Manuscript Number: 2187

Training ν-Support Vector Classifiers: Theory and Algorithms

Chih-Chung Chang and Chih-Jen Lin

Department of Computer Science and

Information Engineering

National Taiwan University

Taipei 106, Taiwan (cjlin@csie.ntu.edu.tw)

Abstract The ν-support vector machine (ν-SVM) for classification proposed by Schölkopf et al. has the advantage of using a parameter ν on controlling the number of support vectors. In this paper, 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. 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. We propose a decomposition method for ν-SVM which is competitive with existing methods for C-SVM. 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:

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

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

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. introduce 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, $Q_{ij} \equiv y_iy_jK(\mathbf{x}_i, \mathbf{x}_j)$, 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.

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 2 we investigate this relation in more detail. The main result (Theorem 5) 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. Based on the work in Section 2, in Section 3 we derive the formulation of ν as a decreasing function of C.

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 4, we propose a decomposition method similar to the software SVM^{light} (Joachims 1998) for C-SVM.

Section 5 presents numerical results. Experiments indicate that several numerical properties on solving (D_C) and (D_ν) are similar. A timing comparison shows that the proposed method for ν-SVM is competitive with existing methods for C-SVM. Finally in Section 6, we give some discussions and conclusions.

2 The Relation Between ν-SVM and C-SVM

In this section we construct a relationship between (D_ν) and (D_C) where the main result is in Theorem 5. 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.

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

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

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

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 $\mathbf{e}^T\boldsymbol{\alpha} \geq \nu$ 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 $\mathbf{e}^T\boldsymbol{\alpha} \geq \nu$ as an equality constraint. From the primal side, it was first shown in (Crisp and Burges 2000) that $\rho \geq 0$ in (P_ν) is redundant. Without $\rho \geq 0$, the dual becomes:

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

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 but (D_ν) may be infeasible. 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 $0 \leq \alpha_i \leq 1/3, [1, 1, -1]\boldsymbol{\alpha} = 0$ and $\mathbf{e}^T\boldsymbol{\alpha} \geq 0.9$. 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 2000).

Lemma 1 *(D_ν) is feasible if and only if $\nu \leq \nu_{max}$, 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 $0 \leq \alpha_i \leq 1/l, i = 1, \ldots, l$, with $\mathbf{y}^T\boldsymbol{\alpha} = 0$, for any $\boldsymbol{\alpha}$ feasible to (D_ν), we have $\mathbf{e}^T\boldsymbol{\alpha} \leq \nu_{max}$. Therefore, if (D_ν) is feasible, $\nu \leq \nu_{max}$. On the other hand, if $0 < \nu \leq \nu_{max}$, $\min(\#y_i = 1, \#y_i = -1) > 0$ 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 $0 \leq \alpha_i \leq 1/l, i = 1, \ldots, l$ and $\mathbf{y}^T\boldsymbol{\alpha} = 0$. If $\nu = 0$, clearly $\boldsymbol{\alpha} = 0$ is a feasible solution of (D_ν). $\square$

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 $\nu_{max} = 1$.

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

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

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

Lemma 2 *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

$$\boldsymbol{\alpha}_1^T\mathbf{Q}\boldsymbol{\alpha}_2 = \frac{1}{2}\boldsymbol{\alpha}_1^T\mathbf{Q}\boldsymbol{\alpha}_1 + \frac{1}{2}\boldsymbol{\alpha}_2^T\mathbf{Q}\boldsymbol{\alpha}_2. \tag{2.3}$$

Since $\mathbf{Q}$ is positive semidefinite, $\mathbf{Q} = L^T L$ so (2.3) 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 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 2, 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 2, 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:

$$\begin{gathered}\mathbf{Q}\boldsymbol{\alpha} - \frac{\mathbf{e}}{Cl} + b\mathbf{y} = \boldsymbol{\lambda} - \boldsymbol{\xi}, \\ \boldsymbol{\lambda}^T\boldsymbol{\alpha} = 0, \boldsymbol{\xi}^T(\frac{\mathbf{e}}{l} - \boldsymbol{\alpha}) = 0, \mathbf{y}^T\boldsymbol{\alpha} = 0 \\ \lambda_i \geq 0, \xi_i \geq 0, 0 \leq \alpha_i \leq 1/l, i = 1, \ldots, l.\end{gathered} \tag{2.4}$$

