Global Convergence of Decomposition Learning Methods for Support Vector Machines

Author(s): Norikazu Takahashi and Tetsuo Nishi

Journal: IEEE Transactions on Neural Networks
Volume: 17
Number: 6
Pages: 1362–1369
Month: November
Year: 2006
Published Version: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4012045

Global Convergence of Decomposition Learning Methods for Support Vector Machines

Norikazu Takahashi, *Member, IEEE* and Tetsuo Nishi, *Fellow, IEEE*

***Abstract*— Decomposition methods are well-known techniques for solving quadratic programming (QP) problems arising in support vector machines (SVMs). In each iteration of a decomposition method, a small number of variables are selected and a QP problem with only the selected variables is solved. Since large matrix computations are not required, decomposition methods are applicable to large QP problems. In this paper, we will make a rigorous analysis of the global convergence of general decomposition methods for SVMs. We first introduce a relaxed version of the optimality condition for the QP problems and then prove that a decomposition method reaches a solution satisfying this relaxed optimality condition within a finite number of iterations under a very mild condition on how to select variables.**

***Index Terms*— Support vector machines, quadratic programming, decomposition method, global convergence, termination**

I. Introduction

Support vector machines (SVMs) have recently attracted great attention in various fields such as pattern recognition, machine learning, neural networks, signal processing, and so on [1]–[4]. Given a set of l training samples, an SVM has to solve a quadratic programming (QP) problem with l variables. If the number of training samples are considerably large, the conventional QP solvers cannot be directly applied to SVM learning because large matrix computations are required. To overcome this difficulty, several techniques called decomposition methods have recently been proposed [5]–[8]. A basic strategy commonly used in decomposition methods is to execute two operations repeatedly until some optimality condition is satisfied; one is to select q variables among l and the other is to minimize the objective function by updating only the the selected q variables. The set of q variables selected for updating is called *the working set*. An example of decomposition methods is the sequential minimal optimization (SMO) algorithm proposed by Platt [5] in which only two variables are selected for the working set in each iteration. Another example is SVM$^{\mathrm{light}}$ [6], the most widely used learning algorithm for SVMs, in which q, the size of the working set, can be set to any even number. Since large matrix computations are not required, decomposition methods are useful for SVMs especially when a large number of training samples are given.

For any decomposition method, it is very important to guarantee that the sequence of solutions converges to an optimal solution within a finite number of iterations. Some theoretical results concerning the global convergence of decomposition methods can be found in the literature [9]–[16]. Keerthi and Gilbert [9] analyzed the behavior of the generalized SMO algorithm and gave a proof that the algorithm terminates within a finite number of iterations under a pre-specified stopping condition and tolerance[1]. Lin [10] proved that the sequence of solutions obtained by an SMO algorithm converges asymptotically to an optimal solution. Recently, Chen *et al.* [12] carried out a comprehensive study on SMO algorithms and gave several results on asymptotic convergence, finite termination, shrinking and caching, and convergence rate in a general setting on the working set selection. However, the results given in [9]–[12] rely heavily on the fact that only two variables are updated in each iteration, and therefore it seems difficult to extend to the general case where q is greater than two.

Chang *et al.* [13] considered a kind of decomposition method in which q is not restricted to two, and proved the convergence of the method. However, the working set selection in their method is formulated in a rather special form, and therefore the result does not hold for other decomposition methods. Lin [14] investigated the properties of the sequence of solutions obtained by SVM$^{\mathrm{light}}$ in detail and proved that any limit point of the sequence is an optimal solution. Lin [15] also showed that a class of decomposition methods including SVM$^{\mathrm{light}}$ stops within a finite number of iterations if a relaxed version of stopping criterion is used. These results are important because the convergence of SVM$^{\mathrm{light}}$ was guaranteed theoretically for the first time. However, on the other hand, these results require a stronger assumption on the Hessian matrix of the objective function than positive semi-definiteness [14, Assumption IV.1], except the special case where $q = 2$ [10]. It is known that this assumption is satisfied if the Gaussian kernel is used and training samples are different from each other, but, it is also known that this assumption does not hold true for any kernel function if training samples contain two or more identical data [10]. List and Simon [16] considered a general class of QP problems which includes the one arising in SVMs, and shown that a decomposition method converges to an optimal solution if its working set selection method satisfies three abstract conditions. This result is general in the sense that no specific working set selection method is

This work was supported in part by the 21st Century COE (Center of Excellence) Program "Reconstruction of Social Infrastructure Related to Information Science and Electrical Engineering".

N. Takahashi is with the Department of Computer Science and Communication Engineering, Kyushu University, Fukuoka, 812-8581 Japan (e-mail: norikazu@csce.kyushu-u.ac.jp).

T. Nishi was with the Department of Computer Science and Communication Engineering, Kyushu University, Fukuoka, 812-8581 Japan. He is now with the Faculty of Science and Engineering, Waseda University, Tokyo, 169-0072 Japan (e-mail: nishi-t@waseda.jp).

[1]The authors of the present paper have recently pointed out the incompleteness of their proof and given a more rigorous proof [11].

assumed, but still requires the same assumption on the Hessian matrix of the objective function as [14] and [15].

In this paper, we will present a new convergence proof for general decomposition methods for SVMs. Unlike the existing results mentioned above, we will neither restrict ourselves to a specific working set selection nor assume any condition on the Hessian matrix of the objective function except for positive semi-definiteness. Instead, we will employ a relaxed version of the Karush-Kuhn-Tucker (KKT) conditions as the optimality condition. The relaxed version contains two positive parameters τ and ϵ, and approaches the strict KKT conditions as both τ and ϵ go to zero. By using the global convergence theorem in optimization theory [17], we will prove that the decomposition method stops within a finite number of iterations after finding an optimal solution for any τ and ϵ if the working set contains at least one pair of indices violating the optimality condition in each iteration. This condition on the working set selection is very mild and thus can be applied to many decomposition methods.

In our convergence proof, closedness of a point-to-set map [17] plays a central role. It is important to note that the point-to-set map defined for the decomposition method considered in this paper is not closed if either the strict KKT condition or the relaxed version which has been often used so far [6], [9], [11], [15] is employed as the optimality condition. Indeed this is the main difficulty on proving the convergence in earlier works. An extensive discussion is given in [14].

This paper is organized as follows. In Section II, the global convergence theorem for general optimization problems and some related results are reviewed for later discussions. In Section III, the QP problem arising in SVMs, two types of optimality conditions, and the algorithm of general decomposition methods are explained. In Section IV, the relationship between two optimality conditions is firstly discussed and then convergence theorems for decomposition methods are proved. In Section V, these convergence theorems are applied to some well-known learning algorithms for SVMs. Finally, concluding remarks are given in Section VI.

II. Preliminaries

First of all, we will review a fundamental result in optimization theory known as the global convergence theorem [17]. This theorem plays a central role in our global convergence analysis of decomposition methods.

Let us consider the following optimization problem.

Problem 1: Find $\boldsymbol{x}$ which minimizes the objective function $f(\boldsymbol{x})$ under the constraint $\boldsymbol{x} \in X$.

An algorithm for solving Problem 1 can be viewed as an iterative process that generates a sequence $\{\boldsymbol{x}(k)\}_{k=0}^{\infty}$ by

$$\boldsymbol{x}(k+1) \in \boldsymbol{A}(\boldsymbol{x}(k)), \quad k = 0, 1, \ldots \tag{1}$$

where $\boldsymbol{x}(0) \in X$ is a given initial point and $\boldsymbol{A}$ is a point-to-set map that assigns to each point in the domain X a subset of X. According to this definition, we can state that developing an algorithm for solving Problem 1 is equivalent to determining the point-to-set map $\boldsymbol{A}$. Apparently, the most desirable property of the map $\boldsymbol{A}$ is that the sequence $\{\boldsymbol{x}(k)\}_{k=0}^{\infty}$ generated by (1) converges to an optimal solution of Problem 1 for any initial point $\boldsymbol{x}(0) \in X$.

