

A Note on the Decomposition Methods for Support Vector Regression

Shuo-Peng Liao

b6506060@csie.ntu.edu.tw

Hsuan-Tien Lin

b6506054@csie.ntu.edu.tw

Chih-Jen Lin

cjlin@csie.ntu.edu.tw

Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan

The dual formulation of support vector regression involves two closely related sets of variables. When the decomposition method is used, many existing approaches use pairs of indices from these two sets as the working set. Basically, they select a base set first and then expand it so all indices are pairs. This makes the implementation different from that for support vector classification. In addition, a larger optimization subproblem has to be solved in each iteration. We provide theoretical proofs and conduct experiments to show that using the base set as the working set leads to similar convergence (number of iterations). Therefore, by using a smaller working set while keeping a similar number of iterations, the program can be simpler and more efficient.

1 Introduction ---

Given a set of data points, $\{(x_1, z_1), \dots, (x_l, z_l)\}$, such that $x_i \in R^n$ is an input and $z_i \in R^1$ is a target output. A major form for solving support vector regression (SVR) is the following optimization problem (Vapnik, 1998):

$$\begin{aligned} \min \quad & \frac{1}{2}(\alpha - \alpha^*)^T Q(\alpha - \alpha^*) + \epsilon \sum_{i=1}^l (\alpha_i + \alpha_i^*) + \sum_{i=1}^l z_i(\alpha_i - \alpha_i^*) \\ & \sum_{i=1}^l (\alpha_i - \alpha_i^*) = 0, 0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, \dots, l, \end{aligned} \quad (1.1)$$

where C is the upper bound, $Q_{ij} \equiv \phi(x_i)^T \phi(x_j)$, α_i and α_i^* are Lagrange multipliers associated with the i th data x_i , and ϵ is the error that users can tolerate. Note that training vectors x_i are mapped into a higher-dimensional

space by the function ϕ . An important property is that for any optimal solution, $\alpha_i \alpha_i^* = 0, i = 1, \dots, l$.

Due to the density of Q , currently the decomposition method is the major method to solve equation 1.1 (Smola & Schölkopf, 1998; Keerthi, Shevade, Bhattacharyya, & Murthy, 2000; Laskov, 2002). It is an iterative process wherein each iteration, the index set of variables is separated to two sets, B and N , where B is the working set. In that iteration, variables corresponding to N are fixed, while a subproblem on variables corresponding to B is minimized.

Following approaches for support vector classification, there are some methods for selecting the working set. Many existing approaches for regression first use these methods to find a set of variables, called the *base set* here; then they expand the base set so all elements are pairs. Here we define the expanded set as the *pair set*. For example, if $\{\alpha_i, \alpha_j^*\}$ are chosen first, they include $\{\alpha_i^*, \alpha_j\}$ in the working set. Then the following subproblem of four variables $(\alpha_i, \alpha_i^*, \alpha_j, \alpha_j^*)$ is solved:

$$\begin{aligned}
\min \quad & \frac{1}{2} \begin{bmatrix} \alpha_i - \alpha_i^* \\ \alpha_j - \alpha_j^* \end{bmatrix}^T \begin{bmatrix} Q_{ii} & Q_{ij} \\ Q_{ji} & Q_{jj} \end{bmatrix} \begin{bmatrix} \alpha_i - \alpha_i^* \\ \alpha_j - \alpha_j^* \end{bmatrix} \\
& + (Q_{iN}(\alpha_N - \alpha_N^*) + z_i)(\alpha_i - \alpha_i^*) \\
& + (Q_{jN}(\alpha_N - \alpha_N^*) + z_j)(\alpha_j - \alpha_j^*) + \epsilon(\alpha_i + \alpha_i^* + \alpha_j + \alpha_j^*) \\
& (\alpha_i - \alpha_i^*) + (\alpha_j - \alpha_j^*) = - \sum_{t \in N} (\alpha_t - \alpha_t^*) \\
& 0 \leq \alpha_i, \alpha_j, \alpha_i^*, \alpha_j^* \leq C.
\end{aligned} \tag{1.2}$$

Note that α_N and α_N^* are fixed elements corresponding to $N = \{t \mid 1 \leq t \leq l, t \neq i, t \neq j\}$.

A reason for selecting pairs is to maintain $\alpha_i \alpha_i^* = 0, i = 1, \dots, l$ throughout all iterations. Hence, the number of nonzero variables during iterations can be kept small. However, from Lin (2001a, Theorem 4.1), it has been shown that for some existing work (e.g., Keerthi et al., 2000; Laskov, 2002), if the base set is used as the working set, the property $\alpha_i \alpha_i^* = 0, i = 1, \dots, l$ still holds. In section 2, we discuss this in more detail.

Recently there have been implementations without using pairs of indices—for example, LIBSVM (Chang & Lin, 2001), SVMtorch (Collobert & Bengio, 2001), and mySVM (Rüping, 2000). A question immediately raised is on the performance of these two approaches—the base approach and the pair approach. The pair approach solves a larger subproblem in each iteration, so there may be fewer iterations; however, a larger subproblem takes more time so the cost of each iteration is higher.

Collobert and Bengio (2001) have stated that working with pairs of variables would force the algorithm to do many computations with null variables until the end of the optimization process. In section 3, we elaborate

on this in a detailed proof. First, we consider approaches with the smallest size of working set (two and four elements for both approaches) where the analytic solution of the subproblem is handily available from sequential minimal optimization (SMO) (Platt, 1998). From mathematical explanations, we show that while solving the subproblems of the pair set containing four variables, in most cases, only those two variables in the base set are updated. Therefore, the number of iterations of both the base and the pair approaches is nearly the same. In addition, for larger working sets, we prove that after some finite number of iterations, the subproblem using only the base set is already optimal for the subproblem using the pair set. These give us theoretical justification that it is not necessary to use pairs of variables.

