Stopping Criteria of Decomposition Methods for Support Vector Machines: a Theoretical Justification

Chih-Jen Lin

Department of Computer Science and

Information Engineering

National Taiwan University

Taipei 106, Taiwan

cjlin@csie.ntu.edu.tw

Abstract

In [8] we prove the convergence of a commonly used decomposition method for SVM. However, there is no theoretical justification about its stopping criterion which is based on the gap of the violation of the optimality condition. It is essential to have that the gap asymptotically approaches zero so we are sure that existing implementations stop in finite iterations after reaching a specified tolerance. Here we prove this result and illustrate it by two extensions: ν-SVM and a multi-class SVM by Crammer and Singer. We then also prove the asymptotic convergence of a decomposition method for this multi-class SVM. Discussions on the difference between this convergence proof and the one in [8] are also included.

I. Introduction

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 machines (SVM) [3], [12] require the solution of the following optimization problem:

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

where e is the vector of all ones, C is the upper bound of all variables, and Q is an 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.

Due to the density of the matrix Q, currently the decomposition method is one of the major methods to solve SVM (e.g. [9], [6], [10]). It is an iterative process where in each iteration the index set of variables are separated to two sets B and N, where B is the working set. Then in that iteration variables corresponding to N are fixed while a sub-problem on variables corresponding to B is minimized.

Practically we need a stopping condition for the decomposition method. Such a criterion usually uses the information of the Karush-Kuhn-Tucker (KKT) condition, that is, the optimality condition of (1): If α is an optimal solution of (1), there is a number b two nonegative vectors λ and μ such that

$$\nabla f(\alpha) + by = \lambda - \mu$$
$$\lambda_i \alpha_i = 0, \mu_i (C - \alpha)_i = 0, \lambda_i \geq 0, \mu_i \geq 0, i = 1, \ldots, l.$$

This can be rewritten as

$$\nabla f(\alpha)_i + by_i \geq 0 \qquad \text{if } \alpha_i = 0, \tag{2a}$$
$$\nabla f(\alpha)_i + by_i \leq 0 \qquad \text{if } \alpha_i = C, \tag{2b}$$
$$\nabla f(\alpha)_i + by_i = 0 \qquad \text{if } 0 < \alpha_i < C. \tag{2c}$$

Since $y_i = \pm 1$, this KKT condition is equivalent to

$$\begin{aligned} m(\alpha) &= \max\Big(\max_{\alpha_i < C, y_i = 1} -\nabla f(\alpha)_i, \max_{\alpha_i > 0, y_i = -1} \nabla f(\alpha)_i\Big) \\ &\leq \min\Big(\min_{\alpha_i < C, y_i = -1} \nabla f(\alpha)_i, \min_{\alpha_i > 0, y_i = 1} -\nabla f(\alpha)_i\Big) = M(\alpha). \end{aligned} \tag{3}$$

If the algorithm for (1) does not stop in finite steps, during iterations $m(\alpha^k) > M(\alpha^k)$, where α^k is the solution of the kth iteration. Hence naturally a stopping criterion can be:

$$m(\alpha^k) \leq M(\alpha^k) + \epsilon, \tag{4}$$

where ϵ is a stopping tolerance. (3) is a much simpler way of describing the KKT condition. Some existing working set selections also follow from identifying elements violating (3). More importantly, unlike earlier approaches where b is calculated using (2c), the condition on free variables, we do not have to worry if there are free variables in the final solution or not. If all variables are at bounds, using (3), b can be simply calculated as $(M(\alpha) + m(\alpha))/2$. Such a stopping criterion has been derived and used in, for example, [7], [1].

In an earlier work [8] on the convergence of the decomposition method proposed in the software SVM^{light} [6], we focused on proving that any limit point of $\{\alpha^k\}$ is an optimal solution of (1). However, such results do not directly support the validity of using (4) as the stopping criterion. To be more precise, even though if we have $\lim_{k\to\infty}\alpha^k = \bar{\alpha}$ which is an optimal solution, directly from the definition of $m(\alpha^k)$ and $m(\bar{\alpha})$, we may not have $\lim_{k\to\infty} m(\alpha^k) = m(\bar{\alpha})$. A similar problem happens for $M(\alpha^k)$ and $M(\bar{\alpha})$. A reason is that it may be possible that $\alpha_i^k > 0, \forall k$ but $\lim_{k\to\infty}\alpha_i^k = \bar{\alpha}_i = 0$. Therefore, we worry about the situation that $\{\alpha^k\}$ converges to an optimal solution but $\lim_{k\to\infty} m(\alpha^k) - M(\alpha^k) > 0$. Then the decomposition implementation never stops by using (4) as the criterion. In Section II we prove that this situation will never happen.

Note that contrary to [8], in [7] when a convergence proof of the decomposition method using two elements as the working set in each iteration was proposed, the authors directly prove the finite stop first and then the asymptotic convergence. Hence they do not face a similar problem on the validity of the stopping criterion.

In Section II we also prove that most bounded variables are identified after finite steps so final iterations focus on a particular set of variables. This analysis supports the use of shrinking techniques in the decomposition method. Section III then shows some extensions on a more complicated optimization formulation. We use two examples to illustrate our results: ν-SVM [11] and a multi-class SVM in [4]. We then also prove the asymptotic convergence of a decomposition method for this multi-class SVM. Discussions on the difference between this convergence proof and the one in [8] are also included.