Definition 1: Let X, Ω and $\boldsymbol{A}$ be a nonempty closed set in $\mathbb{R}^n$, a subset of X, and a point-to-set map from X to X, respectively. If a continuous function Z defined in X satisfies

$$\text{if } \boldsymbol{x} \notin \Omega \text{ and } \boldsymbol{y} \in \boldsymbol{A}(\boldsymbol{x}) \text{ then } Z(\boldsymbol{y}) < Z(\boldsymbol{x})$$

then Z is called a *descent function* for Ω and $\boldsymbol{A}$.

Definition 2: Let X and Y be nonempty closed sets in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. Let $\boldsymbol{A}$ be a point-to-set map from X to Y. The map $\boldsymbol{A}$ is said to be *closed* at $\bar{\boldsymbol{x}}$ if the two assumptions

1) $\boldsymbol{x}(k) \in X$, $\forall k$ and $\lim_{k\to\infty} \boldsymbol{x}(k) = \bar{\boldsymbol{x}}$
2) $\boldsymbol{y}(k) \in \boldsymbol{A}(\boldsymbol{x}(k))$, $\forall k$ and $\lim_{k\to\infty} \boldsymbol{y}(k) = \bar{\boldsymbol{y}}$

imply that $\bar{\boldsymbol{y}} \in \boldsymbol{A}(\bar{\boldsymbol{x}})$. The map $\boldsymbol{A}$ is said to be closed on X if it is closed at each point in X.

In terms of these definitions, the global convergence theorem can be stated as follows.

Theorem 1 ([17]): Let X, Ω and $\boldsymbol{A}$ be a nonempty closed set in $\mathbb{R}^n$, a subset of X, and a point-to-set map from X to X, respectively. Let $\{\boldsymbol{x}(k)\}_{k=0}^{\infty}$ be a sequence generated by (1) with $\boldsymbol{x}(0) \in X$. Every convergent subsequence of $\{\boldsymbol{x}(k)\}_{k=0}^{\infty}$ has a limit in Ω if the following conditions are satisfied.

1) For all k, $\boldsymbol{x}(k)$ belongs to a compact set $S \subseteq X$.
2) There exists a descent function Z for Ω and $\boldsymbol{A}$.
3) The map $\boldsymbol{A}$ is closed on $X \setminus \Omega$.

Note that Ω in Theorem 1 can be any subset of X. In particular, if Ω is set to the set of optimal solutions of Problem 1, Theorem 1 gives a sufficient condition for the sequence $\{\boldsymbol{x}(k)\}_{k=0}^{\infty}$ generated by (1) to converge to an optimal solution. Theorem 1 is hence a very useful tool for proving the global convergence of an algorithm. However, it is difficult in general to develop a map $\boldsymbol{A}$ having the global convergence property for the set of optimal solutions. Thus Ω is often set to, for example, $\{\boldsymbol{x} \mid \boldsymbol{x} \text{ is a local optimal solution}\}$ and $\{\boldsymbol{x} \mid \boldsymbol{x} \in X,\ f(\boldsymbol{x}) \le b\}$ where b is a constant.

The following lemma can easily be obtained.

Lemma 1: Let $\boldsymbol{A} : X \to Y$ and $\boldsymbol{B} : X \to Y$ be point-to-set maps. If both $\boldsymbol{A}$ and $\boldsymbol{B}$ are closed at $\boldsymbol{x}$, then the map $\boldsymbol{C}(\boldsymbol{x}) = \boldsymbol{A}(\boldsymbol{x}) \cup \boldsymbol{B}(\boldsymbol{x})$ is also closed at $\boldsymbol{x}$.

Proof: Let $\{\boldsymbol{x}(k)\}_{k=0}^{\infty}$ be any sequence such that $\boldsymbol{x}(k) \in X$, $\forall k$ and $\lim_{k\to\infty} \boldsymbol{x}(k) = \boldsymbol{x}$. Let $\{\boldsymbol{y}(k)\}_{k=0}^{\infty}$ be any sequence such that $\boldsymbol{y}(k) \in \boldsymbol{C}(\boldsymbol{x}(k))$, $\forall k$ and $\lim_{k\to\infty} \boldsymbol{y}(k) = \boldsymbol{y}$. Then it is obvious that at least one of two statements:

1) $\boldsymbol{y}(k) \in \boldsymbol{A}(\boldsymbol{x}(k))$ for infinitely many k,
2) $\boldsymbol{y}(k) \in \boldsymbol{B}(\boldsymbol{x}(k))$ for infinitely many k

holds. We will assume without loss of generality that the first case holds true. Let the set of all k such that $\boldsymbol{y}(k) \in \boldsymbol{A}(\boldsymbol{x}(k))$ be denoted by K_A. Then the sequence $\{\boldsymbol{x}(k)\}_{k\in K_A}$ satisfies $\boldsymbol{x}(k) \in X$, $\forall k \in K_A$ and $\lim_{k\to\infty, k\in K_A} \boldsymbol{x}(k) = \boldsymbol{x}$. Also, the sequence $\{\boldsymbol{y}(k)\}_{k\in K_A}$ satisfies $\boldsymbol{y}(k) \in \boldsymbol{A}(\boldsymbol{x}(k))$, $\forall k \in K_A$ and $\lim_{k\to\infty, k\in K_A} \boldsymbol{y}(k) = \boldsymbol{y}$. Since the point-to-set map $\boldsymbol{A}$ is closed at $\boldsymbol{x}$, we have $\boldsymbol{y} \in \boldsymbol{A}(\boldsymbol{x})$ which implies that $\boldsymbol{y} \in \boldsymbol{C}(\boldsymbol{x})$. Therefore the point-to-set map $\boldsymbol{C}$ is closed at $\boldsymbol{x}$. ■

A result concerning the closedness of composite maps will also be needed in later discussions.

Definition 3: Let $\boldsymbol{A} : X \to Y$ and $\boldsymbol{B} : Y \to Z$ be point-to-set maps. The composite map $\boldsymbol{C} = \boldsymbol{B}\boldsymbol{A}$ is defined as the

point-to-set map $\boldsymbol{C} : X \to Z$ with

$$\boldsymbol{C}(\boldsymbol{x}) = \cup_{\boldsymbol{y} \in \boldsymbol{A}(\boldsymbol{x})} \boldsymbol{B}(\boldsymbol{y}) .$$

Lemma 2 ([17]): Let $\boldsymbol{A} : X \to Y$ and $\boldsymbol{B} : Y \to Z$ be point-to-set maps. If $\boldsymbol{A}$ is closed at $\boldsymbol{x}$, $\boldsymbol{B}$ is closed on $\boldsymbol{A}(\boldsymbol{x})$ and Y is compact, then the composite map $\boldsymbol{C} = \boldsymbol{B}\boldsymbol{A}$ is closed at $\boldsymbol{x}$.

III. Decomposition Method

A. SVM Dual Problem

Suppose that we are given a set of l training samples $\{(\boldsymbol{p}_i, d_i)\}_{i=1}^{l}$ where $\boldsymbol{p}_i \in \mathbb{R}^n$ is the i-th input pattern and $d_i \in \{1, -1\}$ represents the class to which $\boldsymbol{p}_i$ belongs. The learning of an SVM with the kernel function $K(\cdot, \cdot)$ leads to the following QP problem (for more details on formulation, see for example [1]).

Problem 2: Find $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_l]^T$ which minimizes the objective function

$$W(\boldsymbol{\alpha}) = \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} q_{ij} \alpha_i \alpha_j - \sum_{i=1}^{l} \alpha_i$$

under the constraints

$$\sum_{i=1}^{l} d_i \alpha_i = 0 \tag{2}$$

$$0 \le \alpha_i \le C, \quad i = 1, 2, \ldots, l \tag{3}$$

where $q_{ij} = d_i d_j K(\boldsymbol{p}_i, \boldsymbol{p}_j)$ and C is a user-specified positive constant.

Throughout this paper, we assume that the kernel function $K(\cdot, \cdot)$ satisfies Mercer's condition [1]. In this case, $\boldsymbol{Q} = [q_{ij}] \in \mathbb{R}^{l \times l}$ is a positive semi-definite matrix, and therefore Problem 2 is a convex QP problem. Note that the optimal solution of Problem 2 is not necessarily unique because $W(\boldsymbol{\alpha})$ is not strictly convex. The feasible region of Problem 2, that is, the set of $\boldsymbol{\alpha}$ satisfying (2) and (3), is denoted by S which is apparently a compact set. Also, the set $\{1, 2, \ldots, l\}$ is denoted by L.

