The Analysis of Decomposition Methods for Support Vector Machines*

Chih-Chung Chang, Chih-Wei Hsu, and Chih-Jen Lin
Department of Computer Science and Information Engineering
National Taiwan University
Taipei 106, Taiwan

Abstract

The support vector machine (SVM) is a new and promising technique for pattern recognition. It requires the solution of a large dense quadratic programming problem. Traditional optimization methods cannot be directly applied due to memory restrictions. Up to now, very few methods can handle the memory problem and an important one is the "decomposition method." However, there is no convergence proof so far. In this paper, we connect this method to projected gradient methods and provide theoretical proofs for a version of decomposition methods. An extension to bound-constrained formulation of SVM is also provided.

1 Introduction

The support vector machine is a new and very promising classification technique for pattern recognition. Surveys of SVM are, for example, Cortes and Vapnik [1995], and Vapnik [1998]. Giving training vectors $v_i, i = 1, \ldots, m$ of length k, and a vector a defined as follows

$$a_i = \begin{cases} 1 & \text{if } v_i \text{ in class } 1, \\ -1 & \text{if } v_i \text{ in class } 2, \end{cases}$$

the support vector technique requires the solution of the following quadratic programming problem:

$$\begin{aligned} \min \quad & \frac{1}{2} x^T Q x - e^T x \\ & a^T x = 0, \\ & 0 \leq x_i \leq C, \qquad i = 1, \ldots, n, \end{aligned} \tag{1.1}$$

where e is the vector of all ones, C is the upper bound of all variables, Q is a positive semidefinite matrix. Possible choices of Q_{ij} are, for example, $a_i a_j ((v_i^T v_j)/k)^d$ and $a_i a_j e^{-\|v_i - v_j\|^2/k}$. Note that v_i is considered as a *support vector* if x is the solution of (1.1) and $x_i > 0$.

*This work was supported in part by the National Science Council of Taiwan via the grant NSC-88-2213-E-002-097. (`cjlin@csie.ntu.edu.tw`).

The difficulty of solving (1.1) is the density of Q because Q_{ij} is in general not zero and Q becomes a fully dense matrix. Hence 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 (e.g. [Kaufman, 1998; Mangasarian and Musicant, 1998; Osuna *et al.*, 1997b; Joachims, 1998; Platt, 1998; Saunders *et al.*, 1998]) have proposed methods with successful implementation to conquer this difficulty. Unlike the Newton method which usually involves the whole Hessian matrix Q, these methods only calculate components of Q when they are required in the current iteration. Among them, an important one is the "decomposition method" proposed by Osuna et al. [1997b]. As a variation of the active set methods, they separated the index $\{1, \ldots, n\}$ of the training set to two sets B and N, where $x_i = 0$, for $i \in N$ and B is the working set if x is the current iterate of the algorithm. If we denote x_B and x_N as vectors containing corresponding elements, the objective value is equal to $\frac{1}{2} x_B^T Q_{BB} x_B - (e_B + Q_{BN} x_N)^T x_B + \frac{1}{2} x_N^T Q_{NN} x_N - e_N^T x_N$. At each iteration, x_N is fixed and the following problem with the variable x_B is solved:

$$\begin{aligned} \min \quad & \frac{1}{2} x_B^T Q_{BB} x_B - (e_B - Q_{BN} x_N)^T x_B \\ & 0 \leq x_B \leq C, a_B^T x_B = -a_N^T x_N, \end{aligned} \tag{1.2}$$

where $\begin{bmatrix} Q_{BB} & Q_{BN} \\ Q_{NB} & Q_{NN} \end{bmatrix}$ is a permutation of the matrix Q and q is the size of B. Elements of x_N are kept as zero from the beginning. The KKT condition of (1.2) includes $0 \leq x_B \leq C, a_B^T x_B = -a_N^T x_N$ of (1.2) and

$$\begin{aligned} Q_{BB} x_B + Q_{BN} x_N - e_B = \sigma a_B + \lambda - \mu \qquad (1.3) \\ \lambda_i (x_B)_i = 0, \mu_i (C - (x_B)_i) = 0, \lambda_i \geq 0, \mu_i \geq 0. \end{aligned}$$

If there is one $(x_B)_i$ satisfying $0 < (x_B)_i < C$, from (1.3), σ could be easily obtained. Since all components of x_N are zero, if x is an optimal solution, x_N satisfies the KKT condition of (1.2): $(Qx)_i - 1 - \sigma a_i \geq 0$, for $i \in N$. Therefore, the algorithm moves elements of N which violate the KKT condition of (1.2) to B, and move out zero elements from B to N. If the size of B is bigger than the number of support vectors, because no cycle happens, the algorithm stops in finite steps.

