On the Convergence of the Decomposition Method for Support Vector Machines

Chih-Jen Lin, *Member, IEEE*

***Abstract*—The decomposition method is currently one of the major methods for solving support vector machines (SVMs). Its convergence properties have not been fully understood. The general asymptotic convergence was first proposed by Chang *et al.* However, their working set selection does not coincide with existing implementation. A later breakthrough by Keerthi and Gilbert proved the convergence finite termination for practical cases while the size of the working set is restricted to two. In this paper, we prove the asymptotic convergence of the algorithm used by the software SVM$^{\text{light}}$ and other later implementation. The size of the working set can be any even number. Extensions to other SVM formulations are also discussed.**

***Index Terms*—Classification, decomposition methods, support vector machines (SVMs).**

I. Introduction

The SUPPORT vector machine (SVM) is a new and promising technique for classification. Surveys of SVM are given, for example, by Vapnik [27], [28] and Schölkopf *et al.* [23]. Given training vectors $x_i \in R^n, i = 1, \ldots, l$, in two classes, and a vector $y \in R^l$ such that $y_i \in \{1, -1\}$, the support vector technique requires the solution of the following optimization problem:

$$\begin{aligned} \min \quad & \tfrac{1}{2}\alpha^T Q\alpha - e^T\alpha \\ & 0 \le \alpha_i \le C, \qquad i = 1, \ldots, l \\ & y^T\alpha = 0 \end{aligned} \tag{1}$$

where

- e vector of all ones;
- C upper bound of all variables;
- Q l by l positive semidefinite matrix.

Training vectors x_i are mapped into a higher (maybe infinite) dimensional space by the function ϕ and $Q_{ij} \equiv y_i y_j K(x_i, x_j)$ where $K(x_i, x_j) \equiv \phi(x_i)^T\phi(x_j)$ is the kernel.

The difficulty of solving (1) is the density of Q because Q_{ij} is in general not zero. In this case, Q becomes a fully dense matrix so a prohibitive amount of memory is required to store the matrix. Thus traditional optimization algorithms such as Newton, quasi-Newton, etc., cannot be directly applied. Several authors (for example, Joachims [11], Osuna *et al.* [18], Platt [19], and Saunders *et al.* [22]) have proposed decomposition methods to conquer this difficulty. The basic concept of this method is as follows.

Algorithm I.1—Decomposition Method:

1) Given a number $q \le l$ as the size of the working set. Find α^1 as the initial solution. Set $k = 1$.
2) If α^k is an optimal solution of (1), stop. Otherwise, find a working set $B \subset \{1, \ldots, l\}$ whose size is q. Define $N \equiv \{1, \ldots, l\}\backslash B$ and α_B^k and α_N^k to be subvectors of α^k corresponding to B and N, respectively.
3) Solve the following subproblem with the variable α_B:

$$\begin{aligned} \min \quad & \tfrac{1}{2}\alpha_B^T Q_{BB}\alpha_B - (e_B - Q_{BN}\alpha_N^k)^T\alpha_B \\ & 0 \le (\alpha_B)_i \le C, \qquad i = 1, \ldots, q, \\ & y_B^T\alpha_B = -y_N^T\alpha_N^k \end{aligned} \tag{2}$$

where $\begin{bmatrix} Q_{BB} & Q_{BN} \\ Q_{NB} & Q_{NN} \end{bmatrix}$ is a permutation of the matrix Q.
4) Set α_B^{k+1} to be the optimal solution of (2) and $\alpha_N^{k+1} \equiv \alpha_N^k$. Set $k \leftarrow k + 1$ and go to Step 2).

The basic idea of the decomposition method is that in each iteration, the indexes $\{1, \ldots, l\}$ of the training set are separated to two sets B and N, where B is the working set and $N = \{1, \ldots, l\}\backslash B$. The vector α_N is fixed so the objective value becomes $(1/2)\alpha_B^T Q_{BB}\alpha_B - (e_B - Q_{BN}\alpha_N)^T\alpha_B + (1/2)\alpha_N^T Q_{NN}\alpha_N - e_N^T\alpha_N$. Then a subproblem with the variable α_B, i.e., (2), is solved. Note that B is updated in each iteration. To simplify the notation, we simply use B instead of B^k.

An important issue of the decomposition method is to select the working set B in each iteration [Step 2) of Algorithm I.1]. Among existing methods, Osuna *et al.* [18], and Saunders *et al.* [22] find the working set by choosing elements which violate the Karush–Kuhn–Tucker (KKT) condition. Platt's sequential minimal optimization (SMO) [19] restricts the size of the working set to be two. The advantage is that (2) becomes a small problem so no optimization software is required in practice. His working selection includes some heuristics. A systematic way is proposed by Joachims [11] where he restricts q to be an *even* number. In his software SVM$^{\text{light}}$,[1] the following problem with the variable d is solved:

$$\begin{aligned} \min \quad & \nabla f(\alpha^k)^T d \\ & y^T d = 0, \quad -1 \le d_i \le 1, \qquad i = 1, \ldots, l \qquad \text{(3a)} \\ & d_i \ge 0, \quad \text{if } (\alpha^k)_i = 0, \quad d_i \le 0, \quad \text{if } (\alpha^k)_i = C \qquad \text{(3b)} \\ & |\{d_i | d_i \ne 0\}| \le q \qquad \text{(3c)} \end{aligned}$$

where we represent $f(\alpha) \equiv (1/2)\alpha^T Q\alpha - e^T\alpha$, α^k is the solution at the kth iteration, and $\nabla f(\alpha^k)$ is the gradient of $f(\alpha)$

Manuscript received September 18, 2000. This work was supported in part by the National Science Council of Taiwan under Grant NSC 89-2213-E-002-013.

The author is with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan (e-mail: cjlin@csie.ntu.edu.tw).

Publisher Item Identifier S 1045-9227(01)06456-6.

[1] SVM$^{\text{light}}$ is available at http://www-ais.gmd.de/~thorsten/svm_light.

at α^k. Note that $|\{d_i | d_i \neq 0\}|$ means the number of components of d which are not zero. The constraint (3c) implies that a direction d involving only q variables is obtained. Then components of α^k with nonzero d_i are included in the working set B which is used to construct the subproblem (2). Note that d is only used for identifying B but not as a search direction. In Joachims' original paper, $|\{d_i | d_i \neq 0\}| = q$ instead of (3c) was used. Thus practically the decomposition method always picks q elements in each iteration. It was first pointed out in [3] that in theory q nonzero elements may not be always available so (3c) was proposed.

Joachims [11] used the following procedure to solve (3a)–(3c).

Algorithm I.2—SVM$^{\text{light}}$*'s Working Set Selection:*

1) Sort $y_i \nabla f(\alpha^k)_i$ in the decreasing order.
2) From the top of the sorted list sequentially set $d_i = -y_i$ if $0 < \alpha_i^k < C$ or (3b) is satisfied. If $d_i = -y_i$ violates (3b), set $d_i = 0$ and bypass it. From the bottom of the list sequentially set $d_i = y_i$ if $0 < \alpha_i^k < C$ or (3b) is satisfied. If $d_i = y_i$ violates (3b), set $d_i = 0$ and bypass it. The assignment of $d_i = -y_i$ and y_i is done symmetrically until either
   a) $q/2$ elements of d are assigned to be $-y_i$ from the top and $q/2$ elements of d are assigned to be y_i from the bottom; or
   b) we cannot find $d_i = -y_i$ from the top and $d_i = y_i$ from the bottom at the same time.
3) Elements of d not considered yet are assigned to be zeros.

Algorithm I.2 will be discussed in more detail later. We mention the working set selection here because it is strongly related to the main topic of this paper: the convergence of the decomposition method.

As the decomposition method finds an optimal solution of a subproblem (2), the strict decrease of the objective function holds. However, this does not imply that the sequence $\{\alpha^k\}$ converges to an optimal solution of (1). In fact the convergence issue is not easy and has not been fully understood yet.

The first work on the convergence of the decomposition method is by Chang *et al.* [3]. They proved the convergence of a more generalized algorithm. However, their working set selection is by a different problem:

$$\begin{aligned} \min \quad & \nabla f(\alpha^k)^T d \\ & 0 \le \alpha_i^k + d_i \le C, \qquad i = 1, \ldots, l \\ & y^T d = 0 \\ & |\{d_i | d_i \neq 0\}| \le q. \end{aligned} \tag{4}$$

The main shortcoming is that (4) may not be useful in practice. Unlike Algorithm I.2 for (3a)–(3c), we have not known any comparable method for (4). Note that Algorithm I.2 takes at most $O(l \ln l)$ or $O(lq)$ operations that is acceptable for practical implementation.

Then an important progress is by Keerthi and Gilbert [12], [33] where they proved the finite termination of a decomposition method with $q = 2$. In [13] and [34] the authors showed that the original SMO may not converge so some modifications and improvements were added to the SMO. Then [12] and [33] intended to prove the finite termination of a generalized SMO algorithm. Incidently (3a)–(3c) with $q = 2$ is a special case of the working set selection proposed in [13] and [34] (i.e., modification 2 of SMO in that paper). Thus their proof has covered some existing practical implementation.