B. Optimality Conditions

Since Problem 2 is a convex QP problem, optimal solutions are completely characterized by the KKT conditions [17], that is, $\boldsymbol{\alpha} \in S$ is an optimal solution of Problem 2 if and only if there exist constants λ, $\mu_1, \mu_2, \ldots, \mu_l$, $\nu_1, \nu_2, \ldots, \nu_l$ such that

$$\begin{cases} \partial W(\boldsymbol{\alpha})/\partial \alpha_i + \lambda d_i - \mu_i + \nu_i = 0 \\ \mu_i \alpha_i = 0 \\ \nu_i (\alpha_i - C) = 0 \\ \mu_i \ge 0 \\ \nu_i \ge 0 \end{cases} \tag{4}$$

for all $i \in L$. As shown in [9] and [15], this condition can be rewritten in a more compact form as

$$\min_{i \in I_{\mathrm{up}}(\boldsymbol{\alpha})} F_i(\boldsymbol{\alpha}) \ge \max_{i \in I_{\mathrm{low}}(\boldsymbol{\alpha})} F_i(\boldsymbol{\alpha}) \tag{5}$$

where

$$F_i(\boldsymbol{\alpha}) = d_i \left(\sum_{j=1}^{l} q_{ij} \alpha_j - 1 \right),$$

$$I_{\mathrm{up}}(\boldsymbol{\alpha}) = \{ i \,|\, \alpha_i < C,\, d_i = 1 \} \cup \{ i \,|\, \alpha_i > 0,\, d_i = -1 \} ,$$

$$I_{\mathrm{low}}(\boldsymbol{\alpha}) = \{ i \,|\, \alpha_i < C,\, d_i = -1 \} \cup \{ i \,|\, \alpha_i > 0,\, d_i = 1 \} .$$

Let Ω^* denote the set of optimal solutions of Problem 2. Then, by using the above notations, we can express Ω^* as follows:

$$\Omega^* = \{ \boldsymbol{\alpha} \in S \,|\, \min_{i \in I_{\mathrm{up}}(\boldsymbol{\alpha})} F_i(\boldsymbol{\alpha}) \ge \max_{i \in I_{\mathrm{low}}(\boldsymbol{\alpha})} F_i(\boldsymbol{\alpha}) \} .$$

In a practical situation, the optimality condition (5) is often relaxed as

$$\min_{i \in I_{\mathrm{up}}(\boldsymbol{\alpha})} F_i(\boldsymbol{\alpha}) \ge \max_{i \in I_{\mathrm{low}}(\boldsymbol{\alpha})} F_i(\boldsymbol{\alpha}) - \tau \tag{6}$$

where τ is a positive constant [6], [7], [15].

In this paper, on the other hand, we employ neither (5) nor (6) but the inequality:

$$\min_{i \in I^{\epsilon}_{\mathrm{up}}(\boldsymbol{\alpha})} F_i(\boldsymbol{\alpha}) > \max_{i \in I^{\epsilon}_{\mathrm{low}}(\boldsymbol{\alpha})} F_i(\boldsymbol{\alpha}) - \tau \tag{7}$$

for the optimality condition, where

$$I^{\epsilon}_{\mathrm{up}}(\boldsymbol{\alpha}) = \{ i \,|\, \alpha_i \le C - \epsilon, d_i = 1 \} \cup \{ i \,|\, \alpha_i \ge \epsilon, d_i = -1 \} ,$$

$$I^{\epsilon}_{\mathrm{low}}(\boldsymbol{\alpha}) = \{ i \,|\, \alpha_i \le C - \epsilon, d_i = -1 \} \cup \{ i \,|\, \alpha_i \ge \epsilon, d_i = 1 \} ,$$

and ϵ is any positive constant smaller than $C/2$. Usually ϵ is set to a sufficiently small positive number. The role of $I^{\epsilon}_{\mathrm{up}}(\boldsymbol{\alpha})$ and $I^{\epsilon}_{\mathrm{low}}(\boldsymbol{\alpha})$ is to consider α_i which is sufficiently close to 0 (C, resp.) to be exactly 0 (C, resp.). This is a technique used in the implementation of SVM$^{\mathrm{light}}$ algorithm [6] (see the source code[2] of SVM$^{\mathrm{light}}$ developed by Joachims). In the following, any $\boldsymbol{\alpha} \in S$ satisfying (7) is said to be a (τ, ϵ)-optimal solution. The set of (τ, ϵ)-optimal solutions is denoted by $\Omega^{(\tau,\epsilon)}$, that is,

$$\Omega^{(\tau,\epsilon)} = \{ \boldsymbol{\alpha} \in S \,|\, \min_{i \in I^{\epsilon}_{\mathrm{up}}(\boldsymbol{\alpha})} F_i(\boldsymbol{\alpha}) > \max_{i \in I^{\epsilon}_{\mathrm{low}}(\boldsymbol{\alpha})} F_i(\boldsymbol{\alpha}) - \tau \} .$$

Also, a pair of indices (i, j) such that

$$i \in I^{\epsilon}_{\mathrm{up}}(\boldsymbol{\alpha}),\ j \in I^{\epsilon}_{\mathrm{low}}(\boldsymbol{\alpha}),\ F_i(\boldsymbol{\alpha}) \le F_j(\boldsymbol{\alpha}) - \tau$$

is called a (τ, ϵ)-violating pair at $\boldsymbol{\alpha}$. If a pair (i, j) is not a (τ, ϵ)-violating pair at $\boldsymbol{\alpha}$, the pair is called a (τ, ϵ)-feasible pair at $\boldsymbol{\alpha}$. It is obvious from these definitions that $\boldsymbol{\alpha} \in S$ is a (τ, ϵ)-optimal solution if and only if there is no (τ, ϵ)-violating pair at $\boldsymbol{\alpha}$.

Properties of the two sets Ω^* and $\Omega^{(\tau,\epsilon)}$ as well as the relationship between them will be studied in detail in the next section.

C. Algorithm of Decomposition Method

A basic strategy commonly used in all decomposition methods is to repeat two operations until some optimality condition is satisfied; one is to select q variables among l for the working set and the other is to minimize the objective function $W(\boldsymbol{\alpha})$ by updating only the selected q variables. This is formally expressed as follows:

[2]Source code and binaries of SVM$^{\mathrm{light}}$ are available at http://svmlight.joachims.org/.

Algorithm 1: Given training samples $\{(\boldsymbol{p}_i, d_i)\}_{i=1}^{l}$, a kernel function $K(\cdot,\cdot)$, a positive constant C and an integer $q(\leq l)$, execute the following procedures.

1) Let $\boldsymbol{\alpha}(0) = \mathbf{0}$ and $k = 0$.
2) If $\boldsymbol{\alpha} = \boldsymbol{\alpha}(k)$ satisfies the optimality condition (7) then stop. Otherwise go to Step 3).
3) Select the working set $L_B(k) \subseteq L = \{1, 2, \ldots, l\}$ where $|L_B(k)| \leq q$.
4) Find $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_l]^T$ which minimizes the objective function $W(\boldsymbol{\alpha})$ under the constraints (2), (3) and $\alpha_i = \alpha_i(k),\ \forall i \in L_N(k) = L \setminus L_B(k)$.
5) Set $\boldsymbol{\alpha}(k+1)$ to an optimal solution of the optimization problem in Step 4). Add 1 to k, and go to Step 2).

We will refer to the optimization problem in Step 4) as the subproblem in the following. Since the subproblem has at most q variables, the amount of memory required for Algorithm 1 is linear in l, while l^2 elements of the matrix $\boldsymbol{Q}$ must be stored in memory if Problem 2 is solved at a time. This is the main advantage of decomposition methods. As for computation time, each subproblem can be solved faster as q decreases. In particular, in the case where $q = 2$, subproblems can be solved analytically and hence Algorithm 1 can be implemented without QP solvers [5]. However, it should be noted that the number of iterations increases as q decreases in general.

