# Linear Convergence of a Decomposition Method for Support Vector Machines

**Chih-Jen Lin**
Department of Computer Science and
Information Engineering
National Taiwan University
Taipei 106, Taiwan
cjlin@csie.ntu.edu.tw

**Abstract**

Recently the asymptotic convergence of some commonly used decomposition methods for support vector machines has been established. However, their local convergence rates are still unknown. In this paper, under the assumptions that the kernel matrix is positive definite and the problem is non-degenerate, we prove the linear convergence of a popular decomposition method.

# 1 Introduction

Given training vectors $x_i \in R^n, i = 1, \ldots, l$, in two classes, and a vector $y \in R^l$ such that $y_i \in \{1, -1\}$, the support vector machines (SVM) (Cortes and Vapnik, 1995; Vapnik, 1998) require the solution of the following optimization problem:

$$\begin{aligned} \min_{\alpha} \quad & f(\alpha) = \frac{1}{2}\alpha^T Q \alpha - e^T \alpha \\ \text{subject to} \quad & 0 \le \alpha_i \le C, i = 1, \ldots, l, \\ & y^T \alpha = 0, \end{aligned} \tag{1.1}$$

where $e$ is the vector of all ones, $C$ is the upper bound of all variables, and $Q$ is an $l$ by $l$ positive semidefinite matrix. Training vectors $x_i$ are mapped into a higher (maybe infinite) dimensional space by the function $\phi$ and $Q_{ij} \equiv y_i y_j K(x_i, x_j)$ where $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is the kernel.

Due to the density of the matrix $Q$, currently the decomposition method is one of the major methods to solve SVM (e.g. (Osuna et al., 1997; Joachims, 1998; Platt, 1998)). It is an iterative process and in each iteration the index set of

variables is portioned to two sets $B$ and $N$, where $B$ is the working set. Then in that iteration variables corresponding to $N$ are fixed while a sub-problem on variables corresponding to $B$ is minimized.

Among these decomposition methods, the software $SVM^{light}$ (Joachims, 1998) is a popular one. It has a systematic way for selecting the working set $B$ whose size can be any even number. When the size of $B$ is restricted to sets having two elements, it coincides with a modification of the SMO algorithm by Keerthi et al. (2001). Originally proposed by Platt (1998), the Sequential Minimal Optimization (SMO) algorithm is an extreme of the decomposition method whose working sets are restricted to two elements. The advantage of SMO is that in each iteration the sub-problem can be analytically solved without using an optimization software. Other software which have used the same working set selection as $SVM^{light}$ are, for example, LIBSVM (Chang and Lin, 2001).

The asymptotic convergence of the decomposition method used in $SVM^{light}$ was first proved in (Lin, 2001). More information about existing work on the convergence of decomposition methods can be found in the same paper. Up to now there are no results yet about local convergence of decomposition methods. In this paper we will establish the linear convergence of the method used by $SVM^{light}$. The analysis of convergence rates is very important for optimization methods as it helps to understand how fast an algorithm converges. It can also give more insights on the practical behaviors.

This paper is organized as follows. In section 2 we briefly introduce the algorithm used by $SVM^{light}$, in particular, its working set selection. Section 4 presents the main result of the linear convergence. Using this theoretical result, Section 5 explains some practical behaviors of decomposition methods. Finally in Section 7 we discuss the relation between our proof and some earlier work which focus on general bound-constrained optimization.

## 2 The Method of $SVM^{light}$

In this section we describe the working set selection of $SVM^{light}$ using the Karush-Kuhn-Tucker (KKT) condition, (i.e. the optimality condition) of (1.1): If $\alpha$ is an optimal solution of (1.1), there is a number $b$ and two nonnegative vectors $\lambda$ and

$\mu$ such that

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

where $\nabla f(\alpha) = Q\alpha - e$ is the gradient of $f(\alpha)$. This can be rewritten as

$$\begin{aligned} \nabla f(\alpha)_i + by_i \geq 0 \qquad & \text{if } \alpha_i = 0, \\ \nabla f(\alpha)_i + by_i \leq 0 \qquad & \text{if } \alpha_i = C, \\ \nabla f(\alpha)_i + by_i = 0 \qquad & \text{if } 0 < \alpha_i < C. \end{aligned}$$

Since $y_i = \pm 1$, by defining

$$I_{up}(\alpha) \equiv \{i \mid \alpha_i < C, y_i = 1 \text{ or } \alpha_i > 0, y_i = -1\}, \text{ and}$$
$$I_{low}(\alpha) \equiv \{i \mid \alpha_i < C, y_i = -1 \text{ or } \alpha_i > 0, y_i = 1\},$$

a feasible $\alpha$ is optimal for (1.1) if and only if

$$\max_{i \in I_{up}(\alpha)} -y_i \nabla f(\alpha)_i \leq \min_{i \in I_{low}(\alpha)} -y_i \nabla f(\alpha)_i. \tag{2.1}$$

When $\alpha$ is not an optimal solution, if $i \in I_{up}(\alpha), j \in I_{low}(\alpha)$, and $-y_i \nabla f(\alpha)_i > -y_j \nabla f(\alpha)_j$, following (Keerthi and Gilbert, 2002), we call such $(i, j)$ a "violating pair."