The main shortcoming of the above algorithm is that we do not know the number of support vectors in priori. To handle this issue, the same authors [Osuna *et al.*, 1997a] hence proposed to use a small number q as the size of B. Furthermore, there is no restriction that x_N must only contain zeros. To be more precise, any element in N which violates the KKT condition can enter B. Since q is small in general (less than 100), this method never faces memory problem. Good numerical results were reported so two further improvements are given in [Joachims, 1998; Platt, 1998]. However, even though the strict decrease of the objective function is still hold, there is no theoretical proof to show that the sequence converges to an optimal solution.

The paper by Joachims [1998] has drawn us a lot of attention because of its method of updating B and N. Unlike the method by Osuna et al., where a more "random" selection is used, Joachims solves the following problem in order to select the working set:

$$\begin{aligned} \min \quad & \nabla f(x_k)^T d \\ & a^T d = 0, -1 \le d \le 1 && (1.4a) \\ & d_i \ge 0, \text{ if } (x_k)_i = 0 && (1.4b) \\ & d_i \le 0, \text{ if } (x_k)_i = C && (1.4c) \\ & |\{d_i \mid d_i \ne 0\}| = q. && (1.4d) \end{aligned}$$

Note that $|\{d_i \mid d_i \ne 0\}|$ means the number of components of d which are not zero. The constraint (1.4d) implies that a descent direction involving only q variables is obtained. Then components of x with non-zero d_i are included in the working set B which is used to construct the sub-problem (1.2). Note that d is only used for identifying B but not as a search direction. Using the decomposition method with the new technique for selecting the working set B, Joachims reported promising numerical results. In this paper, we will demonstrate that a variation of his algorithm theoretically converges.

In Section 2, we present a more general algorithm with an analysis on the selection of the working set. Then in Section 3, we prove the convergence of this algorithm using the techniques of projected gradients. Section 4 extends the convergence proofs to bound-constrained formulations of SVM Finally we provide concluding remarks in Section 5.

2 A General Algorithm

To describe the new algorithm, we replace the problem (1.1) by the following form:

$$\min f(x), \quad x \in \Omega, \tag{2.1}$$

where $f(x)$ is a continuously differentiable function from R^n to R and the set Ω is the feasible set of (1.1):

$$\Omega \equiv \{x \mid a^T x = 0, 0 \le x_i \le C, i = 1, \ldots, n\}. \tag{2.2}$$

Therefore, (1.1) is a special case of (2.1) when $f(x) = \frac{1}{2}x^T Q x - e^T x$. In addition, we assume C is a finite positive number so Ω is a bounded set.

In the following we describe a more general algorithm for solving (2.1):

Step 1: Let q be a positive integer with $2 \le q \le n$. Set $k = 1$, choose $\mu_1, \mu_2 \in (0, 1), \gamma_1, \gamma_2 \ge 0$, and find an x_1 as the initial solution.

Step 2: Solve

$$\begin{aligned} \min \quad & \nabla f(x_k)^T d \\ & x_k + d \in \Omega \\ & |\{d_i \mid d_i \ne 0\}| \le q \end{aligned} \tag{2.3}$$

Assume B_k contains indices for which $d_i \ne 0$.

Step 3: If $d = 0$, stop and output x_k as an optimal solution. Otherwise, let

$$\tilde{\nabla} f(x_k)_i \equiv \begin{cases} \nabla f(x_k) & \text{if } i \in B_k, \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

Define

$$\begin{aligned} \Omega_k &\equiv \{x \mid x \in \Omega, x_i = (x_k)_i, i \notin B_k\}, \\ R_k^n &\equiv \{x \mid x \in R^n, x_i = (x_k)_i, i \notin B_k\}. \end{aligned}$$