It is apparent that the sequence $\{\boldsymbol{\alpha}(k)\}_{k=1}^{\infty}$ generated by Algorithm 1 satisfies two conditions:

$$\boldsymbol{\alpha}(k) \in S \tag{8}$$

$$W(\boldsymbol{\alpha}(k+1)) \leq W(\boldsymbol{\alpha}(k)) \tag{9}$$

for all k. Since the objective function $W(\cdot)$ is bounded from below in S, Eq.(9) implies that the sequence $\{W(\boldsymbol{\alpha}(k))\}_{k=0}^{\infty}$ necessarily converges to a certain value. However, on the other hand, it is not clear whether the sequence $\{\boldsymbol{\alpha}(k)\}_{k=0}^{\infty}$ converges to $\Omega^{(\tau,\epsilon)}$ or not.

The optimality condition for the subproblem is expressed in the same form as (5), that is, for given $L_B(k) \subseteq L$ and $\boldsymbol{\alpha}(k) \in S$, $\boldsymbol{\alpha} \in S$ is an optimal solution of the subproblem if and only if the following conditions are satisfied.

$$\begin{cases} \min\limits_{i \in I_{up}(\boldsymbol{\alpha}) \cap L_B(k)} F_i(\boldsymbol{\alpha}) \geq \max\limits_{i \in I_{low}(\boldsymbol{\alpha}) \cap L_B(k)} F_i(\boldsymbol{\alpha}) \\ \alpha_i = \alpha_i(k), \quad \forall i \in L_N(k) = L \setminus L_B(k) \end{cases}$$

IV. Global Convergence Analysis of Decomposition Methods

A. Properties of Ω^* and $\Omega^{(\tau,\epsilon)}$

Before proceeding to the convergence analysis of Algorithm 1, we study properties of Ω^* and $\Omega^{(\tau,\epsilon)}$.

Lemma 3: $I_{\rm up}(\boldsymbol{\alpha}) \supseteq I_{\rm up}^{\epsilon}(\boldsymbol{\alpha})$ and $I_{\rm low}(\boldsymbol{\alpha}) \supseteq I_{\rm low}^{\epsilon}(\boldsymbol{\alpha})$ for any $\boldsymbol{\alpha} \in S$ and $\epsilon \in (0, C/2)$.

Proof: We will prove only the first formula. The second one can be proved similarly. Let i be any member of $I_{\rm up}^{\epsilon}(\boldsymbol{\alpha})$. Then α_i satisfies $0 \leq \alpha_i < C - \epsilon$ if $d_i = 1$ and $\epsilon < \alpha_i \leq C$ if $d_i = -1$. In the former case, i belongs to $I_{\rm up}(\boldsymbol{\alpha})$ because either $0 < \alpha_i < C$ or $\alpha_i = 0$ holds. In the latter case, i belongs to $I_{\rm up}(\boldsymbol{\alpha})$ because either $0 < \alpha_i < C$ or $\alpha_i = C$ holds. Therefore, any member of $I_{\rm up}^{\epsilon}(\boldsymbol{\alpha})$ belongs to $I_{\rm up}(\boldsymbol{\alpha})$. This implies $I_{\rm up}(\boldsymbol{\alpha}) \supseteq I_{\rm up}^{\epsilon}(\boldsymbol{\alpha})$. ■

Proposition 1: $\Omega^{(\tau,\epsilon)} \supseteq \Omega^*$ for any $\tau > 0$ and $\epsilon \in (0, C/2)$.

Proof: Let $\boldsymbol{\alpha}$ be any point in Ω^*. Then $\boldsymbol{\alpha}$ satisfies (5). It follows from Lemma 3 that

$$\min_{i \in I_{\rm up}^{\epsilon}(\boldsymbol{\alpha})} F_i(\boldsymbol{\alpha}) \geq \min_{i \in I_{\rm up}(\boldsymbol{\alpha})} F_i(\boldsymbol{\alpha}),$$

$$\max_{i \in I_{\rm low}(\boldsymbol{\alpha})} F_i(\boldsymbol{\alpha}) \geq \max_{i \in I_{\rm low}^{\epsilon}(\boldsymbol{\alpha})} F_i(\boldsymbol{\alpha}).$$

From these two inequalities and (5), we have

$$\min_{i \in I_{\rm up}^{\epsilon}(\boldsymbol{\alpha})} F_i(\boldsymbol{\alpha}) \geq \max_{i \in I_{\rm low}^{\epsilon}(\boldsymbol{\alpha})} F_i(\boldsymbol{\alpha})$$

which implies $\boldsymbol{\alpha} \in \Omega^{(\tau,\epsilon)}$. ■

Lemma 4: Let $\{\boldsymbol{\alpha}(n)\}_{n=0}^{\infty}$ be any sequence such that $\boldsymbol{\alpha}(n) \in S, \forall n$ and $\lim_{n\to\infty} \boldsymbol{\alpha}(n) = \bar{\boldsymbol{\alpha}}$. Then there exist positive integers n_1 and n_2 such that

$$I_{\rm up}(\boldsymbol{\alpha}(n)) \supseteq I_{\rm up}(\bar{\boldsymbol{\alpha}}),\ \forall n \geq n_1$$

$$I_{\rm low}(\boldsymbol{\alpha}(n)) \supseteq I_{\rm low}(\bar{\boldsymbol{\alpha}}),\ \forall n \geq n_2$$

Proof: We will prove only the first formula. The second one can be proved in the same way. Let i be any member of $I_{\rm up}(\bar{\boldsymbol{\alpha}})$. Then $\bar{\alpha}_i$ satisfies $\bar{\alpha}_i < C$ if $d_i = 1$ and $\bar{\alpha}_i > 0$ if $d_i = -1$. In the former case, since $\alpha_i(n)$ converges to $\bar{\alpha}_i$, there exists a positive integer $n_1(i)$ such that $\alpha_i(n) < C, \forall n \geq n_1(i)$ which implies $i \in I_{\rm up}(\boldsymbol{\alpha}(n)), \forall n \geq n_1(i)$. In the latter case, it is shown in the same way that there exists a positive integer $n_1(i)$ such that $i \in I_{\rm up}(\boldsymbol{\alpha}(n)), \forall n \geq n_1(i)$. Let $n_1 = \max_{i \in I_{\rm up}(\bar{\boldsymbol{\alpha}})} n_1(i)$. Then all members of $I_{\rm up}(\bar{\alpha})$ belong to $I_{\rm up}(\boldsymbol{\alpha}(n)), \forall n \geq n_1$. This completes the proof. ■

This lemma was first given by Lin [14, Lemma IV.4]. We have just restated the result in terms of our notations, and given a proof for the sake of the reader's convenience.

Proposition 2: The set Ω^* is closed.

Proof: Let $\{\boldsymbol{\alpha}(n)\}_{n=1}^{\infty}$ be any sequence such that $\boldsymbol{\alpha}(n) \in \Omega^*, \forall n$ and $\lim_{n\to\infty} \boldsymbol{\alpha}(n) = \bar{\boldsymbol{\alpha}}$. It suffices for us to show that $\bar{\boldsymbol{\alpha}} \in \Omega^*$. Since $\boldsymbol{\alpha}(n) \in \Omega^*, \forall n$, we have

$$\min_{i \in I_{\rm up}(\boldsymbol{\alpha}(n))} F_i(\boldsymbol{\alpha}(n)) \geq \max_{i \in I_{\rm low}(\boldsymbol{\alpha}(n))} F_i(\boldsymbol{\alpha}(n)),\ \forall n.$$

It follows from this inequality and Lemma 4 that there exists a positive integer n_1 such that

$$\min_{i \in I_{\rm up}(\bar{\boldsymbol{\alpha}})} F_i(\boldsymbol{\alpha}(n)) \geq \max_{i \in I_{\rm low}(\bar{\boldsymbol{\alpha}})} F_i(\boldsymbol{\alpha}(n)),\ \forall n \geq n_1.$$

By letting n go to infinity, we have

$$\min_{i \in I_{\rm up}(\bar{\boldsymbol{\alpha}})} F_i(\bar{\boldsymbol{\alpha}}) \geq \max_{i \in I_{\rm low}(\bar{\boldsymbol{\alpha}})} F_i(\bar{\boldsymbol{\alpha}})$$

which implies $\bar{\boldsymbol{\alpha}} \in \Omega^*$. Therefore, Ω^* is a closed set. ■

Lemma 5: Let $\{\boldsymbol{\alpha}(n)\}_{n=0}^{\infty}$ be any sequence such that $\boldsymbol{\alpha}(n) \in S, \forall n$ and $\lim_{n\to\infty} \boldsymbol{\alpha}(n) = \bar{\boldsymbol{\alpha}}$. Then there exist positive integers n_1 and n_2 such that

$$I_{\rm up}^{\epsilon}(\boldsymbol{\alpha}(n)) \subseteq I_{\rm up}^{\epsilon}(\bar{\boldsymbol{\alpha}}),\ \forall n \geq n_1$$

$$I_{\rm low}^{\epsilon}(\boldsymbol{\alpha}(n)) \subseteq I_{\rm low}^{\epsilon}(\bar{\boldsymbol{\alpha}}),\ \forall n \geq n_2$$

for any $\epsilon \in (0, C/2)$.