In section 4, we conduct experiments to demonstrate the validity of our analysis. Then in section 5, we provide some conclusions and a discussion.

There are other decomposition approaches for support vector regression (for example, Flake & Lawrence, 2002). They deal with different situations, which will not be discussed here.

2 Working Set Selection

Here we consider the working set selection from Joachims (1998) and Keerthi, Shevade, Bhattacharyya, and Murthy (2001) which were originally designed for classification. We then apply them for SVR. To make SVR similar to the form of classification, we define the following $2l$ by 1 vectors:

$$\alpha^{(*)} \equiv \begin{bmatrix} \alpha \\ \alpha^* \end{bmatrix}, \text{ and } y_i \equiv \begin{cases} +1 & i = 1, \dots, l, \\ -1, & i = l+1, \dots, 2l. \end{cases} \quad (2.1)$$

Then the regression problem, 1.1, can be reformulated as

$$\begin{aligned} \min f(\alpha^{(*)}) &= \frac{1}{2}(\alpha^{(*)})^T \begin{bmatrix} Q & -Q \\ -Q & Q \end{bmatrix} \alpha^{(*)} + [\epsilon e^T + z^T, \epsilon e^T - z^T] \alpha^{(*)} \\ 0 &\leq \alpha_i^{(*)} \leq C, \quad i = 1, \dots, 2l, \\ y^T \alpha^{(*)} &= 0. \end{aligned} \quad (2.2)$$

Now f is the objective function of equation 2.2.

Practically we can use the Karush-Kuhn-Tucker (KKT) condition to test if a given $\alpha^{(*)}$ is an optimal solution of equation 2.2. If there exists a number b such that for all $i = 1, 2, \dots, 2l$,

$$\nabla f(\alpha^{(*)})_i + by_i \geq 0 \quad \text{if } \alpha_i^{(*)} = 0, \quad (2.3a)$$

$$\nabla f(\alpha^{(*)})_i + by_i \leq 0 \quad \text{if } \alpha_i^{(*)} = C, \quad (2.3b)$$

$$\nabla f(\alpha^{(*)})_i + by_i = 0 \quad \text{if } 0 < \alpha_i^{(*)} < C, \quad (2.3c)$$

a feasible $\alpha^{(*)}$ is optimal for equation 2.2. Note that the range of b can be

determined by

$$m(\alpha^{(*)}) \equiv \max_{1 \leq t \leq 2l} \left(\max_{\alpha_t^{(*)} < C, y_t = 1} -y_t \nabla f(\alpha^{(*)})_t, \right. \\ \left. \max_{\alpha_t^{(*)} > 0, y_t = -1} -y_t \nabla f(\alpha^{(*)})_t \right), \quad (2.4)$$

$$M(\alpha^{(*)}) \equiv \min_{1 \leq t \leq 2l} \left(\min_{\alpha_t^{(*)} > 0, y_t = 1} -y_t \nabla f(\alpha^{(*)})_t, \right. \\ \left. \min_{\alpha_t^{(*)} < C, y_t = -1} -y_t \nabla f(\alpha^{(*)})_t \right). \quad (2.5)$$

That is, a feasible $\alpha^{(*)}$ is an optimal solution if and only if $m(\alpha^{(*)}) \leq b \leq M(\alpha^{(*)})$, or equivalently,

$$M(\alpha^{(*)}) - m(\alpha^{(*)}) \geq 0. \quad (2.6)$$

For convenience, we define the candidates of $m(\alpha^{(*)})$ as the set of all indices t that satisfy $\alpha_t^{(*)} < C$, $y_t = 1$ or $\alpha_t^{(*)} > 0$, $y_t = -1$, where $1 \leq t \leq 2l$. Similarly, we can define the candidates of $M(\alpha^{(*)})$.

At the beginning of iteration k , let $\alpha^{(*),k} = [\alpha^k, (\alpha^*)^k]^T$ be the vector that we are working on. Then we denote $m_k \equiv m(\alpha^{(*),k})$ and $M_k \equiv M(\alpha^{(*),k})$. Also, let $\arg m_k$ be the subset of indices t in the candidates of m_k such that $-y_t \nabla f(\alpha^{(*),k})_t = m(\alpha^{(*),k})$. Similarly, we define $\arg M_k$.

Thus, during iterations of the decomposition method, $\alpha^{(*),k}$ is not optimal yet, so

$$m_k > M_k, \text{ for all } k. \quad (2.7)$$

If we would like to select two elements as the working set, intuitively we tend to choose indices i and j , which satisfy

$$i \in \arg m_k \text{ and } j \in \arg M_k, \quad (2.8)$$

since they cause the maximal violation of the KKT condition.

