Convergence of a Generalized SMO Algorithm for SVM Classifier Design

S.S. Keerthi and *E.G. Gilbert*

Dept. of Mechanical and Production Engineering
National University of Singapore
Singapore 119260

Dept. of Aerospace Engineering
University of Michigan
Ann Arbor, MI 48109-2140

Abstract

Convergence of a generalized version of the modified SMO algorithms given by Keerthi et.al. for SVM classifier design is proved. The convergence results are also extended to modified SMO algorithms for solving ν-SVM classifier problems.

1 Introduction.

Platt's Sequential Minimization Algorithm (SMO) (Platt, 1998) is a simple and efficient algorithm for solving the quadratic programming problem arising in support vector machines. Recently Keerthi et.al. (Keerthi et.al, 1999) pointed out a problem caused by the way SMO maintains and updates a single threshold value and suggested two modified versions of SMO that overcome this problem. Their comparison on benchmark datasets showed that the modified algorithms performed significantly faster than the original SMO in most situations. But convergence results have not been established for these algorithms, thus far. The general convergence results for asymptotic algorithms proved by Chang et.al. (Chang et.al, 1999) do not apply to SMO. This is because, for SMO, the choice of the working set at each iteration is simply based on the 'rate of change' of the objective function, whereas the choice in Chang et.al's algorithm is more complicated due to the inclusion of the 'extent of movement to the actual constraint set boundary'. In this paper we prove convergence of a generalized SMO algorithm, which includes Keerthi et.al.'s modified algorithms as special cases.

The paper is organized as follows. In section 2 we formulate the quadratic programming problem, give the generalized SMO algorithm and state the main convergence result. In section

3, details associated with the main minimization step of the algorithm are discussed. They form the basis for the proof of convergence, which is given in section 4. The generalized SMO algorithm and convergence proof can be easily extended to the ν-SVM formulations of Schölkopf et.al. (Schölkopf et.al, 1998, Schölkopf et.al, 1999). These extensions are discussed in section 5. Some concluding remarks are given in section 6.

2 Generalized SMO and Its Convergence.

Consider the convex quadratic programming problem,

$$\min f(\alpha) = \frac{1}{2}\alpha^T Q\alpha + p^T\alpha \quad \text{s.t. } a_i \le \alpha_i \le b_i \ \forall i; \ \sum_i y_i\alpha_i = c \qquad (QP)$$

where T denotes transpose, Q is symmetric and positive semi definite, $a_i < b_i \ \forall i$ ($a_i = -\infty$ and/or $b_i = \infty$ are allowed) and $y_i \neq 0 \ \forall i$. Let $\mathcal{F}$ denote the feasible set of QP. We will assume that $\mathcal{F}$ is non-empty and f is bounded below on $\mathcal{F}$. These assumptions imply that QP has an optimal solution.

The dual problem arising in SVM classifier design is a special case of QP in which $a_i = 0$, $b_i = C$, $C > 0$, $y_i \in \{+1, -1\}$, $c = 0$, $p_i = -1 \ \forall i$ and $Q_{ij} = y_i y_j K(x_i, x_j) \ \forall i, j$ where x_k is the k-th input training pattern and K is the kernel function satisfying Mercer's condition. Clearly, for this QP, $\mathcal{F}$ is non-empty. Since $\mathcal{F}$ is compact, f is bounded below. Hence the required assumptions hold.

For QP the KKT conditions are both necessary and sufficient. To write down the KKT conditions, let us define the lagrangian

$$L = \frac{1}{2}\alpha^T Q\alpha + p^T\alpha - \sum_i \delta_i(\alpha_i - a_i) + \sum_i \mu_i(\alpha_i - b_i) - \beta(\sum_i \alpha_i y_i - c)$$

Define

$$F_i(\alpha) = ([Q\alpha]_i + p_i)/y_i$$

where $[Q\alpha]_i$ denotes the i-th element of $Q\alpha$. The KKT conditions for QP are:

$$\frac{\partial L}{\partial \alpha_i} = (F_i - \beta)y_i - \delta_i + \mu_i = 0, \quad \delta_i \ge 0, \quad \delta_i(\alpha_i - a_i) = 0, \quad \mu_i \ge 0, \quad \mu_i(\alpha_i - b_i) = 0 \ \ \forall i$$

These conditions can be simplified by considering three cases for each i.

