Manuscript Number: 2187

Training ν-Support Vector Classifiers: Theory and Algorithms

Chih-Chung Chang and Chih-Jen Lin*

Abstract

The ν-support vector machines (ν-SVM) for classification proposed by Schölkopf et al. has the advantage of using a parameter ν on controlling the number of support vectors. However, comparing to regular C-SVM, its formulation is more complicated so up to now there are no effective methods for solving large-scale ν-SVM. In this paper, we modify the ν-SVM to a quadratic program with bound constraints and one simple equality constraint. Then existing decomposition methods can be modified to solve it. We demonstrate a decomposition method similar to the software SVM^{light} for regular C-SVM. Motivated from the possible infeasibility of the dual ν-SVM formulation, we investigate the relation between ν-SVM and C-SVM in detail. We show that in general they are two different problems with the same optimal solution set. Hence we may expect that many numerical aspects on solving them are similar. We also discuss the behavior of ν-SVM by some numerical experiments.

1 Introduction

The ν-support vector classification (Schölkopf et al. 2000; Schölkopf et al. 1999) is a new class of support vector machines (SVM). Given training vectors $\mathbf{x}_i \in \mathbb{R}^n, i = 1, \ldots, l$ in two classes, and a vector $\mathbf{y} \in \mathbb{R}^l$ such that $y_i \in \{1, -1\}$, they consider the following primal problem:

$$(P_\nu) \qquad \begin{aligned} \min \frac{1}{2}\mathbf{w}^T\mathbf{w} - \nu\rho + \frac{1}{l}\sum_{i=1}^{l}\xi_i \\ y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq \rho - \xi_i, \\ \xi_i \geq 0, i = 1, \ldots, l, \ \rho \geq 0. \end{aligned} \tag{1.1}$$

Here $0 \leq \nu \leq 1$ and training vectors $\mathbf{x}_i$ are mapped into a higher (maybe infinite) dimensional space by the function ϕ. This formulation is different from the original

This work was supported in part by the National Science Council of Taiwan via the grant NSC 89-2213-E-002-013.

*Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan (cjlin@csie.ntu.edu.tw).

C-SVM (Vapnik 1998):

$$(P_C) \qquad \min \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \tag{1.2}$$
$$y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i,$$
$$\xi_i \geq 0, i = 1, \ldots, l.$$

In (1.2), a parameter C is used to penalize variables ξ_i. As it is difficult to select an appropriate C, in (P_ν), Schölkopf et al. introduces a new parameter ν which lets one control the number of support vectors and errors. To be more precise, they proved that ν is an upper bound on the fraction of margin errors and a lower bound of the fraction of support vectors. In addition, with probability 1, asymptotically, ν equals to both fractions.

Although (P_ν) has such an advantage, its dual is more complicated than the dual of (P_C):

$$(D_\nu) \qquad \min \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha}$$
$$\mathbf{y}^T\boldsymbol{\alpha} = 0,\ \mathbf{e}^T\boldsymbol{\alpha} \geq \nu,$$
$$0 \leq \alpha_i \leq 1/l, \qquad i = 1, \ldots, l, \tag{1.3}$$

where $\mathbf{e}$ is the vector of all ones, $\mathbf{Q}$ is a positive semidefinite matrix, $Q_{ij} \equiv y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$, and $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ is the kernel.

Remember that the dual of (P_C) is as follows:

$$(D_C) \qquad \min \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha}$$
$$\mathbf{y}^T\boldsymbol{\alpha} = 0, 0 \leq \alpha_i \leq C, i = 1, \ldots, l.$$

Therefore, it can be clearly seen that (D_ν) has one more inequality constraint.

Due to the density of $\mathbf{Q}$, traditional optimization algorithms such as Newton, Quasi Newton, etc., cannot be directly applied to solve (D_C) or (D_ν). Currently major methods on solving large (D_C) (for example, decomposition methods (Osuna et al. 1997; Joachims 1998; Platt 1998; Saunders et al. 1998) and the method of nearest points (Keerthi et al. 2000)) utilize the simple structure of constraints. Because of the additional inequality, these methods cannot be directly used for solving (D_ν). Up to now, there are no implementation for large-scale ν-SVM.

In Section 2, we modify ν-SVM to a quadratic program with bound constraints and one simple equality constraint. Then existing decomposition methods can be modified to solve it. We then demonstrate a decomposition method similar to the software SVM^{light} (Joachims 1998) for C-SVM.

From the discussion in Section 2, we realize that it is possible that (D_ν) is infeasible so we are interested in the relation between (D_ν) and (D_C). Though in (Schölkopf et al. 2000, Proposition 13), this issue has been studied, in Section 3 we investigate this relation in more detail. The main result (Theorem 3.7) shows that solving them is just like solving two different problems with the same optimal solution set. In addition, the increase of C in C-SVM is like the decrease of ν in ν-SVM. Hence we may expect that many numerical aspects on solving them are similar.

Based on the work in Section 3, in Section 4 we derive the formulation of ν as a function of C. Section 5 presents numerical results. Experiments show that several numerical properties on solving (D_C) and (D_ν) are similar. Finally in Section 6, we make conclusions.

2 Modifications of ν-SVM and Decomposition Methods

In this section we try to reformulate and simplify (D_ν). First we show that the inequality $\mathbf{e}^T\boldsymbol{\alpha} \geq \nu$ can be treated as an equality:

Theorem 2.1 *Let $0 \leq \nu \leq 1$. If (D_ν) is feasible, there is at least one optimal solution of (D_ν) which satisfies $\mathbf{e}^T\boldsymbol{\alpha} = \nu$. In addition, if the objective value of (D_ν) is not zero, all optimal solutions of (D_ν) satisfy $\mathbf{e}^T\boldsymbol{\alpha} = \nu$.*

Proof. Since the feasible region of (D_ν) is bounded, if it is feasible, (D_ν) has at least one optimal solution. Assume (1.3) has an optimal solution $\boldsymbol{\alpha}$ such that $\mathbf{e}^T\boldsymbol{\alpha} > \nu$. Since $\mathbf{e}^T\boldsymbol{\alpha} > \nu \geq 0$, by defining

$$\bar{\boldsymbol{\alpha}} \equiv \frac{\nu}{\mathbf{e}^T\boldsymbol{\alpha}}\boldsymbol{\alpha},$$

$\bar{\boldsymbol{\alpha}}$ is feasible to (1.3) and $\mathbf{e}^T\bar{\boldsymbol{\alpha}} = \nu$. Since $\boldsymbol{\alpha}$ is an optimal solution of (D_ν), with $\mathbf{e}^T\boldsymbol{\alpha} > \nu$,

$$\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} \le \bar{\boldsymbol{\alpha}}^T\mathbf{Q}\bar{\boldsymbol{\alpha}} = (\frac{\nu}{\mathbf{e}^T\boldsymbol{\alpha}})^2\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} \le \boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha}. \tag{2.1}$$

Thus $\bar{\boldsymbol{\alpha}}$ is an optimal solution of (D_ν) and $\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} = 0$. This also implies that if the objective value of (D_ν) is not zero, all optimal solutions of (D_ν) satisfy $\mathbf{e}^T\boldsymbol{\alpha} = \nu$. □

Therefore, in general $\mathbf{e}^T\boldsymbol{\alpha} \ge \nu$ in (D_ν) can be written as $\mathbf{e}^T\boldsymbol{\alpha} = \nu$. It has been mentioned in (Schölkopf et al. 2000, Footnote 2) that practically one can alternatively work with $\mathbf{e}^T\boldsymbol{\alpha} \ge \nu$ as an equality constraint. From the primal side, it was first shown in (Crisp and Burges 1999) that $\rho \ge 0$ in (P_ν) is redundant. Without $\rho \ge 0$, the dual becomes:

$$\begin{aligned} \min & \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} \\ & 0 \le \alpha_i \le 1/l, \qquad i = 1, \ldots, l, \\ & \mathbf{y}^T\boldsymbol{\alpha} = 0,\ \mathbf{e}^T\boldsymbol{\alpha} = \nu. \end{aligned} \tag{2.2}$$

Therefore, the equality is naturally obtained. Note that this is an example that two problems have the same optimal solution set but are associated with two duals which have *different* optimal solution sets. It is interesting that here the primal problem which has more restrictions is related to a dual which has a larger feasible region. For our later analysis, we keep on using (D_ν) but not (2.2). Interestingly we will see that the exceptional situation where (D_ν) has optimal solutions such that $\mathbf{e}^T\boldsymbol{\alpha} > \nu$ happens only for those ν which we are not interested in.

Due to the additional inequality, the feasibility of (D_ν) and (D_C) is different. For (D_C), 0 is an trivial feasible point. However, (D_ν) may be infeasible no matter training data are separable or not. An example where (P_ν) is unbounded below and (D_ν) is infeasible is as follows: Given three training data with $y_1 = y_2 = 1$, and $y_3 = -1$. If $\nu = 0.9$, there is no $\boldsymbol{\alpha}$ in (D_ν) which satisfies $0 \le \alpha_i \le 1/3, [1, 1, -1]\boldsymbol{\alpha} = 0$ and $\mathbf{e}^T\boldsymbol{\alpha} \ge 0.9$. Hence (D_ν) is infeasible. When this happens, we can choose $\mathbf{w} = 0, \xi_1 = \xi_2 = 0, b = \rho, \xi_3 = 2\rho$ as a feasible solution of (P_ν). Then the objective value is $-0.9\rho + 2\rho/3$ which goes to $-\infty$ as $\rho \to \infty$. Therefore, (P_ν) is unbounded.

We then describe a lemma which was first proved in (Crisp and Burges 1999).

Lemma 2.2 *(D_ν) is feasible if and only if $\nu \le \nu_{max}$, where*

$$\nu_{max} \equiv \frac{2\min(\#y_i = 1, \#y_i = -1)}{l},$$

and $(\#y_i = 1)$ and $(\#y_i = -1)$ denote the number of elements in the first and second classes, respectively.

Proof. Since $0 \le \alpha_i \le 1/l, i = 1, \ldots, l$, with $\mathbf{y}^T\boldsymbol{\alpha} = 0$, for any $\boldsymbol{\alpha}$ feasible to (D_ν), we have $\mathbf{e}^T\boldsymbol{\alpha} \le \nu_{max}$. Therefore, if (D_ν) is feasible, $\nu \le \nu_{max}$. On the other hand, if $0 < \nu \le \nu_{max}$, $\min(\#y_i = 1, \#y_i = -1) > 0$ so we can define a feasible solution of (D_ν):

$$\alpha_j = \begin{cases} \frac{\nu}{2(\#y_i=1)} & \text{if } y_j = 1, \\ \frac{\nu}{2(\#y_i=-1)} & \text{if } y_j = -1. \end{cases}$$

This $\boldsymbol{\alpha}$ satisfies $0 \le \alpha_i \le 1/l, i = 1, \ldots, l$ and $\mathbf{y}^T\boldsymbol{\alpha} = 0$. If $\nu = 0$, clearly $\boldsymbol{\alpha} = 0$ is a feasible solution of (D_ν). □

Note that the size of ν_{max} depends on how balanced the training set is. If the numbers of positive and negative examples match, then $\nu_{max} = 1$.