II. Main Results on Stopping Criteria

Here we consider a more general problem:

$$\begin{aligned} \min \quad & f(\alpha) \\ & o_i \le \alpha_i \le u_i, i = 1, \dots, l, \\ & y^T\alpha = 0, \end{aligned} \tag{5}$$

where $f(\alpha)$ is any continuously differentiable function, $y_i = \pm 1, i = 1, \ldots, l$, and o and u are lower and upper bounds, respectively. We then define generalized $m(\alpha^k)$ and $M(\alpha^k)$:

$$m(\alpha^k) \equiv \max(\max_{\alpha_t^k < u_t, y_t = 1} -\nabla f(\alpha^k)_t, \max_{\alpha_t^k > o_t, y_t = -1} \nabla f(\alpha^k)_t), \tag{6}$$

and

$$M(\alpha^k) \equiv \min(\min_{\alpha_t^k < u_t, y_t = -1} \nabla f(\alpha^k)_t, \min_{\alpha_t^k > o_t, y_t = 1} -\nabla f(\alpha^k)_t). \tag{7}$$

We denote $\text{arg}m(\alpha^k)$ the set of indices whose $-y_i \nabla f(\alpha^k)_i$ are the same as $m(\alpha^k)$. A similar definition goes for $\text{arg}M(\alpha^k)$. Thus if α^k is not an optimal solution yet, $m(\alpha^k) > M(\alpha^k)$ so $\text{arg}m(\alpha^k) \cap \text{arg}M(\alpha^k) = \emptyset$.

The following theorem shows the validity of the stopping criterion (4):

Theorem II.1 *Assume a decomposition method for solving (5) satisfies the following conditions:*
1. $m(\alpha^k) > M(\alpha^k), \forall k$.
2. At least one of $\text{arg}m(\alpha^k)$ *and one of* $\text{arg}M(\alpha^k)$ *are included in the working set of each iteration.*
3. Consider (7). If α_i^k *satisfies*

$$\alpha_i^k = o_i, y_i = -1 \text{ or } \alpha_i^k = u_i, y_i = 1,$$

α_j^k *satisfies*

$$\alpha_j^k < u_j, y_j = -1 \text{ or } \alpha_j^k > o_j, y_j = 1,$$

and

$$-y_i \nabla f(\alpha^k)_i > -y_j \nabla f(\alpha^k)_j,$$

then if α_i^k *is in the working set,* α_j^k *must be selected as well. For Eq. (6), we assume similar conditions.*
4. $\{\alpha^k\}$ *converges to an optimal solution* $\bar{\alpha}$ *of (5).*
Then

$$\lim_{k \to \infty} (m(\alpha^k) - M(\alpha^k)) = 0. \tag{8}$$

Proof: If the result (8) is wrong, with condition (1) there is an infinite set $\bar{\mathcal{K}}$ and a $\Delta > 0$ such that

$$m(\alpha^k) \geq M(\alpha^k) + \Delta, \forall k \in \bar{\mathcal{K}}. \tag{9}$$

As $\bar{\mathcal{K}}$ has infinitely many elements but the number of pairs of variables is finite, from condition 2 there are indices i and j and an infinite subsequence such that $i \in \arg m(\alpha^k)$ and $j \in \arg M(\alpha^k)$ are both in the working set. Without loss of generality we can consider only the case that there is $\hat{\mathcal{K}} \subset \bar{\mathcal{K}}$ such that $y_i = y_j = -1$ and for all $k \in \hat{\mathcal{K}}$,

$$m(\alpha^k) = \nabla f(\alpha^k)_i \text{ and } M(\alpha^k) = \nabla f(\alpha^k)_j. \tag{10}$$

Thus from the definition of $m(\alpha^k)$ and $M(\alpha^k)$ in (6) and (7),

$$\alpha_i^k > o_i \text{ and } \alpha_j^k < u_j, \forall k \in \hat{\mathcal{K}}. \tag{11}$$

Since f is continuously differentiable,

$$\lim_{k \to \infty} \nabla f(\alpha^k)_i = \nabla f(\bar{\alpha})_i \text{ and } \lim_{k \to \infty} \nabla f(\alpha^k)_j = \nabla f(\bar{\alpha})_j. \tag{12}$$

From (9), (10), and (12), we have

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

If we have an infinite subset of $\hat{\mathcal{K}}$ such that $\alpha_i^{k+1} > o_i$ and $\alpha_j^{k+1} < u_j$, then the KKT condition of the sub-problem implies

$$\nabla f(\alpha^{k+1})_j \geq \nabla f(\alpha^{k+1})_i.$$

Since $\{\alpha^k\}$ is a convergent sequence, taking the limit we have

$$\nabla f(\bar{\alpha})_i \leq \nabla f(\bar{\alpha})_j \tag{14}$$

which contradicts (13).