If $q$, an even number, is the size of the working set $B$ and $\alpha^k$ is the current iterate, $SVM^{light}$ selects the working set in the following way: $q/2$ indices are sequentially selected from elements in $I_{up}(\alpha^k)$ so that

$$-y_{i_1} \nabla f(\alpha^k)_{i_1} \geq -y_{i_2} \nabla f(\alpha^k)_{i_2} \geq \cdots \geq -y_{i_{q/2}} \nabla f(\alpha^k)_{i_{q/2}}. \tag{2.2}$$

The other $q/2$ indices are sequentially selected from $I_{low}(\alpha^k)$ such that

$$-y_{j_1} \nabla f(\alpha^k)_{j_1} \leq \cdots \leq -y_{j_{q/2}} \nabla f(\alpha^k)_{j_{q/2}}. \tag{2.3}$$

Therefore, $SVM^{light}$ essentially finds the $q/2$ most violated pairs into the working set and we call $(i_1, j_1)$ a "maximal violating pair."

We consider only violating pairs so if $-y_{i_{q/2}} \nabla f(\alpha^k)_{i_{q/2}} \leq -y_{j_{q/2}} \nabla f(\alpha^k)_{j_{q/2}}$, we reduce the size of the working set. Note that the working set will not be empty as there is at least one violating pair if $\alpha$ is not optimal yet.

Interestingly this working set selection was originally derived from the concept of feasible directions in constrained optimization though we feel a derivation from the violation of the KKT condition is more intuitive.

# 3 Existing Convergence Results

The asymptotic convergence of an optimization algorithm usually means that any its convergent subsequence goes to a (local) optimum. Note that the strict decrease of the objective value may not imply this property. The asymptotic convergence of decomposition methods was first studied in (Chang et al., 2000). However, the authors were able to consider only some types of decompoeition methods which did not coincide with existing implementations. It was until (Lin, 2002a) that the asymptotic convergence of $SVM^{light}$ was established:

**Theorem 1** *Assume the matrix $Q$ satisfies*

$$\min_I(\min(eig(Q_{II}))) > 0, \tag{3.1}$$

*where $I$ is any subset of $\{1,\ldots,l\}$ with $|I| \leq q$ and $\min(eig(\cdot))$ is the smallest eigenvalue of a matrix. If $\{\alpha^k\}$ is the sequence generated by the decomposition method in Section 2, the limit of any its convergent subsequence is an optimal solution of* (1.1).

If the size of the working set is restriced to two (i.e. $q = 2$), (Lin, 2002a) provides a proof of the above theorem without any assumption.

Another property related to the convergence is the "finite termination" of an algorithm. For a given stopping condition with any pre-specified tolerance, it discusses whether the optimization algorithm terminates in a finite number of iterations. The first such results for the decomposition methods is in (Keerthi and Gilbert, 2002):

**Theorem 2** *If the algorithm in Section 2 is used and $q = 2$, for any given $\epsilon > 0$, after a finite number of iterations,*

$$\max_{i\in I_{up}(\alpha)} -y_i\nabla f(\alpha)_i \leq \min_{i\in I_{low}(\alpha)} -y_i\nabla f(\alpha)_i + \epsilon \tag{3.2}$$

*is satisfied.*

Note that Theorem 2 does not imply Theorem 1 as both sides of (3.2) are not continuous functions of $\alpha$. That is, we cannot take their limits with $\epsilon \to 0$ and claim that any convergent point has already satisfied the KKT condition and hance is an optimum. For the general situation of more than two elements in the working set, (Lin, 2002b) proves Theorem 2 under some minor assumptions.

# 4 Main Results on Linear Convergence

Before proving the main results, we need some assumptions. First we assume that the kernel matrix is positive definite:

**Assumption 1** *$K$ is positive definite.*

Note that $K$ and $Q$, the Hessian of (1.1), have the same eigenvalues so $Q$ is positive definite as well. Then (1.1) is a strictly convex programming problem and hence has a unique global optimum $\alpha^*$.