Since it is still more complicated to handle two linear constraints $\mathbf{y}^T\boldsymbol{\alpha} = 0$ and $\mathbf{e}^T\boldsymbol{\alpha} = \nu$ together, we try to remove one. For C-SVM, (Mangasarian and Musicant 1999), and (Friess et al. 1998) added $b^2/2$ into the objective function and can remove the linear constraint $\mathbf{y}^T\boldsymbol{\alpha} = 0$ in the dual. Here we would like to exploit a similar approach for (P_ν). Consider the following new primal problem:

$$(\bar{P}_\nu) \qquad \begin{aligned} \min & \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{1}{2}b^2 - \nu\rho + \frac{1}{l}\sum_{i=1}^{l}\xi_i \\ & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \ge \rho - \xi_i, \\ & \xi_i \ge 0, i = 1, \ldots, l, \rho \ge 0. \end{aligned} \tag{2.3}$$

The dual of $(\bar{P}_\nu)$ is:

$$(\bar{D}_\nu) \qquad \begin{aligned} \min & \frac{1}{2}\boldsymbol{\alpha}^T(\mathbf{Q} + \mathbf{y}\mathbf{y}^T)\boldsymbol{\alpha} \\ & 0 \le \alpha_i \le 1/l, \qquad i = 1, \ldots, l, \\ & \mathbf{e}^T\boldsymbol{\alpha} \ge \nu. \end{aligned} \tag{2.4}$$

Similar to Theorem 2.1, we can solve $(\bar{D}_\nu)$ using only the equality $\mathbf{e}^T\boldsymbol{\alpha} = \nu$. Hence the new problem has only one simple equality constraint.

In Section 3, we will show that the relation between (D_C) and (D_ν) is similar to the relation between $(\bar{D}_\nu)$ and

$$(\bar{D}_C) \qquad \min \frac{1}{2}\boldsymbol{\alpha}^T(\mathbf{Q} + \mathbf{y}\mathbf{y}^T)\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha}$$
$$0 \le \alpha_i \le C, \qquad i = 1, \ldots, l. \tag{2.5}$$

In particular, by appropriate scaling, the optimal solution set of $(\bar{D}_\nu)$ is the same as that of a $(\bar{D}_C)$. We may worry that after adding $b^2/2$ to the objective function, the generalization properties of $(\bar{D}_\nu)$ may not be as good as those of (D_ν). In (Hsu and Lin 1999), numerical experiments by cross validation show that $(\bar{D}_C)$ generates a comparable model as (D_C). Thus we think it is practically acceptable to solve $(\bar{D}_\nu)$ instead of (D_ν). In addition, the generalization performance by adding $b^2/2$ was discussed in (Cristianini and Shawe-Taylor 2000, Remark 7.4).

Note that the current form of $(\bar{D}_\nu)$ is already very similar to the form of (D_C). They both have l bound constraints and one linear equality constraint. Next we demonstrate an example on modifying existing decomposition methods for ν-SVM. For solving (D_C), the decomposition method separates the index $\{1, \ldots, l\}$ of the training set to two sets B and N, where B is the working set if $\boldsymbol{\alpha}$ is the current iterate of the algorithm. If we denote $\boldsymbol{\alpha}_B$ and $\boldsymbol{\alpha}_N$ as vectors containing corresponding elements, the objective value of (D_C) is equal to $\frac{1}{2}\boldsymbol{\alpha}_B^T\mathbf{Q}_{BB}\boldsymbol{\alpha}_B - (\mathbf{e}_B + \mathbf{Q}_{BN}\boldsymbol{\alpha}_N)^T\boldsymbol{\alpha}_B + \frac{1}{2}\boldsymbol{\alpha}_N^T\mathbf{Q}_{NN}\boldsymbol{\alpha}_N - \mathbf{e}_N^T\boldsymbol{\alpha}_N$. At each iteration, $\boldsymbol{\alpha}_N$ is fixed and the following problem with the variable $\boldsymbol{\alpha}_B$ is solved:

$$\begin{aligned} \min \quad & \frac{1}{2}\boldsymbol{\alpha}_B^T\mathbf{Q}_{BB}\boldsymbol{\alpha}_B - (\mathbf{e}_B - \mathbf{Q}_{BN}\boldsymbol{\alpha}_N)^T\boldsymbol{\alpha}_B \\ & 0 \le (\alpha_B)_i \le C, i = 1, \ldots, q, \\ & \mathbf{y}_B^T\boldsymbol{\alpha}_B = -\mathbf{y}_N^T\boldsymbol{\alpha}_N, \end{aligned} \tag{2.6}$$

where $\begin{bmatrix} \mathbf{Q}_{BB} & \mathbf{Q}_{BN} \\ \mathbf{Q}_{NB} & \mathbf{Q}_{NN} \end{bmatrix}$ is a permutation of the matrix $\mathbf{Q}$ and q is the size of B. The strict decrease of the objective function holds and the theoretical convergence was studied in (Chang et al. 2000).

An important process in the decomposition methods is the selection of the working set B. Using the similarity between $(\bar{D}_\nu)$ and (D_C), with minor modifications, most selections of the working set can be adjusted for solving $(\bar{D}_\nu)$.

In the software SVM^{light} (Joachims 1998), there is a systematic way to find the working set B. In each iteration the following problem is solved:

$$\begin{aligned} \min \quad & \nabla f(\boldsymbol{\alpha}_k)^T \mathbf{d} \\ & \mathbf{y}^T \mathbf{d} = 0,\ -1 \le d_i \le 1, && (2.7) \\ & d_i \ge 0,\ \text{if } (\alpha_k)_i = 0,\ d_i \le 0,\ \text{if } (\alpha_k)_i = C, && (2.8) \\ & |\{d_i \mid d_i \ne 0\}| = q, && (2.9) \end{aligned}$$

where we represent $f(\boldsymbol{\alpha}) \equiv \frac{1}{2}\boldsymbol{\alpha}^T \mathbf{Q}\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha}$, $\boldsymbol{\alpha}_k$ is the iterate at the kth iteration, $\nabla f(\boldsymbol{\alpha}_k)$ is the gradient of $f(\boldsymbol{\alpha})$ at $\boldsymbol{\alpha}_k$. Note that $|\{d_i \mid d_i \ne 0\}|$ means the number of components of d which are not zero. The constraint (2.9) implies that a descent direction involving only q variables is obtained. Then components of $\boldsymbol{\alpha}_k$ with non-zero d_i are included in the working set B which is used to construct the sub-problem (2.6). Note that d is only used for identifying B but not as a search direction.

If q is an even number, (Joachims 1998) showed a simple strategy on solving (2.7)-(2.9). First he sorts $y_i \nabla f(\alpha_k)_i, i = 1, \ldots, l$ in a decreasing order. Then solution is by successively picking the $q/2$ elements from the top of the sorted list which $0 < (\alpha_k)_i < C$ or $d_i = -y_i$ obeys (2.8). Similarly we pick the $q/2$ elements from the bottom of the list for which $0 < (\alpha_k)_i < C$ or $d_i = y_i$ obeys (2.8). Other elements of d are assigned to be zero. Thus these q nonzero elements compose the working set.

To modify the above strategy for $(\bar{D}_\nu)$, we consider the following problem in each iteration:

$$\begin{aligned} \min \quad & \nabla f(\boldsymbol{\alpha}_k)^T \mathbf{d} \\ & \mathbf{e}^T \mathbf{d} = 0,\ -1 \le d_i \le 1, \\ & d_i \ge 0,\ \text{if } (\alpha_k)_i = 0,\ d_i \le 0,\ \text{if } (\alpha_k)_i = 1/l, && (2.10) \\ & |\{d_i \mid d_i \ne 0\}| = q. \end{aligned}$$

Note that now $f(\boldsymbol{\alpha})$ becomes $\frac{1}{2}\boldsymbol{\alpha}^T(\mathbf{Q} + \mathbf{y}\mathbf{y}^T)\boldsymbol{\alpha}$. Then the simple procedure in SVM^{light} becomes as follows: Sort $\nabla f(\alpha_k)_i, i = 1, \ldots, l$ in a decreasing order. Pick the $q/2$ elements from the top of the sorted list which $0 < (\alpha_k)_i < C$ or

$d_i = -1$ obeys (2.10). Then pick the $q/2$ elements from the bottom of the list for which $0 < (\alpha_k)_i < C$ or $d_i = 1$ obeys (2.10). We can easily use KKT condition to check why these q elements are chosen.

Note that the sub-problem (2.6) becomes as follows if decomposition methods are used for solving $(\bar{D}_\nu)$:

$$\begin{aligned} \min \quad & \frac{1}{2}\boldsymbol{\alpha}_B^T(\mathbf{Q}+\mathbf{y}\mathbf{y}^T)_{BB}\boldsymbol{\alpha}_B + ((\mathbf{Q}+\mathbf{y}\mathbf{y}^T)_{BN}\boldsymbol{\alpha}_N)^T\boldsymbol{\alpha}_B \\ & 0 \le (\alpha_B)_i \le 1/l, i = 1, \ldots, q, \\ & \mathbf{e}_B^T\boldsymbol{\alpha}_B = \nu - \mathbf{e}_N^T\boldsymbol{\alpha}_N. \end{aligned}$$

In Section 5, we will conduct some experiments on this new method.

3 The Relation Between ν-SVM and C-SVM

The infeasibility of (D_ν) mentioned in Section 2 gives us a motivation on studying when and why this situation happens. In this section we will investigate this issue by constructing a relationship between (D_ν) and (D_C) where the main result is in Theorem 3.7. The relation between (D_C) and (D_ν) has been discussed in (Schölkopf et al. 2000, Proposition 13) where they show that if (P_ν) leads to $\rho > 0$, then (P_C) with $C = 1/(\rho l)$ leads to the same decision function. Here we will have more complete investigation.