Case 1. $\alpha_i = a_i$

$$(F_i - \beta)y_i \geq 0 \tag{1a}$$

Case 2. $a_i < \alpha_i < b_i$

$$(F_i - \beta)y_i = 0 \tag{1b}$$

Case 3. $\alpha_i = b_i$

$$(F_i - \beta)y_i \leq 0 \tag{1c}$$

Define the following index sets: $I_0(\alpha) = \{i : a_i < \alpha_i < b_i\}$; $I_1(\alpha) = \{i : y_i > 0, \alpha_i = a_i\}$; $I_2(\alpha) = \{i : y_i < 0, \alpha_i = b_i\}$; $I_3(\alpha) = \{i : y_i > 0, \alpha_i = b_i\}$; and, $I_4(\alpha) = \{i : y_i < 0, \alpha_i = a_i\}$. Let us also define:

$$I_{\text{up}}(\alpha) = I_0(\alpha) \cup I_1(\alpha) \cup I_2(\alpha) \; ; \quad I_{\text{low}}(\alpha) = I_0(\alpha) \cup I_3(\alpha) \cup I_4(\alpha)$$

Then the conditions in (1a)-(1c) can be rewritten as

$$\beta \leq F_i(\alpha) \; \forall \, i \in I_{\text{up}}(\alpha) \; ; \; \; \beta \geq F_i(\alpha) \; \forall \, i \in I_{\text{low}}(\alpha) \tag{2}$$

It is easily seen that KKT conditions will hold at $\alpha \in \mathcal{F}$ iff there exists a β satisfying (2).

We will say that (i, j) is a *violating pair at* α if *one* of the following two sets of conditions holds:

$$i \in I_{\text{up}}(\alpha) \; , \; \; j \in I_{\text{low}}(\alpha) \; \text{ and } \; F_i(\alpha) < F_j(\alpha) \tag{3a}$$

$$i \in I_{\text{low}}(\alpha) \; , \; \; j \in I_{\text{up}}(\alpha) \; \text{ and } \; F_i(\alpha) > F_j(\alpha) \tag{3b}$$

Note that optimality conditions will hold at α iff there does not exist a violating pair at α.

Since SMO algorithms generally do not provide an exact solution in a finite number of steps, there is need to define approximate optimality conditions. The condition (2) can be replaced by

$$\beta \leq F_i(\alpha) + \frac{\tau}{2} \; \forall \, i \in I_{\text{up}}(\alpha) \; ; \; \; \beta \geq F_i(\alpha) - \frac{\tau}{2} \; \forall \, i \in I_{\text{low}}(\alpha) \tag{4}$$

where τ is a positive tolerance parameter. If (4) holds we say α is a *τ-optimal solution*. The approximate optimality conditions in (4) are closely related to the stopping conditions used by Platt (Platt, 1998); see Keerthi et.al, 1999 for details.

Let α be a τ-optimal solution and $f^\star$ be the optimal objective function value of QP. Then, for the QP arising from SVM classifier design it can be shown (using duality gap ideas) that $f(\alpha) - f^\star$ is bounded above by $\psi(\tau)$ where ψ is a continuous function with the property that $\psi(\tau) \to 0$ as $\tau \to 0$. Hence, by choosing τ small enough, desired closeness between $f(\alpha)$ and $f^\star$ can be achieved.

Corresponding to (4), the definition of violation can be altered by replacing (3a) and (3b) by:

$$i \in I_{\text{up}}(\alpha)\ ,\ \ j \in I_{\text{low}}(\alpha) \ \text{ and } \ F_i(\alpha) < F_j(\alpha) - \tau \tag{5a}$$

$$i \in I_{\text{low}}(\alpha)\ ,\ \ j \in I_{\text{up}}(\alpha) \ \text{ and } \ F_i(\alpha) > F_j(\alpha) + \tau \tag{5b}$$

If one of (5a) or (5b) holds we will say that (i,j) is a *τ-violating pair at α*. Clearly, α is τ-optimal iff there is no τ-violating pair at α. An easy way of checking (4) is:

$$\min_{i \in I_{\text{up}}(\alpha)} F_i(\alpha) \geq \max_{i \in I_{\text{low}}(\alpha)} F_i(\alpha) - \tau \tag{6}$$

Using this background let us give a general algorithm for solving QP.