Up to now the only available implementation using $q > 2$ with convergence proofs is discussed in [10] and [32]. Instead of using the standard formulation, [10] and [32] solve

$$\begin{aligned} \min \quad & \tfrac{1}{2} \alpha^T (Q + yy^T) \alpha - e^T \alpha \\ & 0 \le \alpha_i \le C, \qquad i = 1, \ldots, l. \end{aligned} \tag{5}$$

This formulation was proposed and studied by, for example, Friess *et al.* [9] and Mangasarian and Musicant [16]. Equation (5) is a bound-constrained problem so the working set selection is by the following problem:

$$\begin{aligned} \min \quad & \nabla \bar{f}(\alpha^k)^T d \\ & 0 \le \alpha_i^k + d_i \le C, \qquad i = 1, \ldots, l \\ & |\{d_i | d_i \neq 0\}| \le q \end{aligned} \tag{6}$$

where $\bar{f}(\alpha) \equiv 1/2 \alpha^T (Q + yy^T) \alpha - e^T \alpha$. The convergence follows from the framework in [3]. An important fact is that because of the simpler constraints, (6) can be solved as efficiently as solving (3a)–(3c). To be more precise, the complexity to solve (6) is similar to Algorithm I.2. However, a direct use of (6) did not perform well so [10] finally used a modified way whose convergence is also not clear.

Furthermore, the use of (5) lacks enough theoretical support on generalization properties. We may worry that by removing the linear constraint and adding $1/2(y^T\alpha)^2$ to the objective function, the generalization property is not as good as solving (1). In addition, as more available software follow the implementation of SVM$^{\text{light}}$ using (3a)–(3c) (e.g., [6] and [21]), the need to prove the convergence with $q > 2$ becomes more emergent. In this paper, we will show that Algorithm I.1 using (3a)–(3c) for the working set selection converges.

Next we discuss some possible obstacles while attempting to prove the convergence. In particular, we think the decomposition method of SVM$^{\text{light}}$ has two major problems.

1) In each iteration, the decomposition method works only on a subset of variables. Popular optimization methods such as Newton or quasi-Newton consider all variables together in each iteration. In fact if q is small, in each iteration only few coordinates of the variable are updated. Hence the algorithm is like the "coordinate search" or "method of alternating variables" in optimization literature. It has been shown by Powell [20] that such methods may not always converge. The work in [3] focused on handling this difficulty and a technique to construct a relationship between (4) and the following problem is utilized:

$$\begin{aligned} \min \quad & \nabla f(\alpha^k)^T d \\ & 0 \le \alpha_i^k + d_i \le C, \qquad i = 1, \ldots, l \\ & y^T d = 0. \end{aligned} \tag{7}$$

2) SVM$^{\text{light}}$ uses (3a)–(3c) for the working set selection. Problem (3a)–(3c) follows from the method of feasible directions by Zoutendijk [30]. The original feasible-direction method of Zoutendijk is to consider (3a)–(3c) without restricting the number of nonzero elements

$$\begin{aligned} \min \quad & \nabla f(\alpha^k)^T d \\ & y^T d = 0, \quad -1 \le d_i \le 1, \qquad i = 1, \ldots, l \\ & d_i \ge 0, \quad \text{if } \alpha_i^k = 0, \quad d_i \le 0, \quad \text{if } \alpha_i^k = C. \end{aligned} \tag{8}$$

The difficulty arises because the convergence of Zoutendijk's method is not generally guaranteed. The main reason is that $\alpha^k + d$ may not be a feasible point of (1) so the map of search directions is not closed. An example showing that Zoutendijk's algorithm may not converge is by Wolfe [29] and more discussions are in [1]. This explains why in [3], (6) instead of (3a)–(3c) is considered because (6) guarantees the feasibility of $\alpha^k + d$. To be more precise, the key difficulty is the problem caused by α_i sitting very close to the boundary and having large violation. Another way to see this is that the objective function of (8) is discontinuous at the boundary. This is a rather peculiar situation not associated with traditional optimization algorithms. That is why methods such as Joachims' decomposition algorithm and SMO by Platt require a very different approach to the proof. The proofs given here and the one in Keerthi and Gilbert use a nontraditional *counting* argument to prove convergences.

In addition, we note that the original Zoutendijk's method directly uses d as the search direction for the optimization algorithm. That is, a step size λ is decided and $\alpha^k + \lambda d$ becomes the next iterate α^{k+1}. This is different from the role of (3a)–(3c) here as d is used only for selecting the working set. Furthermore, in each iteration an exact solution of the subproblem (2) is obtained. This seems to be a nice property which the original Zoutendijk's method lacks of. In [3], such a property was not used as they considered a more general algorithm. For the proof in this paper, we will see that it plays an important role.

The above discussion reveals that the working set selection problem (3a)–(3c) should be deeply investigated. In Section II, we analyze (3a)–(3c) and its solution procedure: Algorithm I.2. Readers who are interested in only the convergence proofs may skip this section. In Section III we sketch the main convergence proof by some figures. Section IV is the main convergence proof. Extensions of the proof to other SVM formulations such as regression and one-class SVM are in Section V. We make conclusions and discussions in Section VI.

II. More Analysis on the Working Set Selection

Though in [11], Joachims has proposed Algorithm I.2 to solve (3a)–(3c), up to now there is no rigorous discussion to justify the use of this algorithm. For example, if the case 2b) of Algorithm I.2 is encountered first, it is not clear what the practical situation looks like. In this section, we will discuss the details of Algorithm I.2 and demonstrate that it really solves (3a)–(3c). For readers who are interested in only the convergence proofs, you can skip this section and directly go to Section III.

First we give a simple assumption.

Assumption II.1—$C > 0$:

If $C = 0$, the only feasible solution of (1) is $\alpha_i = 0$, $i = 1, \ldots, l$. In addition, the constraints of (8) implies $d_i = 0$. If Algorithm I.2 is used without Assumption II.1, for $0 = \alpha_i^k = C$ we may end up with $d_i = y_i(-y_i) \neq 0$ which is not a feasible solution of (8). The assumption looks trivial but in our mind we consider the general situation $l_i \le \alpha_i \le u_i$, where l_i and u_i are lower and upper bounds, respectively. If $l_i = u_i$, then we can remove the ith variable from the original problem easily.

The following theorem shows how Algorithm I.2 solves (8).

Theorem II.2: If the condition 2a) of Algorithm I.2 is not activated (or q is selected large enough), the algorithm will finally stop at i_t (from the top) and i_b (from the bottom) and one of the following will happen.

1) $y_{i_t}\nabla f(\alpha^k)_{i_t}$ is next to $y_{i_b}\nabla f(\alpha^k)_{i_b}$ in the sorted list.
2) There is one element $y_i \nabla f(\alpha^k)_i$ between $y_{i_t}\nabla f(\alpha^k)_{i_t}$ and $y_{i_b}\nabla f(\alpha^k)_{i_b}$ with $0 < \alpha_i^k < C$.

In addition, when the algorithm stops, d is an optimal solution of problem (8).

Proof: When Algorithm I.2 stops at i_t, if the next index in the sorted list of $y_i \nabla f(\alpha^k)_i$, $i = 1, \ldots, l$ is $\bar{i}_t$, there are three possible situations:

$$\begin{aligned} 0 < \alpha_{\bar{i}_t}^k &< C, \quad \text{or} \\ \alpha_{\bar{i}_t}^k &= 0, \quad y_{\bar{i}_t} = -1, \quad \text{or} \\ \alpha_{\bar{i}_t}^k &= C, \quad y_{\bar{i}_t} = 1. \end{aligned} \tag{9}$$

Otherwise, we can move down by assigning $d_{\bar{i}_t} = 0$. Then consider going up from i_b, if the next $\bar{i}_b$ is not $\bar{i}_t$ or i_t, it can not satisfy $0 < \alpha_{\bar{i}_b}^k < C$, or $\alpha_{\bar{i}_b}^k = 0$, $y_{\bar{i}_b} = 1$, or $\alpha_{\bar{i}_b}^k = C$, $y_{\bar{i}_b} = -1$. Otherwise, (9) implies that i_t can move down to $\bar{i}_t$ and then i_b could move up. Hence $\bar{i}_b$ must satisfy $\alpha_{\bar{i}_b}^k = 0$, $y_{\bar{i}_b} = -1$ or $\alpha_{\bar{i}_b}^k = C$, $y_{\bar{i}_b} = 1$. However, for this situation, i_b could move up by setting $d_{\bar{i}_b} = 0$. Hence we are sure that there is at most one element between i_t and i_b. If there is one such element $\bar{i}_t$ and $\alpha_{\bar{i}_t}^k = 0$, $y_i = -1$, or $\alpha_{\bar{i}_t}^k = C$, $y_i = 1$, i_b could move up again by assigning $d_{\bar{i}_t} = 0$. Therefore, from (9) the only possible situation is to have an element i between i_t and i_b with $0 < \alpha_i^k < C$.