First we note that if $C > 0$, by dividing each variable by Cl, (D_C) is equivalent to the following problem:

$$(D'_C) \qquad \begin{aligned} \min & \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - \frac{\mathbf{e}^T\boldsymbol{\alpha}}{Cl} \\ & \mathbf{y}^T\boldsymbol{\alpha} = 0, 0 \le \alpha_i \le 1/l, i = 1, \ldots, l. \end{aligned}$$

It can be clearly seen that (D'_C) and (D_ν) are very similar. We prove the following lemma about (D'_C):

Lemma 3.1 *If (D'_C) has different optimal solutions $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$, then $\mathbf{e}^T\boldsymbol{\alpha}_1 = \mathbf{e}^T\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_1^T\mathbf{Q}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2^T\mathbf{Q}\boldsymbol{\alpha}_2$. Therefore, we can define two functions $\mathbf{e}^T\boldsymbol{\alpha}_C$ and $\boldsymbol{\alpha}_C^T\mathbf{Q}\boldsymbol{\alpha}_C$ on C, where $\boldsymbol{\alpha}_C$ is any optimal solution of (D'_C).*

Proof. Since (D'_C) is a convex problem, if $\boldsymbol{\alpha}_1 \neq \boldsymbol{\alpha}_2$ are both optimal solutions, for all $0 \leq \lambda \leq 1$,

$$\begin{aligned} & \frac{1}{2}(\lambda\boldsymbol{\alpha}_1 + (1-\lambda)\boldsymbol{\alpha}_2)^T\mathbf{Q}(\lambda\boldsymbol{\alpha}_1 + (1-\lambda)\boldsymbol{\alpha}_2) - \mathbf{e}^T(\lambda\boldsymbol{\alpha}_1 + (1-\lambda)\boldsymbol{\alpha}_2)/(Cl) \\ = \; & \lambda(\frac{1}{2}\boldsymbol{\alpha}_1^T\mathbf{Q}\boldsymbol{\alpha}_1 - \mathbf{e}^T\boldsymbol{\alpha}_1/(Cl)) + (1-\lambda)(\frac{1}{2}\boldsymbol{\alpha}_2^T\mathbf{Q}\boldsymbol{\alpha}_2 - \mathbf{e}^T\boldsymbol{\alpha}_2/(Cl)). \end{aligned}$$

This implies

$$\boldsymbol{\alpha}_1^T\mathbf{Q}\boldsymbol{\alpha}_2 = \frac{1}{2}\boldsymbol{\alpha}_1^T\mathbf{Q}\boldsymbol{\alpha}_1 + \frac{1}{2}\boldsymbol{\alpha}_2^T\mathbf{Q}\boldsymbol{\alpha}_2. \tag{3.1}$$

Since $\mathbf{Q}$ is positive semidefinite, $\mathbf{Q} = L^TL$ so (3.1) implies $\|L\boldsymbol{\alpha}_1 - L\boldsymbol{\alpha}_2\| = 0$. Thus $\boldsymbol{\alpha}_2^T\mathbf{Q}\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_1^T\mathbf{Q}\boldsymbol{\alpha}_1$. Therefore, $\mathbf{e}^T\boldsymbol{\alpha}_1 = \mathbf{e}^T\boldsymbol{\alpha}_2$ and the proof is complete. □

Next we prove a theorem on optimal solutions of (D'_C) and (D_ν):

Theorem 3.2 *If (D'_C) and (D_ν) share one optimal solution $\boldsymbol{\alpha}^*$ with $\mathbf{e}^T\boldsymbol{\alpha}^* = \nu$, their optimal solution sets are the same.*

Proof. From Lemma 3.1, any other optimal solution $\boldsymbol{\alpha}$ of (D'_C) also satisfies $\mathbf{e}^T\boldsymbol{\alpha} = \nu$ so $\boldsymbol{\alpha}$ is feasible to (D_ν). Since $\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} = (\boldsymbol{\alpha}^*)^T\mathbf{Q}\boldsymbol{\alpha}^*$ from Lemma 3.1, all (D'_C)'s optimal solutions are also optimal solutions of (D_ν). On the other hand, if $\boldsymbol{\alpha}$ is any optimal solution of (D_ν), it is feasible to (D'_C). With the constraint $\mathbf{e}^T\boldsymbol{\alpha} \geq \nu = \mathbf{e}^T\boldsymbol{\alpha}^*$ and $\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} = (\boldsymbol{\alpha}^*)^T\mathbf{Q}\boldsymbol{\alpha}^*$,

$$\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha}/(Cl) \leq \frac{1}{2}(\boldsymbol{\alpha}^*)^T\mathbf{Q}(\boldsymbol{\alpha}^*) - \mathbf{e}^T\boldsymbol{\alpha}^*/(Cl).$$

Therefore, all optimal solutions of (D_ν) are also optimal to (D'_C). Hence their optimal solution sets are the same. □

If $\boldsymbol{\alpha}$ is an optimal solution of (D'_C), it satisfies the following KKT condition:

$$\begin{gathered} \mathbf{Q}\boldsymbol{\alpha} - \frac{\mathbf{e}}{Cl} + b\mathbf{y} = \boldsymbol{\lambda} - \boldsymbol{\xi}, \\ \boldsymbol{\lambda}^T\boldsymbol{\alpha} = 0, \boldsymbol{\xi}^T(\frac{\mathbf{e}}{l} - \boldsymbol{\alpha}) = 0, \mathbf{y}^T\boldsymbol{\alpha} = 0 \\ \lambda_i \geq 0, \xi_i \geq 0, 0 \leq \alpha_i \leq 1/l, i = 1, \ldots, l. \end{gathered} \tag{3.2}$$

By setting $\rho \equiv 1/(Cl)$ and $\nu \equiv \mathbf{e}^T\boldsymbol{\alpha}$, $\boldsymbol{\alpha}$ also satisfies the KKT condition of (D_ν):

$$\begin{gathered} \mathbf{Q}\boldsymbol{\alpha} - \rho\mathbf{e} + b\mathbf{y} = \boldsymbol{\lambda} - \boldsymbol{\xi}, \\ \lambda^T\boldsymbol{\alpha} = 0, \boldsymbol{\xi}^T(\frac{\mathbf{e}}{l} - \boldsymbol{\alpha}) = 0, \\ \mathbf{y}^T\boldsymbol{\alpha} = 0, \mathbf{e}^T\boldsymbol{\alpha} \geq \nu, \rho(\mathbf{e}^T\boldsymbol{\alpha} - \nu) = 0, \\ \lambda_i \geq 0, \xi_i \geq 0, \rho \geq 0, 0 \leq \alpha_i \leq 1/l, i = 1, \ldots, l. \end{gathered} \tag{3.3}$$

From Theorem 3.2, this implies that for each (D'_C), its optimal solution set is the same as that of (D_ν), where $\nu = \mathbf{e}^T\boldsymbol{\alpha}$. For each (D'_C), such a (D_ν) is unique as from Theorem 2.1, if $\nu_1 \neq \nu_2$, (D_{ν_1}) and (D_{ν_2}) have different optimal solution sets. Therefore, we have the following theorem:

Theorem 3.3 *For each $(D'_C), C > 0$, its optimal solution set is the same as that of one (and only one) (D_ν), where $\nu = \mathbf{e}^T\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}$ is any optimal solution of (D'_C).*

Similarly, we have

Theorem 3.4 *If $(D_\nu), \nu > 0$, has a nonempty feasible set and its objective value is not zero, (D_ν)'s optimal solution set is the same as that of at least one (D'_C).*

Proof. If the objective value of (D_ν) is not zero, from the KKT condition (3.3),

$$\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - \rho\mathbf{e}^T\boldsymbol{\alpha} = -\sum_{i=1}^{l}\xi_i/l.$$

Then $\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} > 0$ and (3.3) imply

$$\rho\mathbf{e}^T\boldsymbol{\alpha} = \boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} + \sum_{i=1}^{l}\xi_i/l > 0,\ \rho > 0, \quad \text{and } \mathbf{e}^T\boldsymbol{\alpha} = \nu.$$

By choosing a $C > 0$ such that $\rho = 1/(Cl)$, $\boldsymbol{\alpha}$ is a KKT point of (D'_C). Hence from Theorem 3.2, the optimal solution set of this (D'_C) is the same as that of (D_ν). □

Next we prove two useful lemmas. The first one deals with the special situation when the objective value of (D_ν) is zero.

Lemma 3.5 *If the objective value of $(D_\nu), \nu \geq 0$, is zero and there is a $(D'_C), C > 0$ such that any its optimal solution $\boldsymbol{\alpha}_C$ satisfies $\mathbf{e}^T\boldsymbol{\alpha}_C = \nu$, then $\nu = \nu_{max}$ and all (D'_C), $C > 0$, have the same optimal solution set as that of (D_ν).*

Proof. For this (D_ν), we can set $\rho = 1/(Cl)$, so $\boldsymbol{\alpha}_C$ is a KKT point of (D_ν). Therefore, since the objective value of (D_ν) is zero, $\boldsymbol{\alpha}_C^T\mathbf{Q}\boldsymbol{\alpha}_C = 0$. Furthermore, we have $\mathbf{Q}\boldsymbol{\alpha}_C = 0$. In this case, (3.2) of (D'_C)'s KKT condition becomes

$$-\frac{\mathbf{e}}{Cl} + \begin{bmatrix} b\mathbf{e}_I \\ -b\mathbf{e}_J \end{bmatrix} = \boldsymbol{\lambda} - \boldsymbol{\xi}, \tag{3.4}$$

where $\lambda_i, \xi_i \geq 0$, and I and J are indices of two different classes. If $b\mathbf{e}_I \geq 0$, there are three situations of (3.4):

$$\begin{bmatrix} > 0 \\ < 0 \end{bmatrix}, \qquad \begin{bmatrix} < 0 \\ < 0 \end{bmatrix}, \qquad \begin{bmatrix} = 0 \\ < 0 \end{bmatrix}.$$

The first case implies $(\boldsymbol{\alpha}_C)_I = 0$ and $(\boldsymbol{\alpha}_C)_J = (\mathbf{e}_J)/l$. Hence if J is nonempty, $\mathbf{y}^T\boldsymbol{\alpha}_C \neq 0$ causes contradiction. Hence all data are in the same class. Therefore, (D_ν) and all (D'_C), $C > 0$, have the unique optimal solution zero due to the constraints $\mathbf{y}^T\boldsymbol{\alpha} = 0$ and $\boldsymbol{\alpha} \geq 0$. Furthermore, $\mathbf{e}^T\boldsymbol{\alpha} = \nu = \nu_{max} = 0$.