Algorithm GSMO. *Let $\tau > 0$ be given.*

0. Choose some $\alpha \in \mathcal{F}$. Set $k = 0$, $\alpha(0) = \alpha$.
1. If $\alpha(k)$ satisfies (6) stop.
2. Choose $(i(k), j(k))$, a τ-violating pair at $\alpha(k)$. Minimize f on $\mathcal{F}$ while varying only $\alpha_{i(k)}$ and $\alpha_{j(k)}$. Let $\alpha(k+1)$ be the point thus obtained. Set $k := k+1$ and go to step 1.

Let us now briefly discuss the two modified SMO algorithms given by Keerthi et.al. (Keerthi et.al, 1999). Since the α_i's that take bound values (i.e., $\alpha_i \in \{a_i, b_i\}$) at optimality are usually identified easily, the main effort of the solution is associated with choosing correct values for α_i, $i \in I_0$. Hence the algorithms choose $\{i(k), j(k)\}$ to be a subset of I_0 in a large fraction of the iterations.[1] The two modified algorithms differ in the way $i(k)$ and $j(k)$ are chosen in those iterations. The first modification sequentially runs through indices in I_0 to choose $i(k)$; then, given $i(k)$ it chooses $j(k)$ greedily by maximizing $|F_{i(k)} - F_{j(k)}|$. On the other hand, the second modification chooses both $i(k)$ and $j(k)$ greedily to maximize $|F_{i(k)} - F_{j(k)}|$. After a stage is reached when no τ-violating pair can be chosen from I_0, all indices are involved in choosing a τ-violating pair. This is done by sequentially going through all indices for choosing $i(k)$ and then

[1]This important heuristic is due to Platt (Platt, 1998). The updates are efficiently done using cache for F_i, $i \in I_0$.

choosing $j(k)$ greedily for each given $i(k)$. The whole process is repeated until no τ-violating pair exists. The implementation, which is done carefully and efficiently using a cache for F_i, $i \in I_0$, is fully explained in Keerthi et.al, 1999. The key point to be noted here is that the two modified SMO algorithms are special instances of the GSMO algorithm.

The main result of this paper concerns the convergence of the GSMO algorithm.

Theorem 1. Algorithm GSMO stops at step 1 after a finite k. ■

To simplify the proof of the theorem we assume hereafter that $y_i > 0$ for all i. This represents no loss of generality. Problem QP satisfies this assumption when it is modified by reversing the signs of both α_i and y_i when $y_i < 0$. Moreover, when GSMO is applied to the original QP and the modified QP it produces equivalent steps.

3 The Minimization Step.

In this section we describe what happens in step 2 of GSMO. The required minimization takes place in the rectangle $S = [a_i, b_i] \times [a_j, b_j]$ along paths where $\alpha_i y_i + \alpha_j y_j$ is constant. In this section we shift our usage of α and $\alpha(\cdot)$; the initial τ-violating $\alpha(k)$ is denoted by α and the parametric change in α is given by $\alpha(t)$. Thus

$$\alpha_i(t) = \alpha_i + t/y_i \ , \quad \alpha_j(t) = \alpha_j - t/y_j \ , \quad \alpha_k(t) = \alpha_k \ \forall k \neq i, j,$$

To further simplify notations let $(\alpha_i, \alpha_j) = \alpha_{ij}$. The objective is to minimize $\phi(t) = f(\alpha(t))$ subject to $\alpha_{ij}(t) \in S$. It is easy to confirm that $\phi(t) = \phi(0) + \phi'(0)t + \phi''(0)t^2/2$ where $\phi'(0) = F_i(\alpha) - F_j(\alpha)$ and $\phi''(0) = \frac{q_{ii}}{y_i^2} + \frac{q_{jj}}{y_j^2} - 2\frac{q_{ij}}{y_i y_j}$. By the positive semi definiteness of Q it follows that $\phi''(0) \geq 0$.

Figure 1 illustrates the main possibilities. We use the word "edge" to denote the set of boundary points of S that lie between corner points of S.[2] The t-paths in S have a negative slope. The $\pm$ signs designate the sign of t required for $\alpha_{ij}(t) \in S$. Descent takes place in the directions indicated by the arrows. Since $i, j \in I_3(\alpha)$ for α_{ij} at B and $i, j \in I_1(\alpha)$ for α_{ij} at C, the corner points B and C cannot be τ-violating. Everywhere else in S, τ-violation is possible.