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

$$\begin{gathered}\mathbf{Q}\boldsymbol{\alpha} - \rho\mathbf{e} + b\mathbf{y} = \boldsymbol{\lambda} - \boldsymbol{\xi}, \\ \lambda^T\boldsymbol{\alpha} = 0, \boldsymbol{\xi}^T(\frac{\mathbf{e}}{l} - \boldsymbol{\alpha}) = 0, \\ \mathbf{y}^T\boldsymbol{\alpha} = 0, \mathbf{e}^T\boldsymbol{\alpha} \geq \nu, \rho(\mathbf{e}^T\boldsymbol{\alpha} - \nu) = 0, \\ \lambda_i \geq 0, \xi_i \geq 0, \rho \geq 0, 0 \leq \alpha_i \leq 1/l, i = 1, \ldots, l.\end{gathered} \tag{2.5}$$

From Theorem 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 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 *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 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 (2.5),

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

Then $\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} > 0$ and (2.5) imply

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

By choosing a $C > 0$ such that $\rho = 1/(Cl)$, $\boldsymbol{\alpha}$ is a KKT point of (D'_C). Hence from Theorem 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 *If the objective value of (D_ν), $\nu \geq 0$, 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, (2.4) of (D'_C)'s KKT condition becomes

$$-\frac{\mathbf{e}}{Cl} + \begin{bmatrix} b\mathbf{e}_I \\ -b\mathbf{e}_J \end{bmatrix} = \boldsymbol{\lambda} - \boldsymbol{\xi}, \tag{2.6}$$

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 (2.6):

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

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 a 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

$$\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - \frac{\mathbf{e}^T\boldsymbol{\alpha}}{Cl} \geq -\frac{\mathbf{e}^T\boldsymbol{\alpha}}{Cl} \geq -\frac{1}{Cl}. \tag{2.7}$$

Now $(\#y_i = 1) = (\#y_i = -1)$ so $\mathbf{e}/l$ is feasible. When $\boldsymbol{\alpha} = \mathbf{e}/l$, the inequality of (2.7) 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, 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. $\Box$

Lemma 4 *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

$$\frac{1}{2}\boldsymbol{\alpha}_1^T\mathbf{Q}\boldsymbol{\alpha}_1 - \frac{\mathbf{e}^T\boldsymbol{\alpha}_1}{C_1 l} \leq \frac{1}{2}\boldsymbol{\alpha}_2^T\mathbf{Q}\boldsymbol{\alpha}_2 - \frac{\mathbf{e}^T\boldsymbol{\alpha}_2}{C_1 l} \tag{2.8}$$

and

$$\frac{1}{2}\boldsymbol{\alpha}_2^T\mathbf{Q}\boldsymbol{\alpha}_2 - \frac{\mathbf{e}^T\boldsymbol{\alpha}_2}{C_2 l} \leq \frac{1}{2}\boldsymbol{\alpha}_1^T\mathbf{Q}\boldsymbol{\alpha}_1 - \frac{\mathbf{e}^T\boldsymbol{\alpha}_1}{C_2 l}. \tag{2.9}$$

Hence

$$\frac{\mathbf{e}^T\boldsymbol{\alpha}_1}{C_2 l} - \frac{\mathbf{e}^T\boldsymbol{\alpha}_2}{C_2 l} \leq \frac{1}{2}\boldsymbol{\alpha}_1^T\mathbf{Q}\boldsymbol{\alpha}_1 - \frac{1}{2}\boldsymbol{\alpha}_2^T\mathbf{Q}\boldsymbol{\alpha}_2 \leq \frac{\mathbf{e}^T\boldsymbol{\alpha}_1}{C_1 l} - \frac{\mathbf{e}^T\boldsymbol{\alpha}_2}{C_1 l}. \tag{2.10}$$

Since $C_2 > C_1 > 0$, (2.10) 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

$$0 \leq \lim_{C\to(C^*)^+}\mathbf{e}^T\boldsymbol{\alpha}_C < \bar{\nu} < \mathbf{e}^T\boldsymbol{\alpha}_{C^*}. \tag{2.11}$$

Hence $\bar{\nu} > 0$. If $(D_{\bar{\nu}})$'s objective value is not zero, from Theorem 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 (2.11) 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, 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, $\mathbf{e}^T\boldsymbol{\alpha}_C = \nu_{max}$, for all C. This contradicts (2.11).