The second case happens only when $\boldsymbol{\alpha}_C = \mathbf{e}/l$. Then $\mathbf{y}^T\boldsymbol{\alpha} = 0$ and $y_i = 1$ or -1 imply that $(\#y_i{=}1) = (\#y_i{=}{-1})$ and $\mathbf{e}^T\boldsymbol{\alpha}_C = 1 = \nu = \nu_{max}$. We then show that $\mathbf{e}/l$ is also an optimal solution of any other (D'_C). Since $0 \leq \alpha_i \leq 1/l, i = 1, \ldots, l$, for any feasible $\boldsymbol{\alpha}$ of $(\bar{D}'_C)$, the objective function satisfies

$$\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - \frac{\mathbf{e}^T\boldsymbol{\alpha}}{Cl} \geq -\frac{\mathbf{e}^T\boldsymbol{\alpha}}{Cl} \geq -\frac{1}{Cl}. \tag{3.5}$$

Now $(\#y_i = 1) = (\#y_i = -1)$ so $\mathbf{e}/l$ is feasible. When $\boldsymbol{\alpha} = \mathbf{e}/l$, the inequality of (3.5) becomes an equality. Thus $\mathbf{e}/l$ is actually an optimal solution of all $(D'_C), C > 0$. Therefore, (D_ν) and all (D_C), $C > 0$ have the same unique optimal solution $\mathbf{e}/l$.

For the third case, $b = 1/(Cl)$, $(\boldsymbol{\alpha}_C)_J = \mathbf{e}_J/l$, $\nu = \mathbf{e}^T\boldsymbol{\alpha}_C = 2\mathbf{e}_J^T(\boldsymbol{\alpha}_C)_J = \nu_{max}$, and J contains elements which have fewer elements. Because there exists such a C and b, for any other C, b can be adjusted accordingly so that the KKT condition is still satisfied. Therefore, from Theorem 3.3, all $(D'_C), C > 0$ have the same optimal solution set as that of (D_ν). The situation when $b\mathbf{e}_I \leq 0$ is similar. □

Lemma 3.6 *Assume $\boldsymbol{\alpha}_C$ is any optimal solution of (D'_C), then $\mathbf{e}^T\boldsymbol{\alpha}_C$ is a continuous decreasing function of C on $(0, \infty)$.*

Proof. If $C_1 < C_2$, and $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are optimal solutions of (D'_{C_1}) and (D'_{C_2}), respectively, we have

$$\frac{1}{2}\boldsymbol{\alpha}_1^T\mathbf{Q}\boldsymbol{\alpha}_1 - \frac{\mathbf{e}^T\boldsymbol{\alpha}_1}{C_1 l} \leq \frac{1}{2}\boldsymbol{\alpha}_2^T\mathbf{Q}\boldsymbol{\alpha}_2 - \frac{\mathbf{e}^T\boldsymbol{\alpha}_2}{C_1 l} \tag{3.6}$$

and

$$\frac{1}{2}\boldsymbol{\alpha}_2^T\mathbf{Q}\boldsymbol{\alpha}_2 - \frac{\mathbf{e}^T\boldsymbol{\alpha}_2}{C_2 l} \leq \frac{1}{2}\boldsymbol{\alpha}_1^T\mathbf{Q}\boldsymbol{\alpha}_1 - \frac{\mathbf{e}^T\boldsymbol{\alpha}_1}{C_2 l}. \tag{3.7}$$

Hence

$$\frac{\mathbf{e}^T\boldsymbol{\alpha}_1}{C_2 l} - \frac{\mathbf{e}^T\boldsymbol{\alpha}_2}{C_2 l} \leq \frac{1}{2}\boldsymbol{\alpha}_1^T\mathbf{Q}\boldsymbol{\alpha}_1 - \frac{1}{2}\boldsymbol{\alpha}_2^T\mathbf{Q}\boldsymbol{\alpha}_2 \leq \frac{\mathbf{e}^T\boldsymbol{\alpha}_1}{C_1 l} - \frac{\mathbf{e}^T\boldsymbol{\alpha}_2}{C_1 l}. \tag{3.8}$$

Since $C_2 > C_1 > 0$, (3.8) implies $\mathbf{e}^T\boldsymbol{\alpha}_1 - \mathbf{e}^T\boldsymbol{\alpha}_2 \geq 0$. Therefore, $\mathbf{e}^T\boldsymbol{\alpha}_C$ is a decreasing function on $(0, \infty)$. From this result, we know that for any $C^* \in (0, \infty)$, $\lim_{C\to(C^*)^+} \mathbf{e}^T\boldsymbol{\alpha}_C$ and $\lim_{C\to(C^*)^-} \mathbf{e}^T\boldsymbol{\alpha}_C$ exist, and

$$\lim_{C\to(C^*)^+} \mathbf{e}^T\boldsymbol{\alpha}_C \leq \mathbf{e}^T\boldsymbol{\alpha}_{C^*} \leq \lim_{C\to(C^*)^-} \mathbf{e}^T\boldsymbol{\alpha}_C.$$

To prove the continuity of $\mathbf{e}^T\boldsymbol{\alpha}_C$, it is sufficient to prove $\lim_{C\to C^*} \mathbf{e}^T\boldsymbol{\alpha}_C = \mathbf{e}^T\boldsymbol{\alpha}_{C^*}$, for all $C^* \in (0, \infty)$.

If $\lim_{C\to(C^*)^+} \mathbf{e}^T\boldsymbol{\alpha}_C < \mathbf{e}^T\boldsymbol{\alpha}_{C^*}$, there is a $\bar{\nu}$ such that

$$0 \leq \lim_{C\to(C^*)^+} \mathbf{e}^T\boldsymbol{\alpha}_C < \bar{\nu} < \mathbf{e}^T\boldsymbol{\alpha}_{C^*}. \tag{3.9}$$

Hence $\bar{\nu} > 0$. If $(D_{\bar{\nu}})$'s objective value is not zero, from Theorem 3.4 and the fact that $\mathbf{e}^T\boldsymbol{\alpha}_C$ is a decreasing function, there exists a $C > C^*$ such that $\boldsymbol{\alpha}_C$ satisfies $\mathbf{e}^T\boldsymbol{\alpha}_C = \bar{\nu}$. This contradicts to (3.9) where $\lim_{C\to(C^*)^+} \mathbf{e}^T\boldsymbol{\alpha}_C < \bar{\nu}$.

Therefore, the objective value of $(D_{\bar{\nu}})$ is zero. Since for all (D_ν), $\nu \leq \bar{\nu}$, their feasible regions include that of $(D_{\bar{\nu}})$, their objective values are also zero. From Theorem 3.3, the fact that $\mathbf{e}^T\boldsymbol{\alpha}_C$ is a decreasing function, and $\lim_{C\to(C^*)^+} \mathbf{e}^T\boldsymbol{\alpha}_C < \bar{\nu}$, each (D'_C), $C > C^*$, has the same optimal solution set as that of one (D_ν), where $\mathbf{e}^T\boldsymbol{\alpha}_C = \nu < \bar{\nu}$. Hence by Lemma 3.5, $\mathbf{e}^T\boldsymbol{\alpha}_C = \nu_{max}$, for all C. This contradicts to (3.9).

Therefore, $\lim_{C\to(C^*)^+} \mathbf{e}^T\boldsymbol{\alpha}_C = \mathbf{e}^T\boldsymbol{\alpha}_{C^*}$. Similarly, $\lim_{C\to(C^*)^-} \mathbf{e}^T\boldsymbol{\alpha}_C = \mathbf{e}^T\boldsymbol{\alpha}_{C^*}$. Thus

$$\lim_{C\to C^*} \mathbf{e}^T\boldsymbol{\alpha}_C = \mathbf{e}^T\boldsymbol{\alpha}_{C^*}.$$

□

Using the above lemmas, we are now ready to prove the main theorem:

Theorem 3.7 *We can define*

$$\lim_{C\to\infty} \mathbf{e}^T\boldsymbol{\alpha}_C = \nu_* \geq 0 \text{ and } \lim_{C\to 0} \mathbf{e}^T\boldsymbol{\alpha}_C = \nu^* \leq 1,$$

where $\boldsymbol{\alpha}_C$ is any optimal solution of (D'_C). Then $\nu^* = \nu_{max}$. For any $\nu > \nu^*$, (D_ν) is infeasible. For any $\nu \in (\nu_*, \nu^*]$, the optimal solution set of (D_ν) is the same as that of either one (D'_C), $C > 0$, or some (D'_C), where C is any number in an interval. For any $0 \le \nu \le \nu_*$, we have that $0 \le \nu \le \nu_*$ if and only if (D_ν) is feasible with zero objective value.

Proof. First from Lemma 3.6 and the fact that $0 \le \mathbf{e}^T\boldsymbol{\alpha} \le 1$, we know ν^* and ν_* can be defined without problems. We then prove $\nu^* = \nu_{max}$ by showing that after C is small enough, all (D'_C)'s optimal solutions $\boldsymbol{\alpha}_C$ satisfy $\mathbf{e}^T\boldsymbol{\alpha}_C = \nu_{max}$.

Assume I includes elements of the class which has fewer elements and J includes elements of the other class. If $\boldsymbol{\alpha}_C$ is an optimal solution of (D'_C), it satisfies the following KKT condition:

$$\begin{bmatrix} \mathbf{Q}_{II} & \mathbf{Q}_{IJ} \\ \mathbf{Q}_{JI} & \mathbf{Q}_{JJ} \end{bmatrix} \begin{bmatrix} (\boldsymbol{\alpha}_C)_I \\ (\boldsymbol{\alpha}_C)_J \end{bmatrix} - \frac{\mathbf{e}}{Cl} + b_C \begin{bmatrix} \mathbf{y}_I \\ \mathbf{y}_J \end{bmatrix} = \begin{bmatrix} (\boldsymbol{\lambda}_C)_I - (\boldsymbol{\xi}_C)_I \\ (\boldsymbol{\lambda}_C)_J - (\boldsymbol{\xi}_C)_J \end{bmatrix},$$

where $\boldsymbol{\lambda}_C \ge 0, \boldsymbol{\xi}_C \ge 0, \boldsymbol{\alpha}_C^T\lambda_C = 0$, and $\boldsymbol{\xi}_C^T(\mathbf{e}/l - \boldsymbol{\alpha}_C) = 0$. When C is small enough, $b_C\mathbf{y}_J > 0$ must hold. Otherwise, since $\mathbf{Q}_{JI}(\boldsymbol{\alpha}_C)_I + \mathbf{Q}_{JJ}(\boldsymbol{\alpha}_C)_J$ is bounded, $\mathbf{Q}_{JI}(\boldsymbol{\alpha}_C)_I + \mathbf{Q}_{JJ}(\boldsymbol{\alpha}_C)_J - \mathbf{e}_J/(Cl) + b_C\mathbf{y}_J < 0$ implies $(\boldsymbol{\alpha}_C)_J = \mathbf{e}_J/l$ which violates the constraint $\mathbf{y}^T\boldsymbol{\alpha} = 0$ if $(\#y_i = 1) \ne (\#y_i = -1)$. Therefore, $b_C\mathbf{y}_J > 0$ so $b_C\mathbf{y}_I < 0$. This implies that $(\boldsymbol{\alpha}_C)_I = \mathbf{e}_I/l$ when C is sufficiently small. Hence $\mathbf{e}^T\boldsymbol{\alpha}_C = \nu_{max} = \nu^*$.