A systematic way to select a larger working set in each iteration is as follows. If q , an even number, is the size of the working set, $q/2$ indices are sequentially selected from the largest $-y_i \nabla f(\alpha^{(*)})_i$ values to the smaller ones in the candidate set of m_k , that is,

$$-y_{i_1} \nabla f(\alpha^{(*),k})_{i_1} \geq \dots \geq -y_{i_{q/2}} \nabla f(\alpha^{(*),k})_{i_{q/2}},$$

where $i_1 \in \arg m_k$. The other $q/2$ indices are sequentially selected from the smallest $-y_i \nabla f(\alpha^{(*)})_i$ values to the larger ones in the candidate set of M_k , that is,

$$-y_{j_{q/2}} \nabla f(\alpha^{(*),k})_{j_{q/2}} \geq \dots \geq -y_{j_1} \nabla f(\alpha^{(*),k})_{j_1},$$

where $j_1 \in \arg M_k$. Also, we have

$$-y_{j_{q/2}} \nabla f(\alpha^{(*)k})_{j_{q/2}} < -y_{i_{q/2}} \nabla f(\alpha^{(*)k})_{i_{q/2}}, \quad (2.9)$$

to ensure that the intersection of these two groups is empty. Thus, if q is large, sometimes the actual number of selected indices may be less than q .

Note that this is the same as the working set selection in Joachims (1998). However, the original derivation in Joachims was from the concept of feasible directions for constrained optimization problems but not from the violation of the KKT condition.

After the base set of q indices is selected, earlier approaches (Keerthi et al., 2000; Laskov 2002) expand the set so all elements in it are pairs. The reason is to keep the property that $\alpha_i^k(\alpha^*)_i^k = 0, i = 1, \dots, l$, for all k . However, if directly using elements in the base set, the following theorem has been proved in (Lin 2001a, theorem 4.1):

Theorem 1. *If the initial solution is zero, then $\alpha_i^k(\alpha^*)_i^k = 0, i = 1, \dots, l$ for all k .*

Hence we know that $\alpha_i^k(\alpha^*)_i^k = 0$ is not a particular advantage of using pairs of indices. Another important issue for the decomposition method is the stopping criterion. From equation 2.6, a natural choice of the stopping criterion is

$$M_k - m_k \geq -\delta, \quad (2.10)$$

where δ , the stopping tolerance, is a small, positive number. For $q = 2$, equation 2.10 is the same as

$$-y_j \nabla f(\alpha^{(*)k})_j - (-y_i \nabla f(\alpha^{(*)k})_i) \geq -\delta, \quad (2.11)$$

where i, j are selected by equation 2.8.

Note that the convergence of the decomposition method under some conditions of the kernel matrix Q is shown in Lin (2001a) for the base approach. Some theoretical justification on the use of the stopping criteria, equation 2.10, for the decomposition method is in Lin (2001b). No particular convergence proof has been made for the pair approach, but we will assume it for our analyses.

3 Number of Iterations

In this section, we discuss the relationship between the solutions of subproblems using the base and the pair approaches. The discussion is divided into two parts. First, we consider approaches with the smallest size of working

set (two and four elements for both approaches). We show that for most iterations, the optimal solution of the subproblem using the base set is already optimal for the subproblem using the pair set, so the difference between the number of iterations of the base and pair approaches should not be large. Then we consider larger working sets. Although the result is not as elegant as the first part, we can still show that after enough iterations, the optimal solution of the subproblem using the base set is the same as the optimal solution of the subproblem using the pair set.

To start our proof, we state an important property on the difference between the i th and $(i + l)$ th gradient elements. Consider α_i and α_i^* , $1 \leq i \leq l$. We have

$$\begin{aligned}\nabla f(\alpha^{(*)})_{i+l} &= -(Q(\alpha - \alpha^*))_i + \epsilon - z_i \\ &= -\nabla f(\alpha^{(*)})_i + 2\epsilon.\end{aligned}$$

Note that $y_i = 1$ and $y_{i+l} = -1$ as defined in equation 2.1, so

$$-y_{i+l}\nabla f(\alpha^{(*)})_{i+l} = -y_i\nabla f(\alpha^{(*)})_i + 2\epsilon. \quad (3.1)$$

We will use this in later analyses.

When $q = 2$, the base set is selected from equation 2.8. It is easy to see that indices i and $i + l$ where $1 \leq i \leq l$ cannot both be chosen at the same time. For example, if indices i and $i + l$ are both selected from equation 2.8, by equations 3.1 and 2.7, we must have $i + l \in \arg m_k$ and $i \in \arg M_k$. By equations 2.4 and 2.5, this means $\alpha_{i+l}^{(*),k} > 0$ and $\alpha_i^{(*),k} > 0$, which violates theorem 1. Therefore, $(i, i + l)$ cannot be chosen together.

Also, if one of $\alpha_i^{(*),k}$ and $\alpha_{i+l}^{(*),k}$ is selected, the other must be zero. We can prove this by contradiction. Without loss of generality, say if $\alpha_i^{(*),k}$ is selected and $\alpha_{i+l}^{(*),k}$ is nonzero, then by theorem 1, $\alpha_i^{(*),k}$ must be zero. So by equations 2.4, 2.5, and 2.8, $i \in \arg m_k$. Moreover, by equation 2.4, both i and $i + l$ are in the candidate set of m_k , and we must have $-y_i\nabla f(\alpha^{(*)},k)_i \geq -y_{i+l}\nabla f(\alpha^{(*)},k)_{i+l}$ since $i \in \arg m_k$. But this contradicts equation 3.1. Therefore, if one of $\alpha_i^{(*),k}$ and $\alpha_{i+l}^{(*),k}$ is selected, the other must be zero.

If any one of (i, j) , $(i, j + l)$, $(i + l, j)$, $(i + l, j + l)$ where $1 \leq i, j \leq l$ is chosen from equation 2.8, our goal is to see the difference in solving the two-variable subproblem and the four-variable subproblem of $(i, j, i + l, j + l)$.