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 5 *We can define*

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

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. In addition, the optimal objective value of (D_ν) is strictly positive. For any $0 \leq \nu \leq \nu_*$, (D_ν) is feasible with zero optimal objective value.

Proof. First from Lemma 4 and the fact that $0 \leq \mathbf{e}^T\boldsymbol{\alpha} \leq 1$, 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 $\boldsymbol{\lambda}_C \geq 0, \boldsymbol{\xi}_C \geq 0, \boldsymbol{\alpha}_C^T \lambda_C = 0$, 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 $(\#y_i = 1) \neq (\#y_i = -1)$. 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 1 we immediately know that (D_ν) is infeasible if $\nu > \nu^*$. From Lemma 4 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, (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, 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 1. It cannot have nonzero objective value due to Theorem 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 $0 \leq \alpha_i \leq 1/l$, i = 1, . . . , l, with Theorem 1, there is a sequence $\{\boldsymbol{\alpha}_{\nu_i}\}, \nu_1 \leq \nu_2 \leq \cdots < \nu_*$ 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 $0 \leq \hat{\boldsymbol{\alpha}} \leq 1/l$ 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 $\nu \leq \nu_*$. From the above discussion, if $\nu \leq \nu_*$, the objective value of (D_ν) is zero. If the objective value of (D_ν) is zero but $\nu > \nu_*$, Theorem 3 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 $\mathbf{w}$ of the primal problem (P_ν) is zero. In (Crisp and Burges 2000, Section 4), they considered such a (P_ν) as a "trivial" problem. Next we present a corollary:

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

Proof. From (Lin 2001a, Theorem 3.3), if data are separable, there is a C^* such that for all $C \geq C^*$, 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, $\mathbf{e}^T\boldsymbol{\alpha}_C \geq 1/l > 0$ for all C. Hence $\nu_* \geq 1/l$.

If $\mathbf{Q}$ is positive definite, the unconstrained problem

$$\min \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha} \tag{2.12}$$

has an unique solution at $\boldsymbol{\alpha} = \mathbf{Q}^{-1}\mathbf{e}$. If we add additional constraints to (2.12),

$$\begin{aligned} \min \quad & \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha} \\ & \mathbf{y}^T\boldsymbol{\alpha} = 0, \alpha_i \geq 0, i = 1, \ldots, l, \end{aligned} \tag{2.13}$$

is a problem with a smaller feasible region. Thus the objective value of (2.13) 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, (2.13) is solvable. From (Lin 2001a, 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. $\square$

In many situations $\mathbf{Q}$ is positive definite. For example, from (Micchelli 1986), if the RBF kernel is used and $\mathbf{x}_i \neq \mathbf{x}_j$, $\mathbf{Q}$ 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. 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:

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

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

$$\mathbf{e}^T\boldsymbol{\alpha}_C = \begin{cases} 1 & \text{if } 0 < C \leq \frac{1}{3}, \\ \frac{1}{3} + \frac{2}{9C} & \text{if } \frac{1}{3} \leq C \leq \frac{4}{3}, \\ \frac{1}{2} & \text{if } \frac{4}{3} \leq C \leq 4, \\ \frac{1}{2C} & \text{if } C \geq 4. \end{cases} \tag{2.14}$$

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.

3 The Relation Between ν and C

A formula like (2.14) 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 (2.14), 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 5 *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. In addition, $\Delta_2 \ge 0$.

