A convergence rate estimate for the SVM Decomposition Method

D.Lai *, A.Shilton†, N.Mani*, M.Palaniswami†
* Department of Electrical and Computer Systems Engineering,
Monash University, Clayton, Vic. 3168, Australia.
† Department of Electrical and Electronic Engineering,
The University of Melbourne Vic. 3010, Australia.
daniel.lai@eng.monash.edu.au

Abstract—The training of Support Vector Machines using the decomposition method has one drawback; namely the selection of working sets such that convergence is as fast as possible. It has been shown by Lin that the rate is linear in the worse case under the assumption that all bounded Support Vectors have been determined. The analysis was done based on the change in the objective function and under a SVMlight selection rule. However, the rate estimate given is independent of time and hence gives little indication as to how the linear convergence speed varies during the iteration.

In this initial analysis, we provide a treatment of the convergence from a gradient contraction perspective. We propose a necessary and sufficient condition which when satisfied provides strict linear convergence of the algorithm. The condition can also be derived. This estimate is shown to monotonically approach this condition, a time dependant rate estimate is then further sets in order to achieve such a convergence rate. Based on the change in the objective function and under a SVMlight selection rule. However, the rate estimate given is independent of time and hence gives little indication as to how the linear convergence speed varies during the iteration.

We propose a necessary and sufficient condition which when satisfied provides strict linear convergence of the algorithm. The condition can also be derived. This estimate is shown to monotonically approach this condition, a time dependant rate estimate is then further sets in order to achieve such a convergence rate. Based on the change in the objective function and under a SVMlight selection rule.

The data to be classified is usually written as and the kernel function which satisfies Mercer’s conditions is:

\[ K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \]

The separating hyperplane has the following form:

\[ f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \]

where \( b \) is the bias of the hyperplane. The SVM problem is usually solved via sequential decomposition, where the following partitioned problem is solved at each iteration.

\[
\begin{align*}
\min_{\alpha_p} \max_b & \sum \alpha_i (y_i \alpha_i - b) \\
\text{subject to:} & \quad \sum \alpha_i = \frac{1}{2} \alpha_p^T H_p \alpha_p + H_s \alpha_s \end{align*}
\]

(3)

subject to:

\[ D_p = \{ \alpha_p | 0 \leq \alpha_p \leq C \} \]

where \( b \in \mathbb{R} \), is treated as a pseudo-Lagrangian multiplier [2]. The partitioned matrices are

\[
G = \begin{bmatrix} G_p & G_g \\ G_g^T & G_s \end{bmatrix}, \quad H_p = \begin{bmatrix} G_p \\ y_p \end{bmatrix}, \quad H_s = \begin{bmatrix} G_s \\ y_s \end{bmatrix}
\]

and the augmented vectors are

\[
\alpha_p' = \begin{bmatrix} \alpha_p \\ b \end{bmatrix}, \quad e_p' = \begin{bmatrix} e_p \\ 0 \end{bmatrix}
\]

Note that the gradient vector of (2) is

\[ v' = H \alpha' - e' \]

while the gradient vector corresponding to the sub-problem is

\[ v_p' = H_p \alpha_p' + H_s \alpha_s' - e_p' \]

(5)

This form of decomposition adds an extra dimension of difficulty. In particular, the sequence of sub-problems solved is central to the speed of the SVM decomposition method [3]. In early SVM decomposition algorithms such as

\[
\begin{align*}
\min_{\alpha_p} \max_b & \sum \alpha_i (y_i \alpha_i - b) \\
\text{subject to:} & \quad \sum \alpha_i = \frac{1}{2} \alpha_p^T H_p \alpha_p + H_s \alpha_s \end{align*}
\]