Thus we have clarified the situation when the algorithm terminates. Next we will show that when the algorithm stops, the following KKT condition is satisfied so d is an optimal solution:

$$\begin{aligned} \nabla f(\alpha^k) &= -by + \lambda_i - \xi_i, \\ y^T d &= 0 \\ \lambda_i(d_i + 1) &= 0, \qquad \text{if } 0 < \alpha_i^k \le C \\ \lambda_i d_i &= 0, \qquad \text{if } \alpha_i^k = 0 \\ \xi_i(1 - d_i) &= 0, \qquad \text{if } 0 \le \alpha_i^k < C \\ \xi_i d_i &= 0, \qquad \text{if } \alpha_i^k = C \\ \lambda_i \ge 0, \quad \xi_i &\ge 0, \qquad i = 1, \ldots, l. \end{aligned} \tag{10}$$

If there is an $\bar{i}_t$ between i_t and i_b, we select b such that

$$y_{\bar{i}_t} \nabla f(\alpha^k)_{\bar{i}_t} + b = 0.$$

Otherwise, we pick b such that

$$y_i \nabla f(\alpha^k)_i + b \geq 0 \text{ for the elements before (include) } i_t \quad (11)$$
$$y_i \nabla f(\alpha^k)_i + b \leq 0 \text{ for the elements after (include) } i_b. \quad (12)$$

Let us consider the case in (11). If $d_i = -y_i$ and $y_i = 1$, by selecting $\xi_i \equiv 0$ and $\lambda_i \equiv \nabla f(\alpha^k)_i + by_i \geq 0$, (10) is satisfied. The situation is similar for $y_i = -1$. If $d_i = 0$, there are two possibilities: $y_i = 1, \alpha_i^k = 0$ or $y_i = -1, \alpha_i^k = C$. For the first case, $\lambda_i \equiv \nabla f(\alpha^k)_i + by_i \geq 0, \lambda_i d_i = 0$ and $\xi \equiv 0$ satisfy (10). The argument for the second case is similar. Furthermore, the same proof can be applied for indexes which satisfy $d_i = y_i$. Thus we have shown that Algorithm I.2 obtains a KKT point. Since (8) is a linear program, a KKT point is an optimal solution. ■

After the procedure of Algorithm I.2 without activating condition 2a) for solving (8), we assume that $i_1, \ldots, i_{m_k/2}$ are indexes of elements with $d_i = -y_i$ (in the decreasing order of $\{y_i \nabla f(\alpha^k)_i\}$) and $j_1, \ldots, j_{m_k/2}$ are indexes of elements with $d_i = y_i$ [in the increasing order of $\{y_i \nabla f(\alpha^k)_i\}$]. Then

$$y_i \nabla f(\alpha^k)_i + b = p_i \geq 0, \qquad i = i_1, \ldots, i_{m_k/2}$$
$$y_i \nabla f(\alpha^k)_i + b = n_i \leq 0, \qquad i = j_1, \ldots, j_{m_k/2}.$$

We have

$$p_{i_1} \geq p_{i_2} \cdots \geq p_{i_{m_k/2}} \geq 0 \geq n_{j_{m_k/2}} \geq \cdots \geq n_{j_1}. \quad (13)$$

Since $d_i = -y_i, i = i_1, \ldots, i_{m_k/2}$ and $d_i = y_i, i = j_1, \ldots, j_{m_k/2}$

$$\nabla f(\alpha^k)_i d_i = (b - p_i), \qquad i = i_1, \ldots, i_{m_k/2}$$
$$\nabla f(\alpha^k)_i d_i = (-b + n_i), \qquad i = j_1, \ldots, j_{m_k/2}.$$

Therefore, the optimal objective value of (8) is

$$\sum_{i=i_1, \ldots, i_{m_k/2}} -p_i + \sum_{i=j_1, \ldots, j_{m_k/2}} n_i.$$

Now we are ready to work on problem (3a)–(3c). We will show that by selecting

$$d_i \equiv \begin{cases} -y_i, & i = i_1, \ldots, i_{\min(q, m_k)/2} \\ y_i, & i = j_1 \ldots, j_{\min(q, m_k)/2} \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

an optimal solution of (3a)–(3c) is obtained. When $q \geq m_k$, the solution we just obtained for (8) is a feasible solution of (3a)–(3c). As (3a)–(3c) has a smaller feasible region than (8), its objective value is not smaller. Thus d defined by (14) is an optimal solution of (3a)–(3c). On the other hand, if $q < m_k$, we consider the following problem:

$$\begin{aligned} \min \quad & \nabla f(\alpha^k)^T d \\ & y^T d = 0, \quad -1 \leq d_i \leq 1, \qquad i = 1, \ldots, l \\ & d_i \geq 0, \quad \text{if } \alpha_i^k = 0, \quad d_i \leq 0, \quad \text{if } \alpha_i^k = C \\ & d_i = 0, \quad \text{if } i \notin \overline{B} \end{aligned} \quad (15)$$

where $\overline{B}$ is any subset of $\{1, \ldots, l\}$ containing $\overline{q}$ elements with $\overline{q} \leq q$. Now $\overline{B}$ is fixed so (15) is reduced to a form of (8) whose number of variables is $\overline{q}$. Hence the same procedure of Algorithm I.2 without Step 2a) could be applied to solve (15). If an optimal solution is $d_i = -y_i, i = r_1, \ldots, r_{\hat{q}/2}, d_i = y_i, i = s_1, \ldots, s_{\hat{q}/2}$ $(\hat{q} \leq \overline{q})$, and $d_i = 0$ otherwise, then the optimal objective value of (15) is

$$\left(-\sum_{i=r_1, \ldots, r_{\hat{q}/2}} y_i \nabla f(\alpha^k)_i + \sum_{i=s_1, \ldots, s_{\hat{q}/2}} y_i \nabla f(\alpha^k)_i \right)$$

which is greater or equal to

$$\left(-\sum_{i=i_1, \ldots, i_{\hat{q}/2}} y_i \nabla f(\alpha^k)_i + \sum_{i=j_1, \ldots, j_{\hat{q}/2}} y_i \nabla f(\alpha^k)_i \right)$$

as we sort $y_i \nabla f(\alpha^k)_i$ in a decreasing order. Since $\hat{q} \leq \overline{q} \leq q$, d defined in (14) is an optimal solution of (3a)–(3c). The following theorem concludes the validity of Algorithm I.2 for (3a)–(3c).

Theorem II.3: If q is an even positive integer, Algorithm I.2 returns an optimal solution of (3a)–(3c) and

$$\begin{aligned} & \frac{l}{q} (\text{optimal objective value of (3a)–(3c)}) \\ & \quad \leq \text{optimal objective value of (8).} \end{aligned} \quad (16)$$

We then show the relation between the working set selection problem (3a)–(3c) and the original optimization problem (1).

Theorem II.4: The optimal objective value of (3a)–(3c) is zero if and only if α is an optimal solution of (1).

Proof: A basic property of Zoutendijk's method is that the optimal objective value of (8) is zero if and only if α is an optimal solution of (1) (see, for example, [1]). Since (3a)–(3c) has a smaller feasible region than (8), if the optimal objective value of (3a)–(3c) is zero, the optimal solution of (8) is also zero. Therefore, α is an optimal solution of (1).

On the other hand, if α is an optimum of (1), with Lemma II.3, the optimal objective value of (3a)–(3c) is zero. ■

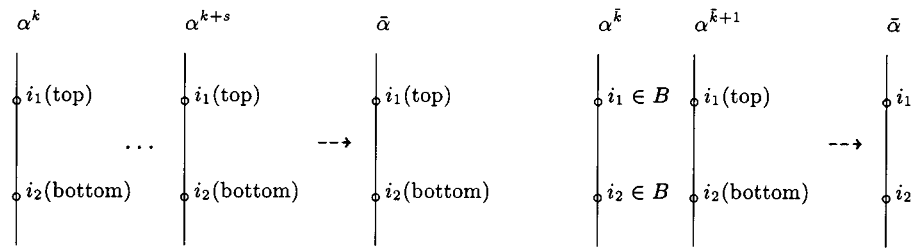
There are different methods for the analysis in this section. For example, in [5] and [31], the authors modified (3a)–(3c) to ν-SVM problems [25] where they used a recursive approach to show the validity of Algorithm I.2.

III. Outline of the Convergence Proof

The convergence of Algorithm I.1 using problem (3a)–(3c) is the main result of this paper. As the proof involves with several complicated lemmas and theorems, in this section we give an outline of the proof. Using some informal terms and figures, we explain some key ideas behind the proof.

First we discuss some observations which help to prove the convergence. If $\hat{\alpha}$ is an optimal solution of (1), it satisfies the following KKT condition: there is a number b such that

$$\begin{aligned} \nabla f(\hat{\alpha})_i + by_i \geq 0 \qquad & \text{if } \hat{\alpha}_i = 0, \\ \nabla f(\hat{\alpha})_i + by_i \leq 0 \qquad & \text{if } \hat{\alpha}_i = C \\ \nabla f(\hat{\alpha})_i + by_i = 0 \qquad & \text{if } 0 < \hat{\alpha}_i < C. \end{aligned} \quad (17)$$



(a) (b)

Fig. 1. $y_i \nabla f(\alpha)_i$ in the order of the sorted list of $y_i \nabla f(\overline{\alpha})_i$, $i = 1, \ldots, l$.