[2] Although Figure 1 has been drawn assuming that a_i, b_i, a_j and b_j are finite, all elements of proof given in this section and the next section also apply to the case where one or more of these four values is not finite.

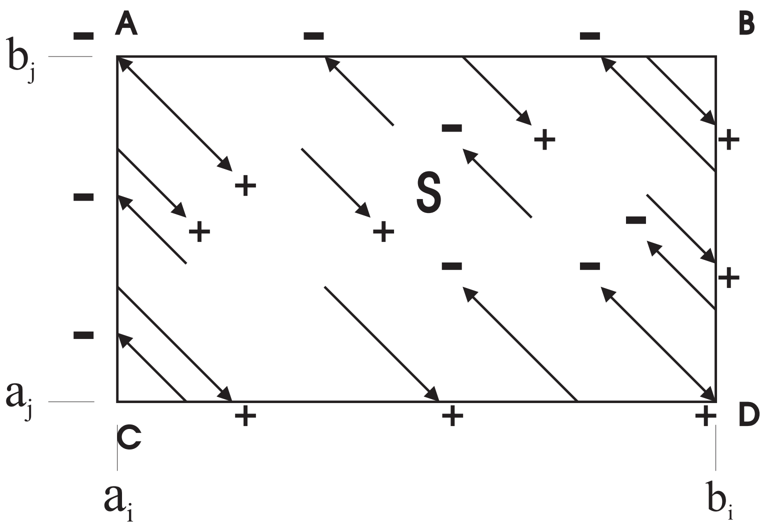

Figure 1: Minimization steps in $S = [a_i, b_i] \times [a_j, b_j]$

Specifically, (5a) can only occur in the interior of S, on edges AB and AC or at corner point A; (5b) can only occur in the interior of S, on edges BD and CD or at corner point D.

Consider, for example, the case where α_{ij} is in the edge AC. It follows that $i \in I_1(\alpha) \subset I_{\rm up}(\alpha)$, $j \in I_0(\alpha) \subset I_{\rm low}(\alpha)$ and $\phi'(0) < -\tau$. The minimum of $\phi(t)$ is reached at $t^\star > 0$ where either $\alpha_{ij}(t^\star) \in \text{int } S$ or $\alpha_{ij}(t^\star) \in \text{bd } S$. The first alternative implies $F_i(\alpha(t^\star)) - F_j(\alpha(t^\star)) = 0$; the second alternative implies $F_i(\alpha(t^\star)) - F_j(\alpha(t^\star)) \leq 0$ and, because the slope of the path is negative, $\alpha_{ij}(t^\star)$ is in the set defined by the union of edges BD and CD and corner D. Thus, after the minimization step the pair (i, j) satisfies $j \in I_{\rm up}(\alpha(t^\star))$, $i \in I_{\rm low}(\alpha(t^\star))$ and it is no longer τ-violating. Of course, $\alpha_{ij}(t^\star) \in \text{int } S$ implies $\phi'(t^\star) = 0$.

Lemma 1. Let $\alpha \in \mathcal{F}$ and (i, j) be a τ-violating pair at α. Let $\alpha_{\rm new}$ be the solution obtained during the minimization step. Then after the minimization step the following results hold: (a) $\alpha_{\rm new} \neq \alpha$; (b) (i, j) is not a τ-violating pair at $\alpha_{\rm new}$; (c) if $\alpha_{ij\ \rm new} \in \text{int } S$, then $F_i(\alpha_{\rm new}) - F_j(\alpha_{\rm new}) = 0$; (d)

$$f(\alpha) - f(\alpha_{\rm new}) \geq \frac{\tau}{2\gamma_{ij}} \|\alpha_{\rm new} - \alpha\| \tag{7}$$

where $\gamma_{ij} = \sqrt{(y_i)^{-2} + (y_j)^{-2}}$ and $\|\cdot\|$ is the euclidean norm. ■