If $(\#y_i = 1) = (\#y_i = -1)$, we can let $\boldsymbol{\alpha}_C = \mathbf{e}/l$ and $b_C = 0$. When C is small enough, this will be a KKT point. Therefore, $\mathbf{e}^T\boldsymbol{\alpha}_C = \nu_{max} = \nu^* = 1$.

From Lemma 2.2 we immediately know that (D_ν) is infeasible if $\nu > \nu^*$. From Lemma 3.6 that $\mathbf{e}^T\boldsymbol{\alpha}_C$ is a continuous function, for any $\nu \in (\nu_*, \nu^*]$, there is a (D'_C) such that $\mathbf{e}^T\boldsymbol{\alpha}_C = \nu$. Then from Theorem 3.3, (D'_C) and (D_ν) have the same optimal solution set.

If (D_ν) has the same optimal solution set as that of (D'_{C_1}) and (D'_{C_2}) where $C_1 < C_2$, since $\mathbf{e}^T\boldsymbol{\alpha}_C$ is a decreasing function, for any $C \in [C_1, C_2]$, its optimal solutions satisfy $\mathbf{e}^T\boldsymbol{\alpha} = \nu$. From Theorem 3.3, its optimal solution set is the same as that of (D_ν). Thus such Cs construct an interval.

If $\nu < \nu_*$, (D_ν) must be feasible from Lemma 2.2. It cannot have nonzero objective value due to Theorem 3.4 and the definition of ν_*. For (D_{ν_*}), if $\nu_* = 0$, the objective value of (D_{ν_*}) is zero as $\boldsymbol{\alpha} = 0$ is a feasible solution. If $\nu_* >$

0, since feasible regions of (D_ν) are bounded by $0 \le \alpha_i \le 1/l$, i = 1, ..., l, with Theorem 2.1, there is a sequence $\{\boldsymbol{\alpha}_{\nu_i}\}, \nu_1 \le \nu_2 \le \cdots < \nu_*$ such that $\boldsymbol{\alpha}_{\nu_i}$ is an optimal solution of (D_{ν_i}), $\mathbf{e}^T\boldsymbol{\alpha}_{\nu_i} = \nu_i$, and $\hat{\boldsymbol{\alpha}} \equiv \lim_{\nu_i \to \nu_*} \boldsymbol{\alpha}_{\nu_i}$ exists. Since $\mathbf{e}^T\boldsymbol{\alpha}_{\nu_i} = \nu_i$, $\mathbf{e}^T\hat{\boldsymbol{\alpha}} = \lim_{\nu_i \to \nu_*} \mathbf{e}^T\boldsymbol{\alpha}_{\nu_i} = \nu_*$. We also have $0 \le \hat{\boldsymbol{\alpha}} \le 1/l$ and $\mathbf{y}^T\hat{\boldsymbol{\alpha}} = \lim_{\nu_i \to \nu_*} \mathbf{y}^T\boldsymbol{\alpha}_{\nu_i} = 0$ so $\hat{\boldsymbol{\alpha}}$ is feasible to (D_{ν_*}). However, $\hat{\boldsymbol{\alpha}}^T\mathbf{Q}\hat{\boldsymbol{\alpha}} = \lim_{\nu_i \to \nu_*} \boldsymbol{\alpha}_{\nu_i}^T\mathbf{Q}\boldsymbol{\alpha}_{\nu_i} = 0$ as $\boldsymbol{\alpha}_{\nu_i}^T\mathbf{Q}\boldsymbol{\alpha}_{\nu_i} = 0$ for all ν_i. Therefore, the objective value of (D_{ν_*}) is always zero.

Next we prove that the objective value of (D_ν) is zero if and only if $\nu \le \nu_*$. From the above discussion, if $\nu \le \nu_*$, the objective value of (D_ν) is zero. If the objective value of (D_ν) is zero but $\nu > \nu_*$, Theorem 3.5 implies $\nu = \nu_{max} = \nu^* = \nu_*$ which causes a contradiction. Hence the proof is complete. □

Note that when the objective value of (D_ν) is zero, the optimal solution w of the primal problem (P_ν) is zero. In (Crisp and Burges 1999, Section 4), they considered such a (P_ν) as a "trivial" problem. Next we present a corollary:

Corollary 3.8 *If training data are separable, $\nu_* = 0$. If training data are non-separable, $\nu_* \ge 1/l > 0$. Furthermore, if $\mathbf{Q}$ is positive definite, training data are separable and $\nu_* = 0$.*

Proof. From (Lin 1999, Theorem 3.3), if data are separable, there is a C^* such that for all $C \ge C^*$, an optimal solution $\boldsymbol{\alpha}_{C^*}$ of (D_{C^*}) is also optimal to (D_C). Therefore, for (D'_C), an optimal solution becomes $\boldsymbol{\alpha}_{C^*}/(Cl)$ and $\mathbf{e}^T\boldsymbol{\alpha}_{C^*}/(Cl) \to 0$ as $C \to \infty$. Thus $\nu_* = 0$. On the other hand, if data are non-separable, no matter how large C is, there are components of optimal solutions at the upper bound. Therefore, $\mathbf{e}^T\boldsymbol{\alpha}_C \ge 1/l > 0$ for all C. Hence $\nu_* \ge 1/l$.

If Q is positive definite, the unconstrained problem

$$\min \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha} \tag{3.10}$$

has an unique solution at $\boldsymbol{\alpha} = \mathbf{Q}^{-1}\mathbf{e}$. If we add additional constraints to (3.10),

$$\begin{aligned} \min \quad & \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha} \\ & \mathbf{y}^T\boldsymbol{\alpha} = 0, \alpha_i \ge 0, i = 1, \ldots, l, \end{aligned} \tag{3.11}$$

is a problem with a smaller feasible region. Thus the objective value of (3.11) is bounded. From Corollary 27.3.1 of (Rockafellar 1970), any bounded finite dimensional space quadratic convex function over a polyhedral attains at least an optimal solution. Therefore, (3.11) is solvable. From (Lin 1999, Theorem 2.2), this implies the following primal problem is solvable:

$$\begin{aligned} \min \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1, i = 1, \ldots, l. \end{aligned}$$

Hence training data are separable. □

In many situations Q is positive definite. For example, from (Micchelli 1986), if the RBF kernel is used, Q is positive definite.

We illustrate the above results by some examples. Given three non-separable training points $\mathbf{x}_1 = 0, \mathbf{x}_2 = 1$, and $\mathbf{x}_3 = 2$ with $\mathbf{y} = [1, -1, 1]^T$, we will show that this is an example of Lemma 3.5. Note that this is a non-separable problem. For all $C > 0$, the optimal solution of (D'_C) is $\boldsymbol{\alpha} = [1/6, 1/3, 1/6]^T$. Therefore, in this case, $\nu^* = \nu_* = 2/3$. For $(D_\nu), \nu \leq 2/3$, an optimal solution is $\boldsymbol{\alpha} = (3\nu/2)[1/6, 1/3, 1/6]^T$ with the objective value

$$(3\nu/2)^2[1/6, 1/3, 1/6] \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -2 \\ 0 & -2 & 4 \end{bmatrix} \begin{bmatrix} 1/6 \\ 1/3 \\ 1/6 \end{bmatrix} = 0.$$

Another example shows that we may have the same value of $\mathbf{e}^T\boldsymbol{\alpha}_C$ for all C in an interval, where $\boldsymbol{\alpha}_C$ is any optimal solution of (D'_C). Given $\mathbf{x}_1 = \left[\begin{smallmatrix} -1 \\ 0 \end{smallmatrix}\right], \mathbf{x}_2 = \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right], \mathbf{x}_3 = \left[\begin{smallmatrix} 0 \\ -1 \end{smallmatrix}\right]$, and $\mathbf{x}_4 = \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right]$ with $\mathbf{y} = [1, -1, 1, -1]^T$, part of the KKT condition of (D'_C) is

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} - \frac{1}{4C} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + b \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix} = \boldsymbol{\lambda} - \boldsymbol{\xi}.$$

Then one optimal solution of (D'_C) is:

$$\begin{array}{lll} \boldsymbol{\alpha}_C & b & \\ = [\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}]^T & \in [1 - \frac{1}{4C}, \frac{1}{4C} - \frac{1}{2}] & \text{if } 0 < C \leq \frac{1}{3}, \\ = \frac{1}{36}[3 + \frac{2}{C}, -3 + \frac{4}{C}, 3 + \frac{2}{C}, 9]^T & = \frac{1}{12C} & \text{if } \frac{1}{3} \leq C \leq \frac{4}{3}, \\ = [\frac{1}{8}, 0, \frac{1}{8}, \frac{1}{4}]^T & = \frac{1}{4C} - \frac{1}{8} & \text{if } \frac{4}{3} \leq C \leq 4, \\ = [\frac{1}{2C}, 0, \frac{1}{2C}, \frac{1}{C}]^T & = \frac{-1}{4C} & \text{if } C \geq 4. \end{array}$$

This is a separable problem. We have $\nu^* = 1$, $\nu_* = 0$, and

$$\mathbf{e}^T\boldsymbol{\alpha}_C = \begin{cases} 1 & \text{if } 0 < C \le \frac{1}{3}, \\ \frac{1}{3} + \frac{2}{9C} & \text{if } \frac{1}{3} \le C \le \frac{4}{3}, \\ \frac{1}{2} & \text{if } \frac{4}{3} \le C \le 4, \\ \frac{1}{2C} & \text{if } C \ge 4. \end{cases} \tag{3.12}$$

We also observe that above results can be extended to $(\bar{D}_\nu)$ and the following problem:

$$(\bar{D}'_C) \qquad \min \frac{1}{2}\boldsymbol{\alpha}^T(\mathbf{Q} + \mathbf{y}\mathbf{y}^T)\boldsymbol{\alpha} - \frac{\mathbf{e}^T\boldsymbol{\alpha}}{Cl}$$
$$0 \le \alpha_i \le 1/l, \qquad i = 1, \ldots, l. \tag{3.13}$$