Proof. If $\{1, \dots, l\}$ 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

$$\begin{bmatrix} \mathbf{Q}_{FF} & \mathbf{Q}_{FA} \\ \mathbf{Q}_{AF} & \mathbf{Q}_{AA} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_F \\ \boldsymbol{\alpha}_A \end{bmatrix} - \frac{\mathbf{e}}{Cl} + b \begin{bmatrix} \mathbf{y}_F \\ \mathbf{y}_A \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{\lambda}_A - \boldsymbol{\xi}_A \end{bmatrix}, \tag{3.1}$$

$$\mathbf{y}_F^T\boldsymbol{\alpha}_F + \mathbf{y}_A^T\boldsymbol{\alpha}_A = 0, \tag{3.2}$$

$$\lambda_i \ge 0, \xi_i \ge 0, i \in A. \tag{3.3}$$

(3.1) and (3.2) 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,

$$\boldsymbol{\alpha}_F = \mathbf{Q}_{FF}^{-1}(\mathbf{e}_F/(Cl) - \mathbf{Q}_{FA}\boldsymbol{\alpha}_A - b\mathbf{y}_F). \tag{3.4}$$

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,

$$\boldsymbol{\alpha}_F = \mathbf{Q}_{FF}^{-1}(\frac{\mathbf{e}_F}{Cl} - \mathbf{Q}_{FA}\boldsymbol{\alpha}_A - \frac{\mathbf{y}_A^T\boldsymbol{\alpha}_A + \mathbf{y}_F^T\mathbf{Q}_{FF}^{-1}(\mathbf{e}_F/(Cl) - \mathbf{Q}_{FA}\boldsymbol{\alpha}_A)}{\mathbf{y}_F^T\mathbf{Q}_{FF}^{-1}\mathbf{y}_F}\mathbf{y}_F). \tag{3.5}$$

We note that for $\underline{C} \le C \le \overline{C}$, if $(\boldsymbol{\alpha}_C)_F$ is defined by (3.5) 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 (3.1) (i.e. the part with right-hand side zero). The sign of the second part is not changed and (3.2) 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 (3.5) 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 (3.4) 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 (3.4) 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 2, $\mathbf{e}^T\boldsymbol{\alpha}_C$ is a well-defined function of C. Hence the representation $\Delta_1 + \Delta_2/C$ is valid for all solutions. From Lemma 4, $\mathbf{e}^T\boldsymbol{\alpha}_C$ is a decreasing function of C so $\Delta_2 \geq 0$. $\square$

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

Theorem 6 *There are $0 < C_1 < \cdots < C_s$ and $A_i, B_i, i = 1, \ldots, s$ such that*

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

where $\boldsymbol{\alpha}_C$ is an optimal solution of $(D_{C'})$. We also have

$$A_i + \frac{B_i}{C_{i+1}} = A_{i+1} + \frac{B_{i+1}}{C_{i+1}}, i = 1, \ldots, s-1. \tag{3.6}$$

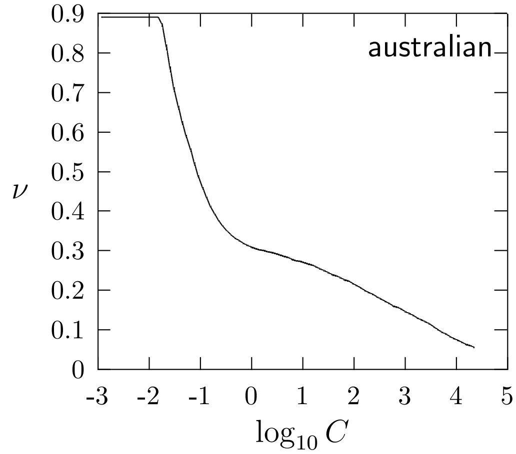
Proof. From Theorem 5, we know $\mathbf{e}^T\boldsymbol{\alpha}_C = \nu^*$ when C is sufficiently small. From Lemma 4, if we gradually increase C, we will reach a C_1 such that if $C > C_1, \mathbf{e}^T\boldsymbol{\alpha}_C < \nu^*$. If for all $C \geq C_1$, $\boldsymbol{\alpha}_C$ are at the same face, from Lemma 5, we have $\mathbf{e}^T\boldsymbol{\alpha}_C = A_1 + B_1/C, \forall C \geq C_1$. Otherwise, from this C_1, we can increase C to a C_2 such that for all intervals $(C_2, C_2+\epsilon), \epsilon > 0$, there is no $\boldsymbol{\alpha}_C$ at the same face as $\boldsymbol{\alpha}_{C_1}$ and $\boldsymbol{\alpha}_{C_2}$. Then from Lemma 5, for $C_1 \leq C \leq C_2$, we can have A_1 and B_1 such that