Theorem 1 implies that the whole sequence $\{\alpha^k\}$ of the decomposition method converges to $\alpha^*$. We can also see that Theorem II.3 of (Lin, 2002b) holds:

1. If the algorithm takes infinite iterations,

$$\max_{i \in I_{up}(\alpha^*)} -y_i \nabla f(\alpha^*)_i = \min_{i \in I_{low}(\alpha^*)} -y_i \nabla f(\alpha^*)_i.$$

   Let us call the above quantity as $b^*$.

2. After $k$ is large enough, only elements whose $-y_i \nabla f(\alpha^*)_i$ are $b^*$ can still be modified. Furthermore, only such elements can still form violating pairs.

Therefore, in final iterations, the algorithm works only on a particular subset of variables. This makes our analysis easier as convergence rates relate to behaviors in final iterations. Moreover, for this particular subset of variables, we need an additional assumption: problem (1.1) is non-degenerate.

**Assumption 2 (Nondegeneracy)** *For the optimal solution $\alpha^*$, we have $\nabla f(\alpha^*)_i + b^* y_i \neq 0$ if $\alpha_i^* = 0$ or $C$.*

This condition is also called strict complementarity in the optimization terminology as it means two values in

$$\alpha_i^*(Q\alpha^* - e + b^* y)_i = 0$$

of the KKT condition cannot be both zeros. The situation is similar for $(C - \alpha_i^*)(Q\alpha^* - e + b^* y)_i = 0$. Therefore, after $k$ is large enough, all bounded variables are fixed and are not included in the working set. By treating bounded variables as constants essentially we are solving a problem with the following form:

$$\begin{aligned} \min_{\alpha} \quad & f(\alpha) = \frac{1}{2}\alpha^T Q \alpha + p^T \alpha \\ \text{subject to} \quad & y^T \alpha = \Delta, \end{aligned} \tag{4.1}$$

where $0 < \alpha_i^k < C$ for all $i$ even though we do not write down inequality constraints explicitly. Then the optimal solution $\alpha^*$ with its Lagrange multiplier $b^*$ can be obtained by the following linear system:

$$\begin{bmatrix} Q & y \\ y^T & 0 \end{bmatrix} \begin{bmatrix} \alpha^* \\ b^* \end{bmatrix} = \begin{bmatrix} -p \\ \Delta \end{bmatrix}. \tag{4.2}$$

In each iteration, we consider minimizing $f(\alpha_B^k + d)$ where $d$ is the direction moving from $\alpha_B^k$ so the sub-problem is

$$\begin{aligned} \min_{d} \quad & \frac{1}{2} d^T Q_{BB} d + \nabla f(\alpha^k)_B^T d. \\ \text{subject to} \quad & y_B^T d = 0, \end{aligned} \tag{4.3}$$

where $\nabla f(\alpha^k) = Q\alpha^k + p$ now. If a solution of (4.3) is $d^k$, then $\alpha_B^{k+1} = \alpha_B^k + d^k$ and $\alpha_N^{k+1} = \alpha_N^k$. With the Lagrange multiplier $b^k$, this sub-problem can be solved by the following equation:

$$\begin{bmatrix} Q_{BB} & y_B \\ y_B^T & 0 \end{bmatrix} \begin{bmatrix} d^k \\ b^k \end{bmatrix} = \begin{bmatrix} -\nabla f(\alpha^k)_B \\ 0 \end{bmatrix}. \tag{4.4}$$

Using (4.2),

$$\begin{aligned} Q(\alpha^k - \alpha^*) &= Q\alpha^k + p + b^* y \\ &= \nabla f(\alpha^k) + b^* y. \end{aligned} \tag{4.5}$$

By defining $Y \equiv \text{diag}(y)$ to be a diagonal matrix with elements of $y$ on the diagonal, with $y_i = \pm 1$, we have

$$-YQ(\alpha^k - \alpha^*) = -Y\nabla f(\alpha^k) - b^* e.$$

Now without inequalities, a "maximal violating pair" is obtained simply by the maximal and the minimal elements of $-Y\nabla f(\alpha^k)$. As simultaneously subtracting a constant $b^*$ does not affect the order of a sequence, we have

$$\begin{aligned}&\text{argmax}_i(-y_i(Q(\alpha^k-\alpha^*))_i) = \text{argmax}_i(-y_i\nabla f(\alpha^k)_i) \text{ and}\\ &\text{argmin}_i(-y_i(Q(\alpha^k-\alpha^*))_i) = \text{argmin}_i(-y_i\nabla f(\alpha^k)_i).\end{aligned} \tag{4.6}$$

The following two theorems are main results on linear convergence. They require two technical lemmas which are left in the end of this section.

