for $k_i \in K_1$ and k_i sufficiently large we have that $j \in S(\alpha^{p(k_i)})$ and $j \in S(\alpha^{p(k_i)+1})$. Therefore, since $\widehat{i} \in S(\alpha^{p(k_i)})$ and $\widehat{i} \in W^{p(k_i)}$, from equation (19) and taking into account the WSS rule, we get that also j belongs to the working set at iteration $p(k_i)$, i.e. $j \in W^{p(k_i)}$. Hence, the pair $(\widehat{i}, j)$ is such that

$$(\widehat{i}, j) \in W^{p(k_i)} \quad \text{and} \quad (\widehat{i}, j) \in R(\alpha^{p(k_i)+1}) \times S(\alpha^{p(k_i)+1}),$$

so that, by Lemma 2 we can write

$$\nabla f(\alpha^{p(k_i)+1})' d^{\widehat{i},j} + 2\tau(\alpha^{p(k_i)+1} - \alpha^{p(k_i)})' d^{\widehat{i},j} \geq 0 \quad \text{for all } k_i \in K_1. \tag{20}$$

Then, taking limits in equation (20), recalling the continuity of ∇f and Proposition 5, we obtain

$$\nabla f(\bar{\alpha})' d^{\widehat{i},j} = \frac{(\nabla f(\bar{\alpha}))_{\widehat{i}}}{y_{\widehat{i}}} - \frac{(\nabla f(\bar{\alpha}))_j}{y_j} \geq 0.$$

Finally, using equation (18) we get

$$-\frac{(\nabla f(\bar{\alpha}))_i}{y_i} \leq -\frac{(\nabla f(\bar{\alpha}))_j}{y_j},$$

which contradicts equation (13). ■

5. On the stopping criterion

In PPD algorithm, we still have to define the termination criterion. A natural way is to use the information on the satisfaction of the necessary and sufficient KKT conditions. Indeed, this was proposed in the original paper [3] on SVM^{light}. Actually, a termination criterion which fits better into the SVM^{light} algorithm has been derived and used in refs. [7,12] and analysed in ref. [13]. In order to describe this stopping criterion, we introduce the following functions $m(\alpha)$, $M(\alpha)$: $\mathcal{F} \to R$:

$$m(\alpha) = \begin{cases} \max_{h \in R(\alpha)} -\frac{(\nabla f(\alpha))_h}{y_h} & \text{if } R(\alpha) \neq \emptyset \\ -\infty & \text{otherwise} \end{cases}$$

$$M(\alpha) = \begin{cases} \min_{h \in S(\alpha)} -\frac{(\nabla f(\alpha))_h}{y_h} & \text{if } S(\alpha) \neq \emptyset \\ +\infty & \text{otherwise} \end{cases}$$

where $R(\alpha)$ and $S(\alpha)$ are the index sets defined in equation (4). By definition of $m(\alpha)$ and $M(\alpha)$, and recalling Proposition 2, it follows that $\bar{\alpha}$ is a global minimum of problem (1) if and only if $m(\bar{\alpha}) \leq M(\bar{\alpha})$.

Now let us consider a sequence of feasible points $\{\alpha^k\}$ convergent to a solution $\bar{\alpha}$. At any iteration k, if α^k is not a solution, it follows (again from Proposition 2) that $m(\alpha^k) > M(\alpha^k)$. Hence, the stopping criterion proposed in refs. [7,12] is

$$m(\alpha^k) \leq M(\alpha^k) + \epsilon, \tag{21}$$

where $\epsilon > 0$ is a stopping tolerance.
We note that the quantities $m(\alpha^k)$ and $M(\alpha^k)$ are evaluated in PPD algorithm (and in SVM^{light} algorithm) in order to identify the working set. Hence, the check of equation (21) does not require any additional computational effort. However, as observed in ref. [13], the functions $m(\alpha)$ and $M(\alpha)$ are not continuous. Indeed, even though if $\alpha^k \to \bar{\alpha}$ for $k \to \infty$, it may happen that $R(\alpha^k) \neq R(\bar{\alpha})$ or $S(\alpha^k) \neq S(\bar{\alpha})$ for k sufficiently large, so that we may not have $\lim_{k\to\infty} m(\alpha^k) = m(\bar{\alpha})$ or $\lim_{k\to\infty} M(\alpha^k) = M(\bar{\alpha})$. Therefore, in general, we may have that the limit point $\bar{\alpha}$ is a solution for problem (1), whereas criterion (21) is never satisfied.