The mapping $P_k : R_k^n \to \Omega_k$ is defined by

$$P_k(x) \equiv \operatorname{argmin}\{\|z - x\| : z \in \Omega_k\}.$$

For $\alpha \ge 0$, define $x_k(\alpha) \equiv P_k(x_k - \alpha\tilde{\nabla} f(x_k))$, and select $\alpha_k > 0$ so that $x_k^C \equiv P_k(x_k - \alpha_k \tilde{\nabla} f(x_k))$,

$$f(x_k^C) \le f(x_k) + \mu_1 \tilde{\nabla} f(x_k)^T (x_k^C - x_k), \tag{2.5}$$

and

$$\alpha_k \ge \gamma_1 \text{ or } \alpha_k \ge \gamma_2 \bar{\alpha}_k > 0,$$

where $\bar{\alpha}_k$ satisfies

$$f(x_k(\bar{\alpha}_k)) > f(x_k) + \mu_2 \nabla f(x_k)^T (x_k(\bar{\alpha}_k) - x_k). \tag{2.6}$$

Step 4: Find a new x_{k+1} such that

$$(x_{k+1})_i = (x_k)_i \text{ if } i \notin B_k \text{ and } f(x_{k+1}) \le f(x_k^C).$$

Go to Step 2.

Joachims requires that $|\{d_i \mid d_i \ne 0\}| = q$. This may not be always possible, so we modify the equality to $|\{d_i \mid d_i \ne 0\}| \le q$. Another difference is that we use $x_k + d \in \Omega$ instead of (1.4a)-(1.4c). Note that (1.4a)-(1.4c) do not ensure the feasibility $x_k + d \in \Omega$ so the solution of (1.4) may not be a feasible direction. However, we use this property for the convergence proof as will be explained in Section 3.

The Step 3 of the algorithm searches for a point x_k^C by following a "partial" projected gradient direction. We called $\tilde{\nabla} f(x_k)$ a "partial" gradient direction since it contains only components of $\nabla f(x_k)$ which are in the working set of x_k. For a standard projected gradient method (e.g. Calamai and Moré [1987]), $x_k^C \equiv P(x_k - \alpha\nabla f(x_k))$, where P is the projection into Ω, and x_k^C is referred as the *Cauchy point.* The sufficient decrease of function values at Cauchy points is guaranteed by conditions (2.5) and (2.6) and is used for the convergence proof. Then by requiring $f(x_{k+1}) \le f(x_k^C)$, we will prove that any limit point of the sequence $\{x_k\}$ is an optimal solution. Note that when $f(x) = \frac{1}{2}x^T Qx - e^T x$, x_k^C in our algorithm is a feasible solution to the subproblem (1.2). Since most existing decomposition methods obtain the optimal solution of (1.2) for x_{k+1}, in Step 4 the requirement that $f(x_{k+1}) \le f(x_k^C)$ is satisfied.

Because the use of $\tilde{\nabla} f(x_k)$ instead of $\nabla f(x_k)$ makes a difference in Step 3, there is a new obstacle proving the convergence. In the rest of this section, we will construct the relation between (2.3) and the following problem:

$$\begin{aligned} \min \quad & \nabla f(x_k)^T d \\ & x_k + d \in \Omega \end{aligned} \tag{2.7}$$

Definition 2.1 *A point x^* is a stationary point of (2.1) if $\nabla f(x^*)^T(x - x^*) \ge 0$, for all $x \in \Omega$.*

Immediately, we have the following result:

Lemma 2.2 *If $OP(x_k)$ is the optimal objective value of (2.7), $OP(x_k) = 0$ if and only if x_k is a stationary point.*

Proof. If x_k is not a stationary point, there is an $x \in \Omega$ such that $\nabla f(x_k)^T(x - x_k) < 0$. Thus $d = x - x_k$ is a feasible solution of (2.7). The objective value $\nabla f(x_k)^T d < 0$ causes a contradiction. The proof of the other direction is similar. □

We will investigate (2.7) in detail. Except $a^T d = 0$, constraints of (2.7) on d can be written as

$$l_i \le d_i \le u_i, i = 1, \ldots, n,$$

where $u_i \equiv C - (x_k)_i \ge 0, l_i \equiv -(x_k)_i \le 0$, and $u_i \ne l_i$, for all i. The KKT condition of (2.7) is

$$\begin{aligned} \nabla f(x_k) - \sigma a &= \lambda - \mu, \\ \lambda_i(d_i - l_i) &= 0, i = 1, \ldots, n, \\ \mu_i(u_i - d_i) &= 0, i = 1, \ldots, n, \end{aligned} \tag{2.8}$$

where σ is a real number. If d is an optimal solution of (2.7), the property $l_i \ne u_i, \forall i$ and (2.8) imply that

$$\begin{aligned} d_i = l_i \quad &\Rightarrow \quad \nabla f(x_k)_i - \sigma a_i = \lambda_i \ge 0, \mu_i = 0, \\ l_i < d_i < u_i \quad &\Rightarrow \quad \nabla f(x_k)_i - \sigma a_i = 0 = \lambda_i = \mu_i, \\ d_i = u_i \quad &\Rightarrow \quad \nabla f(x_k)_i - \sigma a_i = -\mu_i \le 0, \lambda_i = 0. \end{aligned}$$