**Theorem 3** *There is $c<1$ such that after $k$ is large enough,*

$$(\alpha^{k+1}-\alpha^*)^TQ(\alpha^{k+1}-\alpha^*) \le c(\alpha^k-\alpha^*)^TQ(\alpha^k-\alpha^*). \tag{4.7}$$

**Proof.** We directly calculate the difference between the $(k+1)$st and the $k$th iterations:

$$\begin{aligned}&(\alpha^{k+1}-\alpha^*)^TQ(\alpha^{k+1}-\alpha^*) - (\alpha^k-\alpha^*)^TQ(\alpha^k-\alpha^*) &(4.8)\\ =\ & 2(d^k)^T(Q(\alpha^k-\alpha^*))_B + (d^k)^TQ_{BB}d^k\\ =\ & (d^k)^T(2(Q(\alpha^k-\alpha^*))_B - \nabla f(\alpha^k)_B - b^ky_B) &(4.9)\\ =\ & (d^k)^T((Q(\alpha^k-\alpha^*))_B + (b^*-b^k)y_B) &(4.10)\\ =\ & (d^k)^T((Q(\alpha^k-\alpha^*))_B + (b^k-b^*)y_B) &(4.11)\\ =\ & -[-(Q(\alpha^k-\alpha^*))_B + (b^*-b^k)y_B]^TQ_{BB}^{-1}[-(Q(\alpha^k-\alpha^*))_B + (b^*-b^k)y_B],\end{aligned}$$

where (4.9) is from (4.4), (4.10) is from (4.5), (4.11) is obtained by using the fact $y_B^Td^k=0$ from (4.4), and the last equality is from (4.4) and (4.5). If we define

$$\hat{Q} \equiv Y_BQ_{BB}^{-1}Y_B \text{ and } v \equiv -Y(Q(\alpha^k-\alpha^*)), \tag{4.12}$$

where $Y_B \equiv \text{diag}(y_B)$, then $v_B = -Y_B(Q(\alpha^k-\alpha^*))_B$ and (4.8) becomes

$$-[v_B + (b^*-b^k)e_B]^T\hat{Q}[v_B + (b^*-b^k)e_B]. \tag{4.13}$$

Using the fact that at least one "maximal violating pair" is in $B$, with (4.6) we can define

$$v^1 \equiv \max_i(v_i) = \max_{i\in B}(v_i) \text{ and } v^l \equiv \min_i(v_i) = \min_{i\in B}(v_i). \tag{4.14}$$

We denote that min(eig(·)) and max(eig(·)) to be the minimal and maximal eigenvalues of a matrix, respectively. Then

$$\begin{aligned}
& [v_B + (b^* - b^k)e_B]^T \hat{Q} [v_B + (b^* - b^k)e_B] \\
\geq\ & \min(\mathrm{eig}(\hat{Q}))[v_B + (b^* - b^k)e_B]^T [v_B + (b^* - b^k)e_B] \\
\geq\ & \min(\mathrm{eig}(\hat{Q}))\frac{(v^1 - v^l)^2}{2} \qquad (4.15) \\
\geq\ & \frac{\min(\mathrm{eig}(\hat{Q}))}{2}(\frac{y^T Q^{-1} y}{\sum_{i,j} |Q_{ij}^{-1}|})^2 \max(|v^1|, |v^l|)^2 \qquad (4.16) \\
\geq\ & \frac{\min(\mathrm{eig}(\hat{Q}))}{2l}(\frac{y^T Q^{-1} y}{\sum_{i,j} |Q_{ij}^{-1}|})^2 (Q(\alpha^k - \alpha^*))^T Q (\alpha^k - \alpha^*) \qquad (4.17) \\
\geq\ & \frac{\min(\mathrm{eig}(\hat{Q}))}{2l \max(\mathrm{eig}(Q^{-1}))}(\frac{y^T Q^{-1} y}{\sum_{i,j} |Q_{ij}^{-1}|})^2 (Q(\alpha^k - \alpha^*))^T Q^{-1} Q (\alpha^k - \alpha^*) \\
\geq\ & \frac{\min(\mathrm{eig}(\hat{Q}))}{2l \max(\mathrm{eig}(Q^{-1}))}(\frac{y^T Q^{-1} y}{\sum_{i,j} |Q_{ij}^{-1}|})^2 (\alpha^k - \alpha^*)^T Q (\alpha^k - \alpha^*), \qquad (4.18)
\end{aligned}$$

where (4.15) is from (4.14) and Lemma 1, (4.16) is from Lemma 2, and (4.17) follows from (4.14).