(3)

subject to:

\[ D_p = \{ \alpha_p | 0 \leq \alpha_p \leq C \} \]

where \( b \in \mathbb{R} \), is treated as a pseudo-Lagrangian multiplier [2]. The partitioned matrices are

\[
G = \begin{bmatrix} G_p & G_g \\ G_g^T & G_s \end{bmatrix}, \quad H_p = \begin{bmatrix} G_p \\ y_p \end{bmatrix}, \quad H_s = \begin{bmatrix} G_s \\ y_s \end{bmatrix}
\]

and the augmented vectors are

\[
\alpha_p' = \begin{bmatrix} \alpha_p \\ b \end{bmatrix}, \quad e_p' = \begin{bmatrix} e_p \\ 0 \end{bmatrix}
\]

Note that the gradient vector of (2) is

\[ v' = H \alpha' - e' \]

while the gradient vector corresponding to the sub-problem is

\[ v_p' = H_p \alpha_p' + H_s \alpha_s' - e_p' \]

(5)

This form of decomposition adds an extra dimension of difficulty. In particular, the sequence of sub-problems solved is central to the speed of the SVM decomposition method [3]. In early SVM decomposition algorithms such as...
SMO [4] and SVMlight [5] the sequence of sub-problems has consisted of those with the steepest gradients. It has been demonstrated by Lin [7] that sub-problems chosen under a SVMlight rule resulted in a linear convergence rate. In [8], we demonstrated that in addition to the steepest direction, one should also consider the sub-problem constraints and select sub-problems with the largest potential change in the variables. We are interested in the problem of determining an optimal sequence so that the quadratic program is solved in minimal time. The idea then is to look for clues from the theoretical behaviour of the decomposition method.

The purpose of this paper is two-fold. We first investigate the necessary and sufficient conditions for convergence which requires less assumptions than Lin [6] and allows for a more general update rule. The next aim is to derive a relationship between the convergence rates and the choice of sub-problem solved. This ends up being formulated as a rate estimate for the speed of convergence of the decomposition method. We then demonstrate that strict convergence implies a linear rate which approaches unity from below. In the next section, we introduce several notations and definitions. Section III contains our analysis on convergence of the gradient vector under a generalized update rule. Throughout, we adhere to the standard notation where scalars are denoted by italics, column vectors by boldface small letters and we adhere to the standard notation where scalars are denoted gradient vector under a generalized update rule. Throughout, a linear rate which approaches unity from below. In the method. We then demonstrate that strict convergence implies an estimate for the speed of convergence of the decomposition sub-problem solved. This ends up being formulated as a rate relationship between the convergence rates and the choice of a more general update rule. The next aim is to derive a theoretical behaviour of the decomposition method.

Let an arbitrary iteration step be \( t \) where \( t \in T \subset \mathbb{Z}^+ \). It is known that the problem gradients must satisfy the Karush-Kuhn-Tucker (KKT) optimality conditions for the quadratic program to be solved [1], [9]. If \( v' \) denotes the optimal gradient then each individual gradient must satisfy the following KKT conditions

\[
\begin{align*}
\bar{v}_i & \leq 0 \quad \text{if } \alpha_i = C \\
\bar{v}_i & \geq 0 \quad \text{if } \alpha_i = 0 \\
\bar{v}_i & = 0 \quad \text{if } 0 < \alpha_i < C \\
\bar{v}_b & = 0
\end{align*}
\]

Violaters [4] are points with gradients which do not satisfy these conditions. For simplicity, we assume that the optimal gradient \( \bar{v}' \) is unique which implies that the hessian, \( H \) is also assumed to be positive definite. The following is a well known result.

Lemma 2.1 (Cauchy-Schawrz Inequality): For any two vectors \( x, y \in \mathbb{R}^n \) we have the following

\[
| \langle x, y \rangle | \leq \| x \| \| y \|
\]

where \( \langle \cdot, \cdot \rangle \) is the normal dot product and the norm is Euclidean. Equality is achieved if and only if \( x = cy \) for some scalar \( c \). (proofs can be found in e.g. [10], [11]).

The Frobenius norm or Euclidean matrix norm is defined for a \( m \times n \) matrix \( H \) as

\[
\| H \|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} \| H_{ij} \|^2}
\]

When \( n = 1 \) this becomes the standard \( L^2 \)-norm for vectors also known as the Euclidean vector norm.