In ref. [13], it has been proved that, under assumption (8), SVM^{light} algorithm generates a sequence $\{\alpha^k\}$ such that $m(\alpha^k) - M(\alpha^k) \to 0$ for $k \to \infty$. This implies that, for any tolerance ϵ, SVM^{light} algorithm satisfies the stopping criterion (21) in a finite number of iterations. A similar result can be established for PPD algorithm as reported in the following proposition.

PROPOSITION 7 *Let $\{\alpha^k\}$ be the sequence generated by* PPD *algorithm. If $m(\alpha^k) - M(\alpha^k) > 0$ for all k, then*

$$\lim_{k\to\infty} \left(m(\alpha^k) - M(\alpha^k)\right) = 0. \tag{22}$$

Proof The proof is by contradiction. We assume that a subsequence $\{\alpha^k\}_K$ exists such that

$\lim_{k\to\infty, k\in K} \alpha^k = \bar{\alpha}$
$m(\alpha^k) \geq M(\alpha^k) + \epsilon, \quad$ for all $k \in K$, with $\epsilon > 0$.

Thus, from the definition of m and M we have for all $k \in K$

$$-\frac{(\nabla f(\alpha^k))_{i^1(k)}}{y_{i^1(k)}} \geq -\frac{(\nabla f(\alpha^k))_{j^1(k)}}{y_{j^1(k)}} + \epsilon, \tag{23}$$

where $i^1(k) \in I(\alpha^k)$, $j^1(k) \in J(\alpha^k)$, $I(\alpha)$ and $J(\alpha)$ are the sets defined in equation (7).

We claim that there exists a subset of K, which we relabel again with K, such that for all $k \in K$ and for any $s > 0$, we have

$$-\frac{(\nabla f(\alpha^{k+m}))_{i^1(k+m)}}{y_{i^1(k+m)}} \geq -\frac{(\nabla f(\alpha^{k+m}))_{j^1(k+m)}}{y_{j^1(k+m)}} + \frac{\epsilon}{2} \quad \text{for} \quad m \in [0, s]. \tag{24}$$

From equation (23), since both $i^1(k)$ and $j^1(k)$ belong to a finite set, we can individuate a subset of K, relabelled again with K, and two indices $i \in R(\alpha^k)$ and $j \in S(\alpha^k)$ such that for all $k \in K$

$$-\frac{(\nabla f(\alpha^k))_i}{y_i} \geq -\frac{(\nabla f(\alpha^k))_j}{y_j} + \epsilon. \tag{25}$$

Recalling Lemma 1, the continuity of the gradient, and equation (25), we can write for $k \in K$

$$-\frac{(\nabla f(\alpha^{k-p}))_i}{y_i} \geq -\frac{(\nabla f(\alpha^{k-p}))_j}{y_j} + \frac{\epsilon}{2}, \quad \forall p \geq 0. \tag{26}$$

Suppose first that $i \notin R(\alpha^{k-1})$, then, as $i \in R(\alpha^k)$, we must have that $i \in W^{k-1}$. Actually, $j \in W^{k-1}$. Indeed, if $j \notin S(\alpha^{k-1})$, as $j \in S(\alpha^k)$, it follows that it has been included in the working set at iteration $k-1$; otherwise $j \in S(\alpha^{k-1})$, so that equation (26) and the WSS rule imply that it must have been selected too. Hence, we can apply Lemma 2 and write:

$$\nabla f(\alpha^k)' d^{i,j} + 2\tau(\alpha^k - \alpha^{k-1})' d^{i,j} \geq 0 \quad \text{for all } k \in K.$$