Without loss of generality, we consider the case where (i, j) is chosen by equation 2.8. Then $(\alpha_i^{(*),k}, \alpha_j^{(*),k}, \alpha_{i+l}^{(*),k}, \alpha_{j+l}^{(*),k})$ are the corresponding variables at iteration k ; from earlier discussions, we know $\alpha_{i+l}^{(*),k} = \alpha_{j+l}^{(*),k} = 0$. After a two-variable subproblem on $\alpha_i^{(*)}$ and $\alpha_j^{(*)}$ is solved, we assume the new values are $(\bar{\alpha}_i^{(*),k}, \bar{\alpha}_j^{(*),k}, 0, 0)$. From equation 3.1 and the KKT condition, it is easy to see that if $\bar{\alpha}_i^{(*),k} > 0$ and $\bar{\alpha}_j^{(*),k} > 0$, $(\bar{\alpha}_i^{(*),k}, \bar{\alpha}_j^{(*),k}, 0, 0)$ is already an optimal solution of the four-variable problem, equation 1.2.

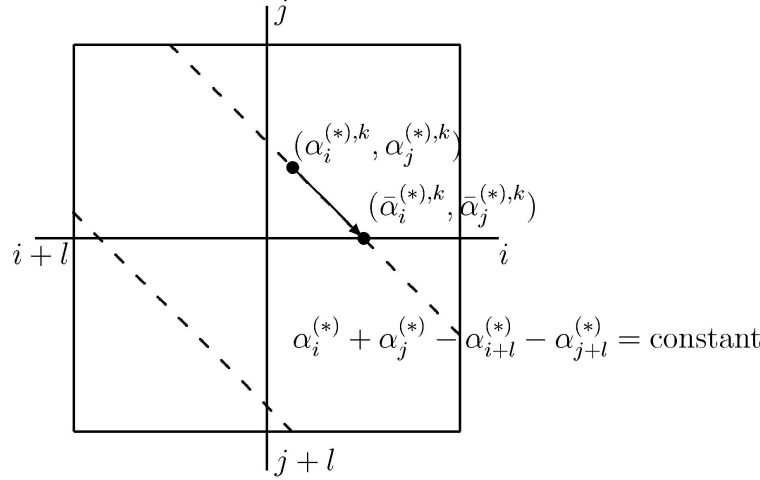


Figure 1: Possible situation of plane changes.

Therefore, the only difference happens when there is a “jump” from the (i, j) plane to another plane of two variables and the objective value of equation 1.2 can be further decreased. We illustrate this in Figure 1.

In this figure, each square represents a plane of two nonzero variables. From the linear constraint

$$\alpha_i^{(*)} + \alpha_j^{(*)} = - \sum_{t \neq i, j} y_t \alpha_t^{(*)}, \quad (3.2)$$

the two dashed parallel lines in Figure 1 show how the solution plane possibly changes. For example, after $\bar{\alpha}_j^{(*)},k$ becomes zero, if $(\bar{\alpha}_i^{(*)},k, \bar{\alpha}_j^{(*)},k, 0, 0)$ is not an optimal solution of equation 1.2, we may further reduce its objective value by entering the (i, j^*) plane. We will check under what conditions $(\bar{\alpha}_i^{(*)},k, \bar{\alpha}_j^{(*)},k, 0, 0)$ is not an optimal solution of equation 1.2.

Since $\alpha_i^{(*)}$ and $\alpha_j^{(*)}$ are adjusted on the line, equation 3.2, we consider the objective value of the subproblem on the (i, j) plane as the following function of a single variable v , where $N = \{t \mid 1 \leq t \leq l, t \neq i, t \neq j\}$ are indices of the fixed variables:

$$g(v) \equiv \frac{1}{2} \begin{bmatrix} \alpha_i^{(*)},k + v & \alpha_j^{(*)},k - v \end{bmatrix} \begin{bmatrix} Q_{ii} & Q_{ij} \\ Q_{ji} & Q_{jj} \end{bmatrix} \begin{bmatrix} \alpha_i^{(*)},k + v \\ \alpha_j^{(*)},k - v \end{bmatrix} \\ + (Q_{i,N}(\alpha_N^k - (\alpha_N^*)^k) + \epsilon + z_i)(\alpha_i^{(*)},k + v)$$

$$\begin{aligned}
& + (Q_{i,N}(\alpha_N^k - (\alpha^*)_N^k) + \epsilon + z_i)(\alpha_j^{(*)k} - v) \\
& = \frac{1}{2}(Q_{ii} - 2Q_{ij} + Q_{jj})v^2 + (\nabla f(\alpha^{(*)k})_i - \nabla f(\alpha^{(*)k})_j)v + \text{constant}.
\end{aligned}$$

Since $\alpha_i^{(*)}$ is increased from $\alpha_i^{(*)k}$ to $\bar{\alpha}_i^{(*)k}$, we know

$$g'(0) = \nabla f(\alpha^{(*)k})_i - \nabla f(\alpha^{(*)k})_j < 0.$$

Now if $(\bar{\alpha}_i^{(*)k}, \bar{\alpha}_j^{(*)k}, 0, 0)$ is not an optimal solution of equation 1.2, we can define a new function $\bar{g}(v)$ similar to $g(v)$ at $(\bar{\alpha}_i^{(*)k}, 0)$ of the $(i, j + l)$ plane. If $\bar{\alpha}_i^{(*)k}$ can be further increased,

$$\bar{g}'(0) = \nabla f(\bar{\alpha}^{(*)k})_i + \nabla f(\bar{\alpha}^{(*)k})_{j+l} < 0. \quad (3.3)$$