Therefore, we have that

$$\alpha_i^{k+1} = o_i \text{ or } \alpha_j^{k+1} = u_j, \text{ after } k \in \hat{\mathcal{K}} \text{ is large enough.} \tag{15}$$

Because of (15) and (11), there is an infinite set $\mathcal{L}$ such that for all $k \in \mathcal{L}$,

$$\alpha_i^k = o_i, \alpha_j^k < u_j, \text{ or } \alpha_i^k > o_i, \alpha_j^k = u_j, \text{ or } \alpha_i^k = o_i, \alpha_j^k = u_j, \tag{16}$$

and

$$\alpha_i^{k+1} > o_i, \alpha_j^{k+1} < u_j. \tag{17}$$

For the first case of (16), α_i^k is selected in the working set and then modified. However, since (13) and condition 4 imply

$$\nabla f(\alpha^k)_i > \nabla f(\alpha^k)_j,$$

after k is large enough, with $\alpha_j^k < u_j$ in (16), from condition 3 this theorem α_j^k is also in the working set of the kth iteration where $k \in \mathcal{L}$. With (17), from the KKT condition of the sub-problem,

$$\nabla f(\alpha^{k+1})_i \leq \nabla f(\alpha^{k+1})_j, \tag{18}$$

The situation for the second case is similar. For the third case, both α_i^k and α_j^k are modified so i and j are in the working set. Hence (18) is also valid. Therefore, we have (18) for all $k \in \mathcal{L}$. As k goes to infinity, we again obtain (14) which contradicts (13).

■

Note that conditions 2 and 3 of Theorem II.1 are requirements on the working set selection. We list conditions instead of focusing on a particular working set selection so that more flexible selections may be used.

We now check that the working set selection of SVM^{light} satisfies the conditions 2 and 3 of Theorem II.1. If q, an even number, is the size of the working set, $q/2$ indices are sequentially selected from elements which satisfy $\alpha_t^k < u_i, y_t = 1$ or $\alpha_t^k > o_i, y_t = -1$ so that

$$-y_{i_1}\nabla f(\alpha^k)_{i_1} \geq -y_{i_2}\nabla f(\alpha^k)_{i_2} \geq \cdots \geq -y_{i_{q/2}}\nabla f(\alpha^k)_{i_{q/2}} \text{ and } i_1 \in \arg m(\alpha^k). \tag{19}$$

The other $q/2$ indices are sequentially selected from elements which satisfy $\alpha_t^k < u_i, y_t = -1$ or $\alpha_t^k > o_i, y_t = 1$ such that

$$-y_{j_{q/2}}\nabla f(\alpha^k)_{j_{q/2}} \geq \cdots \geq -y_{j_1}\nabla f(\alpha^k)_{j_1} \text{ and } j_1 \in \arg M(\alpha^k). \tag{20}$$

Interestingly this working set selection was originally derived from the concept of feasible directions in constrained optimization but not from the violation of the KKT condition.

Regarding the global convergence of $\{\alpha^k\}$ which is the condition 4, unfortunately we prove only weaker results in [8]: under some minor assumptions every limit point of convergent subsequences is an optimal solution. This result does imply the global convergence if (1) has a unique optimal solution. Then Theorem II.1 can be applied. For example, if Q is positive definite, the solution of (1) is unique.

To remedy this problem on the global convergence of $\{\alpha^k\}$, in the following we give a different version of Theorem II.1 which is specific to the working set selection (19) and (20).

Theorem II.2 *The decomposition method using (19) and (20) as the for selecting the working set has*

$$\lim_{k\to\infty}(m(\alpha^k) - M(\alpha^k)) = 0. \tag{21}$$

Proof: Until (13), the proof is similar to that of Theorem II.1.

Remember that we consider the case of $y_i = y_j = -1$. We then claim that for all $k \in \hat{\mathcal{K}}$ large enough,

$$\alpha_i^{k-1} > o_i \text{ and } \alpha_j^{k-1} < u_j. \tag{22}$$

If (22) is wrong,

$$\alpha_i^{k-1} = o_i, \alpha_j^{k-1} < u_j, \text{ or } \alpha_i^{k-1} > o_i, \alpha_j^{k-1} = u_j, \text{ or } \alpha_i^{k-1} = o_i, \alpha_j^{k-1} = u_j. \tag{23}$$

For the first case, α_i^{k-1} is selected in the working set and modified. Since $\{\alpha^k\}, k \in \mathcal{K}$ is a convergent subsequence, Theorem IV.3 of [8] implies that $\{\alpha^{k-1}\}, k \in \mathcal{K}$ also converges to $\bar{\alpha}$. Thus after $k \in \mathcal{K}$ is large enough, (13) implies

$$-y_i \nabla f(\alpha^{k-1})_i > -y_j \nabla f(\alpha^{k-1})_j,$$

Hence (19) and (20) imply that α_j^{k-1} is selected in the working set as well. Therefore, from the KKT condition of the sub-problem at the $(k-1)$st iteration,

$$\nabla f(\alpha^k)_i \leq \nabla f(\alpha^k)_j$$

which is impossible after $k \in \mathcal{K}$ is large enough. Therefore, (22) is correct. Since Theorem IV.3 of [8] shows that for any given s, $\{\alpha^{k-s}\}, k \in \mathcal{K}$ converges to the same point $\bar{\alpha}$ as $\{\alpha^k\}, k \in \mathcal{K}$, using the same argument above we have that for any given s, after $k \in \mathcal{K}$ is large enough,

$$\begin{aligned} &\alpha_i^{k-s} > o_i \text{ and } \alpha_j^{k-s} < u_j, \cdots, \alpha_i^k > o_i \text{ and } \alpha_j^k < u_j, \\ -y_i\nabla f(\alpha^{k-s})_i > &-y_j\nabla f(\alpha^{k-s})_j, \cdots, -y_i\nabla f(\alpha^k)_i > -y_j\nabla f(\alpha^k)_j. \end{aligned} \tag{24}$$

Consider $s = 2l$. Using the same counting procedure in Theorem IV.5 of [8], we can show that at some $k' \in \{k-2l, \dots, k-1\}$, $\alpha_i^{k'}$ and $\alpha_j^{k'}$ are both selected in the working set so

$$f(\alpha^{k'+1})_i \leq f(\alpha^{k'+1})_j,$$

causes a contradiction to (24). The situation for other cases of (23) is similar.