Without the constraint $\mathbf{y}^T\boldsymbol{\alpha} = 0$, all $(\bar{D}_\nu)$, $0 \le \nu \le 1$, and $(\bar{D}'_C), C > 0$ are feasible and solvable. The major difference is that $\nu^* = 1$. Note that when C is small, $\boldsymbol{\alpha} = \mathbf{e}/l$ always satisfies

$$(\mathbf{Q} + \mathbf{y}\mathbf{y}^T)\boldsymbol{\alpha} - \frac{\mathbf{e}}{Cl} = -\boldsymbol{\xi} \le 0,$$

which is the KKT condition of $(\bar{D}'_C)$. Therefore, $\mathbf{e}^T\boldsymbol{\alpha} = 1$ after C is small enough. Regarding ν_*, we can prove that it is the same as that of using (D_ν):

Corollary 3.9 *If $\boldsymbol{\alpha}_C$ is any optimal solution of (D'_C), $\bar{\boldsymbol{\alpha}}_C$ is any optimal solution of $(\bar{D}'_C)$, $\lim_{C\to\infty} \mathbf{e}^T\boldsymbol{\alpha}_C = \nu_*$, and $\lim_{C\to\infty} \mathbf{e}^T\bar{\boldsymbol{\alpha}}_C = \bar{\nu}_*$, we have*

$$\nu_* = \bar{\nu}_*.$$

Proof. Since feasible regions of $(\bar{D}'_C)$ are bounded by $0 \le \alpha_i \le 1/l, i = 1, \ldots, l$, there is a sequence $\{\bar{\boldsymbol{\alpha}}_{C_i}\}, C_1 \le C_2 \le \cdots$ such that $\lim_{C_i\to\infty} \bar{\boldsymbol{\alpha}}_{C_i} = \bar{\boldsymbol{\alpha}}_*$ exists. Thus $\mathbf{e}^T\bar{\boldsymbol{\alpha}}_* = \bar{\nu}_*$. Similar to the result in Theorem 3.7, the objective value of $(D_{\bar{\nu}_*})$ is zero. With the positive semidefiniteness of $\mathbf{Q}$, we know $\frac{1}{2}\bar{\boldsymbol{\alpha}}_*^T\mathbf{Q}\bar{\boldsymbol{\alpha}}_* = 0$ and $\mathbf{y}^T\bar{\boldsymbol{\alpha}}_* = 0$. Therefore, $\bar{\boldsymbol{\alpha}}_*$ is a feasible solution of $(D_{\bar{\nu}_*})$ so the objective value of $(D_{\bar{\nu}_*})$ is zero. From Theorem 3.7, the objective value of (D_ν) is zero if and only if $\nu \le \nu_*$. Thus we have $\bar{\nu}_* \le \nu_*$.

If $\bar{\nu}_* < \nu_*$, there is a ν such that $\bar{\nu}_* < \nu < \nu_*$. For this ν, the objective value of (D_ν) is zero because $\nu < \nu_*$. On the other hand, the objective value of $(\bar{D}_\nu)$ is not zero because $\nu > \bar{\nu}_*$. However, any optimal solution of (D_ν) is feasible to $(\bar{D}_\nu)$. This causes a contradiction. □

In summary this section shows

1. The increase of C in C-SVM is like the decrease of ν in ν-SVM.

2. Solving (D_ν) and (D'_C) is just like solving two different problems with the same optimal solution set. We may expect that many numerical aspects on solving them are similar. However, they are still two different problems so we cannot obtain C without solving (D_ν). Similarly, without solving $((D_C))$, we cannot find ν either.

4 The Relation Between ν and C

A formula like (3.12) motivates us to conjecture that all $\nu = \mathbf{e}^T\boldsymbol{\alpha}_C$ have a similar form. That is, in each interval of C, $\mathbf{e}^T\boldsymbol{\alpha}_C = A + B/C$, where A and B are constants independent of C. The formulation of $\mathbf{e}^T\boldsymbol{\alpha}_C$ will be the main topic of this section.

We note that in (3.12), in each interval of C, $\boldsymbol{\alpha}_C$ are at the same face. Here we say two vectors at the same face if they have the same components which are free, at the lower bound, and at the upper bound. The following lemma deals with the situation when $\boldsymbol{\alpha}_C$ are at the same face:

Lemma 4.1 *If $\underline{C} < \overline{C}$ and there are $\boldsymbol{\alpha}_{\underline{C}}$ and $\boldsymbol{\alpha}_{\overline{C}}$ at the same face, then for each $C \in [\underline{C}, \overline{C}]$, there is at least one optimal solution $\boldsymbol{\alpha}_C$ of (D'_C) which is at the same face as $\boldsymbol{\alpha}_{\underline{C}}$ and $\boldsymbol{\alpha}_{\overline{C}}$. Furthermore,*

$$\mathbf{e}^T\boldsymbol{\alpha}_C = \Delta_1 + \frac{\Delta_2}{C}, \underline{C} \le C \le \overline{C},$$

where Δ_1 and Δ_2 are constants independent of C.

Proof. If $\{1, \ldots, l\}$ are separated to two sets A and F, where A corresponds to bounded variables and F corresponds to free variables of $\boldsymbol{\alpha}_{\underline{C}}$ (or $\boldsymbol{\alpha}_{\overline{C}}$ as they are at the same face), the KKT condition shows

$$\begin{bmatrix} \mathbf{Q}_{FF} & \mathbf{Q}_{FA} \\ \mathbf{Q}_{AF} & \mathbf{Q}_{AA} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_F \\ \boldsymbol{\alpha}_A \end{bmatrix} - \frac{\mathbf{e}}{Cl} + b \begin{bmatrix} \mathbf{y}_F \\ \mathbf{y}_A \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{\lambda}_A - \boldsymbol{\xi}_A \end{bmatrix}, \tag{4.1}$$

$$\mathbf{y}_F^T\boldsymbol{\alpha}_F + \mathbf{y}_A^T\boldsymbol{\alpha}_A = 0, \tag{4.2}$$

$$\lambda_i \ge 0, \xi_i \ge 0, i \in A. \tag{4.3}$$

(4.1) and (4.2) can be rewritten as

$$\begin{bmatrix} \mathbf{Q}_{FF} & \mathbf{Q}_{FA} & \mathbf{y}_F \\ \mathbf{Q}_{AF} & \mathbf{Q}_{AA} & \mathbf{y}_A \\ \mathbf{y}_F^T & \mathbf{y}_A^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_F \\ \boldsymbol{\alpha}_A \\ b \end{bmatrix} - \begin{bmatrix} \mathbf{e}_F/(Cl) \\ \mathbf{e}_A/(Cl) \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{\lambda}_A - \boldsymbol{\xi}_A \\ 0 \end{bmatrix}.$$

If $\mathbf{Q}_{FF}$ is positive definite,

$$\boldsymbol{\alpha}_F = \mathbf{Q}_{FF}^{-1}(\mathbf{e}_F/(Cl) - \mathbf{Q}_{FA}\boldsymbol{\alpha}_A - b\mathbf{y}_F). \tag{4.4}$$

Thus,

$$\mathbf{y}_F^T\boldsymbol{\alpha}_F + \mathbf{y}_A^T\boldsymbol{\alpha}_A = \mathbf{y}_F^T\mathbf{Q}_{FF}^{-1}(\mathbf{e}_F/(Cl) - \mathbf{Q}_{FA}\boldsymbol{\alpha}_A - b\mathbf{y}_F) + \mathbf{y}_A^T\boldsymbol{\alpha}_A = 0$$

implies

$$b = \frac{\mathbf{y}_A^T\boldsymbol{\alpha}_A + \mathbf{y}_F^T\mathbf{Q}_{FF}^{-1}(\mathbf{e}_F/(Cl) - \mathbf{Q}_{FA}\boldsymbol{\alpha}_A)}{\mathbf{y}_F^T\mathbf{Q}_{FF}^{-1}\mathbf{y}_F}.$$

Therefore,

$$\boldsymbol{\alpha}_F = \mathbf{Q}_{FF}^{-1}(\frac{\mathbf{e}_F}{Cl} - \mathbf{Q}_{FA}\boldsymbol{\alpha}_A - \frac{\mathbf{y}_A^T\boldsymbol{\alpha}_A + \mathbf{y}_F^T\mathbf{Q}_{FF}^{-1}(\mathbf{e}_F/(Cl) - \mathbf{Q}_{FA}\boldsymbol{\alpha}_A)}{\mathbf{y}_F^T\mathbf{Q}_{FF}^{-1}\mathbf{y}_F}\mathbf{y}_F). \tag{4.5}$$

We note that for $\underline{C} \le C \le \overline{C}$, if $(\boldsymbol{\alpha}_C)_F$ is defined by (4.5) and $(\boldsymbol{\alpha}_C)_A \equiv (\boldsymbol{\alpha}_{\underline{C}})_A$ (or $(\boldsymbol{\alpha}_{\underline{C}})_A$), then $(\alpha_C)_i \ge 0, i = 1, \ldots, l$. In addition, $\boldsymbol{\alpha}_C$ satisfies the first part of (4.1) (i.e. the part with right-hand side zero). The sign of the second part is not changed and (4.2) is also valid. Thus we have constructed an optimal solution $\boldsymbol{\alpha}_C$ of (D'_C) which is at the same face as $\boldsymbol{\alpha}_{\underline{C}}$ and $\boldsymbol{\alpha}_{\overline{C}}$. Then following from (4.5) and $\boldsymbol{\alpha}_A$ is a constant vector for all $\underline{C} \le C \le \overline{C}$,

$$\begin{aligned}
&e^T\boldsymbol{\alpha}_C \\
&= \mathbf{e}_F^T\mathbf{Q}_{FF}^{-1}(\mathbf{e}_F/(Cl) - \mathbf{Q}_{FA}\boldsymbol{\alpha}_A - b\mathbf{y}_F) + \mathbf{e}_A^T\boldsymbol{\alpha}_A \\
&= \mathbf{e}_F^T\mathbf{Q}_{FF}^{-1}(\mathbf{e}_F/(Cl) - \mathbf{Q}_{FA}\boldsymbol{\alpha}_A - \frac{\mathbf{y}_A^T\boldsymbol{\alpha}_A + \mathbf{y}_F^T\mathbf{Q}_{FF}^{-1}(\mathbf{e}_F/(Cl) - \mathbf{Q}_{FA}\boldsymbol{\alpha}_A)}{\mathbf{y}_F^T\mathbf{Q}_{FF}^{-1}\mathbf{y}_F}\mathbf{y}_F) + \\
&\quad \mathbf{e}_A^T\boldsymbol{\alpha}_A \\
&= (\frac{\mathbf{e}_F^T\mathbf{Q}_{FF}^{-1}\mathbf{e}_F}{l} - \frac{\mathbf{e}_F^T\mathbf{Q}_{FF}^{-1}(\mathbf{y}_F^T\mathbf{Q}_{FF}^{-1}\mathbf{e}_F/l)\mathbf{y}_F}{\mathbf{y}_F^T\mathbf{Q}_{FF}^{-1}\mathbf{y}_F})/C + \Delta_1 \\
&= (\frac{\mathbf{e}_F^T\mathbf{Q}_{FF}^{-1}\mathbf{e}_F}{l} - \frac{(\mathbf{e}_F^T\mathbf{Q}_{FF}^{-1}\mathbf{y}_F)^2}{(\mathbf{y}_F^T\mathbf{Q}_{FF}^{-1}\mathbf{y}_F)l})/C + \Delta_1 \\
&= \Delta_2/C + \Delta_1.
\end{aligned}$$