As $OP(x_k)$ is the optimal objective value of (2.7),

$$\begin{aligned} OP(x_k) &= \sum_{i=1}^{n} \nabla f(x_k)_i d_i \\ &= \sum_{i:d_i=l_i} l_i(\nabla f(x_k)_i - \sigma a_i) + \sum_{i:d_i=u_i} u_i(\nabla f(x_k)_i - \sigma a_i). \end{aligned} \tag{2.9}$$

Since each term of (2.9) is non-positive, if j is the index such that $d_i(\nabla f(x_k)_i - \sigma a_i)$ has the smallest value,

$$d_j(\nabla f(x_k)_j - \sigma a_j) \le \frac{OP(x_k)}{n}. \tag{2.10}$$

Hence if $OP(x_k) < 0$, $d_j \ne 0$. As $a_j d_j \ne 0$, with the constraint $a^T d = 0$, there are some elements i satisfying

$$(a_i d_i)(a_j d_j) < 0. \tag{2.11}$$

Furthermore, since all $a_i, i = 1, \ldots, n$ are 1 or -1, there is at least one index r satisfying (2.11) such that

$$|d_r| \ge \frac{|d_j|}{n}. \tag{2.12}$$

Now we are ready to switch back to (2.3) by considering the following problem:

$$\begin{aligned} \min \quad & \nabla f(x_k)^T d \\ & x_k + d \in \Omega \\ & d_i = 0, \forall i \ne j \text{ and } r. \end{aligned} \tag{2.13}$$

If $|d_r| \le |d_j|$, define $\bar{d}_j \equiv (|d_r|/|d_j|)d_j$. Then $a_j\bar{d}_j + a_r d_r = 0$, and from (2.9), (2.10), and (2.12),

$$\begin{aligned} & d_r \nabla f(x_k)_r + \bar{d}_j \nabla f(x_k)_j \\ = \quad & d_r \nabla f(x_k)_r + \frac{|d_r|}{|d_j|} d_j \nabla f(x_k)_j \\ = \quad & d_r(\nabla f(x_k)_r - \sigma a_r) + \frac{|d_r|}{|d_j|} d_j(\nabla f(x_k)_j - \sigma a_j) \\ \le \quad & \frac{1}{n^2} OP(x_k). \end{aligned}$$

A similar setting can be made for the case when $|d_r| > |d_j|$. Therefore, by assigning other elements of d to be zero, we find a feasible solution of (2.13) with objective value less than $\frac{1}{n^2}OP(x_k)$. As this solution is also a feasible point of (2.3), if $\tilde{OP}(x_k)$ is the optimal objective value of (2.3), we have the following lemma:

Lemma 2.3 *$n^2 \tilde{OP}(x_k) \le OP(x_k)$.*

From Lemma 2.3, the fact that $OP(x_k) \le \tilde{OP}(x_k)$, and Lemma 2.2, we have

Lemma 2.4 *$\tilde{OP}(x_k) = 0$ if and only if x_k is a stationary point.*

3 The Convergence Proof

As the only difference between P_k and P is the restriction on the domain variable x such that $x_i = (x_k)_i, i \notin B_k$, following from Calamai and Moré [1987], we have

Lemma 3.1 *Let P_k be the projection into Ω_k:*

(a) If $z \in \Omega_k$, then

$$(P_k(x) - x)^T(z - P_k(x)) \ge 0, \forall x \in R_k^n. \tag{3.1}$$

(b) P_k is a monotone operator, that is

$$(P_k(y) - P_k(x))^T(y - x) \ge 0, \forall x, y \in R_k^n. \tag{3.2}$$

If $P_k(y) \ne P_k(x)$ then strict inequality holds.

(c) Given $x \in R_k^n$ and $d \in R^n$, $d_i = 0, i \notin B_k$, the function ψ defined by

$$\psi(\alpha) \equiv \frac{\|P_k(x + \alpha d) - x\|}{\alpha}, \; \alpha > 0, \tag{3.3}$$

is non-increasing.