Proof. Parts (a), (b) and (c) are obvious consequences of the preceding discussion. To prove part (d) we first we show that

$$\phi(t^\star) \leq \phi(0) + \frac{1}{2}\phi'(0)t^\star \tag{8}$$

Assume $t^\star > 0$ and $\phi'(0) < 0$; essentially the same argument applies if $t^\star < 0$ and $\phi'(0) > 0$. Suppose $\phi''(0) = 0$. Then $t^\star$ is determined by the point at which the parametric path reaches the boundary of S. Further, (8) holds trivially since $\phi(t)$ is linear with slope $\phi'(0)$. Suppose $\phi''(0) > 0$. Let $t_Q = -\phi'(0)/\phi''(0)$ be the unconstrained minimum of $\phi(t)$. Clearly, $0 < t^\star < t_Q$. This, together with the fact that the line joining the points $(0, \phi(0))$ and $(t_Q, \phi(t_Q))$ in the (t, ϕ) coordinate system is an upper bound for $\phi(t)$ in the interval $[0, t_Q]$, yields (8). Since $\|\alpha - \alpha_{\text{new}}\| = \|\alpha_{ij} - \alpha_{ij}(t^\star)\| = |t^\star|\gamma_{ij}$, $f(\alpha) = \phi(0)$ and $f(\alpha_{\text{new}}) = \phi(t^\star)$, result (8) yields (7). ■

4 Proof of the Convergence Theorem.

We assume that algorithm GSMO proceeds indefinitely, i.e., the pair $(i(k), j(k))$ is τ-violating at $\alpha(k)$ for all $k \geq 0$. We will show that this leads to a contradiction. Since $f(\alpha(k))$ is a decreasing sequence that is bounded from below, there exists $\bar{f}$ such that $f(\alpha(k)) \to \bar{f}$. By (7)

$$\frac{2\gamma}{\tau}[f(\alpha(k)) - f(\alpha(k+1))] \geq \|\alpha(k) - \alpha(k+1)\| \;\; \forall k \geq 0$$

where $\gamma = \max\{\gamma_{ij} : i \neq j\}$. By repeated application of the triangle inequality we get

$$\frac{2\gamma}{\tau}[f(\alpha(k)) - f(\alpha(k+l))] \geq \|\alpha(k) - \alpha(k+l)\| \quad \forall k,\ l \geq 0$$

Thus $\{\alpha(k)\}$ is a cauchy sequence. Since $\mathcal{F}$ is closed, $\{\alpha(k)\}$ converges to some $\bar{\alpha} \in \mathcal{F}$.

In what follows we use notations such as $\{k_t\}$ to denote sequences of integers that have the form $\{k_t : k_{t+1} > k_t \geq 0, t \geq 0\}$. Let

$$I_\infty = \{(\mu, \nu) : \exists \{k_t\} \ni (i(k_t), j(k_t)) = (\mu, \nu)\ \forall t \geq 0\} \tag{9}$$

Clearly, I_∞ is the set of all index pairs that are encountered infinitely many times.

Lemma 2. Assume $(\mu, \nu) \in I_\infty$ and let $\{k_t\}$ be the sequence of all those indices k that satisfy $(i(k), j(k)) = (\mu, \nu)$. Then, the following results hold: (a) given any $\epsilon > 0$ there exists a $\hat{t}$ such that $\|\alpha_{\mu\nu}(k_t) - \bar{\alpha}_{\mu\nu}\| < \epsilon$ and $\|\alpha_{\mu\nu}(k_t + 1) - \bar{\alpha}_{\mu\nu}\| < \epsilon\ \forall\ t \geq \hat{t}$; (b) $|F_\mu(\bar{\alpha}) - F_\nu(\bar{\alpha})| \geq \tau$. ■

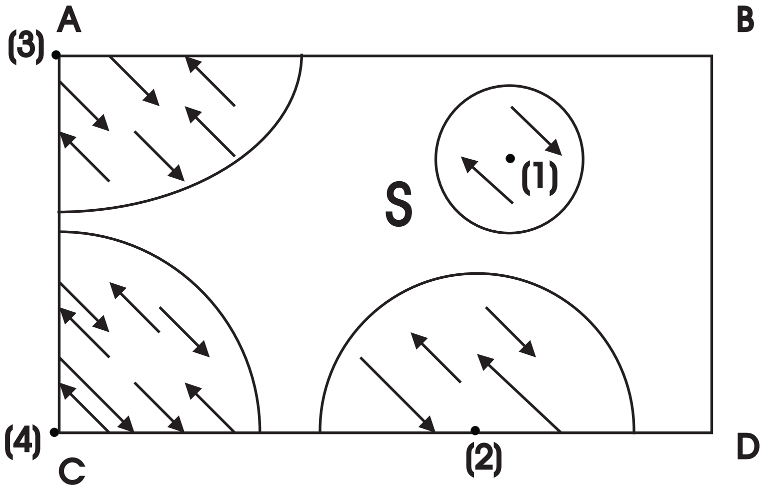

Figure 2: Possible minimization steps in $\bar{B}_\epsilon \cap S$ for cases (1) to (4) corresponding to different placements of $\bar{\alpha}_{\mu\nu}$