If $\mathbf{Q}_{FF}$ is not invertible, it is positive semi-definite so we can have $\mathbf{Q}_{FF} = \hat{\mathbf{Q}}\mathbf{D}\hat{\mathbf{Q}}^T$, where $\hat{\mathbf{Q}}^{-1} = \hat{\mathbf{Q}}^T$ is an orthonormal matrix. Without loss of generality we assume $\mathbf{D} = \begin{bmatrix} \bar{\mathbf{D}} & 0 \\ 0 & 0 \end{bmatrix}$. Then (4.4) can be modified to

$$\mathbf{D}\hat{\mathbf{Q}}^T\boldsymbol{\alpha}_F = \hat{\mathbf{Q}}^{-1}(\mathbf{e}_F/(Cl) - \mathbf{Q}_{FA}\boldsymbol{\alpha}_A - b\mathbf{y}_F).$$

One solution of the above system is

$$\boldsymbol{\alpha}_F = \hat{\mathbf{Q}}^{-T}\begin{bmatrix} \bar{\mathbf{D}}^{-1} & 0 \\ 0 & 0 \end{bmatrix}\hat{\mathbf{Q}}^{-1}(\mathbf{e}_F/(Cl) - \mathbf{Q}_{FA}\boldsymbol{\alpha}_A - b\mathbf{y}_F).$$

Thus a representation similar to (4.4) is obtained and all arguments follow.

Note that due to the positive semi-definiteness of $\mathbf{Q}_{FF}$, $\boldsymbol{\alpha}_F$ may have multiple solutions. From Lemma 3.1, $\mathbf{e}^T\boldsymbol{\alpha}_C$ is a well-defined function of C. Hence the representation $\Delta_1 + \Delta_2/C$ is valid for all solutions. □

The main result on the representation of $\mathbf{e}^T\boldsymbol{\alpha}_C$ is in the following theorem:

Theorem 4.2 *There are* $0 < C_1 < \cdots < C_s$ *and* $A_i, B_i, i = 1, \ldots, s$ *such that*

$$\mathbf{e}^T\boldsymbol{\alpha}_C = \begin{cases} \nu^* & C \le C_1, \\ A_i + \frac{B_i}{C} & C_i \le C \le C_{i+1}, i = 1, \ldots, s-1, \\ A_s + \frac{B_s}{C} & C_s \le C. \end{cases}$$

We also have

$$A_i + \frac{B_i}{C_i} = A_{i+1} + \frac{B_{i+1}}{C_{i+1}}, i = 2, \ldots, s-1.$$

Proof. From Theorem 3.7, we know $\mathbf{e}^T\boldsymbol{\alpha}_C = \nu^*$ when C is sufficiently small. If we gradually increase C, we will reach a C_1 such that if $C > C_1, \mathbf{e}^T\boldsymbol{\alpha}_C < \nu^*$. From this C_1, we can increase C to a C_2 such that if $C > C_2$, no $\boldsymbol{\alpha}_C$ is at the same face as $\boldsymbol{\alpha}_{C_1}$ and $\boldsymbol{\alpha}_{C_2}$. Then from Lemma 4.1, for $C_1 \le C \le C_2$, we can have A_1 and B_1 such that

$$\mathbf{e}^T\boldsymbol{\alpha}_C = A_1 + \frac{B_1}{C}.$$

We can continue this procedure. Since the number of possible faces is finite ($\le 3^l$), we have only finite C_i's. Otherwise, we will have C_i and C_j, $j \ge i+2$, such that there exist $\boldsymbol{\alpha}_{C_i}$ and $\boldsymbol{\alpha}_{C_j}$ at the same face. Then Lemma 4.1 implies that for all $C_i \le C \le C_j$, there are $\boldsymbol{\alpha}_C$ at the same face as $\boldsymbol{\alpha}_{C_i}$ and $\boldsymbol{\alpha}_{C_j}$. This contradicts to the definition of C_{i+1}. □

5 Numerical Experiments

In Section 3, we have shown the relation between (D_ν) and (D'_C) ($(\bar{D}_\nu)$ and $(\bar{D}'_C)$). Here we would like to experiment with their practical behavior. It has been known that when C is large, there may have more numerical difficulties on using decomposition methods for solving (D_C). (see, for example, the discussion in (Hsu and Lin 1999)). Now there is no C in (D_ν) so intuitively we may think that this difficulty no longer exists. In this section, we test the proposed decomposition method on examples with different ν and examine how many iterations are required. Note that because in Section 2 we discuss an algorithm for solving $(\bar{D}_\nu)$, here we work on it but not (D_ν).

Since the constraints $0 \le \alpha_i \le 1/l, i = 1, \ldots, l$, imply α_i are small, the objective value of $(\bar{D}_\nu)$ may be very close to zero. To avoid possible numerical inaccuracy, here we consider the following scaled form of $(\bar{D}_\nu)$:

$$\begin{aligned} \min & \frac{1}{2}\boldsymbol{\alpha}^T(\mathbf{Q} + \mathbf{y}\mathbf{y}^T)\boldsymbol{\alpha} \\ & 0 \le \alpha_i \le 1, i = 1, \ldots, l, \\ & \mathbf{e}^T\boldsymbol{\alpha} = \nu l. \end{aligned} \tag{5.1}$$

The working set selection follows the discussion in Section 2 and here we implement a special case with $q = 2$. The approach of using $q = 2$ adopts the idea in (Platt 1998). Then the working set in each iteration contains only two elements. If they are i_1 and i_2, the solution of (2.10) has a very simple form:

$$\begin{aligned} i_1 &= \text{argmin}_i\{((Q + yy^T)\alpha)_i \mid \alpha_i < 1\}, \\ i_2 &= \text{argmin}_i\{-((Q + yy^T)\alpha)_i \mid \alpha_i > 0\}. \end{aligned}$$

The decomposition method stops if the iterate $\boldsymbol{\alpha}$ satisfies the following condition:

$$-((Q + yy^T)\alpha)_{i_1} + ((Q + yy^T)\alpha)_{i_2} < \epsilon, \tag{5.2}$$

where we set $\epsilon = 10^{-5}$.

We use $\boldsymbol{\alpha}_1 = [1, \ldots, 1, \nu l - \lfloor \nu l \rfloor, 0, \ldots, 0]^T$ as the initial point. Unlike the decomposition method for (D_C), where the zero vector is usually used as the

Table 5.1: All problems, $\nu = 0.8, 0.6, 0.4, 0.2$

Problem	l	ν	C	Iter.	Obj.	BSV	SV	νl	Mis.
australian	690	0.8	0.023253	557	6074.97	546	557	552	99
		0.6	0.051625	1228	1820.64	406	419	413	100
		0.4	0.181072	5706	211.705	257	291	276	100
		0.2	165.386	347530	0.124832	86	232	138	44
diabetes	768	0.8	0.0064	162	6668.46	614	615	614	268
		0.6	0.612632	1564	52.932	457	466	460	168
		0.4	2403.05	2315167	0.010352	249	379	307	120
		0.2	267498	11195886	0.000145	70	343	153	100
german	1000	0.8	0.003473	179	28160.3	797	803	800	300
		0.6	0.323135	956	129.933	579	625	599	245
		0.4	15.4949	34124	5.35281	269	565	400	98
		0.2	170.722	140119	0.421459	74	524	200	21
heart	270	0.8	0.063272	105	652.252	214	220	216	45
		0.6	0.203178	242	134.945	156	166	161	43
		0.4	1.66277	2156	8.8488	94	123	108	35
		0.2	51.5371	19312	0.344443	26	114	54	6
vehicle	846	0.8	0.002197	162	55224.8	675	678	676	212
		0.6	0.007305	336	5442.86	504	511	507	212
		0.4	64.0555	97158	0.61915	313	363	338	106
		0.2	3221.89	2390690	0.012161	108	255	169	35
satimage	4435	0.8	0.000408	857	1.66722e+06	3548	3548	3548	1072
		0.6	0.001244	1421	242457	2658	2663	2660	1072
		0.4	0.026895	2833	20459.8	1767	1780	1774	140
		0.2	0.121746	2730	2191.38	883	892	887	74
letter	15000	0.8	5e-05	2372	1.05765e+08	11998	12001	12000	594
		0.6	7e-05	3633	5.44233e+07	8998	9002	8999	594
		0.4	0.000116	3604	2.02506e+07	5995	6002	6000	594
		0.2	0.000321	2472	2.76423e+06	2996	3004	3000	594
shuttle	43500	0.8	3.1e-05	6981	2.616e+08	34800	34800	34800	9392
		0.6	6.7e-05	10602	5.66317e+07	26099	26101	26099	9392
		0.4	0.004199	11398	670122	17398	17401	17400	6840
		0.2	0.160328	9904	8419.5	8697	8702	8700	2947
a4a	4781	0.8	0.000366	851	1.97605e+06	3821	3827	3824	1188
		0.6	0.001082	1342	228795	2857	2880	2868	1188
		0.4	1.51548	18119	68.1789	1885	1940	1912	757
		0.2	5699.52	14291238	0.034041	571	1747	956	255
w7a	24692	0.8	2.8e-05	6225	3.20665e+08	19752	19755	19753	740
		0.6	3.9e-05	6583	1.6941e+08	14813	14817	14815	740
		0.4	6.2e-05	7005	6.64636e+07	9875	9879	9876	740
		0.2	0.000155	3583	1.10632e+07	4934	4941	4938	740