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

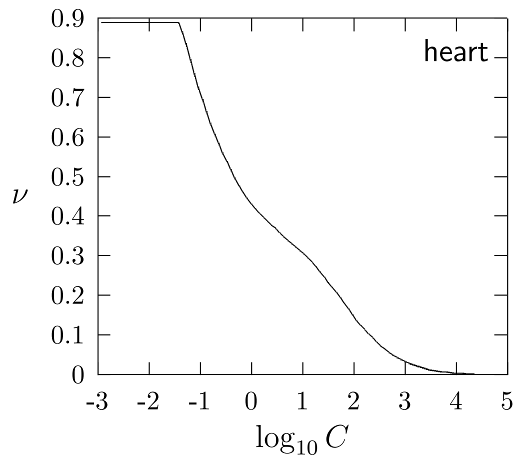
We can continue this procedure. Since the number of possible faces is finite ($\leq 3^l$), we have only finite C_i's. Otherwise, we will have C_i and C_j, $j \geq i+2$, such that there exist $\boldsymbol{\alpha}_{C_i}$ and $\boldsymbol{\alpha}_{C_j}$ at the same face. Then Lemma 5 implies that for all $C_i \leq C \leq C_j$, all $\boldsymbol{\alpha}_C$ are at the same face as $\boldsymbol{\alpha}_{C_i}$ and $\boldsymbol{\alpha}_{C_j}$. This contradicts the definition of C_{i+1}.

From Lemma 4, the continuity of $\mathbf{e}^T\boldsymbol{\alpha}_C$ immediately implies (3.6). $\square$

Finally we provide Figure 1 to demonstrate the relation between ν and C. It clearly indicates that ν is a decreasing function of C. Information about these two test problems australian and heart are in Section 5.



(a)



(b)

Figure 1: The relation between ν and C

4 A Decomposition Method for ν-SVM

Based on existing decomposition methods for C-SVM, in this section we propose a decomposition method for ν-SVM.

For solving (D_C), existing decomposition methods separate 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:

$$
\begin{aligned}
\min \quad & \frac{1}{2}\boldsymbol{\alpha}_B^T\mathbf{Q}_{BB}\boldsymbol{\alpha}_B - (\mathbf{e}_B - \mathbf{Q}_{BN}\boldsymbol{\alpha}_N)^T\boldsymbol{\alpha}_B \\
& \mathbf{y}_B^T\boldsymbol{\alpha}_B = -\mathbf{y}_N^T\boldsymbol{\alpha}_N, \qquad (4.1)\\
& 0 \le (\alpha_B)_i \le C, i = 1, \dots, q,
\end{aligned}
$$

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; Keerthi and Gilbert 2002; Lin 2001b).

An important process in the decomposition methods is the selection of the working set B. 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:

$$
\begin{aligned}
\min \quad & \nabla f(\boldsymbol{\alpha}_k)^T\mathbf{d} \\
& \mathbf{y}^T\mathbf{d} = 0,\ -1 \le d_i \le 1, \qquad (4.2)\\
& d_i \ge 0,\ \text{if } (\alpha_k)_i = 0,\ d_i \le 0,\ \text{if } (\alpha_k)_i = C, \qquad (4.3)\\
& |\{d_i \mid d_i \ne 0\}| = q, \qquad (4.4)
\end{aligned}
$$

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 (4.4) 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 (4.1). 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 (4.2)-(4.4). First he sorts $y_i\nabla f(\boldsymbol{\alpha}_k)_i, i = 1, \dots, l$ 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 (4.3). 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 (4.3). Other elements of d are assigned to be zero. Thus these q nonzero elements compose the working set. A complete analysis of his procedure is in (Lin 2001b, Section 2).