III. THE CONVERGENCE BEHAVIOUR OF THE DECOMPOSITION METHOD

In this section, we investigate the convergence properties of the gradient vector with the aim of arriving at a rate estimate that does not require the assumptions of Lin [6]. The first step is to remove any explicit dependance of the analysis on the type of update rule. Let \( S \) denote the set of mappings (or update rules) where for some \( \alpha'_p \in \mathbb{A} \) and \( t > 0 \), we select a mapping \( s(\alpha'_p) \in S \) such that \( s(\alpha'_p) : \mathbb{R}^{m+1} \rightarrow \mathbb{R}^{m+1} \). The set \( S \) could for example be the set of Newton directions, line search directions [9], [12] or the set of all normalized search directions [13]. In order to generalize this, we define instead the vector of differences

\[
\Delta \alpha'_p = \alpha'^{t+1}_p - \alpha'^t_p
\]

where the choice of mapping is implicitly implied via

\[
\Delta \alpha'_p = s(\alpha'_p) \in S
\]
The decomposition method can then be viewed as the following law for all $t > 0$ and working set $\alpha'_{\beta}$

$$
\alpha_{t+1}^i = \begin{cases} 
\alpha_i^t & \text{if } \alpha_i^t \notin \alpha'_{\beta} \\
\alpha_i^t + s(\alpha_i^t) & \text{if } \alpha_i^t \in \alpha'_{\beta}
\end{cases}
$$

(8)

First we note from (4) the individual gradient through a Lagrangian multiplier $\alpha_i$ is

$$
v_i = y_i \sum_{j=1}^{n} \alpha_j y_j K(x_i, x_j) + b y_i
$$

(9)

The gradient of the pseudo-Lagrangian $b$ is just

$$
v_b = \sum_{j=1}^{n} \alpha_j y_j
$$

(10)

which is the equality constraint in (2).

We begin with two basic results. The first is trivial and is meant to show that the norm of the change in the vector of variables, $\alpha' \in \mathbb{R}^{n+1}$ is the same as the norm of (7). The next result simplifies the notation and allows the analysis to be done in terms of dot products.

**Lemma 3.1:** Let $\alpha'_{\beta} \subseteq \mathbb{A}$ be an arbitrary working set updated at some time $t > 0$. Suppose that $\alpha' \in \mathbb{R}^{n+1}$ is updated in accordance with the law (8) then for all $\alpha'_{\beta} \in \mathbb{R}^{m+1}$ and $t > 0$

$$
\|\alpha'^{t+1} - \alpha^t\| = \|\Delta \alpha'_{\beta}\|
$$

where the norm is taken to be the standard Euclidean norm.

**Proof:** The proof follows from direct calculation of the norm.

$$
\begin{align*}
\|\alpha'^{t+1} - \alpha^t\| &= \sqrt{\sum_{i=1}^{n+1} (\alpha_i^{t+1} - \alpha_i^t)^2} \\
&= \sqrt{\sum_{i \notin I(\alpha'_{\beta})} (\alpha_i^{t+1} - \alpha_i^t)^2 + \sum_{i \in I(\alpha'_{\beta})} (\alpha_i^{t+1} - \alpha_i^t)^2} \\
&= \sqrt{\sum_{i \in I(\alpha'_{\beta})} (\alpha_i^{t+1} - \alpha_i^t)^2} = \|\Delta \alpha'_{\beta}\|
\end{align*}
$$

and the lemma is proved.