Recalling Proposition 5, the continuity of ∇f, and the definition of $d^{i,j}$ in Lemma 2, we can write for $k \in K$ sufficiently large

$$\nabla f(\alpha^k)' d^{i,j} = \frac{(\nabla f(\alpha^k))_i}{y_i} - \frac{(\nabla f(\alpha^k))_j}{y_j} \geq -2\tau(\alpha^k - \alpha^{k-1})' d^{i,j} \geq -\frac{\epsilon}{2},$$

and this contradicts equation (25), so that we must have $i \in R(\alpha^{k-1})$. Assume now that $j \notin S(\alpha^{k-1})$, then, repeating similar reasonings, we obtain again a contradiction.
Hence, by induction, we can conclude that for $k \in K$ and for any $p \geq 1$

$$i, j \in R(\alpha^{k-p}) \times S(\alpha^{k-p}) \quad \text{and} \quad (i, j) \notin W^{k-p}.$$

By the WSS rule, this implies that we must have

$$-\frac{(\nabla f(\alpha^{k-p}))_{i^1(k-p)}}{y_{i^1(k-p)}} \geq -\frac{(\nabla f(\alpha^{k-p}))_i}{y_i} \quad \text{and} \quad -\frac{(\nabla f(\alpha^{k-p}))_{j^1(k-p)}}{y_{j^1(k-p)}} \leq -\frac{(\nabla f(\alpha^{k-p}))_j}{y_j}.$$

Then, recalling equation (26), it follows that equation (24) holds.
Now we are ready to prove equation (22). The proof is similar to the one of Proposition 6 and is divided in two parts.

a. Suppose first that there exists an integer $s \geq 0$ such that for all $k \in K$:

$$i^1(k + m(k)) \in R(\alpha^{k+m(k)+1}) \quad \text{and} \quad j^1(k + m(k)) \in S(\alpha^{k+m(k)+1})$$
$$\text{for some } m(k) \in [0, s]. \tag{27}$$

Since $i^1(k)$ and $j^1(k)$ belong to a finite set, we can extract a further subset relabelled again K such that

$$i^1(k + m(k)) = \widehat{i}, \quad j^1(k + m(k)) = \widehat{j}, \quad \text{for all } k \in K.$$

Lemma 5 implies that $\alpha^{k+m(k)} \to \bar{\alpha}$ for $k \to \infty, k \in K$. Then recalling that $(\widehat{i}, \widehat{j}) \in W^{k+m(k)}$, we can define a subsequence $\{\alpha^k\}_{K_1}$ such that for all $k \in K_1$

$\widehat{i} \in I(\alpha^k), \quad \widehat{j} \in J(\alpha^k)$
$(\widehat{i}, \widehat{j}) \in W^k$
$(\widehat{i}, \widehat{j}) \in R(\alpha^{k+1}) \times S(\alpha^{k+1})$
$\alpha^k \to \bar{\alpha}$ for $k \to \infty, k \in K_1$.
Now we can apply Lemma 2 and write:

$$\nabla f(\alpha^{k+1})' d^{\widehat{i},\widehat{j}} + 2\tau(\alpha^{k+1} - \alpha^k)' d^{\widehat{i},\widehat{j}} \geq 0 \quad \text{for all } k \in K_1. \tag{28}$$

Recalling Proposition 5 and the continuity of ∇f, taking the limit for $k \in K_1$ we get:

$$\nabla f(\bar{\alpha})' d^{\widehat{i},\widehat{j}} = \frac{(\nabla f(\bar{\alpha}))_{\widehat{i}}}{y_{\widehat{i}}} - \frac{(\nabla f(\bar{\alpha}))_{\widehat{j}}}{y_{\widehat{j}}} \geq 0. \tag{29}$$

On the other hand, from equation (24) we have for $k \in K_1$:

$$-\frac{(\nabla f(\alpha^k))_{\widehat{i}}}{y_{\widehat{i}}} \geq -\frac{(\nabla f(\alpha^k))_{\widehat{j}}}{y_{\widehat{j}}} + \frac{\epsilon}{2}, \tag{30}$$

from which, taking limits, we get

$$-\frac{(\nabla f(\bar{\alpha}))_{\widehat{i}}}{y_{\widehat{i}}} > -\frac{(\nabla f(\bar{\alpha}))_{\widehat{j}}}{y_{\widehat{j}}} \tag{31}$$

and this contradicts equation (29).