To modify the above strategy for (D_ν), we consider the following problem in each iteration:

$$\begin{aligned}
\min \quad & \nabla f(\boldsymbol{\alpha}_k)^T \mathbf{d} \\
& \mathbf{y}^{\mathbf{T}}\mathbf{d} = \mathbf{0}, \mathbf{e}^T\mathbf{d} = 0, \ -1 \leq d_i \leq 1, \\
& d_i \geq 0, \text{ if } (\alpha_k)_i = 0, \ d_i \leq 0, \text{ if } (\alpha_k)_i = 1/l, \\
& |\{d_i \mid d_i \neq 0\}| \leq q,
\end{aligned} \tag{4.5}$$

where q is an even integer. Now $f(\boldsymbol{\alpha}) \equiv \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha}$. Here we use "$\leq$" instead of "$=$" because in theory q nonzero elements may not be always available. This was first pointed out in (Chang et al. 2000). Note that the sub-problem (4.1) becomes as follows if decomposition methods are used for solving (D_ν):

$$\begin{aligned}
\min \quad & \frac{1}{2}\boldsymbol{\alpha}_B^T\mathbf{Q}_{BB}\boldsymbol{\alpha}_B + \mathbf{Q}_{BN}\boldsymbol{\alpha}_N^T\boldsymbol{\alpha}_B \\
& \mathbf{y}_B^T\boldsymbol{\alpha}_B = -\mathbf{y}_N^T\boldsymbol{\alpha}_N, \\
& \mathbf{e}_B^T\boldsymbol{\alpha}_B = \nu - \mathbf{e}_N^T\boldsymbol{\alpha}_N, \\
& 0 \leq (\alpha_B)_i \leq 1/l, i = 1, \ldots, q.
\end{aligned} \tag{4.6}$$

Problem (4.5) is more complicated then (4.2) as there is an additional constraint $\mathbf{e}^T\mathbf{d} = 0$. The situation of $q = 2$ has been discussed in (Keerthi and Gilbert 2002). We will describe a recursive procedure for solving (4.5).

We consider the following problem:

$$\begin{aligned}
\min \quad & \sum_{t \in S} \nabla f(\boldsymbol{\alpha}_k)_t d_t \\
& \sum_{t \in S} y_t d_t = 0, \ \sum_{t \in S} d_t = 0, \ -1 \leq d_t \leq 1, \\
& d_t \geq 0, \text{ if } (\alpha_k)_t = 0, \ d_t \leq 0, \text{ if } (\alpha_k)_t = 1/l, \\
& |\{d_t \mid d_t \neq 0, t \in S\}| \leq q,
\end{aligned} \tag{4.7}$$

which is the same as (4.5) if $S = \{1, \ldots, l\}$. We denote the variables $\{d_t | t \in S\}$ as $\mathbf{d}$ and the objective function $\sum_{t \in S} \nabla f(\boldsymbol{\alpha}_k)_t d_t$ as $obj(\mathbf{d})$.

Algorithm 1 *If $q = 0$, the algorithm stops and outputs $\mathbf{d} = 0$. Otherwise choose a pair of indices i and j from either*

$$\begin{aligned} i &= argmin_t\{\nabla f(\boldsymbol{\alpha}_k)_t | y_t = 1, (\alpha_k)_t < 1/l, t \in S\}, \\ j &= argmax_t\{\nabla f(\boldsymbol{\alpha}_k)_t | y_t = 1, (\alpha_k)_t > 0, t \in S\}, \end{aligned} \tag{4.8}$$

or

$$\begin{aligned} i &= argmin_t\{\nabla f(\boldsymbol{\alpha}_k)_t | y_t = -1, (\alpha_k)_t < 1/l, t \in S\}, \\ j &= argmax_t\{\nabla f(\boldsymbol{\alpha}_k)_t | y_t = -1, (\alpha_k)_t > 0, t \in S\}, \end{aligned} \tag{4.9}$$

depending on which one gives a smaller $\nabla f(\boldsymbol{\alpha}_k)_i - \nabla f(\boldsymbol{\alpha}_k)_j$. If there are no such i and j, or $\nabla f(\boldsymbol{\alpha}_k)_i - \nabla f(\boldsymbol{\alpha}_k)_j \geq 0$, the algorithm stops and outputs a solution $\mathbf{d} = 0$. Otherwise we assign $d_i = 1, d_j = -1$ and determine values of other variables by recursively solving a smaller problem of (4.7):

$$\begin{aligned} \min \quad & \sum_{t \in S'} \nabla f(\boldsymbol{\alpha}_k)_t d_t \\ & \sum_{t \in S'} y_t d_t = 0, \ \sum_{t \in S'} d_t = 0, \ -1 \leq d_t \leq 1, \\ & d_t \geq 0, \ \text{if } (\alpha_k)_t = 0, \ d_t \leq 0, \ \text{if } (\alpha_k)_t = 1/l, \\ & |\{d_t \mid d_t \neq 0, t \in S'\}| \leq q', \end{aligned} \tag{4.10}$$

where $S' = S \backslash \{i, j\}$ and $q' = q - 2$.