**Lemma 3.2:** Let $v'$ denote the gradient of the constrained quadratic problem defined as in (4) and let $\bar{v'}$ denote the optimal gradient so that $v', \bar{v'} \in \mathbb{R}^{n+1}$. Denote further the problem Hessian as $H \in \mathbb{R}^{n+1 \times n+1}$ and $\Delta v' = v' - \bar{v'}$. For $t > 0$ and $\alpha'_{\beta} \subseteq \mathbb{A}$

$$
\|v'^{t+1} - \bar{v'}\| \leq \gamma \|v'^t - \bar{v'}\|
$$

where $\gamma \geq 0$ can be written in dot product form as

$$
\sqrt{\langle H' \Delta \alpha'_{\beta}, H' \Delta \alpha'_{\beta} \rangle} + 2 \langle \Delta v', H' \Delta \alpha'_{\beta} \rangle + \langle \Delta v', \Delta v' \rangle \leq \gamma \sqrt{\langle \Delta v', \Delta v' \rangle}
$$

(11)

Here, the norm is the Euclidean norm and $H'$ denotes the $\mathbb{R}^{n+1 \times m+1}$ composite matrix defined as

$$
H' = \begin{bmatrix} H_p & \bar{H} \end{bmatrix}
$$

**Proof:** The proof is through direct calculation. Using (4) we have

$$
v'^{t+1} - \bar{v'} = v'^{t+1} - v'^t + v'^t - \bar{v'} = H \Delta \alpha' + v'^t - \bar{v'}
$$

Let $H_i$ denote the $i^{th}$ row of the matrix $H$. Then using the fact that $\Delta \alpha' = 0$ for $i \notin I(\alpha'_{\beta})$ (the indice set is taken to mean the indice for the current working set $\alpha'_{\beta}$) from Lemma 3.1, we compute the following

$$
\begin{align*}
\|v'^{t+1} - \bar{v'}\|^2 &= \|H \Delta \alpha' + v'^t - \bar{v'}\|^2 \\
&= \sum_{i=1}^{n+1} \left( \sum_{j \in I(\alpha'_{\beta})} H_{ij} \Delta \alpha_j + v_i'^t - \bar{v}_i' \right)^2 \\
&= \sum_{i=1}^{n+1} \left( \sum_{j \in I(\alpha'_{\beta})} H_{ij} \Delta \alpha_j \right)^2 + 2 \sum_{i=1}^{n+1} (v_i'^t - \bar{v}_i')^2 \\
&\geq 2 \left( \sum_{i \in I(\alpha'_{\beta})} (v_i'^t - \bar{v}_i') \sum_{j \in I(\alpha'_{\beta})} H_{ij} \Delta \alpha_j \right) \\
&\geq 2 \langle H' \Delta \alpha'_{\beta}, H' \Delta \alpha'_{\beta} \rangle + 2 \langle \Delta v', H' \Delta \alpha'_{\beta} \rangle + \langle \Delta v', \Delta v' \rangle
\end{align*}
$$

Using this result, we can rewrite

$$
\|v'^{t+1} - \bar{v'}\|^2 \leq \gamma^2 \|v'^t - \bar{v'}\|^2
$$

as

$$
\langle H' \Delta \alpha'_{\beta}, H' \Delta \alpha'_{\beta} \rangle + 2 \langle \Delta v', H' \Delta \alpha'_{\beta} \rangle + (1 - \gamma^2) \langle \Delta v', \Delta v' \rangle \leq 0
$$

(12)

Then taking the square root and noting that $\gamma > 0$ we obtain the required result

$$
\sqrt{\langle H' \Delta \alpha'_{\beta}, H' \Delta \alpha'_{\beta} \rangle} + 2 \langle \Delta v', H' \Delta \alpha'_{\beta} \rangle + \langle \Delta v', \Delta v' \rangle \leq \gamma \sqrt{\langle \Delta v', \Delta v' \rangle}
$$

This completes the proof.

Convergence of the gradient to the optimal point can be seen in Fig 1 where a desired trajectory of the gradient vector is depicted in Fig 1(b). We are interested in the condition that should be imposed on the working set, so that the gradient vector which starts in an area usually defined by a ball, $B_X$ stays in $B_X$ for all time $t$ and converges strictly to the center of the ball. Strict in our sense means that at every iteration
The following lemma proposes a sufficient and necessary condition for this to occur.