We define $x_k(\alpha) \equiv P_k(x_k - \alpha\tilde{\nabla} f(x_k))$, a function of α, which yields the following inequalities: From (3.1),

$$\tilde{\nabla} f(x_k)^T(x_k - x_k(\alpha)) \geq \frac{\|x_k(\alpha) - x_k\|^2}{\alpha}, \alpha > 0. \quad (3.4)$$

Therefore,

$$\|(x_k - \alpha\tilde{\nabla} f(x_k)) - x_k\|\|x_k(\alpha) - x_k\| \geq$$
$$((x_k - \alpha\tilde{\nabla} f(x_k)) - x_k)^T(x_k(\alpha) - x_k) \geq \|x_k(\alpha) - x_k\|^2,$$

and consequently

$$\|\alpha\tilde{\nabla} f(x_k)\| \geq \|x_k(\alpha) - x_k\|. \quad (3.5)$$

From (3.2), if $\alpha \geq \beta$, then

$$\tilde{\nabla} f(x_k)^T(x_k - x_k(\alpha)) \geq \tilde{\nabla} f(x_k)^T(x_k - x_k(\beta)). \quad (3.6)$$

The following lemma ensures that there is an x_k^C satisfying the requirements in Step 3 of the algorithm.

Lemma 3.2 *If x_k is not a stationary point of (2.1), then*

(a) For all $\alpha > 0$, $\tilde{\nabla} f(x_k)^T(x_k(\alpha) - x_k) < 0$.

(b) There is a μ in $(0, 1)$, and an $\hat{\alpha} > 0$ such that for all $0 < \alpha < \hat{\alpha}$,

$$f(x_k(\alpha)) \leq f(x_k) + \mu\tilde{\nabla} f(x_k)^T(x_k(\alpha) - x_k), \quad (3.7)$$

and

(c) There are positive constants γ_1 and γ_2, and constants μ_1 and μ_2 in $(0, 1)$, and $\alpha_k > 0$ such that (2.5) and (2.6) are satisfied.

Proof. If (a) is false, we can find an $\bar{\alpha}$ such that

$$\tilde{\nabla} f(x_k)^T(x_k(\bar{\alpha}) - x_k) \geq 0.$$

Then from (3.6), for all $0 \leq \alpha \leq \bar{\alpha}$, $\tilde{\nabla} f(x_k)^T(x_k(\alpha) - x_k) \geq 0$. However, from (3.4),

$$\tilde{\nabla} f(x_k)^T(x_k(\alpha) - x_k) \leq -\frac{\|x_k(\alpha) - x_k\|^2}{\alpha} \leq 0.$$

Therefore, $x_k(\alpha) = x_k$, for all $0 \leq \alpha \leq \bar{\alpha}$. Because x_k is not a stationary point, from Lemma 2.4, (2.3) has a solution $d^* \neq 0$ and there exists an $\hat{\alpha} < \bar{\alpha}$ such that $\|\hat{\alpha}\tilde{\nabla} f(x_k)\| \leq \|d^*\|$. Because $\tilde{\nabla} f(x_k)^T d^* < 0$, we can find an $r \neq 0$ such that $(-\hat{\alpha}\tilde{\nabla} f(x_k) - \frac{d^*}{r})^T d^* = 0$. Then

$$\|(x_k + \frac{d^*}{r}) - (x_k - \hat{\alpha}\tilde{\nabla} f(x_k))\| < \|\hat{\alpha}\tilde{\nabla} f(x_k)\|.$$

This contradicts the property that $\operatorname{argmin}\{\|z - (x_k - \hat{\alpha}\tilde{\nabla} f(x_k))\| : z \in \Omega_k\} = x_k$.