Table 5.2: Large problems, $\nu = 0.1, 0.05, 0.02, 0.01$

Problem	l	ν	C	Iter.	Obj.	BSV	SV	νl	Mis.
satimage	4435	0.1	0.615475	4783	177.702	436	448	443	59
		0.05	5.92071	31903	6.28106	210	235	221	50
		0.02	498.303	403162	0.053866	54	149	88	11
		0.01	2226.34	639353	0.007708	10	135	44	5
letter	15000	0.1	0.002034	1363	76699.7	1496	1504	1500	594
		0.05	0.377606	6462	450.214	742	757	750	121
		0.02	18.461	53977	2.96639	276	324	300	76
		0.01	306.438	543764	0.155342	111	205	150	15
shuttle	43500	0.1	8.96413	303531	89.0582	4346	4355	4350	501
		0.05	179.898	1052288	2.19431	2168	2184	2175	98
		0.02	8282.67	27736662	0.036382	850	896	870	66
		0.01	27146.1	52522524	0.005238	386	531	435	50
a4a	4781	0.1	59058.6	34585532	0.00201	208	1682	478	110
		0.05	205498	28088782	0.000282	55	1570	239	211
		0.02	245831	9061246	0.000138	7	1600	95	233
		0.01	312420	4465335	6.4e-05	1	1563	47	480
w7a	24692	0.1	0.000556	1633	883735	2455	2477	2469	740
		0.05	8.39917	385817	19.306	1150	1370	1234	417
		0.02	2347.03	8060630	0.031724	340	1137	493	174
		0.01	210977	19881862	7.1e-05	202	939	246	2350

initial solution so $\nabla f(\boldsymbol{\alpha}_1) = -\mathbf{e}$, now $\boldsymbol{\alpha}_1$ contains $\lfloor \nu l \rfloor + 1$ nonzero components. In order to obtain $\nabla f(\boldsymbol{\alpha}_1) = (\mathbf{Q} + \mathbf{y}\mathbf{y}^T)\boldsymbol{\alpha}_1$ of (2.10), in the beginning of the decomposition procedure, we must compute $\lfloor \nu l \rfloor + 1$ columns of $\mathbf{Q}$. This might be a disadvantage of using ν-SVM. Further investigations are needed on this issue.

We test the RBF kernel with $Q_{ij} = y_i y_j e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/n}$, where n is the number of attributes of a training data. Our implementation is part of the software LIBSVM* (version 2.0) which is an integrated package for SVM classification and regression.

We test problems from various collections with different ν. Problems australian to shuttle are from the Statlog collection (Michie et al. 1994). Problems adult4 and web7 are compiled by Platt (1998) from the UCI Machine Learning Repository (Murphy and Aha 1994). Note that all problems from Statlog are with real numbers so we scale them to $[-1, 1]$. Problems adult4 and web7 are with binary representation so we do not conduct any scaling. Some of these problems have more than 2 classes so we treat all data not in the first class as in the second class.

Table 5.1 lists results using $\nu = 0.8, 0.6, 0.4$, and 0.2. Since we expect that large

*LIBSVM is available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

problems might only need a smaller proportion of training data as support vectors, we also tried smaller ν (0.1, 0.05, 0.02, 0.01) on large problems ($l \geq 4,000$). The results are in Table 5.2.

Tables 5.1 and 5.2 report number of training data in each class, the corresponding C in $(\bar{D}'_C)$, number of iterations, objective values, number of bounded support vectors, number of support vectors, νl, and number of misclassified training data. Here C is calculated by

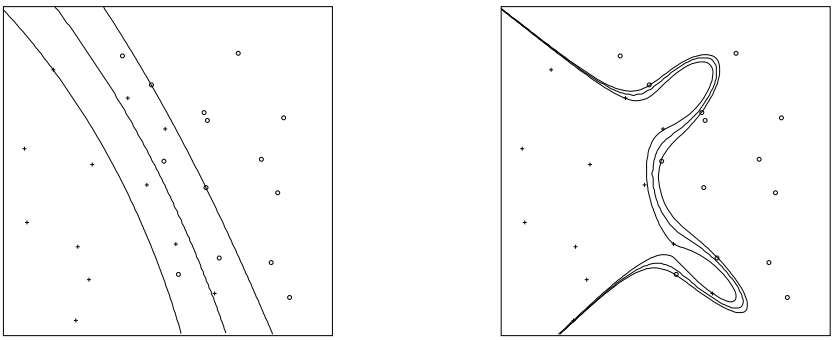
$$C = 1/\rho,$$

where ρ is the Lagrange multiplier of the equality constraint of (5.1). This is different from the formulation $C = 1/(\rho l)$ in Section 3 because of the use of the scaled form (5.1). Results in both tables confirm the theoretical understanding that ν and C change in opposite directions.

From (Schölkopf et al. 2000), we know that νl is a lower bound of the number of support vectors and an upper bound of the number of bounded support vectors (also number of misclassified training data). Note that for (5.1), this property on νl still holds. It can be clearly seen from Tables 5.1 and 5.2 that νl lies between the number of support vectors and bounded support vectors. Furthermore, we can see that if ν becomes smaller, the total number of support vectors decreases. This is consistent with the situation of using (D_C), where the increase of C decreases the number of support vectors.

We also observe that though the total number of support vectors decreases as ν becomes smaller, the number of free support vectors increases. When ν is decreased (C is increased), the separating hyperplane tries to to fit as many training data as possible. Hence more points (that is, more free α_i) tend to be at two planes $\mathbf{w}^T\phi(\mathbf{x}) + b = \pm\rho$. We illustrate this in Figures 5.1(a) and (b), where $\nu = 0.5$ and 0.2, respectively, are used on the same problem. Since the weakest part of the decomposition method is that it cannot consider all variables together in each iteration (only q elements are selected), a larger number of free variables may cause more difficulty.

This gives an explanation why a lot more iterations are required when ν are small. Therefore, here we have given an example that for solving (D_C) and $(\bar{D}_\nu)$, the decomposition method faces a similar difficulty.



(a) $\nu = 0.5$ (b) $\nu = 0.2$

Figure 5.1: Training data and separating hyperplanes

However, an interesting exception is that for the problem adult4, when ν is decreased from 0.1 to 0.05, 0.02, and 0.01, the number of iterations does decrease. We also observe that the number C using $\nu = 0.1$ is larger than those of using smaller ν. In theory C should increase as ν decreases so it turns out that when $\nu \leq 0.05$, results are erroneous. As the objective values become so small, the 10^{-5} stopping criterion is too loose. That is, the decomposition method stops too early and returns a wrong solution. This situation suggests that a relative stopping criterion might be better than an absolute one. How to choose an appropriate stopping criteria will be a future research issue.

Even though we expect that many numerical properties of solving (D_C) and (D_ν) are similar, ν-SVM may still have the advantage of using a more reasonable parameter ν. In addition to problems mentioned above, we think some other numerical issues of ν-SVM should be investigated in the future:

1. Better working set selections of the decomposition method for solving $(\bar{D}_\nu)$.

2. The issue on how to choose a reasonable ν. The selection of ν not only affects the error rate but also the number of iterations of the decomposition method. For example, the number of iterations does not change much for w7a when ν decreases from 0.8 to 0.2, but for a4a the number of iterations increases more than 15,000-fold.

3. The comparison of using the decomposition method on C-SVM and ν-SVM.

6 Conclusions

In this paper, we modify ν-SVM to a different form where existing decomposition methods can be adapted to solve it. The relation between ν-SVM and C-SVM is also investigated. In particular, we show that solving them is just like solving two different problems with the same optimal solution set. We have mentioned several issues for future investigation which can lead ν-SVM to be a practical tool.

Acknowledgments

The second author thanks Craig Saunders for bringing him to the attention of ν-SVM. He also thanks an anonymous referee of (Lin 1999) whose comments lead him to think about the infeasibility of (D_ν). The authors also thank Dr. Bernhard Schölkopf for some helpful comments.

References

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

Crisp, D. J. and C. J. C. Burges (1999). A geometric interpretation of ν-SVM classifiers. In *NIPS99*.

Cristianini, N. and J. Shawe-Taylor (2000). *An Introduction to Support Vector Machines*. Cambridge, UK: Cambridge University Press.

Friess, T.-T., N. Cristianini, and C. Campbell (1998). The kernel adatron algorithm: a fast and simple learning procedure for support vector machines. In *Proceeding of 15th Intl. Conf. Machine Learning*. Morgan Kaufman Publishers.

Hsu, C.-W. and C.-J. Lin (1999). A simple decomposition method for support vector machines. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.

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. Shevade, C. Bhattacharyya, and K. Murthy (1999). Improvements to Platt's SMO algorithm for SVM classifier design. Technical report, Department of Mechanical and Production Engineering, National University of Singapore.

Keerthi, S. S., C. B. S. K. Shevade, and K. R. K. Murthy (2000). A fast iterative nearest point algorithm for support vector machine classifier design. *IEEE Trans. Neural Networks 11*(1), 124–136.

Lin, C.-J. (1999). Formulations of support vector machines: a note from an optimization point of view. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan. To appear in *Neural Computation*.

Mangasarian, O. L. and D. R. Musicant (1999). Successive overrelaxation for support vector machines. *IEEE Trans. Neural Networks 10*(5), 1032–1037.

Micchelli, C. A. (1986). Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation 2*, 11–22.

Michie, D., D. J. Spiegelhalter, and C. C. Taylor (1994). *Machine Learning, Neural and Statistical Classification*. Englewood Cliffs, N.J.: Prentice Hall. Data available at anonymous ftp: `ftp.ncc.up.pt/pub/statlog/`.

Murphy, P. M. and D. W. Aha (1994). UCI repository of machine learning databases. Available at `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Osuna, E., R. Freund, and F. Girosi (1997). Training support vector machines: An application to face detection. In *Proceedings of CVPR'97*.

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.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton, NJ: Princeton University Press.

Saunders, C., M. O. Stitson, J. Weston, L. Bottou, B. Schölkopf, and A. Smola (1998). Support vector machine reference manual. Technical Report CSD-TR-98-03, Royal Holloway, University of London, Egham, UK.

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

Schölkopf, B., A. Smola, R. C. Williamson, and P. L. Bartlett (2000). New support vector algorithms. *Neural Computation 12*, 1083 – 1121.

Schölkopf, B., A. J. Smola, and R. Williamson (1999). Shrinking the tube: A new support vector regression algorithm. In M. S. Kearns, S. A. Solla, and D. A. Cohn (Eds.), *Advances in Neural Information Processing Systems*, Volume 11, Cambridge, MA. MIT Press.

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