Lemma 3.3: Let \( \mathbf{v}' \) denote the gradient of the constrained quadratic problem defined as in (4) and \( \alpha' \) the corresponding variables of the problem. For an arbitrary working set, \( \alpha'_p \in A \), let \( \mathbf{H}' \) be the \( \mathbb{R}^{n+1 \times m+1} \) composite matrix defined as

\[
\mathbf{H}' = \begin{bmatrix}
\mathbf{H}_v' \\
\mathbf{H}_i'
\end{bmatrix}
\]

Then the gradient vector \( \mathbf{v}' \) converges strictly at each iteration to \( \bar{\mathbf{v}}' \) i.e.

\[
\|\mathbf{v}'^{t+1} - \bar{\mathbf{v}}'\| \leq \gamma_t \|\mathbf{v}'^t - \bar{\mathbf{v}}'\|
\]

where \( 0 \leq \gamma_t \leq 1 \) if and only if \( \forall i \in I(\alpha') \)

\[
-2 \leq \frac{\mathbf{H}'_i \Delta \alpha'_p}{\mathbf{v}'^t_i - \bar{\mathbf{v}}'_i} \leq 0
\]

Proof: Let us denote \( \Delta \mathbf{v}' = \mathbf{v}'^t - \bar{\mathbf{v}}' \) for some \( t > 0 \). Then working with the square of (14) we have from Lemma 3.2 equation (12) written element wise as

\[
\sum_{i=1}^{n+1} \left( \mathbf{H}'_i \Delta \alpha'_p \right)^2 + 2 \mathbf{H}'_i \Delta \alpha'_p + (1 - \gamma_t^2) \Delta \mathbf{v}'^2 \leq 0
\]

where \( \mathbf{H}'_i \) is the \( i^{th} \) row of the matrix \( \mathbf{H}' \). Then for each summand of \( i \) we can generally write

\[
|\left( \mathbf{H}'_i \Delta \alpha'_p \right)^2 + 2 \mathbf{H}'_i \Delta \alpha'_p + (1 - \gamma_t^2) \Delta \mathbf{v}'^2| \geq 0
\]

which indicates that we allow for the summand to be either positive or negative. For any \( t > 0 \), the summand is a standard quadratic with \( \Delta \mathbf{v}'_i \) being the \( i^{th} \) element of the constant vector \( \Delta \mathbf{v}' \) and \( \mathbf{H}'_i \Delta \alpha'_p \) the variable dependant on the choice of working set \( \alpha'_p \). We are going to assume first that \( 0 \leq \gamma_t \leq 1 \) and show that this can only hold under condition (14). More over if condition (14) is not satisfied for some \( i \) then \( \gamma_t > 1 \). Using the standard quadratic formula, the roots of the equation are real otherwise the discriminant gives

\[
4 \Delta \mathbf{v}'^2 - 4(1 - \gamma_t^2) \Delta \mathbf{v}'^2 < 0
\]

\[
\Rightarrow \gamma_t^2 < 0
\]

Then applying the quadratic formula to (15) we get

\[
\mathbf{H}'_i \Delta \alpha'_p = -2 \mathbf{v}'^t_i \pm \sqrt{4(1 - \gamma_t^2) \Delta \mathbf{v}'^2}
\]

and obtain the factorized form

\[
|\left( \mathbf{H}'_i \Delta \alpha'_p + (1 + \gamma_t) \Delta \mathbf{v}'_i \right) \left( \mathbf{H}'_i \Delta \alpha'_p + (1 - \gamma_t) \Delta \mathbf{v}'_i \right) |
\]