For (b), because x_k is not a stationary point and f is continuously differentiable, from (a) of this Lemma and (3.4), we have

$$|\frac{-f(x_k) + f(x_k(\alpha))}{\tilde{\nabla} f(x_k)^T(x_k(\alpha) - x_k)} - 1| \leq \frac{o(\|x_k(\alpha) - x_k\|)}{-\tilde{\nabla} f(x_k)^T(x_k(\alpha) - x_k)}$$
$$\leq \frac{\alpha o(\|x_k(\alpha) - x_k\|)}{\|x_k(\alpha) - x_k\|^2}. \quad (3.8)$$

From (3.3), $\|x_k(\alpha) - x_k\|/\alpha$ is non-increasing. On the other hand, from (3.5), $\|x_k(\alpha) - x_k\| \leq \alpha\|\tilde{\nabla} f(x_k)\|$, so

$$\frac{\alpha}{\|x_k(\alpha) - x_k\|} = O(1). \quad (3.9)$$

Thus, (3.8) implies

$$\lim_{\alpha \to 0^+} \frac{-f(x_k) + f(x_k(\alpha))}{\tilde{\nabla} f(x_k)^T(x_k(\alpha) - x_k)} = 1.$$

Hence the result (3.7) immediately follows.

We prove (c) by giving an example from Bertsekas [1976]: Define $\alpha_k = \gamma_2^{m_k}\gamma_1$, where m_k is the smallest integer such that (3.7) is satisfied. From (b), we know this α_k exists. Then by using this $x_k(\alpha_k)$ as x_k^C, both (2.5) and (2.6) are satisfied. □

The following theorem from Calamai and Moré [1987] is very important to our convergence proof. Because we are now using $\tilde{\nabla} f(x_k)$ instead of $\nabla f(x_k)$ as the search direction of the Cauchy step, to demonstrate its validity, here we present the whole proof.

Theorem 3.3 *Let $f : R^n \to R$ be continuously differentiable on Ω.*

$$\lim_{k \to \infty} \frac{\|x_k^C - x_k\|}{\alpha_k} = 0.$$

Proof. Assume there is an infinite subset K_0 such that

$$\frac{\|x_k^C - x_k\|}{\alpha_k} \geq \epsilon, k \in K_0 \text{ and } \lim_{k \to \infty, k \in K_0} x_k = x^*. \quad (3.10)$$

We will prove that this assumption leads to a contradiction. From (3.10), for all $k \in K_0$,

$$\frac{\|x_k^C - x_k\|^2}{\alpha_k} \geq \epsilon \max\{\epsilon\alpha_k, \|x_k^C - x_k\|\}. \quad (3.11)$$

Since Ω is bounded and $x_k \in \Omega, \forall k$, $\{x_k\}$ is bounded, $\{f(x_k)\}_{k \in K_0}$ converges, and condition (2.5) implies $\{\tilde{\nabla} f(x_k)^T(x_k - x_k^C)\}_{k \in K_0} \to 0$. Hence (3.4) and (3.11) show that

$$\lim_{k \in K_0, k \to \infty} \alpha_k = 0 \text{ and } \lim_{k \in K_0, k \to \infty} \|x_k^C - x_k\| = 0.$$

In particular, we have shown that eventually $\alpha_k < \gamma_1$ and hence $\alpha_k \geq \gamma_2\bar{\alpha}_k$ where $\bar{\alpha}_k$ satisfies (2.6). Now assign $\bar{x}_k^C = x_k(\bar{\alpha}_k)$ and $\beta_k = \min(\alpha_k, \bar{\alpha}_k)$. With this, (3.3) implies

$$\frac{\|x_k(\beta_k) - x_k\|^2}{\beta_k} \geq \beta_k(\frac{\|x_k^C - x_k\|}{\alpha_k})(\frac{\|\bar{x}_k^C - x_k\|}{\bar{\alpha}_k}).$$

Since $\|x_k^C - x_k\| \geq \epsilon\alpha_k$ and $\alpha_k \geq \gamma_2\bar{\alpha}_k$ for $k \in K_0$, we obtain

$$\frac{\|x_k(\beta_k) - x_k\|^2}{\beta_k} \geq \epsilon \min\{1, \gamma_2\}\|\bar{x}_k^C - x_k\|. \quad (3.12)$$

Together with (3.4) and (3.6), (3.12) implies that for $k \in K_0$,

$$\min\{\tilde{\nabla} f(x_k)^T(x_k - x_k^C), \tilde{\nabla} f(x_k)^T(x_k - \bar{x}_k^C)\}$$
$$\geq \epsilon \min\{1, \gamma_2\}\|\bar{x}_k^C - x_k\|. \quad (3.13)$$