b. Thus, we can assume that condition (27) does not hold, so that, we must have for all $k \in K$ and for all $m \geq 0$

$$i^1(k+m) \in R(\alpha^{k+m}) \quad \text{and} \quad j^1(k+m) \in S(\alpha^{k+m})$$

and

$$i^1(k+m) \notin R(\alpha^{k+m+1}) \quad \text{and/or} \quad j^1(k+m) \notin S(\alpha^{k+m+1}).$$

Without loss of generality, we consider only the case that $i^1(k+m) \notin R(\alpha^{k+m+1})$.
Then we have

$$\begin{aligned} &i^1(k) \in R(\alpha^k) && \text{and} \quad i^1(k) \notin R(\alpha^{k+1}) \\ &i^1(k+1) \in R(\alpha^{k+1}) && \text{and} \quad i^1(k+1) \notin R(\alpha^{k+2}) \\ &\vdots && \vdots \end{aligned}$$

As $i^1(k)$ belongs to $\{1, \ldots, l\}$, we can extract a subset of K (that we relabel K) such that for all $k \in K$ we can write

$$i^1(k + h(k)) = i^1(k + n(k)) = \widehat{i}, \quad \text{with} \quad 0 \leq h(k) < n(k) \leq l.$$

Then, we can define a subset K_1 such that for all $k_i \in K_1$,

$$i^1(k_i) = i^1(k_{i+1}) = \widehat{i}, \quad \text{with} \quad k_i < k_{i+1} \leq k_i + l,$$

and $\alpha^{k_i} \to \bar{\alpha}$ for $k_i \to \infty$ and $k_i \in K_1$. Hence, we can write

$$\widehat{i} \in R(\alpha^{k_i}), \quad \text{and} \quad \widehat{i} \notin R(\alpha^{k_i+1}) \quad \text{and} \quad \widehat{i} \in R(\alpha^{k_{i+1}}), \tag{32}$$

which means that index $\widehat{i}$ must have been inserted in the working set and modified by the optimization process between the iterates $k_i + 1$ and $k_{i+1} \leq k_i + l$.

Regarding $j^1(k_i)$, since it belongs to a finite set, we can extract a further subsequence, which we relabel again K_1, such that $j^1(k_i) = \widehat{j}$ for all $k_i \in K_1$. Since for all $k_i \in K_1$, a $k \in K$ exists such that $k_i - k \leq l$, we get from equation (24) that for all $k_i \in K_1$

$$-\frac{(\nabla f(\alpha^{k_i}))_{\widehat{i}}}{y_{\widehat{i}}} \geq -\frac{(\nabla f(\alpha^{k_i}))_{\widehat{j}}}{y_{\widehat{j}}} + \frac{\epsilon}{2}, \tag{33}$$

which is analogous to equation (30), so that taking limits we get equation (31). The continuity of the gradient allows us to state also that for all $m \geq 0$

$$-\frac{(\nabla f(\alpha^{k_i+m}))_{\widehat{i}}}{y_{\widehat{i}}} > -\frac{(\nabla f(\alpha^{k_i+m}))_{\widehat{j}}}{y_{\widehat{j}}}. \tag{34}$$

Now consider the integer $p(k_i)$ such that $k_i < p(k_i) \leq k_{i+1} \leq k_i + l$, and for which

$$\widehat{i} \in S(\alpha^{p(k_i)}), \quad \widehat{i} \in W^{p(k_i)}, \quad \widehat{i} \in R(\alpha^{p(k_i)+1}) \cdots \widehat{i} \in R(\alpha^{k_{i+1}}). \tag{35}$$

The existence of $p(k_i)$ follows from equation (32).

Assume first that

$$\widehat{j} \in S(\alpha^{p(k_i)}), \quad \widehat{j} \in S(\alpha^{p(k_i)+1}). \tag{36}$$

Since $\widehat{i} \in S(\alpha^{p(k_i)})$ and $\widehat{j} \in S(\alpha^{p(k_i)})$, and $\widehat{i} \in W^{p(k_i)}$, then the WWS rule with equation (34) imply that also the index $\widehat{j}$ must be in the working set at iteration $p(k_i)$; moreover, from equation (35) and (36), we have that $(\widehat{i}, \widehat{j}) \in R(\alpha^{p(k_i)+1}) \times S(\alpha^{p(k_i)+1})$.