We now check the four possible situations that arise. Suppose that

\[
\left( \mathbf{H}'_i \Delta \alpha'_p + (1 + \gamma_t) \Delta \mathbf{v}'_i \right) \left( \mathbf{H}'_i \Delta \alpha'_p + (1 - \gamma_t) \Delta \mathbf{v}'_i \right) \geq 0
\]

which gives the following inequalities for each factor

Case 1:

\[
\frac{\mathbf{H}'_i \Delta \alpha'_p}{\Delta \mathbf{v}'_i} \geq -(1 + \gamma_t) \quad \frac{\mathbf{H}'_i \Delta \alpha'_p}{\Delta \mathbf{v}'_i} \leq -(1 - \gamma_t)
\]

which is possible only for \( \gamma_t = 0 \). It then follows that we must have \( \frac{\mathbf{H}'_i \Delta \alpha'_p}{\Delta \mathbf{v}'_i} = -1 \) for this to hold.

Case 2:

\[
\frac{\mathbf{H}'_i \Delta \alpha'_p}{\Delta \mathbf{v}'_i} \leq -(1 + \gamma_t) \quad \frac{\mathbf{H}'_i \Delta \alpha'_p}{\Delta \mathbf{v}'_i} \geq -(1 - \gamma_t)
\]

which again is only possible if \( \gamma_t = 0 \) and we have the same result as in Case 1. Now if instead we have

\[
\left( \mathbf{H}'_i \Delta \alpha'_p + (1 + \gamma_t) \Delta \mathbf{v}'_i \right) \left( \mathbf{H}'_i \Delta \alpha'_p + (1 - \gamma_t) \Delta \mathbf{v}'_i \right) \leq 0
\]

then

Case 3:

\[
\frac{\mathbf{H}'_i \Delta \alpha'_p}{\Delta \mathbf{v}'_i} \geq -(1 + \gamma_t) \quad \frac{\mathbf{H}'_i \Delta \alpha'_p}{\Delta \mathbf{v}'_i} \leq -(1 - \gamma_t)
\]

or

\[
-(1 + \gamma_t) \leq \frac{\mathbf{H}'_i \Delta \alpha'_p}{\Delta \mathbf{v}'_i} \leq -(1 - \gamma_t)
\]
In this case, having $0 \leq \gamma_t \leq 1$ implies that

$$-2 \leq \frac{H'_i \Delta \alpha'_p}{\Delta v'_i} \leq 0$$

Case 4:

$$\frac{H'_i \Delta \alpha'_p}{\Delta v'_i} \leq -(1 + \gamma_t) \quad \frac{H'_i \Delta \alpha'_p}{\Delta v'_i} \geq -(1 - \gamma_t)$$

or

$$- (1 - \gamma_t) \leq \frac{H'_i \Delta \alpha'_p}{\Delta v'_i} \leq -(1 + \gamma_t)$$

The inequality itself holds only if $\gamma_t = 0$ and like cases 1-2, this means that $H'_i \Delta \alpha'_p = 1$. Combining the results, we deduce then that $\forall i$

$$-2 \leq \frac{H'_i \Delta \alpha'_p}{\Delta v'_i} \leq 0$$

is sufficient for us to ensure $0 \leq \gamma_t \leq 1$. Now to show that it is also necessary, note that cases 1-4 mean that for (13) and $0 < \gamma_t \leq 1$ to hold simultaneously requires that for every $i$

$$(H'_i \Delta \alpha'_p)^2 + 2 \Delta v'_i H'_i \Delta \alpha'_p + (1 - \gamma_t^2) \Delta v'_i^2 < 0$$

otherwise, we have the case of $\gamma_t = 0$ which implies possible convergence. Suppose for some $i$, we have instead

$$\frac{H'_i \Delta \alpha'_p}{\Delta v'_i} > 0 \quad \text{or} \quad \frac{H'_i \Delta \alpha'_p}{\Delta v'_i} < -2$$

then

$$\gamma_t^2 > 1 + \frac{(H'_i \Delta \alpha'_p)^2}{\Delta v'_i^2} + 2 \frac{H'_i \Delta \alpha'_p}{\Delta v'_i}$$

which contradicts our requirement $0 \leq \gamma_t \leq 1$. This completes the proof.