Here we give more details about the derivation of (4.16): If $v^1 v^l \leq 0$, then of course

$$|v^1 - v^l| \geq \max(|v^1|, |v^l|).$$

With $y_i = \pm 1$, $\frac{y^T Q^{-1} y}{\sum_{i,j} |Q_{ij}^{-1}|} \leq 1$ so (4.16) follows. On the other hand, if $v^1 v^l \geq 0$, we consider $v = (YQY)(-Y(\alpha^k - \alpha^*))$ from (4.12). Since $-e^T Y(\alpha^k - \alpha^*) = -y^T(\alpha^k - \alpha^*) = 0$, we can apply Lemma 2: With

$$\begin{gathered}
|(YQY)_{ij}^{-1}| = |Q_{ij}^{-1} y_i y_j| = |Q_{ij}^{-1}| \text{ and} \\
e^T (YQY)^{-1} e = y^T Q^{-1} y,
\end{gathered}$$

we have

$$\begin{aligned}
|v^1 - v^l| &\geq \max(|v^1|, |v^l|) - \min(|v^1|, |v^l|) \\
&\geq (\frac{y^T Q^{-1} y}{\sum_{i,j} |Q_{ij}^{-1}|}) \max(|v^1|, |v^l|)
\end{aligned}$$

which implies (4.16).

Then we can define a constant $c$ as follows:

$$c \equiv 1 - \min_B \left( \frac{\min(\mathrm{eig}(Q_{BB}^{-1}))}{2l \max(\mathrm{eig}(Q^{-1}))}(\frac{y^T Q^{-1} y}{\sum_{i,j} |Q_{ij}^{-1}|})^2 \right) < 1.$$

Combining (4.13) and (4.18), after $k$ is large enough, (4.7) holds. □

The linear convergence of the objective function is as follows:

**Theorem 4** *There is $c < 1$ such that after $k$ is large enough,*

$$f(\alpha^{k+1}) - f(\alpha^*) \leq c(f(\alpha^k) - f(\alpha^*)).$$

**Proof.** We will show that for any $k$,

$$f(\alpha^k) - f(\alpha^*) = \frac{1}{2}(\alpha^k - \alpha^*)^T Q(\alpha^k - \alpha^*)$$

so the proof immediately follows from Theorem 3. Using (4.2),

$$\begin{aligned} & f(\alpha^k) - f(\alpha^*) \\ = \; & \frac{1}{2}(\alpha^k)^T Q\alpha^k + p^T\alpha^k - \frac{1}{2}(\alpha^*)^T Q\alpha^* - p^T\alpha^* \\ = \; & \frac{1}{2}(\alpha^k)^T Q\alpha^k + (-Q\alpha^* - b^*y)^T\alpha^k - \frac{1}{2}(\alpha^*)^T Q\alpha^* - (-Q\alpha^* - b^*y)^T\alpha^* \\ = \; & \frac{1}{2}(\alpha^k)^T Q\alpha^k - (\alpha^*)^T Q\alpha^k + \frac{1}{2}(\alpha^*)^T Q\alpha^* \qquad (4.19) \\ = \; & \frac{1}{2}(\alpha^k - \alpha^*)^T Q(\alpha^k - \alpha^*). \end{aligned}$$

Since we always keep the feasibility of $\alpha^k$, we can use $y^T\alpha^k = \Delta$ to cancel out the term $y^T\alpha^*$ and have (4.19). □

Next we present two technical lemmas used earlier.

**Lemma 1** *If $v_1 \geq \cdots \geq v_l$,*

$$\sum_{i=1}^{l} v_i^2 \geq \frac{(v_1 - v_l)^2}{2}.$$

**Proof.**

$$\sum_{i=1}^{l} v_i^2 \geq v_1^2 + v_l^2 \geq \frac{(v_1 - v_l)^2}{2}.$$

□

**Lemma 2** *If $Q$ is invertible, then for any $x$ such that*

1. $e^T x = 0$,

*2.* $v \equiv Qx$, $\max_i((Qx)_i) = v^1 > v^l = \min_i((Qx)_i)$, *and* $v^1 v^l \geq 0$,

*we have*

$$\min(|v^1|, |v^l|) \leq (1 - \frac{e^T Q^{-1} e}{\sum_{i,j} |Q^{-1}_{ij}|}) \max(|v^1|, |v^l|).$$