Proof. Since $\{k_t\}$ is an infinite sequence, it is possible to choose $\hat{t}$ so that $k_{\hat{t}}$ is as large as we please. Choose it so that $\|\alpha(k) - \bar{\alpha}\| \le \epsilon$ for all $k \ge k_{\hat{t}}$. This implies (a) is satisfied. Result (b) follows from (5), the continuity of F_μ and F_ν and $\alpha(k_t) \to \bar{\alpha}$ as $t \to \infty$. ■

Figure 2 shows the set $S = [a_\mu, b_\mu] \times [a_\nu, b_\nu]$ and illustrates some possibilities that represent the behavior of the minimization step at large k. Below we will refer to A and D as τ-violating corners, and B and C as non τ-violating corners. There are four cases for the location of $\bar{\alpha}_{\mu\nu}$: (1) $\bar{\alpha}_{\mu\nu} \in$ int S; (2) $\bar{\alpha}_{\mu\nu} \in$ edge of S; (3) $\bar{\alpha}_{\mu\nu} \in \tau$-violating corner of S; (4) $\bar{\alpha}_{\mu\nu} \in$ non τ-violating corner of S. Let $\bar{B}_\epsilon = \{\alpha_{\mu\nu} : \|\alpha_{\mu\nu} - \bar{\alpha}_{\mu\nu}\| < \epsilon\}$. For each case we assume ϵ is chosen sufficiently small so that $\bar{B}_\epsilon \cap S$ includes only the geometric features (interior points, edges, corner points) of S that are associated with $\bar{\alpha}_{\mu\nu}$. The minimization step generates transitions from $\alpha_{\mu\nu}(k_t) \in \bar{B}_\epsilon \cap S$ to $\alpha_{\mu\nu}(k_t + 1) \in \bar{B}_\epsilon \cap S$. The possible transitions for each case are indicated in figure 2 by directed line segments. We use a simplifying locution to describe how the transitions take place between the sets int S and bd S. For example, "k_t is int $\to$ bd" means $\alpha_{\mu\nu}(k_t) \in$ int S, $\alpha_{\mu\nu}(k_t + 1) \in$ bd S.

Lemma 3. Let (μ, ν), $\{k_t\}$, $\hat{t}$, S and $\epsilon > 0$ be determined as described in Lemma 2 and the preceding paragraph. Then, there exists a $\tilde{t} \ge \hat{t}$ such that for each $t \ge \tilde{t}$, k_t is either int $\to$ bd or bd $\to$ bd. ■

Proof. Using the results of Figure 2 we consider, for each case of $\bar{\alpha}_{\mu\nu}$, all possible transitions and their implications. In case (1), k_t is int $\to$ int for all $t \ge \hat{t}$. Thus, by part (c) of

Lemma 1 and $\alpha(k_t+1) \to \bar{\alpha}$, it follows that $F_\mu(\bar{\alpha}) - F_\nu(\bar{\alpha}) = 0$. This contradicts part (b) of Lemma 2. Hence, case (1) can not occur. In cases (2) and (3) there are three alternatives for each $t \geq \hat{t}$: k_t is either int $\to$ int, bd $\to$ int, or int $\to$ bd. Suppose there exists a subsequence $\{l_s\} \subset \{k_t : t \geq \hat{t}\}$ such that either l_s is int $\to$ int for all $s \geq 0$ or l_s is bd $\to$ int for all $s \geq 0$. Repeating the argument used in case (1), with $\{l_s\}$ replacing $\{k_t\}$, leads to a contradiction. Thus, cases (2) and (3) can only occur if there exists a $\tilde{t} \geq \hat{t}$ such that k_t is int $\to$ bd for all $t \geq \tilde{t}$. The argument in case (4) proceeds as the one in cases (2) and (3), except we are left with the additional possibility that k_t is bd $\to$ bd. ∎