Proof: We will prove only the first formula. The second one can be proved similarly. Let i be any nonmember of $I_{\rm up}^{\epsilon}(\bar{\boldsymbol{\alpha}})$. Then $\bar{\alpha}_i$ satisfies $\bar{\alpha}_i > C - \epsilon$ if $d_i = 1$ and $\bar{\alpha}_i < \epsilon$

if $d_i = -1$. In the former case, since $\alpha_i(n)$ converges to $\bar{\alpha}_i$, there exists a positive integer $n_1(i)$ such that $\alpha_i(n) > C - \epsilon$, $\forall n \geq n_1(i)$ which implies $i \notin I^{\epsilon}_{\rm up}(\boldsymbol{\alpha}(n))$, $\forall n \geq n_1(i)$. In the latter case, it is shown in the same way that there exists a positive integer $n_1(i)$ such that $i \notin I^{\epsilon}_{\rm up}(\boldsymbol{\alpha}(n))$, $\forall n \geq n_1(i)$. Let $n_1 = \max_{i \notin I^{\epsilon}_{\rm up}(\bar{\boldsymbol{\alpha}})} n_1(i)$. Then all nonmembers of $I_{\rm up}(\bar{\alpha})$ do not belong to $I^{\epsilon}_{\rm up}(\boldsymbol{\alpha}(n))$, $\forall n \geq n_1$. This is equivalent to the first formula. ■

Proposition 3: The set $S \setminus \Omega^{(\tau,\epsilon)}$ is closed for any $\tau > 0$ and $\epsilon \in (0, C/2)$.

Proof: Let $\{\boldsymbol{\alpha}(n)\}_{n=1}^{\infty}$ be any sequence such that $\boldsymbol{\alpha}(n) \in S \setminus \Omega^{(\tau,\epsilon)}$, $\forall n$ and $\lim_{n\to\infty} \boldsymbol{\alpha}(n) = \bar{\boldsymbol{\alpha}}$. Then we have

$$\min_{i \in I^{\epsilon}_{\rm up}(\boldsymbol{\alpha}(n))} F_i(\boldsymbol{\alpha}(n)) \leq \max_{i \in I^{\epsilon}_{\rm low}(\boldsymbol{\alpha}(n))} F_i(\boldsymbol{\alpha}(n)) - \tau, \ \forall n.$$

It follows from this inequality and Lemma 5 that there exists a positive integer n_1 such that

$$\min_{i \in I^{\epsilon}_{\rm up}(\bar{\boldsymbol{\alpha}})} F_i(\boldsymbol{\alpha}(n)) \leq \max_{i \in I^{\epsilon}_{\rm low}(\bar{\boldsymbol{\alpha}})} F_i(\boldsymbol{\alpha}(n)) - \tau, \ \forall n \geq n_1.$$

By letting n go to infinity in both sides, we have

$$\min_{i \in I^{\epsilon}_{\rm up}(\bar{\boldsymbol{\alpha}})} F_i(\bar{\boldsymbol{\alpha}}) \leq \max_{i \in I^{\epsilon}_{\rm low}(\bar{\boldsymbol{\alpha}})} F_i(\bar{\boldsymbol{\alpha}}) - \tau$$

which means $\bar{\boldsymbol{\alpha}} \in S \setminus \Omega^{(\tau,\epsilon)}$. Therefore, $\Omega^{(\tau,\epsilon)}$ is a closed set for any $\tau > 0$ and $\epsilon \in (0, C/2)$. ■

Proposition 4: The set $\Omega^{(\tau,\epsilon)}$ converges to Ω^* as the positive constants τ and ϵ approach 0.

Proof: One can easily see that $\lim_{\epsilon\to 0+} I^{\epsilon}_{\rm up}(\boldsymbol{\alpha}) = I_{\rm up}(\boldsymbol{\alpha})$ and $\lim_{\epsilon\to 0+} I^{\epsilon}_{\rm low}(\boldsymbol{\alpha}) = I_{\rm low}(\boldsymbol{\alpha})$, where $\epsilon \to 0+$ means ϵ approaches 0 from right. Thus we have

$$\lim_{\epsilon\to 0+} \Omega^{(\tau,\epsilon)} = \{\boldsymbol{\alpha} \in S \mid \min_{i \in I_{up}(\boldsymbol{\alpha})} F_i(\boldsymbol{\alpha}) > \max_{i \in I_{low}(\boldsymbol{\alpha})} F_i(\boldsymbol{\alpha}) - \tau\}.$$

Furthermore, if τ approaches 0 from right, then the right-hand side of the above inequality converges to Ω^*, that is, $\lim_{\tau\to 0+} \lim_{\epsilon\to 0+} \Omega^{(\tau,\epsilon)} = \Omega^*$. ■

B. Convergence Proof

Let $V_q(\boldsymbol{\alpha})$ be the family of sets $M \subseteq L$ such that $|M| \leq q$ and M contains at least one (τ,ϵ)-violating pair at $\boldsymbol{\alpha} \in S$. Then the following lemma holds.

Lemma 6: Let $\{\boldsymbol{\alpha}(n)\}_{n=0}^{\infty}$ be any sequence such that $\boldsymbol{\alpha}(n) \in S$, $\forall n$ and $\lim_{n\to\infty} \bar{\boldsymbol{\alpha}}$. If $\bar{\boldsymbol{\alpha}} \in S \setminus \Omega^{(\tau,\epsilon)}$ then

$$V_q(\boldsymbol{\alpha}(n)) \ \subseteq \ V_q(\bar{\boldsymbol{\alpha}}) \tag{10}$$

for sufficiently large n.

Proof: Let (i,j) be any (τ,ϵ)-feasible pair at $\bar{\boldsymbol{\alpha}}$. Then at least one of the following three conditions holds.

1) $i \notin I^{\epsilon}_{\rm up}(\bar{\boldsymbol{\alpha}})$
2) $j \notin I^{\epsilon}_{\rm low}(\bar{\boldsymbol{\alpha}})$
3) $i \in I^{\epsilon}_{\rm up}(\bar{\boldsymbol{\alpha}})$, $j \in I^{\epsilon}_{\rm low}(\bar{\boldsymbol{\alpha}})$, $F_i(\bar{\boldsymbol{\alpha}}) > F_j(\bar{\boldsymbol{\alpha}}) - \tau$

In Case 1), it is easily seen from Lemma 5 that $i \notin I^{\epsilon}_{\rm up}(\boldsymbol{\alpha}(n))$ for sufficiently large n, which means (i,j) is a (τ,ϵ)-feasible pair at $\boldsymbol{\alpha}(n)$ for sufficiently large n. In Case 2), we can draw the same conclusion as Case 1). In Case 3), it follows from Lemma 5 and the continuity of $F_i(\cdot)$ that

$$i \in I^{\epsilon}_{\rm up}(\boldsymbol{\alpha}(n)), j \in I^{\epsilon}_{\rm low}(\boldsymbol{\alpha}(n)), F_i(\boldsymbol{\alpha}(n)) > F_j(\boldsymbol{\alpha}(n)) - \tau$$

holds for sufficiently large n. This means (i,j) is a (τ,ϵ)-feasible pair at $\boldsymbol{\alpha}(n)$ for sufficiently large n. Therefore, in all cases, the set of (τ,ϵ)-feasible pairs at $\bar{\boldsymbol{\alpha}}$ is included in that at $\boldsymbol{\alpha}(n)$ for sufficiently large n. Conversely, the set of (τ,ϵ)-violating pairs at $\boldsymbol{\alpha}(n)$ is included in that at $\bar{\boldsymbol{\alpha}}$ for sufficiently large n. Eq.(10) is immediately derived from this fact and the definition of $V_q(\cdot)$. ■