Therefore, the assumption (9) is wrong so

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

■

Based on the above results next we show that after k is large enough, only elements whose $-y_i\nabla f(\bar{\alpha})_i$ are $m(\bar{\alpha})$ or $M(\bar{\alpha})$ can still be modified. For simplification, we only show results extended from Theorem II.1.

Theorem II.3 *Under the same assumptions as Theorem II.1, we have:*
1. For any $\bar{\alpha}_i$ whose corresponding $-y_i\nabla f(\bar{\alpha})_i$ is neither $m(\bar{\alpha})$ nor $M(\bar{\alpha})$, after k is large enough, α_i^k is at a bound and is equal to $\bar{\alpha}_i$.
2. After k is large enough, only elements in

$$\{t \mid -y_t\nabla f(\bar{\alpha})_t = m(\bar{\alpha}) = M(\bar{\alpha})\} \tag{25}$$

can still be possibly modified.

Proof: First we know that from the KKT condition, if $\bar{\alpha}_i$'s $-y_i\nabla f(\bar{\alpha})_i$ is neither $m(\bar{\alpha})$ nor $M(\bar{\alpha})$, $\bar{\alpha}_i$ is at a bound. Without loss of generality we consider an

$$\bar{\alpha}_i = o_i \text{ with } y_i = -1 \text{ and } \nabla f(\bar{\alpha})_i > M(\bar{\alpha}). \tag{26}$$

If the result of this theorem is wrong, $\alpha_i^k > o_i$ happens infinitely many times. Therefore, from (26), there is an infinite set $\bar{\mathcal{K}}$ and a $\Delta > 0$ such that

$$m(\alpha^k) \geq \nabla f(\alpha^k)_i > M(\bar{\alpha}) + \Delta, \forall k \in \bar{\mathcal{K}}. \tag{27}$$

Now for any $j \in \arg M(\bar{\alpha})$, we have $\bar{\alpha}_j < u_j, y_j = -1$ or $\bar{\alpha}_j > o_j, y_j = 1$. Therefore, $\alpha_j^k < u_j, y_j = -1$ or $\alpha_j^k > o_j, y_j = 1$ after k is large enough. Since $\{\alpha^k\}$ is a convergent sequence, $\{\alpha^k\}$ is in a compact region. With $M(\alpha^k) \leq -y_j \nabla f(\alpha^k)_j$, there is an infinite subset $\hat{\mathcal{K}}$ of $\bar{\mathcal{K}}$ such that $\lim_{k\in\hat{\mathcal{K}}} M(\alpha^k)$ exists and

$$\lim_{k\in\hat{\mathcal{K}}} M(\alpha^k) \leq M(\bar{\alpha}). \tag{28}$$

Hence (27) and (28) imply

$$\lim_{k\in\hat{\mathcal{K}}} m(\alpha^k) - M(\alpha^k) \neq 0, \tag{29}$$

which contradicts Theorem II.1. This completes the first part of the proof. The second part also immediately follows. ■

The above analysis supports the use of the shrinking technique in the decomposition method as in final iterations most variables are not changed.

Note that (5) is a more general formulation so results in this section apply to different formulations discussed in [8] such as support vector regression and one-class SVM.

III. Extensions

In this section we consider a more general problem:

$$\begin{aligned} \min \quad & f(\alpha) \\ & \sum_{m=1}^{r_i} y_i^m \alpha_i^m = \Delta_i, \\ & o_i^m \leq \alpha_i^m \leq u_i^m, m = 1, \ldots, r_i, i = 1, \ldots, l, \end{aligned}$$

where $y_i^m = \pm 1$. Therefore, there are $\sum_{i=1}^{l} r_i$ variables and l linear equality constraints. Hence it is like that there are l group of variables where each one satisfies a linear constraint. The KKT condition requires that there are $b_1, \ldots, b_l$ such that for all $i =$

$1, \ldots, l, m = 1, \ldots, r_i$,

$$\begin{aligned} \nabla f(\alpha)_i^m + b_i y_i^m &\geq 0 \quad \text{if } \alpha_i^m < u_i^m, \\ &\leq 0 \quad \text{if } \alpha_i^m > o_i^m, \end{aligned}$$

where $\nabla f(\alpha)_i^m$ means $\partial f(\alpha)/\partial \alpha_i^m$. We can rewrite the KKT condition as

$$\begin{aligned} m(\alpha)_i = \max(\max_{\alpha_i^m < u_i^m, y_i^m = 1} -\nabla f(\alpha)_i^m, \max_{\alpha_i^m > o_i^m, y_i^m = -1} \nabla f(\alpha)_i^m) \\ \leq \min(\min_{\alpha_i^m < u_i^m, y_i^m = -1} \nabla f(\alpha)_i^m, \min_{\alpha_i^m > o_i^m, y_i^m = 1} -\nabla f(\alpha)_i^m) = M(\alpha)_i, i = 1, \ldots, l. \end{aligned} \tag{30}$$

By defining

$$m(\alpha) \equiv \max_i m(\alpha)_i \text{ and } M(\alpha) \equiv \max_i M(\alpha)_i,$$

the stopping criterion can be

$$m(\alpha^k) - M(\alpha^k) \leq \epsilon, \tag{31}$$

where ϵ is the stopping tolerance.