We are now in a position to complete the proof of Theorem 1. There exists a $\hat{k}$ such that for all $k \geq \hat{k}$, $(i(k), j(k)) \in I_\infty$ and, by Lemma 3, the minimization transitions are either int $\to$ bd or bd $\to$ bd. The transition int $\to$ bd causes the number of components of $\alpha(k)$ that are active (on a constraint boundary) to increase by one or two. The transition bd $\to$ bd occurs only in case (4) and it moves from one edge of S to another edge S. Hence, the transition bd $\to$ bd causes the number of components of $\alpha(k)$ that are active to remain constant. Since the number of active constraints can not increase without bound, it follows that int $\to$ bd transitions cannot occur infinitely many times. Let (μ, ν) be any pair in I_∞. From the preceding result and Lemma 3, there exists $\bar{t} \geq \tilde{t}$ such that k_t is bd $\to$ bd for all $t \geq \bar{t}$. Since bd $\to$ bd transitions occur only when $\bar{\alpha}_{\mu\nu}$ is at a non τ-violating corner of S, $\alpha(k_t)$ alternates between the two edges of S that are adjacent to $\bar{\alpha}_{\mu\nu}$. Hence, there exist subsequences $\{l_s^a\} \subset \{k_t : t \geq \bar{t}\}$ and $\{l_s^b\} \subset \{k_t : t \geq \bar{t}\}$ such that: for all $s \geq 0$, $\alpha = \alpha(l_s^a)$ satisfies (5a) and $\alpha = \alpha(l_s^b)$ satisfies (5b) . Letting $s \to \infty$ in these two results provides our contradiction: $F_\mu(\bar{\alpha}) - F_\nu(\bar{\alpha}) \leq -\tau$, $F_\mu(\bar{\alpha}) - F_\nu(\bar{\alpha}) \geq \tau$.

5 Extensions.

The GSMO algorithm as well as the convergence result in Theorem 1 can be easily extended to other SVM classification formulations. First consider the ν-SVM formulation for estimating the support of a distribution, as given by Schölkopf et.al. (Schölkopf et.al, 1999). Since the dual formulations given in equations (12) and (15)-(16) of Schölkopf et.al, 1999 are directly in the form of QP, GSMO and Theorem 1 apply to them. It is easy to give an efficient practical algorithm for this special QP, along the lines of the modified SMO algorithms in Keerthi et.al, 1999. Such an algorithm is expected to perform even better than the SMO algorithm implemented in

Schölkopf et.al, 1999.

Now consider the ν-SVM classification formulation given in Schölkopf et.al, 1998. The problem can be written as (Crisp and Burges, 1999)

$$\min \frac{1}{2}\alpha^T Q\alpha \;\; \text{s.t.} \;\; 0 \leq \alpha_i \leq \gamma\,, \; \sum_{i\in L} \alpha_i = 1\,, \; \sum_{i\in M} \alpha_i = 1 \qquad (QP1)$$

where $\gamma > 0$, and, L and M are disjoint index sets with $L \cup M = \{1, \dots, m\}$. (Here m is the number of α_i variables, i.e., $\alpha \in R^m$.) Q is symmetric and positive semi definite, as usual. We will briefly explain how GSMO can be extended for this problem.

The lagrangian is

$$L = \frac{1}{2}\alpha^T Q\alpha - \sum_i \delta_i \alpha_i + \sum_i \mu_i(\alpha_i - \gamma) - \mu_L(\sum_{i\in L} \alpha_i - 1) - \mu_M(\sum_{i\in M} \alpha_i - 1)$$

Define: $F_i(\alpha) = [Q\alpha]_i$; $L_0(\alpha) = \{i \in L : 0 < \alpha_i < \gamma\}$; $L_1(\alpha) = \{i \in L : \alpha_i = 0\}$; $L_2(\alpha) = \{i \in L : \alpha_i = \gamma\}$; $M_0(\alpha) = \{i \in M : 0 < \alpha_i < \gamma\}$; $M_1(\alpha) = \{i \in M : \alpha_i = 0\}$; $M_2(\alpha) = \{i \in M : \alpha_i = \gamma\}$; $L_{\rm up}(\alpha) = L_0(\alpha) \cup L_1(\alpha)$; $L_{\rm low}(\alpha) = L_0(\alpha) \cup L_2(\alpha)$; $M_{\rm up}(\alpha) = M_0(\alpha) \cup M_1(\alpha)$; and, $M_{\rm low}(\alpha) = M_0(\alpha) \cup M_2(\alpha)$.