Algorithm 1 assigns nonzero values to at most $q/2$ pairs. The indices of nonzero elements in the solution $\mathbf{d}$ are used as B in the sub-problem (4.6). Note that Algorithm 1 can be implemented as an iterative procedure by selecting $q/2$ pairs sequentially. Then the computational complexity is similar to Joachim's strategy. Here for the convenience of writing proofs, we describe it in a recursive way. Next we prove that Algorithm 1 solves (4.5).

Lemma 6 *If there is an optimal solution $\mathbf{d}$ of (4.7), there exists an optimal integer solution $\mathbf{d}^*$ with $d_t^* \in \{-1, 0, 1\}$, for all $t \in S$.*

Proof. Because $\sum_{t \in S} d_t = 0$, if there are some non-integer elements in $\mathbf{d}$, there must be at least two. Furthermore, from the linear constraints

$$\sum_{t \in S} y_t d_t = 0 \text{ and } \sum_{t \in S} d_t = 0,$$

we have

$$\sum_{t \in S, y_t = 1} y_t d_t = 0 \text{ and } \sum_{t \in S, y_t = -1} y_t d_t = 0. \tag{4.11}$$

Thus if there are only two non-integer elements d_i and d_j, they must satisfy $y_i = y_j$.

Therefore, if $\mathbf{d}$ contains some non-integer elements , there must be two of them d_i and d_j which satisfy $y_i = y_j$. If $d_i + d_j = c$,

$$\nabla f(\boldsymbol{\alpha}_k)_i d_i + \nabla f(\boldsymbol{\alpha}_k)_j d_j = (\nabla f(\boldsymbol{\alpha}_k)_i - \nabla f(\boldsymbol{\alpha}_k)_j) d_i + c \nabla f(\boldsymbol{\alpha}_k)_j. \tag{4.12}$$

Since $d_i, d_j \notin \{-1, 0, 1\}$ and $-1 < d_i, d_j < 1$, if $\nabla f(\boldsymbol{\alpha}_k)_i \neq \nabla f(\boldsymbol{\alpha}_k)_j$, we can pick a sufficiently small $\epsilon > 0$ and shift d_i and d_j by $-\epsilon(\nabla f(\boldsymbol{\alpha}_k)_i - \nabla f(\boldsymbol{\alpha}_k)_j)$ and $\epsilon(\nabla f(\boldsymbol{\alpha}_k)_i - \nabla f(\boldsymbol{\alpha}_k)_j)$, respectively, without violating their feasibility. Then the decrease of the objective value contradicts the assumption that $\mathbf{d}$ is an optimal solution. Hence we know $\nabla f(\boldsymbol{\alpha}_k)_i = \nabla f(\boldsymbol{\alpha}_k)_j$.

Then we can eliminate at least one of the non-integers by shifting d_i and d_j by $\text{argmin}_v\{|v| : v \in \{d_i - \lfloor d_i \rfloor, \lceil d_i \rceil - d_i, d_j - \lfloor d_j \rfloor, \lceil d_j \rceil - d_j\}\}$. The objective value is the same because of (4.12) and $\nabla f(\boldsymbol{\alpha}_k)_i = \nabla f(\boldsymbol{\alpha}_k)_j$. We can repeat this process until an integer optimal solution $\mathbf{d}^*$ is obtained. $\square$

Lemma 7 *If there is an optimal integer solution* $\mathbf{d}$ *of (4.7) which is not all zero and* (i, j) *can be chosen from (4.8) or (4.9), then there is an optimal integer solution* $\mathbf{d}^*$ *with* $d_i^* = 1$ *and* $d_j^* = -1$.