However, from equation 3.1 and $\bar{\alpha}_i^{(*)k} - \alpha_i^{(*)k} = -(\bar{\alpha}_j^{(*)k} - \alpha_j^{(*)k})$,

$$\begin{aligned}
& \nabla f(\bar{\alpha}^{(*)k})_i + \nabla f(\bar{\alpha}^{(*)k})_{j+l} \\
& = \nabla f(\bar{\alpha}^{(*)k})_i - \nabla f(\bar{\alpha}^{(*)k})_j + 2\epsilon \\
& = \nabla f(\alpha^{(*)k})_i + Q_{ii}(\bar{\alpha}_i^{(*)k} - \alpha_i^{(*)k}) + Q_{ij}(\bar{\alpha}_j^{(*)k} - \alpha_j^{(*)k}) \\
& \quad - (\nabla f(\alpha^{(*)k})_j + Q_{ji}(\bar{\alpha}_i^{(*)k} - \alpha_i^{(*)k}) + Q_{jj}(\bar{\alpha}_j^{(*)k} - \alpha_j^{(*)k})) + 2\epsilon \\
& = (\nabla f(\alpha^{(*)k})_i - \nabla f(\alpha^{(*)k})_j) \\
& \quad + (\bar{\alpha}_i^{(*)k} - \alpha_i^{(*)k})(Q_{ii} - 2Q_{ij} + Q_{jj}) + 2\epsilon.
\end{aligned} \quad (3.4)$$

Since Q is positive semidefinite, $Q_{ii}Q_{jj} - Q_{ij}^2 \geq 0$ implies $Q_{ii} - 2Q_{ij} + Q_{jj} \geq 0$. With $\bar{\alpha}_i^{(*)k} - \alpha_i^{(*)k} \geq 0$ and equation 3.3, we know if

$$(\nabla f(\alpha^{(*)k})_i - \nabla f(\alpha^{(*)k})_j) + (\bar{\alpha}_i^{(*)k} - \alpha_i^{(*)k})(Q_{ii} - 2Q_{ij} + Q_{jj}) + 2\epsilon \geq 0, \quad (3.5)$$

it is impossible to move $(\bar{\alpha}_i^{(*)k}, 0)$ on $(i, j + l)$ plane further. That is, $(\bar{\alpha}_i^{(*)k}, \bar{\alpha}_j^{(*)k}, 0, 0)$ is already an optimal solution of equation 1.2. For other cases, that is, $(i, j + l)$, $(i + l, j)$, and $(i + l, j + l)$, results are the same. Note that now, $1 \leq i, j \leq l$ so $\nabla f(\alpha^{(*)k})_i - \nabla f(\alpha^{(*)k})_j$ is actually the value obtained in equation 2.11. That is, it is the number used for checking the stopping criterion. Therefore, we have the following theorem:

Theorem 2. *For all iterations with the violation on the stopping criterion, equation 2.11, no more than 2ϵ , an optimal solution of the two-variable subproblem is already an optimal solution of the corresponding four-variable subproblem.*

If ϵ is not small, the stopping tolerance in most iterations is smaller than 2ϵ . In addition, as most decomposition iterations are spent in the final stage

due to slow convergence, this theorem has shown a conclusive result that no matter whether the two-variable or four-variable approach is used, the difference on the number of iterations should not be much.

For larger working set ($q > 2$), we may not be able to get results as elegant as theorem 2. When $q = 2$, for example, we know exactly the relation on the changes of $\alpha_i^{(*),k}$ and $\alpha_j^{(*),k}$ in one iteration, as $\bar{\alpha}_i^{(*),k} - \alpha_i^{(*),k} = -(\bar{\alpha}_j^{(*),k} - \alpha_j^{(*),k})$. However, when $q > 2$, the change on each variable can be different. In the following, we will show that if $\{\alpha^{(*),k}\}$ is an infinite sequence, during final iterations, after k is large enough, solving the subproblem of the base set is the same as solving the larger subproblem of the pair set. Next, we describe some properties that will be used for the proof.

Assume that the sequence $\{\alpha^{(*),k}\}$ of the base approach converges to an optimal solution $\hat{\alpha}^{(*)}$. Then we can define

$$\hat{M} \equiv M(\hat{\alpha}^{(*)}) \text{ and } \hat{m} \equiv m(\hat{\alpha}^{(*)}). \quad (3.6)$$

We also note that equation 2.9 implies that for any index $1 \leq i \leq 2l$ in the working set of the k th iteration,

$$M_k \leq -y_i \nabla f(\alpha^{(*),k})_i \leq m_k. \quad (3.7)$$

Now we describe two theorems from Lin (2001b) that are needed for the main proof. These theorems deal with a general framework of decomposition methods for different SVM formulations. We can easily check that the base approach satisfies the required conditions of these two theorems so they can be applied:

Theorem 3.

$$\lim_{k \rightarrow \infty} m_k - M_k = 0. \quad (3.8)$$

Theorem 4. For any $\hat{\alpha}_i^{(*)}$, $1 \leq i \leq 2l$, whose corresponding $-y_i \nabla f(\hat{\alpha}^{(*)})_i$ is neither \hat{m} nor \hat{M} , after k is large enough, $\alpha_i^{(*),k}$ is at a bound and is equal to $\hat{\alpha}_i^{(*)}$.

Immediately, we have a corollary of theorem 3, which is specific to SVR:

Corollary 1. After k is large enough, for all $i = 1, 2, \dots, l$, $\alpha_i^{(*),k}$ and $\alpha_{i+l}^{(*),k}$ would not be both selected in the base working set.