We now use (3.13) to obtain the desired contradiction. Since $\{\tilde{\nabla} f(x_k)^T(x_k - x_k^C)\}$ converges to zero, (3.13) implies that $\{\|\bar{x}_k^C - x_k\|\}$ converges to zero. Thus the continuity of ∇f, the fact $\nabla f(x_k)^T(x_k - \bar{x}_k^C) = \tilde{\nabla} f(x_k)^T(x_k - \bar{x}_k^C)$, and the boundedness of $\{x_k\}$ show that

$$|f(\bar{x}_k^C) - f(x_k) - \tilde{\nabla} f(x_k)^T(\bar{x}_k^C - x_k)| = o(\|\bar{x}_k^C - x_k\|).$$

Therefore, if

$$\rho_k(\alpha) = \frac{f(x_k) - f(x_k(\alpha))}{\tilde{\nabla} f(x_k)^T(x_k - x_k(\alpha))},$$

then

$$|\rho_k(\bar{\alpha}) - 1| \le \frac{o(\|\bar{x}_k^C - x_k\|)}{\tilde{\nabla} f(x_k)^T(x_k - \bar{x}_k^C)}.$$

Hence (3.13) establishes that $\rho_k(\bar{\alpha}) > \mu_2$ for all $k \in K_0$ sufficiently large. This is the desired contradiction because (2.6) guarantees that $\rho(\bar{\alpha}_k) < \mu_2$. $\square$

We are now in a position to prove the main theorem of this paper.

Theorem 3.4 *Let $f : R^n \to R$ be continuously differentiable on Ω. Any limit point of $\{x_k\}$ is a stationary point of (2.1).*

Proof. For any $z \in \Omega_k$, from (3.1),

$$\begin{aligned} & \alpha_k \tilde{\nabla} f(x_k)^T(x_k^C - z) \le -(x_k^C - x_k)^T(x_k^C - z) \\ \le\ & -(x_k^C - x_k)^T(x_k - z) \le \|x_k^C - x_k\|\|x_k - z\|. \end{aligned}$$

Hence,

$$\begin{aligned} & \tilde{\nabla} f(x_k)^T(x_k - z) \\ \le\ & \tilde{\nabla} f(x_k)^T(x_k - x_k^C) + \frac{\|x_k^C - x_k\|}{\alpha_k}\|x_k - z\|. \end{aligned}$$

If d is a direction at x_k with $x_k + d \in \Omega_k$, then setting $z = x_k + d$ yields

$$-\tilde{\nabla} f(x_k)^T d \le \tilde{\nabla} f(x_k)^T(x_k - x_k^C) + \frac{\|x_k^C - x_k\|}{\alpha_k}\|d\|.$$

Since this inequality is true for all such d, with Lemma 2.3, for any $x \in \Omega$,

$$\begin{aligned} & \nabla f(x_k)^T(x_k - x) \le -OP(x_k) \le -n^2 \tilde{OP}(x_k) \\ =\ & n^2 \max_{x_k + d \in \Omega_k} -\tilde{\nabla} f(x_k)^T d \\ \le\ & n^2(\tilde{\nabla} f(x_k)^T(x_k - x_k^C) + \frac{\|x_k^C - x_k\|}{\alpha_k} \max_{x_k + d \in \Omega} \|d\|). \end{aligned}$$

Theorem 3.3 implies that $\|x_k^C - x_k\|/\alpha_k$ converges to zero. Since $\{f(x_k)\}$ converges, the inequality (2.5) implies that $\tilde{\nabla} f(x_k)^T(x_k - x_k^C)$ converges to zero. Since $f(x)$ is continuously differentiable on Ω and $\|d\|$ is bounded, if x^* is a limit point of a subsequence $\{x_k\}_{k \in K_0}$,

$$\lim_{k \in K_0, k \to \infty} \nabla f(x_k)^T(x_k - x) = \nabla f(x^*)^T(x^* - x) \le 0.$$

Since this inequality is true for all $x \in \Omega$, from Definition 2.1, x^* is a stationary point of (2.1). $\square$

4 Bound-constrained Formulations of SVM

Recently several authors [Mangasarian and Musicant, 1998; Friess *et al.*, 1998] have proposed different methods for formulating support vector machines. The resulting quadratic program is a bound-constrained problem:

$$\begin{aligned} \min\ & \frac{1}{2} x^T Q x - e^T x \\ & 0 \le x_i \le C, \qquad i = 1, \dots, n. \end{aligned}$$

Note that the matrix Q here is not the same as the matrix Q in (1.1). In this section we would like to demonstrate that if the proposed algorithm in Section 2 is applied to this bound-constrained problem, the convergence proof can still follow.