**Proof.** Since $v^1 > v^l$ and $v^1 v^l \geq 0$, we have either $v^1 > v^l \geq 0$ or $0 \geq v^1 > v^l$. For the first case, if the result is wrong,

$$v^l > (1 - \frac{e^T Q^{-1} e}{\sum_{i,j} |Q^{-1}_{ij}|}) v^1,$$

so for $j = 1, \ldots, l$,

$$\begin{aligned} v^1 - v_j &\leq v^1 - v^l \\ &< (\frac{e^T Q^{-1} e}{\sum_{i,j} |Q^{-1}_{ij}|}) v^1. \end{aligned} \tag{4.20}$$

With $x = Q^{-1} v$ and (4.20),

$$\begin{aligned} e^T x &= e^T Q^{-1} v \\ &= \sum_{i,j} Q^{-1}_{ij} v_j \\ &= \sum_{i,j} Q^{-1}_{ij} (v^1 - (v^1 - v_j)) \\ &\geq v^1 e^T Q^{-1} e - (v^1 - v^l) \sum_{i,j} |Q^{-1}_{ij}| \\ &> v^1 \left( e^T Q^{-1} e - (\frac{e^T Q^{-1} e}{\sum_{i,j} |Q^{-1}_{ij}|}) \sum_{i,j} |Q^{-1}_{ij}| \right) \\ &= 0 \end{aligned}$$

causes a contradiction. The case of $0 \geq v^1 > v^l$ is similar. □

# 5 Some Practical Considerations

Earlier experiments have pointed out that if the kernel matrix is well conditioned, the decomposition method converges more quickly. This has been mentioned in, for example, (Hsu and Lin, 2002, Section 5).

Results in this paper provide more insights about this observation. Here, we discuss the situation when the RBF kernel is used (i.e., $K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$).

When $\gamma$ is large, $Q \to I$ is well conditioned. We show that for larger $\gamma$, the linear convergence rate is higher:

Since $Q_{ii} = 1, i = 1, \ldots, l$ for the RBF kernel,

$$\sum_{i=1}^{l} \lambda_i = \text{trace}(Q) = l,$$

where $\lambda_1, \ldots, \lambda_l$ are eigenvalues of $Q$. Therefore,

$$\min(\text{eig}(Q_{BB}^{-1})) \le 1 \text{ and } \max(\text{eig}(Q^{-1})) \ge 1.$$

With $(y^T Q^{-1} y)/\sum_{i,j} |Q_{ij}^{-1}| \le 1$,

$$\min_B \left( \frac{\min(\text{eig}(Q_{BB}^{-1}))}{l \max(\text{eig}(Q^{-1}))} (\frac{y^T Q^{-1} y}{2\sum_{i,j} |Q_{ij}^{-1}|})^2 \right) \le \frac{1}{4l}.$$

When $\gamma$ is large, $Q \to I$ so

$$\min_B \left( \frac{\min(\text{eig}(Q_{BB}^{-1}))}{l \max(\text{eig}(Q^{-1}))} (\frac{y^T Q^{-1} y}{2\sum_{i,j} |Q_{ij}^{-1}|})^2 \right) \to \frac{1}{4l},$$

its largest possible value. Therefore, the convergence seems faster when the kernel matrix is well-conditioned.

On the other hand, when $Q$ is very ill-conditioned, $1/\max(\text{eig}(Q^{-1})) = \min(\text{eig}(Q))$ can be very small. Then the rate constant $c$ is close to 1 so the convergence is very slow. For linear SVM with the number of training samples greater than the number of attributes, $Q$ is only positive semi-definite so $\min(\text{eig}(Q)) = 0$. Practically decomposition methods converge very slowly for such cases so indeed people consider that SMO might not be very suitable for linear SVM (Chung et al., 2002). Though results in this paper assume the positive definiteness of the kernel matrix, if we consider such linear SVM as ill-conditioned problems, our results also helps to explain the slow convergence. We think that theoretical properties of decomposition methods for linear SVM are worth for further investigation.

# 6 An Example

We have shown that under some general conditions, the decomposition method discussed here is at least linearly convergent. However, it is still not clear whether

the convergence is actually better than linear or not. Here, we present a simple example which exactly has the linear convergence. Hence, in theory, the linear convergence is already the best worst-case analysis.