Suppose that equation (36) does not hold; hence, recalling that $\widehat{j} \in S(\alpha^{k_{i+1}})$, consider the integer $q(k_i)$ such that $p(k_i) \leq q(k_i) < k_{i+1} \leq k_i + l$, and for which

$$\widehat{j} \in R(\alpha^{q(k_i)}), \quad \widehat{j} \in W^{q(k_i)}, \quad \widehat{j} \in S(\alpha^{q(k_i)+1}) \cdots \widehat{j} \in S(\alpha^{k_{i+1}}). \tag{37}$$

If $p(k_i) = q(k_i)$ then, from equations (35) and (37) we have $\widehat{i}, \widehat{j} \in W^{q(k_i)}$ and $(\widehat{i}, \widehat{j}) \in R(\alpha^{q(k_i)+1}) \times S(\alpha^{q(k_i)+1})$.

If $p(k_i) < q(k_i)$, then $\widehat{i} \in R(\alpha^{q(k_i)})$, so that, as $\widehat{j} \in R(\alpha^{q(k_i)})$ and $\widehat{j} \in W^{q(k_i)}$, from the WSS rule and equation (34) we get that $\widehat{i} \in W^{q(k_i)}$; moreover, from equations (35) and (37) we have that $(\widehat{i}, \widehat{j}) \in R(\alpha^{q(k_i)+1}) \times S(\alpha^{q(k_i)+1})$.

Summarizing, we can define a subsequence $\{\alpha^k\}_{K_2} \to \bar{\alpha}$ such that for all $k \in K_2$ the pair $(\widehat{i}, \widehat{j})$ is such that

$$(\widehat{i}, \widehat{j}) \in W^k \quad \text{and} \quad (\widehat{i}, \widehat{j}) \in R(\alpha^{k+1}) \times S(\alpha^{k+1}),$$

so that, using equation (33) and proceeding as in part 'a', we get the contradiction. ■

6. Conclusion and remarks

The main contribution of this paper is the definition of a decomposition method for SVM problem (1), whose convergence can be guaranteed without any further assumption on the Hessian matrix Q.

The core of the convergence analysis stays in the fact that, thanks to the presence of the proximal point modification, we can assure that the distance between successive iterates goes to zero. We note that in the case of dimension of the working set fixed to $q = 2$, which corresponds to SMO algorithm, this property holds without the need of the proximal point term

modification, as shown in ref. [9]. However, we stress that the property stated in Proposition 5 does not depend on the fact that the objective function is quadratic and convex, so it remains true in the case of generic continuous function $f(\alpha)$. By slight changes of the proof, also the compactness of the feasible set can be relaxed, thus allowing that some bounds take the value $\pm\infty$. Of course, without the compactness hypothesis on $\mathcal{F}$, some other assumption is needed to ensure the existence of limit points. Thus, the decomposition approach proposed here can be applied also to problems of the type

$$\begin{aligned} \min \quad & f(\alpha) \\ \text{s.t.} \quad & b'\alpha = c \\ & l \leq \alpha \leq u, \end{aligned}$$

where $f(\alpha)$ is a (possibly nonconvex) smooth function, $b, l, u \in R^l$, $c \in R$ and $-\infty \leq l < u \leq \infty$. Obviously, in the nonconvex case, it is possible to guarantee convergence only to stationary points, i.e. points satisfying the first-order necessary KKT conditions.

The algorithm model proposed here requires at each iteration the computation of the exact solution of the quadratic programming subproblem (10). In the case of $q = 2$, the analytical solution of the subproblem is known, so that the introduction of the proximal point modification is neither theoretically nor practically motivated. When $q > 2$, the solution of the subproblem is not available in closed form, and hence, an iterative method must be used. We expect that, in general, the presence of the proximal point term may improve the rate of convergence of the iterative method, since it may make the Hessian matrix of the subproblem better conditioned. Future work will be devoted to the definition of convergent decomposition methods based on inexact minimization of the subproblems. This will require the study of efficient minimization techniques for quadratic programming and the definition of suitable truncating criteria for ensuring convergence properties.