For any $M \subseteq L$ and $\boldsymbol{\alpha} \in S$, we define the point-to-set map $\Gamma_M(\boldsymbol{\alpha})$ as

$$\Gamma_M(\boldsymbol{\alpha}) \triangleq \{\boldsymbol{y} \in S \mid y_i = \alpha_i \ \ \forall i \in L \setminus M, \min_{i \in I_{\rm up}(\boldsymbol{y}) \cap M} F_i(\boldsymbol{y}) \geq \max_{i \in I_{\rm low}(\boldsymbol{y}) \cap M} F_i(\boldsymbol{y})\} .$$

By using this definition, the set of optimal solutions of the subproblem in Step 4) can be expressed as $\Gamma_{L_B(k)}(\boldsymbol{\alpha}(k))$. We also define a point-to-set map $\boldsymbol{A}$ from S to itself as follows:

$$\boldsymbol{A}(\boldsymbol{\alpha}) = \begin{cases} \cup_{M \in V_q(\boldsymbol{\alpha})} \Gamma_M(\boldsymbol{\alpha}), & \text{if } \boldsymbol{\alpha} \notin \Omega^{(\tau,\epsilon)} \\ \boldsymbol{\alpha}, & \text{if } \boldsymbol{\alpha} \in \Omega^{(\tau,\epsilon)} \end{cases} \tag{11}$$

Lemma 7: For any $M \subseteq L$, the point-to-set map $\Gamma_M(\boldsymbol{\alpha})$ is closed on S.

Proof: Let $\{\boldsymbol{\alpha}(n)\}_{n=0}^{\infty}$ be any sequence such that $\boldsymbol{\alpha}(n) \in S$, $\forall n$ and $\lim_{n\to\infty} \boldsymbol{\alpha}(n) = \bar{\boldsymbol{\alpha}} \in S$. Let $\{\boldsymbol{\beta}(n)\}_{n=0}^{\infty}$ be any sequence such that $\boldsymbol{\beta}(n) \in \Gamma_M(\boldsymbol{\alpha}(n))$, $\forall n$ and $\lim_{n\to\infty} \boldsymbol{\beta}(n) = \bar{\boldsymbol{\beta}}$. Then $\boldsymbol{\beta}(n)$ satisfies

$$\beta_i(n) = \alpha_i(n), \quad \forall i \in L \setminus M \tag{12}$$

$$\min_{i \in I_{\rm up}(\boldsymbol{\beta}(n)) \cap M} F_i(\boldsymbol{\beta}(n)) \geq \max_{i \in I_{\rm low}(\boldsymbol{\beta}(n)) \cap M} F_i(\boldsymbol{\beta}(n)) \tag{13}$$

for all n. It is obvious from (12) that

$$\bar{\beta}_i = \bar{\alpha}_i, \quad \forall i \in L \setminus M \, .$$

Also, by applying the argument used in the proof of Proposition 2 to (13), we have

$$\min_{i \in I_{\rm up}(\bar{\boldsymbol{\beta}}) \cap M} F_i(\bar{\boldsymbol{\beta}}) \geq \max_{i \in I_{\rm low}(\bar{\boldsymbol{\beta}}) \cap M} F_i(\bar{\boldsymbol{\beta}}) \, .$$

Therefore $\bar{\boldsymbol{\beta}}$ belongs to $\Gamma_M(\bar{\boldsymbol{\alpha}})$ which implies that $\Gamma_M(\boldsymbol{\alpha})$ is closed at $\bar{\boldsymbol{\alpha}}$. Since $\bar{\boldsymbol{\alpha}}$ can be any point in S, we can conclude that $\Gamma_M(\boldsymbol{\alpha})$ is closed on S. ■

Lemma 8: The point-to-set map $\boldsymbol{A}(\boldsymbol{\alpha})$ defined by (11) is closed on $S \setminus \Omega^{(\tau,\epsilon)}$.

Proof: Let $\{\boldsymbol{\alpha}(n)\}_{n=0}^{\infty}$ be any sequence such that $\boldsymbol{\alpha}(n) \in S \setminus \Omega^{(\tau,\epsilon)}$, $\forall n$ and $\lim_{n\to\infty} \boldsymbol{\alpha}(n) = \bar{\boldsymbol{\alpha}} \in S \setminus \Omega^{(\tau,\epsilon)}$. Let $\{\boldsymbol{\beta}(n)\}_{n=0}^{\infty}$ be any sequence such that $\boldsymbol{\beta}(n) \in \boldsymbol{A}(\boldsymbol{\alpha}(n))$, $\forall n$ and $\lim_{n\to\infty} \boldsymbol{\beta}(n) = \bar{\boldsymbol{\beta}}$. Then it follows from Lemma 6 that there exists a positive integer n_1 such that

$$\boldsymbol{\beta}(n) \in \cup_{M \in V_q(\bar{\boldsymbol{\alpha}})} \Gamma_M(\boldsymbol{\alpha}(n)), \quad \forall n \geq n_1 \, . \tag{14}$$

As shown in Lemma 7, the point-to-set map $\Gamma_M(\boldsymbol{\alpha})$ is closed at $\bar{\boldsymbol{\alpha}}$ for each $M \in V_q(\bar{\boldsymbol{\alpha}})$. Moreover, we easily see from Lemma 1 that $\cup_{M \in V_q(\bar{\boldsymbol{\alpha}})} \Gamma_M(\boldsymbol{\alpha})$ is closed at $\bar{\boldsymbol{\alpha}}$. This result, together with (14), indicates that $\bar{\boldsymbol{\beta}} \in \cup_{M \in V_q(\bar{\boldsymbol{\alpha}})} \Gamma_M(\bar{\boldsymbol{\alpha}})$. Therefore $\boldsymbol{A}(\boldsymbol{\alpha})$ is closed at $\bar{\boldsymbol{\alpha}} \in S \setminus \Omega^{(\tau,\epsilon)}$. Since $\bar{\boldsymbol{\alpha}}$ can be

any point in $S \setminus \Omega^{(\tau,\epsilon)}$, we can conclude that the point-to-set map $\boldsymbol{A}(\boldsymbol{\alpha})$ is closed on $S \setminus \Omega^{(\tau,\epsilon)}$. ■

Lemma 9: The objective function $W(\boldsymbol{\alpha})$ of Problem 2 is a descent function for the set of (τ,ϵ)-optimal solutions $\Omega^{(\tau,\epsilon)}$ and the point-to-set map $\boldsymbol{A}(\boldsymbol{\alpha})$ defined by (11).

Proof: Let $\boldsymbol{\beta}$ be any point belonging to $\boldsymbol{A}(\boldsymbol{\alpha})$. If $\boldsymbol{\alpha} \notin \Omega^{(\tau,\epsilon)}$, there exists an $M \in V_q(\boldsymbol{\alpha})$ such that $\boldsymbol{\alpha} \notin \Gamma_M(\boldsymbol{\alpha})$ and $\boldsymbol{\beta} \in \Gamma_M(\boldsymbol{\alpha})$. This implies that $W(\boldsymbol{\beta}) < W(\boldsymbol{\alpha})$. Therefore $W(\boldsymbol{\alpha})$ is a descent function for $\Omega^{(\tau,\epsilon)}$ and $\boldsymbol{A}(\boldsymbol{\alpha})$. ■

Now we are ready for giving the global convergence theorem for Algorithm 1, which is the main result of this paper.

Theorem 2: Let $\{\boldsymbol{\alpha}(k)\}_{k=0}^{\infty}$ be the sequence generated by Algorithm 1. If the working set $L_B(k)$ contains at least one (τ,ϵ)-violating pair at $\boldsymbol{\alpha}(k)$ for all k, then any convergent subsequence of $\{\boldsymbol{\alpha}(k)\}_{k=0}^{\infty}$ has a limit in $\Omega^{(\tau,\epsilon)}$.