Similar to the situation in Section II, we can define two sets arg$m(\alpha^k)$ and arg$M(\alpha^k)$. With some modifications on conditions 2 and 3, we can have results similar to Theorem II.1:

Theorem III.1 *Assume all conditions of Theorem II.1 hold with the following modifications:*
2'. In each iteration, if the ith group has the largest $m(\alpha^k)_i - M(\alpha^k)_i$, then at least one of argm$(\alpha^k)_i$ and one of argM$(\alpha^k)_i$ are included in the working set,
3'. In each iteration, variables of the group with the largest $m(\alpha^k)_i - M(\alpha^k)_i$ satisfy condition 3.
Then

$$\lim_{k \to \infty} m(\alpha^k) - M(\alpha^k) = 0. \tag{32}$$

Some SVM formulations are of this form. In the following we will give two examples: ν-SVM and a multi-class SVM by Crammer and Singer. The ν-SVM [11] can be written as the following form [2]:

$$
\begin{aligned}
\min \quad & \frac{1}{2}\alpha^T Q\alpha \\
& \sum_{m=1}^{r_1} \alpha_1^m = \frac{\nu}{2}, \sum_{m=1}^{r_2} \alpha_2^m = \frac{\nu}{2}, \\
& 0 \leq \alpha_1^m \leq \frac{1}{l}, m = 1, \ldots, r_1, \\
& 0 \leq \alpha_2^m \leq \frac{1}{l}, m = 1, \ldots, r_2,
\end{aligned}
$$

where $\nu \in [0, 1]$ is a parameter to adjust the number of support vectors and training errors, r_1 and r_2 are number of training data in two classes, and $l = r_1 + r_2$. A stopping criterion like (31) has been used in the experiment of [2] which implemented a modified decomposition method from the one in [8]. We can easily check that conditions 2' and 3' of Theorem III.1 are satisfied. Regarding the convergence of $\{\alpha^k\}$, though we have not explicitly written down the proof, we conjecture that the same results in [8] that every limit point is an optimal solution should still apply.

Another example is a formulation for multi-class SVM by Crammer and Singer [4]:

$$
\begin{aligned}
\min_{\alpha} \quad & \frac{1}{2}\alpha^T (K \otimes I)\alpha + e^T\alpha \\
& \sum_{m=1}^{r} \alpha_i^m = 0, && \text{(33a)} \\
& \alpha_i^m \leq C_{\bar{y}_i}^m, && \text{(33b)} \\
& i = 1, \ldots, l, m = 1, \ldots, r,
\end{aligned}
$$

where $\otimes$ is the Kronecker product, r is the number of classes, K is an l by l kernel matrix, I is a r by r identity matrix, $\bar{y}_i \in \{1, \ldots, r\}$ is the label of the ith data, and

$$
C_{\bar{y}_i}^m = \begin{cases} 0 & \text{if } \bar{y}_i \neq m, \\ C & \text{if } \bar{y}_i = m. \end{cases}
$$

Hence the stopping criterion can be

$$
\max_i \Big(\max_{\alpha_i^m < C_{\bar{y}_i}^m} -\nabla f(\alpha)_i^m - \min_{\alpha_i^m \leq C_{\bar{y}_i}^m} -\nabla f(\alpha)_i^m \Big) < \epsilon. \tag{34}
$$

Note that since there are no lower bounds and all coefficients in (33a) are +1, the condition $\alpha_i^m > o_i^m, y_i^m = 1$ in (30) becomes $\alpha_i^m \leq C_{\bar{y}_i}^m$.

Implementations of decomposition methods for (33) have been discussed in [4], [5]. Basically the ith group of variables which has the maximal violation in (31) becomes the working set. Thus the sub-problem at the kth iteration is:

$$\begin{aligned} \min_{\alpha_i} \quad & \frac{1}{2} K_{ii} \alpha_i^T \alpha_i + (e + \sum_{j \neq i} K_{ij} (\alpha^k)_j)^T \alpha_i \\ & e^T \alpha_i = 0, \\ & \alpha_i^m \leq C_{\bar{y}_i}^m, m = 1, \ldots, r. \end{aligned} \tag{35}$$

where $(\alpha^k)_j$ means $[(\alpha^k)_j^1, \ldots, (\alpha^k)_j^r]^T$, and e is an $r \times 1$ vector of all ones. In general we denote

$$\alpha_j \equiv [\alpha_j^1, \ldots, \alpha_j^r]^T.$$

Note that (41) is a very simple problem. In [4], two methods were proposed for it: one is an $O(r \log r)$ algorithm while the other is an iterative procedure.

Since a whole group of variables $\alpha_1^1, \ldots, \alpha_1^r$ is selected, condition 3' of Theorem III.1 holds. If the sequence converges to an optimal solution, we have (32). In the next section we will prove the convergence of this decomposition method.