For any scalar α_i, we can consider two situations

$$0 < \alpha_i < C \text{ or } (\alpha_i = C \text{ and } y_i = 1) \text{ or } (\alpha_i = 0 \text{ and } y_i = -1) \quad (18)$$

$$0 < \alpha_i < C \text{ or } (\alpha_i = C \text{ and } y_i = -1) \text{ or } (\alpha_i = 0 \text{ and } y_i = 1). \quad (19)$$

Then the KKT condition (17) can be rewritten as

$$\begin{aligned} y_i \nabla f(\hat{\alpha})_i + b \geq 0 \qquad & \text{if } \hat{\alpha}_i \text{ satisfies (19)} \\ y_i \nabla f(\hat{\alpha})_i + b \leq 0 \qquad & \text{if } \hat{\alpha}_i \text{ satisfies (18).} \end{aligned} \quad (20)$$

Note that (18) [(19)] is the condition in Algorithm I.2 where α_i^k can be a candidate for selection from the top (bottom) of the sorted list of $y_i \nabla f(\alpha^k)_i$, $i = 1, \ldots, l$. In the following, we shall refer a variable α_i as a

- "top" candidate: if it satisfies (18);
- "top only" candidate: if it satisfies ($\alpha_i = C$ and $y_i = 1$) or ($\alpha_i = 0$ and $y_i = -1$);
- "bottom" candidate: if it satisfies (19);
- "bottom only" candidate: if it satisfies ($\alpha_i = C$ and $y_i = -1$) or ($\alpha_i = 0$ and $y_i = 1$).

From Assumption II.1, $C > 0$ so the following two statements are equivalent:

$$\alpha_i \text{ is a "top only" candidate} \equiv \alpha_i \text{ is not a "bottom" candidate.}$$

Therefore, once α_i^k is a "top only" candidate, next time when it is selected, in Algorithm I.2, it must be picked from the top of the sorted list.

It can be clearly seen that the KKT condition (20) implies that all "top" candidates have the same or smaller $y_i \nabla f(\hat{\alpha})_i$ than "bottom" candidates. Therefore, when applying Algorithm I.2 to problem (3a)–(3c) of an optimal solution, except those elements with $y_i \nabla f(\hat{\alpha}) + b = 0$, we cannot do any selection.

For free variables, their $y_i \nabla f(\hat{\alpha})_i$ are equal. On the other hand, if $\hat{\alpha}_i$ are at bounds, their $y_i \nabla f(\hat{\alpha})_i$ are usually different. This leads us to suspect that in final iterations, for all bounded α_is, their associated $y_i \nabla f(\alpha)_i$ are already in correct places of the sorted list of $y_i \nabla f(\hat{\alpha})_i$, $i = 1, \ldots, l$.

Of course it is possible that $\nabla f(\hat{\alpha})_i + by_i = 0$ even if $\hat{\alpha}_i$ is at a bound. This is the so called "degenerate" case in optimization terminology. For degenerate or free variables, $y_i \nabla f(\hat{\alpha})_i$ are all equal. We will focus on analyzing this group of variables. Indeed we will show that for any limit point of a convergent subsequence, only indexes from this particular group are still under consideration.

Next we outline the proof. We consider any convergent subsequence $\{\alpha^k\}$, $k \in \mathcal{K}$ and $\overline{\alpha} \equiv \lim_{k \to \infty, k \in \mathcal{K}} \alpha^k$. We prove that for any given positive integer s, the sequence $\{\alpha^{k+s}\}$, $k \in \mathcal{K}$ converges to $\overline{\alpha}$. Therefore, if $\overline{\alpha}_i$ is a "top" ("bottom") candidate, then after $k \in \mathcal{K}$ is large enough, $\alpha_i^k, \alpha_i^{k+1}, \ldots, \alpha_i^{k+s}$ are all "top" ("bottom") candidates. These results will be proved in Lemmas IV.3 and IV.4a). An illustration is in Fig. 1(a), where the vertical line on the right represents the sorted list $y_i \nabla f(\overline{\alpha})_i$, $i = 1, \ldots, l$ (in the decreasing order) and lines on the left are corresponding values of $y_i \nabla f(\alpha^k)_i$ to $y_i \nabla f(\alpha^{k+s})_i$.

We then prove that a situation like Fig. 1(b) cannot happen. That is, if

$$y_{i_1} \nabla f(\overline{\alpha})_{i_1} > y_{i_2} \nabla f(\overline{\alpha})_{i_2}$$

then after $k \in \mathcal{K}$ is large enough, for any $\overline{k} \in \{k, k+1, \ldots, k+s-1\}$, if i_1 and i_2 are both in the working set of the $\overline{k}$th iteration, it is impossible to have $\alpha_{i_1}^{\overline{k}+1}$ and $\alpha_{i_2}^{\overline{k}+1}$ are "top" and "bottom" candidates at the same time.

The final part of the proof comes from Fig. 2 (Theorem IV.5). Assume $i_1(i_2)$ is the first "top" ("bottom") candidate of the sorted list of $y_i \nabla f(\overline{\alpha})_i$, $i = 1, \ldots, l$ and $I_1(I_2)$ is the set not lower (higher) than $i_1(i_2)$. If

$$y_{i_1} \nabla f(\overline{\alpha})_{i_1} > y_{i_2} \nabla f(\overline{\alpha})_{i_2} \quad (21)$$

we prove that after $k \in \mathcal{K}$ is large enough, there is a $\overline{k} \in \{k, k+1, \ldots, k+2l\}$ such that $\alpha_{I_1}^{\overline{k}}$ ($\alpha_{I_2}^{\overline{k}}$) has only "bottom only" ("top only") elements. Therefore, at the $\overline{k}$th iteration, both i_1 and i_2 must be selected as they are the first "top" and "bottom" elements, respectively [from Fig. 1(a)]. However, at the $(\overline{k}+1)$st iteration, from Fig. 1(a), we know that i_1 and i_2 are "top" and "bottom" elements again. This violates the results in Fig. 1(b). Therefore, the assumption (21) is wrong so at $\overline{\alpha}$, elements selected by (3a)–(3c) have the same $y_i \nabla f(\overline{\alpha})_i$. With this result we can show that the working set selection problem at $\overline{\alpha}$ has zero optimal objective value. Then from Theorem II.4, $\overline{\alpha}$ is an optimal solution of (1).

Thus the main effort of Theorem IV.5 is on a counting process to show that in at most $2l$ iterations, all elements in I_1 and I_2 become "bottom only" and "top only," respectively.

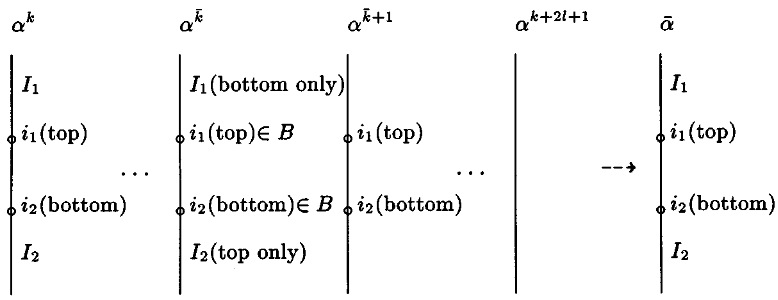


Fig. 2. A counting process on I_1 and I_2 (Theorem IV.5).

IV. Convergence Proofs

In this section we prove the convergence of Algorithm I.1 using problem (3a)–(3c) for the working set selection (i.e., the algorithm used by SVM$^{\text{light}}$). If Algorithm I.1 stops in finite number of iterations, from Step 2), α^k is already an optimum. Hence here we consider the case where Algorithm I.1 takes infinite iterations. First we make an assumption.

Assumption IV.1: The matrix Q satisfies

$$\min_I(\min(eig(Q_{II}))) > 0$$

where

- I any subset of $\{1, \ldots, l\}$ with $|I| \le q$;
- Q_{II} square submatrix of Q;
- $\min(eig(\cdot))$ smallest eigenvalue of a matrix.

If Q is positive definite, then Assumption IV.1 is true. For example, if the RBF kernel $K(x_i, x_j) = e^{\|x_i - x_j\|^2}$ is used and all $x_i \ne x_j$, from [17], Q is positive definite. Since practically q is selected as a small number ($\le$100), if data are mapped into higher dimensional spaces, Q tends to be positive definite so in general Assumption IV.1 holds.

The following lemma shows the sufficient decrease of $f(\alpha)$.

Lemma IV.2:

$$f(\alpha^{k+1}) \le f(\alpha^k) - \frac{\sigma}{2}\|\alpha^{k+1} - \alpha^k\|^2 \tag{22}$$

where $\sigma = \min_I(\min(eig(Q_{II})))$.