Consider $x_1, x_2, x_3$ with $\|x_1 - x_2\| = \|x_1 - x_3\| = \|x_2 - x_3\|$, $y = [1, 1, -1]^T$, and $C = \infty$. If the RBF kernel is used, the dual SVM problem is

$$\begin{aligned} \min_{\alpha_1,\alpha_2,\alpha_3} \quad & \frac{1}{2}\begin{bmatrix}\alpha_1 & \alpha_2 & \alpha_3\end{bmatrix}\begin{bmatrix}1 & a & -a \\ a & 1 & -a \\ -a & -a & 1\end{bmatrix}\begin{bmatrix}\alpha_1 \\ \alpha_2 \\ \alpha_3\end{bmatrix} - (\alpha_1 + \alpha_2 + \alpha_3) \\ \text{subject to} \quad & \alpha_1 + \alpha_2 - \alpha_3 = 0, \\ & 0 \le \alpha_1, \alpha_2, \alpha_3, \end{aligned}$$

where $a = e^{-\gamma\|x_i - x_j\|^2}$. We assume $C$ is large so is not needed here. At the optimal solution,

$$\alpha^* = \begin{bmatrix}\frac{2}{3(1-a)} & \frac{2}{3(1-a)} & \frac{4}{3(1-a)}\end{bmatrix}^T. \tag{6.1}$$

We will show that after $k$ is large enough,

$$(\alpha^{k+1} - \alpha^*)^T Q(\alpha^{k+1} - \alpha^*) = \frac{1}{4}(\alpha^k - \alpha^*)^T Q(\alpha^k - \alpha^*). \tag{6.2}$$

Now $q$, the size of the working set, must be two so the three possible sets are $\{1, 2\}, \{1, 3\}$, and $\{2, 3\}$. We can see that Assumptions 1 and 2 are easily satisfied. Thus, after $k$ is large enough, $\alpha_i^k, i = 1, \ldots, 3$ are strictly positive. Then, in each iteration, after solving the sub-problem the two variables are positive so they have the same $y_i \nabla f(\alpha)_i$. Hence, under the rules of (2.2) and (2.3), any one for the other two possible sets can be the working set of the next iteration. For example, if $\{1, 3\}$ is the working set of the $k$th iteration, then for the $(k + 1)$st iteration, either $\{2, 3\}$ or $\{1, 2\}$ can be used.

For convenience, we define

$$e_i^k \equiv \alpha_i^k - \alpha_i^*, i = 1, \ldots, 3.$$

We claim that at the $k$th iteration:

1. If $\{1, 3\}$ is the working set, then

$$2e_1^{k+1} + e_2^{k+1} = 0. \tag{6.3}$$

2. If $\{2,3\}$ is the working set, then

$$e_1^{k+1} + 2e_2^{k+1} = 0. \tag{6.4}$$

3. If $\{1,2\}$ is the working set, then

$$e_1^{k+1} - e_2^{k+1} = 0. \tag{6.5}$$

For the first case, using (4.4),

$$\begin{bmatrix} 1 & -a \\ -a & 1 \end{bmatrix} \begin{bmatrix} \alpha_1^{k+1} \\ \alpha_3^{k+1} \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} + b^{k+1} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = -\alpha_2^k \begin{bmatrix} a \\ -a \end{bmatrix}. \tag{6.6}$$

With

$$\alpha_1^{k+1} - \alpha_3^{k+1} = -\alpha_2^k,$$

we have

$$\alpha_1^{k+1} = \frac{1}{1-a} - \frac{\alpha_2^k}{2}.$$

Therefore, using (6.1) and $\alpha_2^{k+1} = \alpha_2^k$,

$$2(\alpha_1^{k+1} - \alpha_1^*) + (\alpha_2^{k+1} - \alpha_2^*) = \frac{2}{1-a} - 3\alpha_2^* = 0.$$

The second case, (6.4), can be derived by a similar way. For the third case, it is easy to see that if $\{1,2\}$ is the working set, $\alpha_1^{k+1} = \alpha_2^{k+1}$. With $\alpha_1^* = \alpha_2^*$,

$$e_1^{k+1} = e_2^{k+1} = \frac{e_1^k + e_2^k}{2} = \frac{e_3^k}{2}. \tag{6.7}$$

Using $e_2^{k+1} = e_2^k$, $e_1^{k+1} = e_1^k$, and $e_1^{k+1} = e_2^{k+1} = e_3^k/2$, for the three respective cases, by induction, if $e_i^1 \neq 0, i = 1, \ldots, 3$, then

$$e_i^k \neq 0, i = 1, \ldots, 3, \text{ for all } k. \tag{6.8}$$

Now we are ready to prove (6.2). With $\alpha_3^k = \alpha_1^k + \alpha_2^k$,

$$\begin{aligned} & (\alpha^k - \alpha^*)^T Q (\alpha^k - \alpha^*) \\ = \;& 2(1-a)((e_1^k)^2 + (e_2^k)^2 + e_1^k e_2^k). \end{aligned}$$