We note that the result above assumes nothing about the condition on the Hessian matrix, $H$ nor the sequence of working sets. However, if the working sets are selected according to the condition (14) then we are assured of strict linear convergence. This is the essence of the following corollary which demonstrates that if the condition holds for all time $t$, we have a strict norm contraction on a neighbourhood defined by a ball, $B_X$.

**Corollary 3.1:** Assume that for all $t \in T$ that we select $\alpha'_p \in \mathbb{A}$ and update it according to (8) such that condition (14) holds. Then

$$\|v^{t+1} - \bar{v}\| \leq C_t \|v^0 - \bar{v}\|$$

where $\forall t > 0$ we have $0 \leq C_t \leq 1$. More over, if the gradient $v'$ starts initially in some neighbourhood, $B_X$ then it remains in $B_X$ for all $t$.

**Proof:** We make use of Lemma 3.3 and induction. Condition (14) allows us to write at $t = 1$,

$$\|v^1 - \bar{v}\| \leq \gamma_0 \|v^0 - \bar{v}\|$$

where $0 \leq \gamma_0 \leq 1$. This can also be equivalently written as $\gamma_0 = \frac{1}{k_0}$ for some $k_0 \geq 1$. Hence we can now write (17) as

$$\|v^{t+1} - \bar{v}\| \leq \frac{1}{k_0} \|v^0 - \bar{v}\|$$

Following similarly, we deduce at $t = t' + 1$

$$\|v^{t'+1} - \bar{v}\| \leq \frac{1}{k_{t'}} \|v^{t'} - \bar{v}\|$$

$$\leq \prod_{i=0}^{t'} \frac{1}{k_i} \|v^0 - \bar{v}\|$$

$$= C_{t'} \|v^0 - \bar{v}\|$$

Since $\forall i$ we had $k_i \geq 1$, then it can be seen that $0 \leq C_{t'} \leq 1$. The final statement of the corollary holds trivially by defining $B_X$ as

$$B_X[v', r] = \{v' \in \mathbb{R}^{n+1} ||v' - \bar{v}| | \leq \|v^0 - \bar{v}\|\}$$

and the proof is complete.

We now provide the main result of this work, that is an initial convergence rate estimate under the previous condition. The difference between this result and standard rate estimates is that it is time dependant.

**Lemma 3.4:** Let $v'$ denote the gradient of the constrained quadratic problem defined as in (4) and $\alpha'$ the corresponding variables of the problem. Let $\alpha'_p \in \mathbb{A}$ and $H'$ be the corresponding $\mathbb{R}^{n+1 \times m+1}$ composite matrix defined as in Lemma 3.3. Assume that for all $t > 0$ and all $i \in I(\alpha')$ we have

$$|\alpha'_i - \bar{\alpha}_i| \leq k |\alpha'_i - \bar{\alpha}_i|$$

where $k > 0$ and $\bar{\alpha}_i$ are the optimal values of the multipliers respectively. Then if $\forall i$ and $\forall t > 0$ condition (14) holds and $k \to \infty$ we have

$$\|v^{t+1} - \bar{v}\| \leq \gamma_t \|v^t - \bar{v}\|$$

where

$$\gamma_t \geq \frac{1 - \max_{\alpha'_p \in \mathbb{A}} \frac{H'_i \Delta \alpha'_p}{\Delta v'_i}}{\|v^t - \bar{v}\|}$$

Furthermore,

$$\lim_{t \to \infty} \gamma_t = 1$$

with $\gamma_t \to 1$ monotonically from below as $t \to \infty$.