Proof. By the convergence of $m_k - M_k$ to 0, after k is large enough, $m_k - M_k < \epsilon$. If there exists $1 \leq i \leq l$ such that $\alpha_i^{(*),k}$ and $\alpha_{i+l}^{(*),k}$ are both selected in the base working set, from equation 3.7, $M_k \leq -\nabla f(\alpha^{(*),k})_i \leq m_k$ and

$M_k \leq \nabla f(\alpha^{(*)k})_{i+l} \leq m_k$. However, equation 3.1 shows $\nabla f(\alpha^{(*)k})_{i+l} = -\nabla f(\alpha^{(*)k})_i + 2\epsilon$ so $m_k - M_k \geq 2\epsilon$, and there is a contradiction.

Next, we describe the main proof of this section, which is an analysis on the infinite sequence $\{\alpha^{(*)k}\}$.

Theorem 5. *We assume that $\hat{M} \neq \hat{m} + 2\epsilon$. After k is large enough, any optimization subproblem of the base set is already optimal for the larger subproblem of the pair set.*

Proof. If the result is wrong, there is an index $1 \leq i \leq l$ and an infinite set \mathcal{K} such that for all $k \in \mathcal{K}$, $\alpha_i^{(*)k}$ (or $\alpha_{i+l}^{(*)k}$) is selected in the working set, but then $\alpha_{i+l}^{(*)k}$ (or $\alpha_i^{(*)k}$) is also modified. Without loss of generality, we assume that $\alpha_i^{(*)k}$ is selected in the working set, but $\alpha_{i+l}^{(*)k}$ is modified infinite times. So by theorem 4,

$$\nabla f(\hat{\alpha}^{(*)})_{i+l} = \hat{m}, \text{ or } \nabla f(\hat{\alpha}^{(*)})_{i+l} = \hat{M}.$$

By equation 3.1,

$$-\nabla f(\hat{\alpha}^{(*)})_i = \hat{m} - 2\epsilon < \hat{m}, \text{ or } -\nabla f(\hat{\alpha}^{(*)})_i = \hat{M} - 2\epsilon < \hat{M}.$$

For the second case, by the assumption that $\hat{m} \neq \hat{M} - 2\epsilon$, we have $\hat{m} < -\nabla f(\hat{\alpha}^{(*)})_i$ or $\hat{m} > -\nabla f(\hat{\alpha}^{(*)})_i$. But if $\hat{m} < -\nabla f(\hat{\alpha}^{(*)})_i$, we have $\hat{m} < -\nabla f(\hat{\alpha}^{(*)})_i < \hat{M}$, which is impossible for an optimal solution. Hence,

$$-y_i \nabla f(\hat{\alpha}^{(*)})_i < \hat{m} \quad (3.9)$$

holds for both cases. Therefore, we can define

$$\Delta \equiv \min(\epsilon/2, (\hat{m} - (-y_i \nabla f(\hat{\alpha}^{(*)})_i))/3) > 0.$$

By the convergence of the sequence $\{-y_j \nabla f(\alpha^{(*)k})_j\}$ to $-y_j \nabla f(\hat{\alpha}^{(*)})_j$, for all $j = 1, \dots, 2l$, after k is large enough,

$$|y_j \nabla f(\alpha^{(*)k})_j - y_j \nabla f(\alpha^{(*)k+1})_j| \leq \Delta \text{ and} \quad (3.10)$$

$$|y_j \nabla f(\alpha^{(*)k})_j - y_j \nabla f(\hat{\alpha}^{(*)})_j| \leq \Delta. \quad (3.11)$$

Suppose that at the k th iteration, $j \in \arg M_k$ is selected in the working set and

$$-y_j \nabla f(\alpha^{(*)k})_j = M_k.$$

By equations 3.7, 3.10, 3.11, and 3.9,

$$-y_j \nabla f(\hat{\alpha}^{(*)})_j \leq -y_j \nabla f(\alpha^{(*)k})_j + \Delta = M_k + \Delta$$

$$\begin{aligned}
&\leq -y_i \nabla f(\alpha^{(*)k})_i + \Delta \leq -y_i \nabla f(\hat{\alpha}^{(*)})_i + 2\Delta \\
&\leq -y_i \nabla f(\hat{\alpha}^{(*)})_i + 2(\hat{m} - (-y_i \nabla f(\hat{\alpha}^{(*)})_i))/3 \\
&< \hat{m} \leq \hat{M}.
\end{aligned} \tag{3.12}$$

From theorem 4 and equation 3.12, after k is large enough, $\alpha_j^{(*)k}$ is at a bound and is equal to $\hat{\alpha}_j^{(*)}$. That is, $\alpha_j^{(*)k} = \alpha_j^{(*)k+1} = \hat{\alpha}_j^{(*)}$. Since $\alpha_j^{(*)k+1} = \alpha_j^{(*)k}$ and $\alpha_j^{(*)k}$ is in the candidates of M_k , by equations 3.7, 3.10, and 3.11,

$$\begin{aligned}
M_{k+1} &\leq -y_j \nabla f(\alpha^{(*)k+1})_j \\
&\leq -y_j \nabla f(\alpha^{(*)k})_j + \Delta = M_k + \Delta \\
&\leq -y_i \nabla f(\alpha^{(*)k})_i + \Delta \\
&\leq -y_i \nabla f(\alpha^{(*)k+1})_i + 2\Delta.
\end{aligned}$$

Hence, we get

$$M_{k+1} \leq -y_i \nabla f(\alpha^{(*)k+1})_i + \epsilon. \tag{3.13}$$

On the other hand, $\alpha_i^{(*)k}$ is modified to $\alpha_i^{(*)k+1}$, so at least one of them is strictly positive. By the definition of m_k ,

$$\nabla f(\alpha^{(*)k})_{i+l} \leq m_k, \text{ or } \nabla f(\alpha^{(*)k+1})_{i+l} \leq m_{k+1}. \tag{3.14}$$

From equations 3.1, 3.7, 3.13, and 3.14, for all large enough $k \in \mathcal{K}$,

$$M_k \leq m_k - 2\epsilon, \text{ or } M_{k+1} \leq m_{k+1} - \epsilon.$$

Therefore,

$$\lim_{k \rightarrow \infty} m_k - M_k \neq 0$$

which contradicts theorem 3.