IV. Convergence of a Decomposition Method for Multi-class SVM by Crammer and Singer

The decomposition method mentioned in the previous section for multi-class SVM is not in the category of decomposition methods considered in [8]. Hence proofs in [8] cannot be directly used. However, since the working set selection as well as the sub-problem are quite special where each time one of the r groups of variables is considered, we will show that the convergence proof is even simpler. First we prove a simple lemma:

Lemma IV.1 *Consider the following problem:*

$$\begin{aligned} \min_{s} \quad & \frac{1}{2} s^T Q s + p^T s \\ & y^T s = 0, \\ & o_i \leq s_i \leq u_i, i = 1, \ldots, l, \end{aligned} \tag{36}$$

where $y_i = \pm 1, u_i \geq 0,$ *and* $o_i \leq 0, i = 1, \ldots, l.$ *Let* $\min(eig(\cdot))$ *be the smallest eigenvalue of a matrix. If there is a* $\sigma > 0$ *such that* $\min(eig(Q)) > \sigma$*, then at an optimal solution* s *of (36),*

$$\frac{1}{2}s^T Q s + p^T s \leq -\frac{\sigma}{2}\|s\|^2.$$

Proof: From the KKT condition of (36), if s is an optimal solution, there is a b such that

$$\begin{aligned} (Qs)_i + p_i + by_i &= 0 && \text{if } o_i < s_i < u_i, \\ (Qs)_i + p_i + by_i &\geq 0 && \text{if } s_i = o_i, \\ (Qs)_i + p_i + by_i &\leq 0 && \text{if } s_i = u_i. \end{aligned}$$

Since $o_i \leq 0$ and $u_i \geq 0$,

$$s^T Q s + p^T s + b y^T s \leq 0.$$

With $y^T s = 0$,

$$\frac{1}{2}s^T Q s + p^T s \leq -\frac{1}{2}s^T Q s \leq -\frac{\sigma}{2}\|s\|^2.$$

■

From now on we consider any convergent subsequence $\{\alpha^k\}, k \in \mathcal{K}$ and $\lim_{k\to\infty, k\in\mathcal{K}} \alpha^k = \bar{\alpha}$.

To use the above lemma, we make the following assumption:

Assumption IV.2 *The kernel matrix* K *satisfies*

$$K_{ii} > 0, i = 1, \ldots, l.$$

Thus we can define

$$\sigma \equiv \min_{i=1,\ldots,l} K_{ii}.$$

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 $\bar{\alpha}$.*

Proof: If the ith group is selected as the working set, we have

$$f(\alpha) - f(\alpha^k) = \frac{1}{2}K_{ii}d^Td + p^Td,$$

where

$$d = [\alpha_i^1 - (\alpha^k)_i^1, \ldots, \alpha_i^r - (\alpha^k)_i^r]^T = \alpha_i - (\alpha^k)_i \text{ and } p = e + \sum_{j=1}^{l} K_{ij}(\alpha^k)_j.$$

Hence an equivalent form of the sub-problem (41) is

$$\begin{aligned} \min_d \quad & \frac{1}{2}K_{ii}d^Td + p^Td \\ & e^Td = 0, \\ & d^m \leq C_{\bar{y}_i}^m - (\alpha^k)_i^m, m = 1, \ldots, r, \end{aligned} \tag{37}$$

where d is the variable. Since α^k is a feasible point of (33), $C_{\bar{y}_i}^m - (\alpha^k)_i^m \geq 0$.

As the smallest eigenvalue of the Hessian of (37) is K_{ii} and (37) is in the form of (36), from Assumption IV.2 and Lemma IV.1, we have

$$f(\alpha^{k+1}) - f(\alpha^k) \leq -\frac{\sigma}{2}\|\alpha^{k+1} - \alpha^k\|^2. \tag{38}$$

Next we show that $\{f(\alpha^k)\}$ is a convergent sequence. First we know that $\{f(\alpha^k)\}$ is decreasing. From (33a) and (33b), we actually have

$$\begin{aligned} 0 \leq \alpha_i^m \leq C, & \text{ if } \bar{y}_i = m, \\ -C \leq \alpha_i^m \leq 0, & \text{ if } \bar{y}_i \neq m. \end{aligned} \tag{39}$$

Hence the feasible region of (33) is a compact set so $\lim_{k\to\infty} f(\alpha^k)$ exists and

$$\lim_{k\to\infty} f(\alpha^k) - f(\alpha^{k+1}) = 0. \tag{40}$$

Then for the subsequence $\{\alpha^{k+1}\}, k \in \mathcal{K}$, from (38) and (40) we have

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

Thus

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

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

We then need a technical lemma:

Lemma IV.4 *If $m(\bar{\alpha})_i > M(\bar{\alpha})_i$, then after $k \in \mathcal{K}$ is large enough, $(\alpha^k)_i^m, m = 1, \ldots, r$ are not changed.*

Proof: Assume $(\alpha^k)_i, k \in \mathcal{K}$ is selected and changed infinitely many times. At any of these $(\alpha^k)_i$, we solve the sub-problem (35) to obtain $(\alpha^{k+1})_i$. Now consider the following problem with variable α_i:

$$\begin{aligned} \min_{\alpha_i} \quad & \frac{1}{2} K_{ii} \alpha_i^T \alpha_i + (e + \sum_{j\neq i} K_{ij}(\bar{\alpha}_j))^T \alpha_i \\ & e^T \alpha_i = 0, \\ & \alpha_i^m \leq C_{\bar{y}_i}^m, m = 1, \ldots, r. \end{aligned} \tag{41}$$

Since $m(\bar{\alpha})_i > M(\bar{\alpha})_i$, $\bar{\alpha}_i$ is not an optimal solution of (41). Assume an optimal solution of (41) is $\tilde{\alpha}_i$.