**Proof:** First note that if $\gamma_t = 0$ then $v^t = \bar{v}$ and the problem is solved. So we consider the case which gives $0 < \gamma_t \leq 1$. From the proof of Lemma 3.3, one sees that this applies for Case 3 where the condition can be rewritten in terms of $\gamma_t$ as

$$\gamma_t \geq - \left(1 + \frac{H'_i \Delta \alpha'_p}{\Delta v'_i} \right)$$

$$\gamma_t \geq 1 + \frac{H'_i \Delta \alpha'_p}{\Delta v'_i}$$
The monotonic behaviour of the lower bound of which can be simplified to

$$\gamma_t \geq 1 + \frac{H_i' \Delta \alpha'_p}{\Delta v'_i}$$

for all $\forall i \in I(\alpha')$. Then we deduce that

$$\gamma_t \geq 1 + \frac{H_i' \Delta \alpha'_p}{\Delta v'_i} \geq 1 - \frac{H_i' \Delta \alpha'_p}{\Delta v'_i}$$

$$\geq 1 - \max_{\alpha'_p \in A} \frac{H_i' \Delta \alpha'_p}{\Delta v'_i}$$

(21)

This proves (19). Now using (4), Lemma 2.1 and Lemma 3.1 we get

$$\max_{\alpha'_p \in A} \frac{H_i' \Delta \alpha'_p}{\Delta v'_i} = \max_{\alpha'_p \in A} \left| \frac{\sum_{j=1}^{n+1} H_{ij} \Delta \alpha'_p(j)}{v'_i - v_i} \right|$$

$$= \max_{\alpha'_p \in A} \left| \frac{\sum_{j=1}^{n+1} H_{ij} (\alpha'_j^{(t+1)} - \alpha'_j^{(t)})}{v'_i - v_i} \right|$$

$$= \max_{\alpha'_p \in A} \left| \frac{\sum_{j=1}^{n+1} H_{ij} |\alpha'_j^{(t+1)} - \alpha'_j^{(t)}|}{v'_i - v_i} \right|$$

Then (18) and $k \to \infty$ implies further

$$\lim_{t \to \infty} \max_{\alpha'_p \in A} \frac{H_i' \Delta \alpha'_p}{\Delta v'_i} = \frac{1}{k} = 0$$

(22)

and finally taking limits of (21) we get

$$\lim_{t \to \infty} \gamma_t \geq 1 - \lim_{t \to \infty} \max_{\alpha'_p \in A} \frac{H_i' \Delta \alpha'_p}{\Delta v'_i}$$

(23)

$$\lim_{t \to \infty} \gamma_t = 1$$

The monotonic behaviour of the lower bound of $\gamma_t$ captured by (23) is due to (22) approaching zero from above i.e. some positive starting value. We note that one should regard $\gamma_t \to 1$ as $t \to \infty$ as a slowing down in convergence speed.

IV. DISCUSSION

There are several immediate implications that emerge from this investigation. Firstly, the necessary and sufficient condition for strict linear convergence of the decomposition method is restrictive. For efficient implementations, one might have to construct search heuristics in order to ensure it is met at every iteration. Instead of searching for a working set that satisfies the condition, it may be more practical to construct a family of update rules $S$ consisting of normalized feasible directions (in the sense of Zoutendijk [13]) such that it holds for the entire duration of the iteration or at least for a larger part of the iteration. It is still unclear how to do this, but we believe that it would necessitate several computations at each iteration. This requires us to be mindful of other important optimization factors such as computational complexity and ease of practical implementation. We note that if the rate of convergence slows down, perhaps it may not be so useful to enforce strict convergence. Further investigation is required to ascertain if superlinear or quadratic convergence is possible with the decomposition technique.

V. CONCLUSION

In this paper, we study the convergence properties of the problem gradients when the SVM is trained using a decomposition method. We provide a necessary and sufficient condition which has to be satisfied by any working set selection rule in order to guarantee strict linear convergence. A time-dependant rate estimate is then proposed and shown to monotonically tend to unity as we approach the optimal solution. Convergence rate estimates may be hard to satisfy in a practical algorithm, however they give a theoretical perspective of the potential speed of the decomposition method.

REFERENCES


Author/s: Lai, D.; Shilton, A.; Palaniswami, M.

Title: A convergence rate estimate for the SVM decomposition method

Date: 2005


Persistent Link: http://hdl.handle.net/11343/34064

Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.