Proof: Assume B is the working set at the kth iteration and $N \equiv \{1, \ldots, l\}\backslash B$. If we define $s \equiv \alpha^{k+1} - \alpha^k$, then $s_N = 0$ and

$$\begin{aligned} f(\alpha^{k+1}) &- f(\alpha^k) \\ &= \tfrac{1}{2}s^TQs + s^TQ\alpha^k - e^Ts \\ &= \tfrac{1}{2}s_B^TQ_{BB}s_B + s_B^T(Q\alpha^k)_B - e_B^Ts_B. \end{aligned} \tag{23}$$

That is, in the kth iteration, we solve the following problem with the variable s_B:

$$\begin{aligned} \min \quad & \tfrac{1}{2}\, s_B^TQ_{BB}s_B + s_B^T(Q\alpha^k)_B - e_B^Ts_B \\ & 0 \le (\alpha^k + s)_i \le C, \qquad i \in B \\ & y_B^Ts_B = 0, \end{aligned} \tag{24}$$

which is a different representation of (2). The KKT condition of (24) shows that there is a b^{k+1} such that

$$(Q(\alpha^k + s))_i - 1 + b^{k+1}y_i = 0 \quad \text{if } 0 < \alpha_i^k + s_i < C,\ i \in B \tag{25}$$

$$(Q(\alpha^k + s))_i - 1 + b^{k+1}y_i \ge 0 \quad \text{if } \alpha_i^k + s_i = 0,\ i \in B \tag{26}$$

$$(Q(\alpha^k + s))_i - 1 + b^{k+1}y_i \le 0 \quad \text{if } \alpha_i^k + s_i = C,\ i \in B. \tag{27}$$

Define $F \equiv \{i | 0 < \alpha_i^k + s_i < C,\ i \in B\}$ and $A \equiv \{i | \alpha_i^k + s_i = 0 \text{ or } C,\ i \in B\}$. We have $B = F \cup A$ and from (27)

$$\begin{aligned} (Q\alpha^k)_F &= -(Qs)_F + e_F - b^{k+1}y_F \\ &= -Q_{FF}s_F - Q_{FA}s_A + e_F - b^{k+1}y_F. \end{aligned} \tag{28}$$

With (28), the last two terms of (23) become

$$\begin{aligned} s_B^T(Q\alpha^k)_B &- e_B^Ts_B \\ &= s_F^T(Q\alpha^k)_F - e_F^Ts_F + s_A^T((Q(\alpha^k + s))_A + b^{k+1}y_A - e_A) \\ &\quad - s_A^T(Qs)_A - b^{k+1}y_A^Ts_A \\ &= s_F^T(Q\alpha^k)_F - e_F^Ts_F + s_A^T((Q(\alpha^k + s))_A + b^{k+1}y_A - e_A) \\ &\quad - s_A^T(Q_{AF}s_F + Q_{AA}s_A) - b^{k+1}y_A^Ts_A \\ &= -s_F^TQ_{FF}s_F - s_F^TQ_{FA}s_A - b^{k+1}y_B^Ts_B \\ &\quad + s_A^T((Q(\alpha^k + s))_A + b^{k+1}y_A - e_A) \\ &\quad - s_A^T(Q_{AF}s_F + Q_{AA}s_A). \end{aligned} \tag{29}$$

If $\alpha_i^k + s_i = 0$, then $s_i \le 0$ and if $\alpha_i^k + s_i = C$, then $s_i \ge 0$. Hence from (26) and (27)

$$s_A^T((Q(\alpha^k + s))_A + b^{k+1}y_A - e_A) \le 0. \tag{30}$$

With (29), (30)

$$\tfrac{1}{2}s_B^TQ_{BB}s_B = \tfrac{1}{2}s_F^TQ_{FF}s_F + s_F^TQ_{FA}s_A + \tfrac{1}{2}s_A^TQ_{AA}s_A$$

and $y_B^Ts_B = 0$, (23) becomes

$$\begin{aligned} -\tfrac{1}{2}s_F^TQ_{FF}s_F &- \tfrac{1}{2}s_A^TQ_{AA}s_A - s_FQ_{FA}s_A \\ &+ s_A^T((Q(\alpha^k + s))_A + b^{k+1}y_A - e_A) \\ \le& -\tfrac{1}{2}[s_F^T \quad s_A^T]\begin{bmatrix} Q_{FF} & Q_{FA} \\ Q_{AF} & Q_{AA} \end{bmatrix}\begin{bmatrix} s_F \\ s_A \end{bmatrix} \\ \le& -\frac{\sigma}{2}\|s_B\|^2 = -\frac{\sigma}{2}\|s\|^2. \end{aligned} \tag{31}$$

■

From now on we consider any convergent subsequence $\{\alpha^k\}$, $k \in \mathcal{K}$ and $\lim_{k\to\infty,\, k\in\mathcal{K}} \alpha^k = \overline{\alpha}$. We then have the following lemma.

Lemma IV.3: For any given positive integer s, the sequence $\{\alpha^{k+s}\}$, $k \in \mathcal{K}$ converges to $\overline{\alpha}$. In addition, $\{y_i \nabla f(\alpha^{k+s})_i\}$ converges to $y_i \nabla f(\overline{\alpha})_i$, for $i = 1, \ldots, l$.

Proof: First we know that $\{f(\alpha^k)\}$ is a decreasing sequence. Since $0 \le \alpha_i \le C$, $i = 1, \ldots, l$, the feasible region of (1) is a compact set. Thus we know that $\{f(\alpha^k)\}$ converges to a finite number.

Then for the subsequence $\{\alpha^{k+1}\}$, $k \in \mathcal{K}$, from Lemma IV.2 we have

$$\begin{aligned}
&\lim_{k\to\infty} \|\alpha^{k+1} - \overline{\alpha}\| \\
&\quad \le \lim_{k\to\infty} (\|\alpha^{k+1} - \alpha^k\| + \|\alpha^k - \overline{\alpha}\|) \\
&\quad \le \lim_{k\to\infty} \left(\sqrt{\frac{2}{\sigma}(f(\alpha^k) - f(\alpha^{k+1}))} + \|\alpha^k - \overline{\alpha}\| \right) \\
&\quad = 0.
\end{aligned}$$

Thus

$$\lim_{k\to\infty,\, k\in\mathcal{K}} \alpha^{k+1} = \overline{\alpha}.$$

From $\{\alpha^{k+1}\}$ we can prove $\lim_{k\to\infty,\, k\in\mathcal{K}} \alpha^{k+2} = \overline{\alpha}$ too. Therefore, $\lim_{k\to\infty,\, k\in\mathcal{K}} \alpha^{k+s} = \overline{\alpha}$ for any given s.

The results on $\{y_i \nabla f(\alpha^{k+s})_i\}$ follows from the continuity of $\nabla f(\alpha)$. ■

We then need a technical lemma.

Lemma IV.4: Let $\overline{\alpha}$ be as in Lemma IV.3.

a) If $\overline{\alpha}_i$ satisfies (18) [(19)], then for any given positive integer s, after $k \in \mathcal{K}$ is large enough, $\alpha_i^k, \alpha_i^{k+1}, \ldots, \alpha_i^{k+s}$ all satisfy (18) [(19)]. In other words, if $\overline{\alpha}_i$ is a "top" ("bottom") candidate, then after $k \in \mathcal{K}$ is large enough, $\alpha_i^k, \alpha_i^{k+1}, \ldots, \alpha_i^{k+s}$ are all "top" ("bottom") candidates.

b) In addition, if

$$y_{i_1} \nabla f(\overline{\alpha})_{i_1} > y_{i_2} \nabla f(\overline{\alpha})_{i_2} \tag{32}$$

then after $k \in \mathcal{K}$ is large enough, for any $\overline{k} \in \{k, k+1, \ldots, k+s-1\}$, if i_1 and i_2 are both in the working set of the $\overline{k}$th iteration, it is impossible to have $\alpha_{i_1}^{\overline{k}+1}$ and $\alpha_{i_2}^{\overline{k}+1}$ satisfy (18) and (19), respectively. In other words, $\alpha_{i_1}^{\overline{k}+1}$ and $\alpha_{i_2}^{\overline{k}+1}$ cannot be top and bottom candidates at the same time.

Proof: The first result immediately follows from Assumption II.1, Lemma IV.3, and the definition of (18) and (19).

For the second result of this lemma, we assume that it is possible that both $\alpha_{i_1}^{\overline{k}+1}$ and $\alpha_{i_2}^{\overline{k}+1}$ satisfy (18) and (19), respectively. Since $\alpha_B^{\overline{k}+1}$ is an optimal solution of (2), from the KKT condition of the subproblem (2) and a similar form of (20), if $\alpha_{i_1}^{\overline{k}+1}$ satisfies (18), there is a $b^{\overline{k}+1}$ such that

$$y_{i_1} \nabla f(\alpha^{\overline{k}+1})_{i_1} + b^{\overline{k}+1} \le 0. \tag{33}$$

On the other hand, if $\alpha_{i_2}^{\overline{k}+1}$ satisfies (19), then

$$y_{i_2} \nabla f(\alpha^{\overline{k}+1})_{i_2} + b^{\overline{k}+1} \ge 0. \tag{34}$$

Thus (33) and (34) imply

$$y_{i_1} \nabla f(\alpha^{\overline{k}+1})_{i_1} \le y_{i_2} \nabla f(\alpha^{\overline{k}+1})_{i_2}$$

which contradicts to (32) when $\overline{k}$ is large enough. ■

Finally, the main theorem is as follows.