Proof: From the definition of the point-to-set map $\boldsymbol{A}(\boldsymbol{\alpha})$ in (11) and the assumption on the working set $L_B(k)$, it is apparent that $\boldsymbol{\alpha}(k+1) \in \boldsymbol{A}(\boldsymbol{\alpha}(k))$ for all k. The sequence $\{\boldsymbol{\alpha}(k)\}_{k=0}^{\infty}$ belongs to S which is compact. As shown in Lemma 8, the map $\boldsymbol{A}(\boldsymbol{\alpha})$ is closed on $S \setminus \Omega^{(\tau,\epsilon)}$. Also, as shown in Lemma 9, the objective function $W(\boldsymbol{\alpha})$ is a descent function for $\Omega^{(\tau,\epsilon)}$ and $\boldsymbol{A}(\boldsymbol{\alpha})$. Therefore, we can conclude from Theorem 1 that any convergent subsequence of $\{\boldsymbol{\alpha}(k)\}_{k=0}^{\infty}$ has a limit in $\Omega^{(\tau,\epsilon)}$. ■

The following theorem is immediately derived from Theorem 2 and Proposition 3.

Theorem 3: If the working set $L_B(k)$ contains at least one (τ,ϵ)-violating pair at $\boldsymbol{\alpha}(k)$ for all k, then Algorithm 1 stops at $\Omega^{(\tau,\epsilon)}$ within a finite number of iterations for any $\tau > 0$ and $\epsilon \in (0, C/2)$.

Remark 1: According to Proposition 4, we can make $\Omega^{(\tau,\epsilon)}$ as close as we want to Ω^* by setting τ and ϵ to sufficiently small positive numbers. In other words, we can make the limit of the sequence of solutions generated by Algorithm 1 as close as we want to an optimal solution of Problem 2.

Remark 2: In the above discussion, we have assumed that the subproblems can be solved exactly. However, this assumption is not necessarily required. In fact, if it is guaranteed that $W(\boldsymbol{\alpha}(k+1))$ is less than $W(\boldsymbol{\alpha}(k))$ as far as $L_B(k)$ contains at least one (τ,ϵ)-violating pair at $\boldsymbol{\alpha}(k)$, then Theorems 2 and 3 still hold. This weaker condition will be useful in practical situations where the subproblems are solved numerically and thus only approximate solutions can be obtained.

Remark 3: If we employ (6) instead of (7) for the optimality condition, the global convergence of Algorithm 1 cannot be proved as above because the point-to-set map corresponding to (11) is not closed in this case.

Let us next consider the case where (τ,ϵ)-violating pairs are not always contained in $L_B(k)$. Theorem 2 does not hold in this case because $W(\boldsymbol{\alpha})$ is not a descent function for $\Omega^{(\tau,\epsilon)}$ and $\boldsymbol{A}(\boldsymbol{\alpha})$ defined by (11). However, if at least one (τ,ϵ)-violating pair is selected for the working set within a certain period of iterations, Algorithm 1 still has the convergence property. This is formally stated as follows.

Theorem 4: Let $\{\boldsymbol{\alpha}(k)\}_{k=0}^{\infty}$ be the sequence generated by Algorithm 1. If there exists a positive integer m such that one of m sets $L_B(k), L_B(k+1), \ldots, L_B(k+m-1)$ contains at least one (τ,ϵ)-violating pair at $\boldsymbol{\alpha}(k)$ for all k, then Algorithm 1 stops at $\Omega^{(\tau,\epsilon)}$ within a finite number of iterations for any $\tau > 0$ and $\epsilon \in (0, C/2)$.

Proof: Let us define the point-to-set map $\boldsymbol{A}^m(\boldsymbol{\alpha})$ as

$$\boldsymbol{A}^m(\boldsymbol{\alpha}) = \begin{cases} \cup_{(M_1,\ldots,M_m)\in V_q^m(\boldsymbol{\alpha})} \Gamma_{M_m} \cdots \Gamma_{M_1}(\boldsymbol{\alpha}), & \text{if } \boldsymbol{\alpha} \notin \Omega^{(\tau,\epsilon)} \\ \boldsymbol{\alpha}, & \text{if } \boldsymbol{\alpha} \in \Omega^{(\tau,\epsilon)} \end{cases}$$

where $V_q^m(\boldsymbol{\alpha})$ is the set of all sequences of the sets $(M_1, M_2, \ldots, M_m)$ such that 1) $M_i \subseteq L$, $i = 1, 2, \ldots, m$, 2) $|M_i| \le q$, $i = 1, 2, \ldots, m$, and 3) at least one M_i contains at least one (τ,ϵ)-violating pair at $\boldsymbol{\alpha}$. Let $\{\tilde{\boldsymbol{\alpha}}(k)\}_{k=0}^{\infty}$ be the sequence defined by $\tilde{\boldsymbol{\alpha}}(k) = \boldsymbol{\alpha}(mk)$. Then the sequence $\{\tilde{\boldsymbol{\alpha}}(k)\}_{k=0}^{\infty}$ satisfies

$$\tilde{\boldsymbol{\alpha}}(k+1) \in \boldsymbol{A}^m(\tilde{\boldsymbol{\alpha}}(k)) \subseteq S, \quad \forall k\,.$$

By applying Lemmas 1, 2, 6 and 7 we can easily show that $\boldsymbol{A}^m(\boldsymbol{\alpha})$ is closed on $S \setminus \Omega^{\tau,\epsilon}$. Moreover, it follows from Property 3) of $V_q^m(\boldsymbol{\alpha})$ mentioned above that the objective function $W(\boldsymbol{\alpha})$ is a descent function for $\Omega^{(\tau,\epsilon)}$ and $\boldsymbol{A}^m(\boldsymbol{\alpha})$. Thus, by Theorem 1, any convergent subsequence of $\{\tilde{\boldsymbol{\alpha}}(k)\}_{k=0}^{\infty}$ has a limit in $\Omega^{(\tau,\epsilon)}$. Since $S \setminus \Omega^{(\tau,\epsilon)}$ is closed for any $\tau > 0$ and $\epsilon \in (0, C/2)$, the convergent subsequence enters $\Omega^{(\tau,\epsilon)}$ within a finite number of iterations. ■

V. Application to Existing Learning Algorithms

In this section, we will discuss the global convergence of some existing decomposition methods by applying the results obtained in the previous section.

A. Generalized SMO Algorithm

Generalized SMO algorithm [9] is a special type of decomposition methods in which the optimality condition is given by (6) and the working set is composed of a pair of indices (i, j) violating (6) for $\boldsymbol{\alpha} = \boldsymbol{\alpha}(k)$, that is,

$$i \in I_{\rm up}(\boldsymbol{\alpha}(k)),\, j \in I_{\rm low}(\boldsymbol{\alpha}(k)),\, F_i(\boldsymbol{\alpha}(k)) < F_j(\boldsymbol{\alpha}(k)) - \tau\,.$$

It has already been proved that this algorithm always stops within a finite number of iterations [9], [11]. We will now show that almost the same result can be derived by using a theorem in the previous section. To do so, we need to modify the generalized SMO algorithm slightly; the optimality condition is given by (7) instead of (6), and the working set is composed of a (τ,ϵ)-violating pair at $\boldsymbol{\alpha}(k)$ where τ and ϵ are sufficiently small positive numbers. Then it is easily seen from Theorem 3 that the algorithm stops within a finite number of iterations.

B. SVM^{light}

SVM$^{\rm light}$ proposed by Joachims [6] is one of the most widely used decomposition methods for SVMs. The optimality condition used in SVM$^{\rm light}$ is given by (6) and the working set selection is done in a systematic way as follows:

Algorithm 2: Given an even number $q(\le l)$ and the current solution $\boldsymbol{\alpha}(k) \in S$, execute the following procedures.

1) Sort $\{F_i(\boldsymbol{\alpha}(k))\}_{i=1}^{l}$ in decreasing order. Let the list of subscripts of the sorted list be $i_1, i_2, \ldots, i_l$.