Let us say that α is τ-optimal if

$$\min_{i\in L_{\rm up}(\alpha)} F_i(\alpha) \geq \max_{i\in L_{\rm low}(\alpha)} F_i(\alpha) - \tau \quad \text{and} \quad \min_{i\in M_{\rm up}(\alpha)} F_i(\alpha) \geq \max_{i\in M_{\rm low}(\alpha)} F_i(\alpha) - \tau \qquad (10)$$

As in section 2 it is easily checked that KKT conditions hold at α iff (10) holds for $\tau = 0$. Hence, for $\tau > 0$, (10) is a relaxation of the KKT conditions. We will say (i, j) is a τ-violating pair at α if the following conditions hold: (i) either ($i \in L$ and $j \in L$) or ($i \in M$ and $j \in M$); and (ii) either (5a) or (5b) holds with the I_p sets replaced by the corresponding L_p or M_p sets. Clearly, α is τ-optimal iff there is no τ-violating pair at α.

With these definitions in place, we can use algorithm GSMO to solve QP1. (In step 1, (10) should be used instead of (6).) Theorem 1 holds for this GSMO algorithm too. Except for some rewriting associated with replacing the I_p sets by the M_p and L_p sets the proof of this result is very much similar to that in sections 3 and 4. The key factor that makes the proof go through easily is the disjointness of L and M. For efficiency GSMO can be implemented in a way similar to the algorithms in Keerthi et.al, 1999. We are currently implementing and testing this algorithm.

6 Conclusion.

In this paper we have established convergence results for the modified SMO algorithms related to classification. We believe that extension of the ideas to similar algorithms for regression problems (Shevade et.al, 1999, Schölkopf et.al, 1998) is possible. We are currently working on the details.

Apart from SMO, the decomposition algorithm of Joachims (Joachims, 1998) is another very efficient algorithm for SVMs. Recently Lin (Lin, 2000) has used ideas similar to those in this paper to prove asymptotic convergence of Joachims' algorithm.

Acknowledgements.

We thank Chong Jin Ong for valuable discussions. He read the paper very carefully and made useful suggestions for improving the presentation.

References

Chang,C.C., Hsu,C.W., & Lin,C.J. (2000). The analysis of decomposition methods for support vector machines. *IEEE Transactions on Neural Networks*, Vol. 11, No. 4, pp.1003-1008.

Crisp,D. & Burges,C. (1999). A geometric interpretation of ν-SVM classifiers, *Neural Information Processing Systems Conference*, Denver, CO, USA, 1999.

Joachims,T. (1998). Making large-scale support vector machine learning practical, in B. Schölkopf, C. Burges, A. Smola. *Advances in Kernel Methods: Support Vector Machines*, MIT Press, Cambridge, MA, December 1998.

Keerthi,S.S., Shevade,S.K., Bhattacharyya,C., & Murthy,K.R.K. (1999). Improvements to Platt's SMO algorithm for SVM classifier design, Tech. Report CD-99-14, Control Division, Dept. of Mechanical and Production Engineering, National University of Singapore.

Lin,C.J. (2000). On the convergence of the decomposition method for support vector machines, Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, Sept. 2000.

Platt,J.C. (1998). Fast training of support vector machines using sequential minimal optimization, in B. Schölkopf, C. Burges, A. Smola. *Advances in Kernel Methods: Support Vector Machines*, MIT Press, Cambridge, MA.

Schölkopf,B., Smola,A.J., Williamson,R., & Bartlett,P. (1998). New support vector algorithms, Neuro COLT Technical Report TR-1998-031, Royal Holloway College.

Schölkopf,B., Platt,J., Shawe-Taylor,J., Smola,A.J., & Williamson,R.C. (1999). Estimating the support of a high-dimensional distribution, TR 99-87, Microsoft Research.

Shevade,S.K., Keerthi,S.S., Bhattacharyya,C., & Murthy,K.R.K. (1999). Improvements to the SMO algorithm for SVM regression, Tech. Report CD-99-16, Control Division, Dept. of Mechanical and Production Engineering, National University of Singapore.