Consider a more general bound-constrained problem in the following form:

$$\min f(x), \quad x \in \Omega, \tag{4.1}$$

where $f(x)$ is a continuously differentiable function from R^n to R and:

$$\Omega \equiv \{x \mid 0 \le x_i \le C, i = 1, \dots, n\}. \tag{4.2}$$

If $\tilde{OP}(x_k)$ and $OP(x_k)$ are defined by the same way, now it is much easier to establish a relation between them. Instead of Lemma 2.3, we have:

Lemma 4.1 $\lceil \frac{n}{q} \rceil \tilde{OP}(x_k) \le OP(x_k)$.

Proof. First we assume d^* is an optimal solution of (2.7). We then divide $\{1, \dots, n\}$ to $\lceil \frac{n}{q} \rceil$ subsets: $I_1, \dots, I_{\lceil n/q \rceil}$. Let $I_1 = B_k$ and we distribute the rest elements to $I_2, \dots, I_{\lceil n/q \rceil}$ such that each subset has no more than q elements. If we define vectors $d^i, i = 1, \dots, \lceil \frac{n}{q} \rceil$ as

$$d_j^i = \begin{cases} d_j^* & j \in I_i \\ 0 & \text{otherwise}, \end{cases}$$

then all $d^i, i = 1, \dots, \lceil \frac{n}{q} \rceil$, are feasible solutions of (2.3). Thus,

$$\nabla f(x_k)^T d^i \ge \tilde{OP}(x_k), \forall i = 1, \dots, \lceil \frac{n}{q} \rceil.$$

Therefore,

$$\begin{aligned} & OP(x_k) = \nabla f(x_k)^T d^* \\ =\ & \sum_{i=1}^{\lceil n/q \rceil} \nabla f(x_k)^T d^i \ge \lceil \frac{n}{q} \rceil \tilde{OP}(x_k). \end{aligned}$$

$\square$

It can be clearly seen that Lemmas 3.1 and 3.2, and Theorem 3.3 are still true without any modification. Then in the proof of Theorem 3.4, we use Lemma 4.1 to obtain the convergence result.

5 Conclusion

In this paper we discuss the convergence of decomposition methods for support vector machines. The proof is based on the connection between a more general algorithm and the techniques of projected gradients. We also demonstrate that the proof can be easily applied to general bound-constrained formulations of SVM. Investigating the rate of convergence and practical numerical behaviors are topics for future research.

Acknowledgments

The third author thanks Dr. Jorge Moré for bringing him to the subject of support vector machines and some very helpful discussions.

References

[Bertsekas, 1976] D. P. Bertsekas. On the Goldstein-Levitin-Polyak gradient projection method. *IEEE Trans. Automat. Control*, 21:174–184, 1976.

[Calamai and Moré, 1987] P. H. Calamai and J. J. Moré. Projected gradient methods for linearly constrained problems. *Math. Programming*, 39:93–116, 1987.

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

[Friess *et al.*, 1998] T. T. Friess, N. Cristianini, and C. Campbell, The kernel adatron algorithm: a fast and simple learning procedure for support vector machine. In *Proc. 15th International Conference on Machine Learning,* Morgan Kaufman Publishers, 1998.

[Joachims, 1998] T. Joachims. *Making large-scale SVM learning practical*, chapter 11. In Schölkopf et al. [1998], 1998.

[Kaufman, 1998] L. Kaufman. *Solving the quadratic programming problem arising in support vector classification*, chapter 10. In Schölkopf et al. [1998], 1998.

[Mangasarian and Musicant, 1998] O. L. Mangasarian and D. R. Musicant. Successive overrelaxation for support vector machines. Technical report, University of Wisconsin, 1998.

[Osuna *et al.*, 1997a] E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In *IEEE NNSP'97*, 1997.

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

[Platt, 1998] J. C. Platt. *Fast training of support vector machines using sequential minimal optimization*, chapter 12. In Schölkopf et al. [1998], 1998.

[Saunders *et al.*, 1998] C. Saunders, M. O. Stitson, J. Weston, L. Bottou, B. Schölkopf, and A. J. Smola, Support Vector Machine Reference Manual. Technical report, Royal Holloway College, London, UK, 1998.

[Schölkopf *et al.*, 1998] B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors. *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.

[Vapnik, 1998] V. Vapnik. *Statistical Learning Theory*. John Wiley, 1998.