Acknowledgements

We wish to thank Chih-Jen Lin for his suggestions that lead to improve the paper. This paper was partially supported by CNR-Agenzia2000, National Research Program 'Optimization methods for Support Vector Machines training'.

References

[1] Cortes, C. and Vapnik, V., 1995, Support-vector network. *Machine Learning*, **20**, 273–297.
[2] Vapnik, V., 1995, *The Nature of Statistical Learning Theory* (New York: Springer-Verlag).
[3] Joachims, T., 1998, Making large scale SVM learning practical. In: C.B.B Schölkopf and A. Smola (Eds) *Advances in Kernel Methods – Support Vector Learning* (MA: MIT Press).
[4] Lucidi, S., Palagi, L. and Sciandrone, M., 2002, Convergent decomposition techniques for linearly constrained optimization. Tech. Rep. 16-02, Department of Computer and System Sciences, University of Rome 'La Sapienza', Rome, Italy.
[5] Platt, J.C. 1998, Fast training of support vector machines using sequential minimal optimization. In: C.B.B. Schölkopf and A. Smola (Eds) *Advances in Kernel Methods – Support Vector Learning* (Cambridge, MA: MIT Press).
[6] Bertsekas, D. and Tsitsiklis, J., 1989, *Parallel and Distributed Computation* (Englewood Cliffs, NJ: Prentice-Hall International Editions).
[7] Keerthi, S. and Gilbert, E., 2002, Convergence of a generalized SMO algorithm for SVM. *Machine Learning*, **46**, 351–360.
[8] Lin, C.-J., 2001, On the convergence of the decomposition method for Support Vector Machines. *IEEE Transactions on Neural Networks*, **12**, 1288–1298.
[9] Lin, C.-J., 2002, Asymptotic convergence of an SMO algorithm without any assumptions. *IEEE Transactions on Neural Networks*, **13**, 248–250.
[10] Grippo, L. and Sciandrone, M., 1999, Globally convergent block-coordinate techniques for unconstrained optimization. *Optimization Methods and Software*, **10**, 587–637.

[11] Grippo, L. and Sciandrone, M., 2000, On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. *Operations Research Letters*, **26**, 127–136.
[12] Chang, C.-C. and Lin, C.-J., 2001, *LIBSVM: A Library for Support Vector Machines* (Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm).
[13] Lin, C.-J., 2002, A formal analysis of stopping criteria of decomposition methods for support vector machines. *IEEE Transactions on Neural Networks*, **13**, 1045–1052.
[14] Chang, C.-C., Hsu, C.-W. and Lin, C.-J., 2000, The analysis of decomposition methods for support vector machines. *IEEE Transactions on Neural Networks*, **11**, 1003–1008.
[15] Auslender, A., 1992, Asymptotic properties of the Fenchel dual functional and applications to decomposition problems. *Journal of Optimisation Theory and Applications*, **73**, 427–449.
[16] Bertsekas, D. and Tseng, P., 1994, Partial proximal minimization algorithm for convex programming. *SIAM Journal of Optimization*, **4**, 551–572.

Appendix A

Propositions 2, 4 have been proved in ref. [4]. We report here the proofs for sake of completeness.

Proof of Proposition 2 First we assume that the feasible point α^* is a solution of problem (1). If one of the sets $R(\alpha^*)$ and $S(\alpha^*)$ is empty, then the assertion of the proposition is obviously true. If both the sets $R(\alpha^*)$ and $S(\alpha^*)$ are not empty, Proposition 1 implies the existence of a multiplier λ^* such that the pair (α^*, λ^*) satisfies conditions (3) which can be written as follows:

$$\max_{i \in L^+(\alpha^*) \cup U^-(\alpha^*)} \left\{ -\frac{(\nabla f(\alpha^*))_i}{y_i} \right\} \leq \lambda^* \leq \min_{i \in L^-(\alpha^*) \cup U^+(\alpha^*)} \left\{ -\frac{(\nabla f(\alpha^*))_i}{y_i} \right\}$$

$$\lambda^* = -\frac{(\nabla f(\alpha^*))_i}{y_i} \quad \forall\, i \notin L(\alpha^*) \cup U(\alpha^*).$$

Then recalling the definition of the sets $R(\alpha^*)$ and $S(\alpha^*)$, we can write:

$$\max_{h \in R(\alpha^*)} -\frac{(\nabla f(\alpha^*))_h}{y_h} \leq \min_{h \in S(\alpha^*)} -\frac{(\nabla f(\alpha^*))_h}{y_h},$$

which implies that there exists no pair of indices i and j, with $i \in R(\alpha^*)$ and $j \in S(\alpha^*)$, satisfying equation (5).