Theorem IV.5: Any limit point of $\{\alpha^k\}$ is a global minimum of (1).

Proof: Assume $\overline{\alpha}$ is the limit point of any convergent subsequence $\{\alpha^k\}$, $k \in \mathcal{K}$. If $\overline{\alpha}$ is not an optimal solution of (1), from Theorem II.4, the following problem has a nonzero solution:

$$\begin{aligned}
\min \quad & \nabla f(\overline{\alpha})^T d \\
& -1 \le d_i \le 1, \quad y^T d = 0 \\
& d_i \ge 0, \quad \text{if } \overline{\alpha}_i = 0, \qquad d_i \le 0, \quad \text{if } \overline{\alpha}_i = C \\
& |\{d_i | d_i \ne 0\}| \le q.
\end{aligned} \tag{35}$$

If we can prove that only elements with the same $y_i \nabla f(\overline{\alpha})_i$ can have nonzero d_i, by assuming B contains such indexes, then

$$\nabla f(\overline{\alpha})^T d = \sum_{i\in B} y_i^2 \nabla f(\overline{\alpha})_i d_i = (y_i \nabla f(\overline{\alpha})_i) y_B^T d_B = 0.$$

This contradicts to the assumption that (35) has a nonzero solution. Hence $\overline{\alpha}$ is an optimal solution.

Therefore, in the rest of this proof we will show that when solving (35), only elements with the same $y_i \nabla f(\overline{\alpha})_i$ can have nonzero d_i. Assume i_1 (i_2) is the first element selected from the top (bottom) of the sorted list of $y_i \nabla f(\overline{\alpha})_i$, $i = 1, \ldots, l$. We claim that

$$y_{i_1} \nabla f(\overline{\alpha})_{i_1} = y_{i_2} \nabla f(\overline{\alpha})_{i_2}.$$

If the result is wrong, then

$$y_{i_1} \nabla f(\overline{\alpha})_{i_1} > y_{i_2} \nabla f(\overline{\alpha})_{i_2}. \tag{36}$$

Define

$$I_1 \equiv \{i | y_i \nabla f(\overline{\alpha})_i \ge y_{i_1} \nabla f(\overline{\alpha})_{i_1}\}$$

and

$$I_2 \equiv \{i | y_i \nabla f(\overline{\alpha})_i \le y_{i_2} \nabla f(\overline{\alpha})_{i_2}\}.$$

From (36), $I_1 \cap I_2 = \emptyset$.

Since i_1 (i_2) is the first element selected from the top (bottom) of the sorted list, $\overline{\alpha}_{i_1}$ ($\overline{\alpha}_{i_2}$) satisfies (18) [(19)]. After $k \in \mathcal{K}$ is large enough, from Lemma IV.4, $\alpha_{i_1}^k, \alpha_{i_1}^{k+1}, \ldots, \alpha_{i_1}^{k+2l}$ are all "top" candidates, where l is the length of each vector α [i.e., the number of variables of (1)]. In addition, $\alpha_{i_2}^k, \alpha_{i_2}^{k+1}, \ldots, \alpha_{i_2}^{k+2l}$ are all "bottom" candidates. Then in each $\overline{k}$ of kth, $(k+1)$st, $\ldots$, $(k+2l-1)$st iterations, i_1 and i_2 can not both be selected because of (36) and Lemma IV.4.

We then claim that if i_1 is not selected at the $\bar{k}$th iteration, then all "top" candidates selected are from I_1. Since $\bar{k}$ is large enough, for any $\alpha_i^{\bar{k}}$, $i \notin I_1$, which is a "top" candidate

$$y_i \nabla f(\overline{\alpha})_i < y_{i_1} \nabla f(\overline{\alpha})_{i_1}$$

implies that i can not be chosen earlier then i_1. Similarly, if i_2 is not selected at the $\bar{k}$th iteration, then all "bottom" candidates selected are from I_2.

Now for the $\bar{k}$th iteration, we consider three situations.

Case 1) Neither i_1 nor i_2 is selected. Then all "top" ("bottom") candidates selected are in I_1 (I_2). For any $i \in I_1$ and $j \in I_2$ selected in the $\bar{k}$th iteration, from Lemma IV.4b), at the next iteration, either $\alpha_i^{\bar{k}+1}$ becomes a "bottom only" element or $\alpha_j^{\bar{k}+1}$ becomes a "top only" element. Therefore, there are two cases to consider.

Case 1-1) All elements selected from I_1 become "bottom only." Then the number of "bottom only" variables in I_1 is increased by at least one. On the other hand, since $I_1 \cap I_2 = \emptyset$ and from the assumption of case 1, all variables selected from I_2 are "bottom" elements. Hence the number of "top only" variables in I_2 is at least the same.

Case 1-2) All elements selected from I_2 become "top only." Similarly, the number of "top only" variables in I_2 is increased by at least one, while the number of "bottom only" variables in I_1 is at least the same.

Case 2) Only i_1 is selected: As i_2 is not selected, all "bottom" elements selected are in I_2. Since i_1 is selected and $\alpha_{i_1}^{\bar{k}+1}$ is a "top" candidate, all "bottom" elements selected in I_2 become "top only." Therefore, the number of "top only" variables in I_2 increases by at least one. On the other hand, the number of "bottom only" variables in I_1 is at least the same.

Case 3) Only i_2 is selected: Similar to case 2), the number of "bottom only" variables in I_1 increases at least one and the number of "top only" variables in I_2 is at least the same.

Therefore, in at most l iterations, either all elements in I_1 become "bottom only" or all elements in I_2 become "top only." If I_1 reaches "bottom only" first, from Assumption II.1, for later iterations, elements in I_1 are not "top" candidates so i_1 must be selected. Therefore, we only have case 2) left. Then after at most another l iterations, all I_2 are "top only." Therefore, we must have a $\bar{k} \in \{k, k+1, \ldots, k+2l\}$ such that both i_1 and i_2 are selected. This contradicts to Lemma IV.4. Hence the proof is complete. ■

Under some conditions (for example, Q is positive definite), (1) has a unique solution. Hence $\{\alpha^k\}$ is a globally convergent sequence whose limit point is this unique solution. In Burges and Crisp [2], there are discussions on conditions under which the SVM solution is unique.

V. Extensions

Consider a general problem with the following form:

$$\begin{aligned} \min \quad & \tfrac{1}{2}\alpha^T Q \alpha + p^T \alpha \\ & y^T \alpha = \Delta, \\ & l_i \le \alpha_i \le u_i, \qquad i = 1, \ldots, l \end{aligned} \tag{37}$$

where $-\infty < l_i < u_i < \infty$, $i = 1, \ldots, l$, Q is any symmetric positive semidefinite matrix satisfying Assumption IV.1, and $y_i = \pm 1$, $i = 1, \ldots, l$. The convergence proof described in the previous section is still valid if Algorithm I.1 with the following generalized working set selection is used for solving (37)

$$\begin{aligned} \min \quad & \nabla f(\alpha^k)^T d \\ & y^T d = 0, \quad -1 \le d_i \le 1, \qquad i = 1, \ldots, l \\ & d_i \ge 0, \quad \text{if } (\alpha^k)_i = l_i, \quad d_i \le 0, \quad \text{if } (\alpha^k)_i = u_i \\ & |\{d_i | d_i \ne 0\}| \le q. \end{aligned} \tag{38}$$

It can be seen that $y_i = \pm 1$ plays an important role here. Algorithm I.2 is not valid for solving (38) if this condition does not hold. In addition, in the convergence proof we specifically utilize many properties of Algorithm I.2 [e.g., we consider the sorted list of $y_i \nabla f(\overline{\alpha})_i$] so the condition $y_i = \pm 1$ is also used. In [12] and [33], the authors handled a more generalized problem where the only restriction on y_i is $y_i \ne 0$.

Problem (37) covers most SVM formulations. For example, given a set of data points $\{(x_1, z_1), \ldots, (x_l, z_l)\}$ such that $x_i \in R^n$ is an input and $z_i \in R^1$ is a target output, the usual form of support vector regression is as follows:

$$\begin{aligned} \min \ & \tfrac{1}{2}(\alpha - \alpha^*)^T Q (\alpha - \alpha^*) + \epsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) \\ & + \sum_{i=1}^{l} z_i(\alpha_i - \alpha_i^*) \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0 \\ & 0 \le \alpha_i, \ \alpha_i^* \le C, \qquad i = 1, \ldots, l \end{aligned} \tag{39}$$

where $Q_{ij} = \phi(x_i)^T \phi(x_j)$.