2) Set $L_B(k) = \emptyset$, $v = 0$, $m = 1$ and $n = l$.
3) While $i_m \notin I_{\rm low}(\boldsymbol{\alpha}(k))$ and $m \leq l$, add 1 to m.
4) While $i_n \notin I_{\rm up}(\boldsymbol{\alpha}(k))$ and $n \geq 1$, subtract 1 from n.
5) If $m \geq n$ then stop. Otherwise, add 2 to v and add $\{m, n\}$ to the set $L_B(k)$.
6) If $v = q$ then stop. Otherwise, go to Step 3).

For more details on the working set selection of SVM$^{\rm light}$, please consult [14].

As well as in case of the generalized SMO algorithm, we need to modify both the optimality condition and the working set selection in SVM$^{\rm light}$ in order to apply the results in the previous section; the optimality condition is given by (7) instead of (6) and the working set selection is done by the following algorithm:

Algorithm 3: Given an even number $q(\leq l)$ and the current solution $\boldsymbol{\alpha}(k) \in S$, execute the following procedures.

1) Sort $\{F_i(\boldsymbol{\alpha}(k))\}_{i=1}^{l}$ in decreasing order. Let the the list of subscripts of the sorted list be $i_1, i_2, \ldots, i_l$.
2) Set $L_B(k) = \emptyset$, $v = 0$, $m = 1$ and $n = l$.
3) While $i_m \notin I^{\epsilon}_{\rm low}(\boldsymbol{\alpha}(k))$ and $m \leq l$, add 1 to m.
4) While $i_n \notin I^{\epsilon}_{\rm up}(\boldsymbol{\alpha}(k))$ and $n \geq 1$, subtract 1 from n.
5) If $m \geq n$ then stop. Otherwise, add 2 to v and add $\{m, n\}$ to the set $L_B(k)$.
6) If $v = q$ then stop. Otherwise, go to Step 3).

We will show that $L_B(k)$ obtained by Algorithm 3 contains at least one (τ, ϵ)-violating pair if $\boldsymbol{\alpha}(k) \in S \backslash \Omega^{(\tau,\epsilon)}$. Recall that $\boldsymbol{\alpha} \in \Omega^{(\tau,\epsilon)}$ if and only if (7) holds. Thus if $\boldsymbol{\alpha}(k) \in S \setminus \Omega^{(\tau,\epsilon)}$, the following inequality holds.

$$\min_{i \in I^{\epsilon}_{\rm up}(\boldsymbol{\alpha}(k))} F_i(\boldsymbol{\alpha}(k)) + \tau \leq \max_{i \in I^{\epsilon}_{\rm low}(\boldsymbol{\alpha}(k))} F_i(\boldsymbol{\alpha}(k)) \qquad (15)$$

Let m_1 be the value of m at the instance when Algorithm 3 first exits Step 3). Then $i_{m_1} \in I^{\epsilon}_{\rm low}(\boldsymbol{\alpha}(k))$ and $F_{i_{m_1}}(\boldsymbol{\alpha}(k)) = \max_{i \in I^{\epsilon}_{\rm low}(\boldsymbol{\alpha}(k))} F_i(\boldsymbol{\alpha}(k))$. Also, let n_1 be the value of n at the instance when Algorithm 3 first exits Step 4). Then $i_{n_1} \in I^{\epsilon}_{\rm up}(\boldsymbol{\alpha}(k))$ and $F_{i_{n_1}}(\boldsymbol{\alpha}(k)) = \min_{i \in I^{\epsilon}_{\rm up}(\boldsymbol{\alpha}(k))} F_i(\boldsymbol{\alpha}(k))$. It follows from (15) that $F_{n_1}(\boldsymbol{\alpha}(k)) + \tau \leq F_{m_1}(\boldsymbol{\alpha}(k))$ which implies that (n_1, m_1) is a (τ, ϵ)-violating pair at $\boldsymbol{\alpha}(k)$ and that $m_1 < n_1$. Therefore $L_B(k)$ obtained by Algorithm 3 contains at least one (τ, ϵ)-violating pair at $\boldsymbol{\alpha}(k)$.

By applying Theorem 3, we can conclude that SVM$^{\rm light}$ with the optimality condition (7) and the working set selection described by Algorithm 3 stops within a finite number of iterations after finding a (τ, ϵ)-optimal solution.

Remark 4: In the software package SVM$^{\rm light}$ the optimality condition and the working set selection are not implemented exactly as described in (6) and Algorithm 2, respectively, for practical reason. Difference is in the definition of $I_{\rm up}(\boldsymbol{\alpha})$ and $I_{\rm low}(\boldsymbol{\alpha})$. In the software, the following definitions are used:

$$I_{\rm up}(\boldsymbol{\alpha}) = \{i \,|\, \alpha_i < C - \epsilon,\, d_i = 1\} \cup \{i \,|\, \alpha_i > \epsilon,\, d_i = -1\},$$

$$I_{\rm low}(\boldsymbol{\alpha}) = \{i \,|\, \alpha_i < C - \epsilon,\, d_i = -1\} \cup \{i \,|\, \alpha_i > \epsilon,\, d_i = 1\}$$

where ϵ is a sufficiently small positive number. It is worth noting that these definitions are same as $I^{\epsilon}_{\rm up}(\boldsymbol{\alpha})$ and $I^{\epsilon}_{\rm low}(\boldsymbol{\alpha})$ except for the equal sign. The results of this paper show that the relaxation of the optimality condition has not only practical but also theoretical significance.

VI. Concluding Remarks

Global convergence property of decomposition methods for SVMs are studied. We have first introduced a relaxed optimality condition, and then proved a decomposition method stops within a finite number of iterations after finding an optimal solution if the working set selection satisfies a certain condition. We have also shown that the generalized SMO algorithm and SVM$^{\rm light}$ satisfy this condition and thus have the global convergence property. Since our new convergence theorems require little restriction on the working set selection method, the authors believe that they can be applied to a wide class of decomposition methods.

References

[1] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.
[2] B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds., *Advances in Kernel Methods: Support Vector Learning*. Cambridge, Massachusetts: MIT Press, 1999.
[3] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK: Cambridge University Press, 2000.
[4] S. Haykin, *Neural Networks*. Upper Saddle River, NJ: Prentice Hall, 1999.
[5] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods: Support Vector Machines*, B. Schölkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1998.
[6] T. Joachims, "Making large-scale support vector machine learning practical," in *Advances in Kernel Methods: Support Vector Machines*, B. Schölkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1998.
[7] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to Platt's SMO algorithm for SVM classifier design," *Neural Computation*, vol. 13, pp. 637–649, 2001.
[8] C.-W. Hsu and C.-J. Lin, "A simple decomposition method for support vector machines," *Machine Learning*, vol. 46, pp. 291–314, 2002.
[9] S. S. Keerthi and E. G. Gilbert, "Convergence of a generalized SMO algorithm for SVM classifier design," *Machine Learning*, vol. 46, pp. 351–360, 2002.
[10] C.-J. Lin, "Asymptotic convergence of an SMO algorithm without any assumption," *IEEE Trans. Neural Networks*, vol. 13, pp. 248–250, Jan. 2002.
[11] N. Takahashi and T. Nishi, "Rigorous proof of termination of SMO algorithm for support vector machines," *IEEE Trans. Neural Networks*, vol. 16, no. 3, pp. 774–776, May 2005.
[12] P.-H. Chen, R.-E. Fan, and C.-J. Lin, "A study on SMO-type decomposition methods for support vector machines," to appear in *IEEE Trans. Neural Networks*.
[13] 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, no. 4, pp. 1003–1008, 2000.
[14] C.-J. Lin, "On the convergence of the decomposition method for support vector machines," *IEEE Trans. Neural Networks*, vol. 12, pp. 1288–1298, Nov. 2001.
[15] ——, "A formal analysis of stopping criteria of decomposition methods for support vector machines," *IEEE Trans. Neural Networks*, vol. 13, pp. 1045–1052, Sept. 2002.
[16] N. List and H. U. Simon, "A general convergence theorem for the decomposition method," in *Proceedigns of the 17th Annual Conference on Learning Theory*, 2004, pp. 363–377.
[17] D. G. Luenberger, *Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley Publishing, 1989.