If $\{1,3\}$ is the working set, then with (6.3) and $\alpha_2^{k+1} = \alpha_2^k$,

$$\begin{aligned} & (\alpha^{k+1} - \alpha^*)^T Q (\alpha^{k+1} - \alpha^*) \\ = \;& 2(1-a)((e_1^{k+1})^2 + (e_2^{k+1})^2 + e_1^{k+1} e_2^{k+1}) \\ = \;& \frac{3}{2}(1-a)(e_2^k)^2. \end{aligned} \tag{6.9}$$

The validity of (6.2) requires

$$\frac{\frac{3}{2}(1-a)(e_2^k)^2}{2(e_1^k)^2+2(e_2^k)^2+2e_1^k e_2^k}=\frac{1}{4}$$

which, under (6.8), is equivalent to

$$(e_1^k-e_2^k)(e_1^k+2e_2^k)=0. \tag{6.10}$$

Since $\{1,3\}$ is the current working set, in the previous iteration, the set must be $\{1,2\}$ or $\{2,3\}$. Thus, (6.10) follows from (6.5) and (6.4).

The proof for the case that $\{2,3\}$ is the working set is very similar. If $\{1,2\}$ is the working set, putting (6.7) into (6.9), (6.10) becomes

$$(2e_1^k+e_2^k)(e_1^k+2e_2^k)=0,$$

so the result also follows.

Indeed by a more detailed description, we can show that if the initial solution is zero, (6.2) holds for all $k=1,2,\ldots$

# 7 Discussion

The decomposition method has been an old optimization technique which is also called, for example, "coordinate search," "method of alternating variables," or "coordinate descent method." However, in most cases only bound-constrained or unconstrained optimization problems are considered where the linear convergence (without the non-degeneracy assumption) has been established in, for example, (Luo and Tseng, 1992) and references therein. With the additional linear constraint $y^T\alpha=0$ and differences on the working set selection, we have not been able to get similar proofs without the non-degeneracy assumption. How to fill this gap is a further research issue.

On the other hand, after using the non-degeneracy assumption and (Lin, 2002b, Theorem II.3) to remove inequalities, (4.1) is a very simple problem. Hence we essentially follow the structure of proving the linear convergence of the steepest descent method for unconstrained convex quadratic programming problems (see, for example, (Nocedal and Wright, 1999, Chapter 3.3)). Two news things we have to take care of are:

1. Using the property that the "maximal violation pair" is selected so (4.16), an expression only on the variables of the working set, can be connected to (4.17) which is related to all variables.

2. Handling the linear constraint $y^T\alpha = \Delta$. For the unconstrained case there is no $b^*$ and $b^k$ so (4.15) can directly imply (4.16). Here we need Lemma 2 to connect them.

# Acknowledgments

This work was supported in part by the National Science Council of Taiwan via the grant NSC 90-2213-E-002-111.

# References

Chang, C.-C., C.-W. Hsu, and C.-J. Lin (2000). The analysis of decomposition methods for support vector machines. *IEEE Transactions on Neural Networks 11*(4), 1003–1008.

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

Chung, K.-M., W.-C. Kao, C.-L. Sun, and C.-J. Lin (2002). Decomposition methods for linear support vector machines. Technical report, Department of Computer Science and Information Engineering, National Taiwan University.

Cortes, C. and V. Vapnik (1995). Support-vector network. *Machine Learning 20*, 273–297.

Hsu, C.-W. and C.-J. Lin (2002). A simple decomposition method for support vector machines. *Machine Learning 46*, 291–314.

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

Keerthi, S. S. and E. G. Gilbert (2002). Convergence of a generalized SMO algorithm for SVM classifier design. *Machine Learning 46*, 351–360.

Keerthi, S. S., S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation 13*, 637–649.

Lin, C.-J. (2001). On the convergence of the decomposition method for support vector machines. *IEEE Transactions on Neural Networks 12*(6), 1288–1298.

Lin, C.-J. (2002a). Asymptotic convergence of an SMO algorithm without any assumptions. *IEEE Transactions on Neural Networks 13*(1), 248–250.

Lin, C.-J. (2002b). A formal analysis of stopping criteria of decomposition methods for support vector machines. *IEEE Transactions on Neural Networks 13*(5), 1045–1052.

Luo, Z.-Q. and P. Tseng (1992). On the convergence of coordinate descent method for convex dierentiable minimization. *Journal of Optimization Theory and Applications 72*(1), 7–35.

Nocedal, J. and S. J. Wright (1999). *Numerical Optimization*. New York, NY: Springer-Verlag.

Osuna, E., R. Freund, and F. Girosi (1997). Training support vector machines: An application to face detection. In *Proceedings of CVPR'97*, New York, NY, pp. 130–136. IEEE.

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

Vapnik, V. (1998). *Statistical Learning Theory*. New York, NY: Wiley.