Since α_i^{k+1} is an optimal solution of (35), we have

$$\begin{aligned} & \frac{1}{2} K_{ii} (\alpha^{k+1})_i^T (\alpha^{k+1})_i + (e + \sum_{j\neq i} K_{ij}(\alpha^k)_j)^T (\alpha^{k+1})_i \\ \leq \; & \frac{1}{2} K_{ii} \tilde{\alpha}_i^T \tilde{\alpha}_i + (e + \sum_{j\neq i} K_{ij}(\alpha^k)_j)^T \tilde{\alpha}_i. \end{aligned}$$

As $k \in \mathcal{K}$ goes to infinity, from Lemma IV.3,

$$\frac{1}{2} K_{ii} \bar{\alpha}_i^T \bar{\alpha}_i + (e + \sum_{j\neq i} K_{ij}\bar{\alpha}_j)^T \bar{\alpha}_i \leq \frac{1}{2} K_{ii} \tilde{\alpha}_i^T \tilde{\alpha}_i + (e + \sum_{j\neq i} K_{ij}\bar{\alpha}_j)^T \tilde{\alpha}_i$$

contradicts that $\bar{\alpha}_i$ is not an optimal solution of (41). Therefore, $(\alpha^k)_i$ is not selected after k is large enough. Hence $(\alpha^k)_i$ remains the same. ■

The main result on the convergence is the following:

Theorem IV.5 *Any limit point of convergent subsequences of $\{\alpha^k\}$ is a global minimum of (33).*

Proof: Using (39) we know the feasible region of (33) is compact. Hence $\{\alpha^k\}$ has convergent subsequences. Assume $\bar{\alpha}$ is the limit point of one convergent subsequence $\{\alpha^k\}, k \in \mathcal{K}$. If $\bar{\alpha}$ is not an optimal solution of (33), from the KKT condition (30), there are some groups j such that

$$m(\bar{\alpha})_j > M(\bar{\alpha})_j.$$

We define

$$\delta \equiv \min\{m(\bar{\alpha})_j - M(\bar{\alpha})_j \mid m(\bar{\alpha})_j - M(\bar{\alpha})_j > 0\}. \tag{42}$$

Using Lemma IV.4 and the continuity of $\nabla f(\alpha)$, we consider all $k \in \mathcal{K}$ large enough such that those $(\alpha^k)_j$ satisfying $m(\bar{\alpha})_j > M(\bar{\alpha})_j$ are not changed any more,

$$|\nabla f(\alpha^{k_1})_i^m - \nabla f(\bar{\alpha})_i^m| < \delta/8, \forall i = 1, \ldots, l, m = 1, \ldots, r, k \leq k_1 \leq k + l, \text{ and} \tag{43}$$

$$|\nabla f(\alpha^{k_1})_i^m - \nabla f(\alpha^{k_2})_i^m| < \delta/(8l), \forall i = 1, \ldots, l, m = 1, \ldots, r, k \leq k_1, k_2 \leq k + l. \tag{44}$$

Consider the kth iteration where a group i_1 is selected and modified. Thus

$$m(\alpha^k)_{i_1} > M(\alpha^k)_{i_1} \text{ but } m(\alpha^{k+1})_{i_1} \leq M(\alpha^{k+1})_{i_1}. \tag{45}$$

Using (34) we assume that

$$-\nabla f(\alpha^{k+1})_{i_1}^{m_1} = m(\alpha^{k+1})_{i_1}. \tag{46}$$

Then at the $(k+1)$st iteration if $i_2 \neq i_1$ is selected, then $(\alpha^{k+1})_{i_1} = (\alpha^{k+2})_{i_1}$ is not changed. Thus if

$$-\nabla f(\alpha^{k+2})_{i_1}^{m_2} = m(\alpha^{k+2})_{i_1}, \tag{47}$$

using (44), (46), (47), and the definition of $m(\alpha)$, we have

$$\begin{aligned}
-\nabla f(\alpha^{k+1})_{i_1}^{m_1} - \delta/(8l) &\leq -\nabla f(\alpha^{k+2})_{i_1}^{m_1} \\
&\leq -\nabla f(\alpha^{k+2})_{i_1}^{m_2} \\
&\leq -\nabla f(\alpha^{k+1})_{i_1}^{m_2} + \delta/(8l) \\
&\leq -\nabla f(\alpha^{k+1})_{i_1}^{m_1} + \delta/(8l).
\end{aligned}$$

Thus,

$$\begin{aligned} & |-\nabla f(\alpha^{k+1})^{m_1}_{i_1} + \nabla f(\alpha^{k+2})^{m_2}_{i_1}| \\ = \;& |m(\alpha^{k+1})_{i_1} - m(\alpha^{k+2})_{i_1}| \leq \delta/(8l). \end{aligned} \quad (48)$$