4 Experiments

We consider two regression problems, *abalone* (4177 data) and *add10* (9792 data), from Blake and Merz (1998) and Friedman (1988), respectively. The RBF kernel is used:

$$Q_{ij} \equiv e^{-\gamma \|x_i - x_j\|^2}.$$

Since our purpose is not to examine the quality of the solutions, we do not perform model selection on the value of γ . Instead, we fix it to $1/n$, where n is the number of attributes in each data. For these two problems, n is

Table 1: Problem abalone.

| Parameters | Iteration (2-variable) | Iteration (4-variable) | SV | Candidates ^a | Jumps ^b |
|---------------------------|---------------------------|---------------------------|------|-------------------------|--------------------|
| $C = 10, \epsilon = 0.1$ | 18,930 | 18,790 | 3967 | 8438 | 14 |
| $C = 10, \epsilon = 1$ | 10,173 | 10,173 | 2183 | 1705 | 0 |
| $C = 10, \epsilon = 10$ | 17 | 17 | 7 | 0 | 0 |
| $C = 100, \epsilon = 0.1$ | 142,190 | 140,686 | 3938 | 63,057 | 207 |
| $C = 100, \epsilon = 1$ | 122,038 | 119,981 | 2147 | 10,187 | 48 |
| $C = 100, \epsilon = 10$ | 261 | 261 | 9 | 0 | 0 |

Notes: ^aThe number of iterations that violate conditions of Theorem 2.

^bThe number of plane changes as illustrated in Figure 1.

Table 2: Problem add10.

| Parameters | Iteration (2-variable) | Iteration (4-variable) | SV | Candidates | Jumps |
|---------------------------|---------------------------|---------------------------|------|------------|-------|
| $C = 10, \epsilon = 0.1$ | 28,625 | 28,955 | 9254 | 13,918 | 1 |
| $C = 10, \epsilon = 1$ | 22,067 | 21,913 | 4997 | 2795 | 1 |
| $C = 10, \epsilon = 10$ | 116 | 116 | 14 | 0 | 0 |
| $C = 100, \epsilon = 0.1$ | 350,344 | 350,325 | 9158 | 109,644 | 12 |
| $C = 100, \epsilon = 1$ | 227,604 | 227,604 | 4265 | 14,628 | 0 |
| $C = 100, \epsilon = 10$ | 105 | 105 | 11 | 0 | 0 |

eight and ten, respectively. Based on our experience, we think that it is an appropriate value when data are scaled to $[-1, 1]$.

Tables 1 and 2 present results using different ϵ and C on two problems. We consider ϵ to be greater than 0.1 because for smaller ϵ , the number of support vectors approaches the number of training data. On the other hand, we consider ϵ up to 10 where the number of support vectors is close to zero. For each parameter set, we present the number of iterations by both two-variable and four-variable approaches, number of support vectors, number of iterations that violate conditions of theorem 2 (i.e., possible candidates of jumps), and the number of real jumps as illustrated in Figure 1 when using the four-variable approach.

The solution of the four-variable approach is obtained as follows. First a two-variable problem obtained from equation 2.8 is solved. If there is at least one variable that goes to zero, another two-variable problem has to be solved. As indicated in Figure 1, at most three two-variable problems are needed. For our experiments, both versions of the code are directly modified from LIBSVM (version 2.03).

It can be clearly seen that both approaches take nearly the same number of iterations. In addition, the number of jumps while using the four-variable

Table 3: Problem abalone (First 200 Data).

| Parameters | $q = 10$ | | | $q = 20$ | | |
|---------------------------|---------------------|----------------------|-------------------|---------------------|----------------------|-------------------|
| | Iteration (base) | Iteration (pairs) | NAVM ^a | Iteration (base) | Iteration (pairs) | NAVM ^a |
| $C = 10, \epsilon = 0.1$ | 132 | 173 | 77 | 100 | 114 | 111 |
| $C = 10, \epsilon = 1$ | 49 | 46 | 3 | 42 | 40 | 10 |
| $C = 10, \epsilon = 10$ | 1 | 1 | 0 | 1 | 1 | 0 |
| $C = 100, \epsilon = 0.1$ | 1641 | 1983 | 184 | 1168 | 1086 | 250 |
| $C = 100, \epsilon = 1$ | 807 | 552 | 29 | 258 | 310 | 46 |
| $C = 100, \epsilon = 10$ | 1 | 1 | 0 | 1 | 1 | 0 |

Note: ^aNumber of added variables that are modified for the pair approach.

approach is very small, especially when ϵ is larger. Furthermore, the number of “candidates” is much larger than the number of real jumps. This means that $(\bar{\alpha}_i^{(*),k} - \alpha_i^{(*),k})(Q_{ii} - 2Q_{ij} + Q_{jj})$ of equation 3.4 is large enough so equation 3.5 is usually satisfied.