Now we assume that there exists no pair of indices i and j, with $i \in R(\alpha^*)$ and $j \in S(\alpha^*)$ satisfying equation (5). First, we consider the case that one of the sets $R(\alpha^*)$ and $S(\alpha^*)$ is empty. Suppose, without loss of generality, that $R(\alpha^*) = \emptyset$. Hence $\{i\colon 0 < \alpha_i^* < C\} = \emptyset$ and $S(\alpha^*) = L^-(\alpha^*) \cup U^+(\alpha^*) = \{1, \ldots, l\}$. Therefore conditions (3) are satisfied by choosing any λ^* such that

$$\lambda^* \leq \min_{1 \leq i \leq l} -\frac{(\nabla f(\alpha^*))_i}{y_i}.$$

In case that both the sets $R(\alpha^*)$ and $S(\alpha^*)$ are not empty, we have that

$$\max_{h \in R(\alpha^*)} -\frac{(\nabla f(\alpha^*))_h}{y_h} \leq \min_{h \in S(\alpha^*)} -\frac{(\nabla f(\alpha^*))_h}{y_h}.$$

Therefore, we can define a multiplier λ^* such that

$$\max_{h \in R(\alpha^*)} -\frac{(\nabla f(\alpha^*))_h}{y_h} \leq \lambda^* \leq \min_{h \in S(\alpha^*)} -\frac{(\nabla f(\alpha^*))_h}{y_h}, \tag{A1}$$

so that the first and second sets of inequalities of equation (3) are satisfied. Then the definition of the sets $R(\alpha^*)$ and $S(\alpha^*)$ and the choice of the multiplier λ^* [satisfying equation (A1)]

imply that

$$\max_{\{i:\ 0<\alpha_i<C\}} -\frac{(\nabla f(\alpha^*))_i}{y_i} \leq \lambda^* \leq \min_{\{i:\ 0<\alpha_i<C\}} -\frac{(\nabla f(\alpha^*))_i}{y_i},$$

so that the set of equalities of equation (3) is verified. ■

Proof of Proposition 3 The proof is by contradiction. Assume that an integer $\bar{j}$ exists, such that $\bar{j} \in R(\bar{\alpha})$ and $\bar{j} \notin R(\alpha^k)$ for each $k \geq \bar{k}$. We can assume without loss of generality that $y_{\bar{j}} > 0$ so that, by definition of $R(\bar{\alpha})$, we get $\bar{\alpha}_{\bar{j}} < C$. By assumption $\bar{j} \notin R(\alpha^k)$, which implies that $\alpha^k_{\bar{j}} = C$ for $k \geq \bar{k}$. Since $\alpha^k \to \bar{\alpha}$ for $k \to \infty$, this implies $\bar{\alpha}_{\bar{j}} = C$, which leads to a contradiction. ■

Proof of Proposition 4 We show that the defined direction d is such that

$$y'd = 0 \quad \text{and} \quad d_i \geq 0 \ \ \forall i \in L(\hat{\alpha}) \quad \text{and} \quad d_j \leq 0 \ \ \forall j \in U(\hat{\alpha}).$$

Indeed, the definition of d yields that $y'd = y_i d_i + y_j d_j = 0$. Moreover, we have $i \in R(\hat{\alpha})$, so that, if $i \in L(\alpha)$, then, by equation (4), we must have $i \in L^+(\hat{\alpha})$, and hence $d_i = 1/y_i > 0$. Analogously, since $j \in S(\hat{\alpha})$, if $j \in U(\hat{\alpha})$ then $j \in U^+(\hat{\alpha})$ and hence $d_j = -1/y_j < 0$. The same conclusion can be drawn for the other two cases. ■