Similarly,

$$|M(\alpha^{k+1})_{i_1} - M(\alpha^{k+2})_{i_1}| \leq \delta/(8l),$$

so with (45),

$$m(\alpha^{k+2})_{i_1} \leq M(\alpha^{k+2})_{i_1} + \delta/(4l). \quad (49)$$

Thus in the next l iterations, no matter i_1 is selected or not, we have

$$m(\alpha^{k+1})_{i_1} \leq M(\alpha^{k+1})_{i_1} + \delta/4, \ldots, m(\alpha^{k+l})_{i_1} \leq M(\alpha^{k+l})_{i_1} + \delta/4. \quad (50)$$

On the other hand, from (42), (43) and the fact that

$$\alpha^k_j = \cdots = \alpha^{k+l}_j = \bar{\alpha}_j, \quad (51)$$

some group j with $m(\bar{\alpha})_j > M(\bar{\alpha})_j$ satisfies

$$\begin{aligned} m(\alpha^{k+1})_j - M(\alpha^{k+1})_j \geq m(\bar{\alpha})_j - M(\bar{\alpha})_j - \delta/4 \geq 3\delta/4, \ldots, \\ m(\alpha^{k+l})_j - M(\alpha^{k+l})_j \geq 3\delta/4. \end{aligned} \quad (52)$$

Note that the first inequality of (52) comes from a similar derivation of (49). For (49), in (48) the difference between $m(\alpha^{k+1})_{i_1}$ and $m(\alpha^{k+2})_{i_1}$ is estimated. For (52), since (51), we can directly measure the difference between $m(\alpha^{k+1})_j$ and $m(\bar{\alpha})_j$.

Hence (50) and (52) imply that i_1 should not be selected in the $(k+1), \ldots, (k+l)$th iterations. Therefore, as totally there are l groups of variables, in l iterations, some group j with $m(\bar{\alpha})_j > M(\bar{\alpha})_j$ must be selected. This contradicts Lemma IV.4 which shows that this jth group of variables should not be selected.

Therefore, any limit point of $\{\alpha^k\}$ is a KKT point of (33). As (33) is a convex optimization problem, any limit point is a global minimum. ■

If the kernel matrix K is positive definite, $K \otimes I$ is also positive definite. Then (33) is a strictly convex problem so there is a unique optimal solution. Thus $\{\alpha^k\}$ is a globally convergent sequence if K is positive definite.

V. Conclusions and Discussions

Originally we tried to prove Theorem II.1 by using as few conditions on the decomposition methods as possible. Surprisingly we finally need most properties of an existing working set selection. After finishing the convergence proof [8], an open question left is whether more flexible working set selections still lead to convergence. So far in many scenarios unfortunately properties of a systematic working selection are always needed. This seems to hint that proving more generalized convergence may not be an easy task.

Next we discuss the convergence proof for the formulation by Crammer and Singer. We can see that Assumption IV.2 is generally true. For example, for the polynomial kernel $K(x, y) = (x^T y)^d$, if all data are not zero vectors, $K_{ii} = (x_i^T x_i)^d > 0$. On the other hand, in [8] it assumes $\min_I(\min(\text{eig}(K_{II}))) > 0$, where I is any subset of $\{1, \ldots, l\}$ with $|I| \leq q$, K_{II} is a square sub-matrix of K, and $\min(\text{eig}(\cdot))$ is the smallest eigenvalue of a matrix. We have to consider any I as there are no restrictions on the working set. On the other hand, the reason why Assumption IV.2 is simpler is that now in each iteration one of the l groups is selected. Each group has r variables $\alpha_i^1, \ldots, \alpha_i^r$. Hence the square sub-matrix of $K \otimes I$ is reduced to a small diagonal matrix $K_{ii} \otimes I$. Then all eigenvalues of $K_{ii} \otimes I$ is K_{ii}.

The proof in this paper also shows the importance of Lemma IV.3. For both proofs here and in [8], as we are not able to prove the global convergence of $\{\alpha^k\}$, instead we prove that $\{\alpha^{k+1}\}, k \in \mathcal{K}$ converges if $\{\alpha^k\}, k \in \mathcal{K}$ is a convergent subsequence. Then this property is used to link several sub-problems in subsequent iterations.

References

[1] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[2] C.-C. Chang and C.-J. Lin. Training ν-support vector classifiers: Theory and algorithms. *Neural Computation*, (9), 2001. To appear.

[3] C. Cortes and V. Vapnik. Support-vector network. *Machine Learning*, 20:273–297, 1995.

[4] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. Technical report, School of Computer Science and Engineering, Hebrew University, 2001.

[5] C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, 2001.

[6] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, Cambridge, MA, 1998. MIT Press.

[7] S. S. Keerthi and E. G. Gilbert. Convergence of a generalized SMO algorithm for SVM classifier design. *Machine Learning*, 2001. To appear.

[8] C.-J. Lin. On the convergence of the decomposition method for support vector machines. *IEEE Transactions on Neural Networks*, 2001. To appear.

[9] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: An application to face detection. In *Proceedings of CVPR'97*, 1997.

[10] J. C. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, Cambridge, MA, 1998. MIT Press.

[11] B. Schölkopf, A. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207 – 1245, 2000.

[12] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, 1998.