Next, we experiment with using larger working sets. We use a simple implementation written in MATLAB so only small problems are tested. We consider the first 200 data points of abalone and add10. Results are in Tables 3 and 4, where we show the number of iterations and the number of added variables that are modified (NAVM) for the pair approach. Note that we use “number of added variables that are modified” instead of “number of jumps” since for larger subproblems, we cannot model the change of variables as jumps between planes.

The NAVM column is defined as follows. If the pair approach is used, the working set in the k th iteration is the union of two sets: B_k , which is the base set, and its extension, \bar{B}_k . We check the value of variables in $\alpha_{\bar{B}_k}^{(*)}$ before and after solving the subproblem. NAVM is the sum of the count for those modified variables in $\alpha_{\bar{B}_k}^{(*)}$ throughout all iterations, so it is at most the sum of $|\bar{B}_k|$ throughout all iterations, which is roughly the number of iterations multiplied by the maximal size of the base set, q .

In Tables 3 and 4, the column NAVM is relatively small compared to the number of iterations multiplied by q . That means that in nearly all the optimization steps, only variables corresponding to the base set rather than the extended set are changed. In addition, from Tables 1 through 4, we found that the pair approach may not lead to fewer iterations, so it is not necessary to use the pair approach for solving SVR.

Table 4: Problem add10 (First 200 Data).

| Parameters | $q = 10$ | | | $q = 20$ | | |
|---------------------------|---------------------|----------------------|-------------------|---------------------|----------------------|-------------------|
| | Iteration (base) | Iteration (pairs) | NAVM ^a | Iteration (base) | Iteration (pairs) | NAVM ^a |
| $C = 10, \epsilon = 0.1$ | 59 | 50 | 13 | 27 | 26 | 33 |
| $C = 10, \epsilon = 1$ | 55 | 60 | 0 | 18 | 18 | 0 |
| $C = 10, \epsilon = 10$ | 2 | 2 | 0 | 2 | 2 | 0 |
| $C = 100, \epsilon = 0.1$ | 2519 | 2094 | 112 | 1317 | 1216 | 135 |
| $C = 100, \epsilon = 1$ | 944 | 956 | 12 | 236 | 278 | 28 |
| $C = 100, \epsilon = 10$ | 2 | 2 | 0 | 2 | 2 | 0 |

Note: ^aNumber of added variables modified for the pair approach.

5 Conclusion and Discussion

From our theoretical proofs, we show that in the final iterations of the decomposition methods, the solution of the subproblem for the base approach is the same as that for the pair approach. This means that extending the base working set to the pair set will not benefit much, so it is not necessary to use the pair method.

Experiments confirmed our analysis. The difference between number of iterations of the two approaches is negligible. Moreover, the pair approach solves a larger optimization subproblem in each iteration, which costs more time, so the program using the base approach is more efficient.

We mentioned in equation 2.2 that the regression problem can be reformulated to have the same structure as the classification problem, so if we solve SVR using the base approach, it is possible to use the same program for classification with little modification. For example, LIBSVM used this strategy. However, if equation 2.2 is directly applied without using as many regression properties as possible, our experience shows that the performance may be a little worse than software specially designed for regression.

Acknowledgments

This work was supported in part by the National Science Council of Taiwan, grant NSC 89-2213-E-002-106.

References

- Blake, C. L., & Merz, C. J. (1998). *UCI repository of machine learning databases* (Tech. Rep.). Irvine: University of California, Department of Information and Computer Science. Available on-line: <http://www.ics.uci.edu/~mllearn/MLRepository.html>.

- Chang, C.-C., & Lin, C.-J. (2001). LIBSVM: A library for support vector machines. [Computer software]. Available on-line: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- Collobert, R., & Bengio, S. (2001). SVMtorch: A support vector machine for large-scale regression and classification problems. *Journal of Machine Learning Research*, 1, 143–160. Available on-line: <http://www.idiap.ch/learning/SVMtorch.html>.
- Flake, G. W., & Lawrence, S. (2002). Efficient SVM regression training with SMO. *Machine Learning*, 46, 271–290.
- Friedman, J. (1988). *Multivariate adaptive regression splines* (Tech. Rep. No. 102). Stanford, CA: Laboratory for Computational Statistics, Department of Statistics, Stanford University.
- Joachims, T. (1998). Making large-scale SVM learning practical. In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in kernel methods—support vector learning*. Cambridge, MA: MIT Press.
- Keerthi, S. S., Shevade, S., Bhattacharyya, C., & Murthy, K. (2000). Improvements to SMO algorithm for SVM regression. *IEEE Transactions on Neural Networks*, 11, 1188–1193.
- Keerthi, S. S., Shevade, S., Bhattacharyya, C., & Murthy, K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, 13, 637–649.
- Laskov, P. (2002). An improved decomposition algorithm for regression support vector machines. *Machine Learning*, 46, 315–350.
- Lin, C.-J. (2001a). On the convergence of the decomposition method for support vector machines. *IEEE Transactions on Neural Networks*, 12, 1288–1298.
- Lin, C.-J. (2001b). *Stopping criteria of decomposition methods for support vector machines: a theoretical justification* (Tech. Rep.) Taipei, Taiwan: Department of Computer Science and Information Engineering, National Taiwan University. To appear in *IEEE Transactions on Neural Networks*.
- Platt, J. C. (1998). Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in kernel methods—support vector learning*. Cambridge, MA: MIT Press.
- Rüping, S. (2000). mySVM—another one of those support vector machines. [Computer software]. Available on-line: <http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/>.
- Smola, A. J., & Schölkopf, B. (1998). *A tutorial on support vector regression*. (Neuro COLT Tech. Rep. TR-1998-030). Egham, Surrey: Royal Holloway College.
- Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.