We can rewrite (39) as

$$\begin{aligned} \min \ & \tfrac{1}{2}[\alpha^T, (\alpha^*)^T] \begin{bmatrix} Q & -Q \\ -Q & Q \end{bmatrix} \begin{bmatrix} \alpha \\ \alpha^* \end{bmatrix} \\ & + [\epsilon e^T + z^T, \epsilon e^T - z^T] \begin{bmatrix} \alpha \\ \alpha^* \end{bmatrix} y^T \begin{bmatrix} \alpha \\ \alpha^* \end{bmatrix} = 0 \\ & 0 \le \alpha_i, \ \alpha_i^* \le C, \qquad i = 1, \ldots, l \end{aligned} \tag{40}$$

where y is a $2l$ by 1 vector with $y_i = 1$, $i = 1, \ldots, l$ and $y_i = -1$, $i = l+1, \ldots, 2l$. Equation (40) is in the form of (37) so Algorithms I.1 and I.2 can be applied.

However, using Algorithms I.1 and I.2 for (40) is a little different from existing decomposition methods for regression. Note that though (39) is a problem with $2l$ variables, it has very special structures. For example, the KKT condition implies that at an optimal solution of (39), $\alpha_i\alpha_i^* = 0$. Early work on SVM regression (e.g., [8], [14], [15], [26], and [35]) all tried to take advantage of these structures and focused on problem (39). Except [15] they mainly consider selecting two elements as the working set in each iteration. Some characteristics of their methods are the following.

1) In each iteration, two indexes i_1 and i_2 are selected from $\{1, \ldots, l\}$.
2) To keep $\alpha_i^k(\alpha^*)_i^k = 0$, they solve a subproblem with four variables α_{i_1}, α_{i_2}, $\alpha_{i_1}^*$, and $\alpha_{i_2}^*$.

That is, it is like that they use $q = 4$ in Algorithm I.1 with a different working set selection from Algorithm I.2.

If Algorithms I.1 and I.2 with $q = 2$ are directly used for (40), two indexes are selected from $\{1, \ldots, 2l\}$ and a subproblem (2) with two variables is solved. The advantage of working on (40) is that a generalized implementation can be directly used for both classification and regression. However, a possible shortcoming is that special structures of (39) are not considered so there may have computational overheads. Surprisingly we will show that $\alpha_i^k(\alpha^*)_i^k = 0$ still holds if Algorithms I.1 and I.2 are directly applied to solve (40). Another issue is on the Hessian $\overline{Q} = \left[\begin{smallmatrix} Q & -Q \\ -Q & Q \end{smallmatrix}\right]$ of the objective function of (40). Now $\overline{Q}$ is only positive semidefinite so it is unlikely that Assumption IV.1 can be true. In the following theorem we will show that if only Q instead of $\overline{Q}$ satisfies Assumption IV.1, the convergence for solving (40) follows.

Theorem V.1: If Algorithms I.1 and I.2 are used for solving (40) and the initial solution is zero, then $\alpha_i^k(\alpha^*)_i^k = 0$, $i = 1, \ldots, l$ for all k. In addition, if Q satisfies Assumption IV.1, $\left[\begin{smallmatrix} \alpha^k \\ (\alpha^*)^k \end{smallmatrix}\right]$ converges to an optimal solution of (39).

Proof: We prove the first result by the mathematical induction. It is true that if the initial solution is zero, for the first iteration, $\alpha_i^1(\alpha^*)_i^1 = 0$, $i = 1, \ldots, l$. Assume the result is true for the kth iteration and we will prove $\alpha_i^{k+1}(\alpha^*)_i^{k+1} = 0$, $i = 1, \ldots, l$.

We consider three situations in the kth iteration.

1) In Algorithm I.2, both i and $i + l$ are selected in the working set: Then from the KKT condition of the subproblem (2), $\alpha_i^{k+1}(\alpha^*)_i^{k+1} = 0$.
2) Only i but not $i + l$ is selected in the working set and $\alpha_i^k = 0$: For this case, $d_i = 1$ after solving (3a)–(3c). Since $y_i d_i = 1$, we realize that in Algorithm I.2, the index i is selected from the bottom of the sorted list of $y_i\nabla f(\alpha^k, (\alpha^*)^k)_i$, $i = 1, \ldots, 2l$. However, we also have

$$\begin{aligned} y_i\nabla f(\alpha^k, (\alpha^*)^k)_i &= (Q(\alpha^k - (\alpha^*)^k))_i + \epsilon + z_i \\ &\geq (Q(\alpha^k - (\alpha^*)^k))_i - \epsilon + z_i \\ &= y_{i+l}\nabla f(\alpha^k, (\alpha^*)^k)_{i+l}. \end{aligned}$$

In other words, index $i+l$ is closer than i to the bottom of the sorted list. Therefore, if $i+l$ is not selected, index $i+l$ is not a "bottom" candidate so $(\alpha^*)_i^k$ does not satisfy (19). As $y_{i+l} = -1$, $(\alpha^*)_i^k = 0$. Since $i+l$ is not selected in the kth iteration, $(\alpha^*)_i^{k+1} = (\alpha^*)_i^k = 0$ so $\alpha_i^{k+1}(\alpha^*)_i^{k+1} = 0$.
3) Only i but not $i+l$ is selected in the working set and $\alpha_i^k > 0$: Then $\alpha_i^k(\alpha^*)_i^k = 0$ implies $(\alpha^*)_i^k = 0$. Therefore, $(\alpha^*)_i^{k+1} = 0$ and $\alpha_i^{k+1}(\alpha^*)_i^{k+1} = 0$.

When only $i + l$ but not i is selected, the situation is similar. Thus we have finished the proof that $\alpha_i^k(\alpha^*)_i^k = 0$, $i = 1, \ldots, l$, for all k.

Next we switch to the second goal of this theorem. Now the Hessian of the objective function of (40) is $\overline{Q} = \left[\begin{smallmatrix} Q & -Q \\ -Q & Q \end{smallmatrix}\right]$. We remember that Assumption IV.1 is needed near (31) in the proof of Lemma IV.2. If we can prove

$$-\frac{1}{2}s_B^T\overline{Q}_{BB}s_B \leq -\frac{\sigma}{2}\|s_B\|^2 \tag{41}$$

where σ is related only to Q, then a condition on Q instead of $\overline{Q}$ is sufficient for the convergence. Thus the main task is to prove that (41) is true.

We define the following disjoint index sets

$$\begin{aligned} B_1 &= \{i \mid 1 \leq i \leq l,\ i \in B \text{ and } i + l \in B\} \\ B_1^* &= \{i + l \mid i \in B_1\} \\ B_2 &= \{i \mid 1 \leq i \leq l,\ i \in B \text{ and } i + l \notin B\} \\ B_3^* &= \{i + l \mid 1 \leq i \leq l,\ i \notin B \text{ and } i + l \in B\} \\ B_3 &= \{i \mid i + l \in B_3^*\}. \end{aligned}$$

Then

$$B = B_1 \cup B_1^* \cup B_2 \cup B_3^*.$$

Thus

$$\begin{aligned} s_B^T\overline{Q}_{BB}s_B = &[s_{B_1} - s_{B_1^*},\ s_{B_2},\ s_{B_3^*}] \\ &\cdot \begin{bmatrix} Q_{B_1B_1} & Q_{B_1B_2} & -Q_{B_1B_3} \\ Q_{B_2B_1} & Q_{B_2B_2} & -Q_{B_2B_3} \\ -Q_{B_3B_1} & -Q_{B_3B_2} & Q_{B_3B_3} \end{bmatrix} \\ &\cdot \begin{bmatrix} s_{B_1} - s_{B_1^*} \\ s_{B_2} \\ s_{B_3^*} \end{bmatrix}. \end{aligned} \tag{42}$$

Since

$$\begin{bmatrix} Q_{B_1B_1} & Q_{B_1B_2} & -Q_{B_1B_3} \\ Q_{B_2B_1} & Q_{B_2B_2} & -Q_{B_2B_3} \\ -Q_{B_3B_1} & -Q_{B_3B_2} & Q_{B_3B_3} \end{bmatrix}$$

has the same eigenvalues as

$$\begin{bmatrix} Q_{B_1B_1} & Q_{B_1B_2} & Q_{B_1B_3} \\ Q_{B_2B_1} & Q_{B_2B_2} & Q_{B_2B_3} \\ Q_{B_3B_1} & Q_{B_3B_2} & Q_{B_3B_3} \end{bmatrix}$$

which is a square submatrix of Q, and $|B_1 \cup B_2 \cup B_3| \leq q$, we have

$$\begin{aligned} -\frac{1}{2}s_B^T\overline{Q}_{BB}s_B &\leq -\frac{\sigma}{2}\left\|\begin{bmatrix} s_{B_1} - s_{B_1^*} \\ s_{B_2} \\ s_{B_3^*} \end{bmatrix}\right\|^2 - \frac{\sigma}{2}\left\|\begin{bmatrix} s_{B_1} \\ s_{B_1^*} \\ s_{B_2} \\ s_{B_3^*} \end{bmatrix}\right\|^2 \\ &= -\frac{\sigma}{2}\|s_B\|^2 \end{aligned}$$

where $\sigma = \min_I(\min(eig(Q_{II})))$, and I is any subset of $\{1, \ldots, l\}$ with $|I| \leq q$. Note that $\|s_{B_1} - s_{B_1^*}\|^2 \geq \|\left[\begin{smallmatrix} s_{B_1} \\ s_{B_1^*} \end{smallmatrix}\right]\|^2$

is because of the following reasons: Since $\alpha_i^{k+1}(\alpha^*)_i^{k+1} = 0$, we consider two situations:

1) $\alpha_i^{k+1} = 0$: Then if $\alpha_i^k = 0$, $s_i = 0$ so $s_i s_{i+l} \leq 0$. On the other hand, if $\alpha_i^k > 0$, $(\alpha^*)_i^k = 0$. Hence $s_i \leq 0$ and $s_{i+l} \geq 0$ imply $s_i s_{i+l} \leq 0$.
2) $(\alpha^*)_i^{k+1} = 0$: Similarly, $s_i s_{i+l} \leq 0$.

Therefore, we have $-s_{B_1}^T s_{B_1^*} \geq 0$ so

$$\|s_{B_1} - s_{B_1^*}\|^2 = \left\| \begin{bmatrix} s_{B_1} \\ s_{B_1^*} \end{bmatrix} \right\|^2 - 2 s_{B_1}^T s_{B_1^*} \geq \left\| \begin{bmatrix} s_{B_1} \\ s_{B_1^*} \end{bmatrix} \right\|^2.$$

■

Recent implementation using Algorithms I.1 and I.2 with $q = 2$ for (40) are `LIBSVM` [4] and `SVMTorch` [6].

Note that it may be possible to extend convergence results in this section for algorithms used in [14], [15], [26], and [35], but here we will not get into details.

We then briefly discuss two other SVM formulations: one-class SVM and ν-SVM. For one-class SVM [24], the formulation is already in the form of (37). For ν-SVM [25], it has two linear constraints so is not covered by the algorithm and proof here.

The last formulation which we will consider is the support vector classification with quadratic penalty functions [7]. The required optimization problem is as follows:

$$\begin{aligned} \min \quad & \tfrac{1}{2}\alpha^T (Q + I/C)\alpha - e^T\alpha \\ & 0 \leq \alpha_i, \qquad i = 1, \ldots, l \\ & y^T\alpha = 0 \end{aligned} \tag{43}$$

where I is the identity matrix.

As the upper bound of α is ∞ so the feasible region of (43) seems to be unbounded. This is a difficulty since we need the bounded property in order to have convergent subsequences.

For any feasible point α of (43), since Q is positive semidefinite

$$\frac{1}{2C}\alpha^T\alpha - e^T\alpha \leq \frac{-1}{2}\alpha^T Q\alpha \leq 0.$$

Therefore

$$\frac{1}{2C}\sum_{i=1}^{l}(\alpha_i - C)^2 \leq \frac{l}{2C}\cdot C^2 = \frac{lC}{2}$$

implies that

$$\alpha_i - C \leq \sqrt{l}C.$$

Thus all feasible α are in fact in a compact set. Then the convergence proof follows.

VI. Conclusion and Discussion

In this section we give some notes about the convergence proof. The property that (2) is exactly solved is used both in Lemmas IV.2 and IV.4. This confirms the conjectures in Section I where we think that an optimal solution of (2) makes a difference from the original Zoutendijk's method.

It is unfortunate that we need Assumption IV.1 for the proof. We hope that this gap can be filled sometime in the future.

The convergence proof also suggests a possible way to improve the implementation. In final iterations, as the order of sorting $y_i \nabla f(\alpha^*)_i$, $i = 1, \ldots, l$ is about fixed, it might be possible to consider fewer elements on the working set selection.

Acknowledgment

The author thanks S. Keerthi for many helpful comments.

References

[1] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 2nd ed. New York: Wiley, 1993.
[2] C. J. C. Burges and D. J. Crisp, "Uniquness of the SVM solution," in *Proc. NIPS99*, 1999.
[3] C.-C. Chang, C.-W. Hsu, and C.-J. Lin, "The analysis of decomposition methods for support vector machines," *IEEE Trans. Neural Networks*, vol. 11, pp. 1003–1008, July 2000.
[4] C.-C. Chang and C.-J. Lin. (2001) LIBSVM: A library for support vector machines. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm
[5] *Training ν-support vector classifiers: Theory and algorithms*, vol. 13, no. 9, pp. 2119–2147, 2001.
[6] R. Collobert and S. Bengio. (2001) SVMTorch: A support vector machine for large-scale regression and classification problems. *J. Machine Learning Res.* [Online], pp. 143–160. Available: http://www.idiap.ch/learning/SVMTorch.html
[7] C. Cortes and V. Vapnik, "Support-vector network," *Machine Learning*, vol. 20, pp. 273–297, 1995.
[8] G. W. Flake and S. Lawrence, "Efficient SVM regression training with SMO," *Machine Learning*, vol. 46, pp. 271–290, 2002.
[9] T.-T. Friess, N. Cristianini, and C. Campbell, "The kernel adatron algorithm: A fast and simple learning procedure for support vector machines," in *Proc. 15th Int. Conf. Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1998.
[10] C.-W. Hsu and C.-J. Lin, "A simple decomposition method for support vector machines," *Machine Learning*, vol. 46, pp. 291–314, 2002.
[11] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods—Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA: MIT Press, 1998.
[12] S. Keerthi and E. G. Gilbert, "Convergence of a generalized SMO algorithm for SVM classifier design," *Machine Learning*, vol. 46, pp. 351–360, 2002.
[13] S. Keerthi, S. Shevade, C. Bhattacharyya, and K. Murthy, "Improvements to Platt's SMO algorithm for SVM regression," *IEEE Trans. Neural Networks*, vol. 11, pp. 1188–1193, 2000.
[14] ——, "Improvements to SMO algorithm for SVM classifier," *Neural Comput.*, vol. 13, pp. 637–649, 2001.
[15] P. Laskov, "An improved decomposition algorithm for regression support vector machines," *Machine Learning*, vol. 46, pp. 315–350, 2002.
[16] O. L. Mangasarian and D. R. Musicant, "Successive overrelaxation for support vector machines," *IEEE Trans. Neural Networks*, vol. 10, pp. 1032–1037, Sept. 1999.
[17] C. A. Micchelli, "Interpolation of scattered data: Distance matrices and conditionally positive definite functions," *Constructive Approximation*, vol. 2, pp. 11–22, 1986.
[18] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: An application to face detection," in *Proc CVPR'97*, 1997.
[19] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA: MIT Press, 1998.
[20] M. J. D. Powell, "On search directions for minimization," *Math. Programming*, vol. 4, pp. 193–201, 1973.
[21] S. Rüping. (2000) mySVM—Another one of those support vector machines. [Online]. Available: http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/
[22] C. Saunders, M. O. Stitson, J. Weston, L. Bottou, B. Schölkopf, and A. Smola, "Support vector machine reference manual," Royal Holloway, Univ. London, Egham, U.K., Tech. Rep. CSD-TR-98-03, 1998.
[23] B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds., *Advances in Kernel Methods—Support Vector Learning*. Cambridge, MA: MIT Press, 1998.
[24] *Estimating the support of a high-dimensional distribution*, vol. 13, no. 7, pp. 1443–1471, 2001.

[25] B. Schölkopf, A. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Comput.*, vol. 12, pp. 1207–1245, 2000.

[26] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," Royal Holloway College, Neuro COLT Tech. Rep. TR-1998-030, 1998.

[27] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[28] ——, *Statistical Learning Theory*. New York: Wiley, 1998.

[29] P. Wolfe, "On the convergence of gradient methods under constraint," *IBM J. Res. Development*, vol. 16, pp. 407–411, 1972.

[30] G. Zoutendijk, *Methods of Feasible Directions*. Amsterdam, The Netherlands: Elsevier, 1960.

[31] C.-C. Chang and C.-J. Lin, "Training ν-support vector classifiers: Theory and algorithms," Neural Comput., vol. 13, no. 7, pp. 2119–2147, 2000, to be published.

[32] C.-W. Hsu and C.-J. Lin, "A simple decomposition method for support vector machines," Machine Learning, vol. 46, pp. 291–314, 1999, to be published.

[33] S. Keerthi and E. G. Gilbert, "Convergence of a generalized SMO algorithm for SVM classifier design," Machine Learning, vol. 46, pp. 351–360, 2002, to be published.

[34] S. Keerthi, S. Shevade, C. Bhattacharyya, and K. Murthy, "Improvements to Platt's SMO algorithm for SVM classifier design," Neural Comput., vol. 13, pp. 637–649, 2002, to be published.

[35] S. Keerthi, S. Shevade, C. Bhattacharyya, and K. Murthy, "Improvements to SMO algorithm for SVM regression," *IEEE Trans. Neural Networks*, vol. 11, pp. 1188–1193, Sept. 2000.

Chih-Jen Lin (S'91–M'98) received the B.S. degree in mathematics from National Taiwan University, Taipei, Taiwan, in 1993. He received the M.S. and Ph.D. degrees from the Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor.

Since September 1998, he has been an Assistant Professor in the Department of Computer Science and Information Engineering at National Taiwan University. His research interests include machine learning, numerical optimization, and various applications of operations research.