Proof. As (i, j) can be chosen from (4.8) or (4.9), we know $(\alpha_k)_i < 1/l$ and $(\alpha_k)_j > 0$. We will show that if $d_i \neq 1$ and $d_j \neq -1$, we can construct an optimal integer solution $\mathbf{d}^*$ from $\mathbf{d}$ such that $d_i^* = 1$ and $d_j^* = -1$.

We first note that for any nonzero integer element $d_{i'}$, from (4.11), there is a nonzero integer element $d_{j'}$ such that

$$d_{j'} = -d_{i'} \text{ and } y_{j'} = y_{i'}.$$

We define $p(i') \equiv j'$.

If $d_i = -1$, we can find $i' = p(i)$ such that $d_{i'} = 1$ and $y_i = y_{i'}$. Since $d_{i'} = 1$, $(\alpha_k)_{i'} < 1/l$. By the definition of i and the fact that $(\alpha_k)_i < 1/l$, $\nabla f(\boldsymbol{\alpha}_k)_i \leq \nabla f(\boldsymbol{\alpha}_k)_{i'}$. Let $d_i^* = 1$, $d_{i'}^* = -1$, and $d_t^* = d_t$ otherwise. Then

$obj(\mathbf{d}^*) \leq obj(\mathbf{d})$ so $\mathbf{d}^*$ is also an optimal solution. Similarly, if $d_j = 1$, we can have an optimal solution $\mathbf{d}^*$ with $d_j^* = -1$.

Therefore, if the above transformation has been done, we have only three cases left: $(d_i, d_j) = (0, -1), (1, 0)$, and $(0, 0)$. For the first case, we can find an $i' = p(j)$ such that $d_{i'} = 1$ and $y_{i'} = y_i = y_j$. From the definition of i and the fact that $(\alpha_k)_{i'} < 1/l$ and $(\alpha_k)_i < 1/l$, $\nabla f(\boldsymbol{\alpha}_k)_i \leq \nabla f(\boldsymbol{\alpha}_k)_{i'}$. We can define $d_i^* = 1, d_{i'}^* = 0$, and $d_t^* = d_t$ otherwise. Then $obj(\mathbf{d}^*) \leq obj(\mathbf{d})$ so $\mathbf{d}^*$ is also an optimal solution. If $(d_i, d_j) = (1, 0)$, the situation is similar.

Finally we check the case where d_i and d_j are both zero. Since $\mathbf{d}$ is a nonzero integer vector, we can consider a $d_{i'} = 1$ and $j' = p(i')$. From (4.8) and (4.9), $\nabla f(\boldsymbol{\alpha}_k)_i - \nabla f(\boldsymbol{\alpha}_k)_j \leq \nabla f(\boldsymbol{\alpha}_k)_{i'} - \nabla f(\boldsymbol{\alpha}_k)_{j'}$. Let $d_i^* = 1$, $d_j^* = -1$, $d_{i'}^* = d_{j'}^* = 0$, and $d_t^* = d_t$ otherwise. Then $\mathbf{d}^*$ is feasible to (4.7) and $obj(\mathbf{d}^*) \leq obj(\mathbf{d})$. Thus $\mathbf{d}^*$ is an optimal solution. □

Lemma 8 *If there is an integer optimal solution of (4.7) and Algorithm 1 outputs a zero vector* $\mathbf{d}$*, then* $\mathbf{d}$ *is already an optimal solution of (4.7).*

Proof. If the result is wrong, there is an integer optimal solution $\mathbf{d}^*$ of (4.7) such that

$$obj(\mathbf{d}^*) = \sum_{t \in S} \nabla f(\boldsymbol{\alpha}_k)_t d_t^* < 0.$$

Without loss of generality, we can consider only the case of

$$\sum_{t \in S, y_t = 1} \nabla f(\boldsymbol{\alpha}_k)_t d_t^* < 0. \tag{4.13}$$

From (4.11) and $d_t^* \in \{-1, 0, 1\}$, the number of indices satisfying $d_t^* = 1, y_t = 1$ is the same as those of $d_t^* = -1, y_t = 1$. Therefore, we must have

$$\min_{d_t^* = 1, y_t = 1} \nabla f(\boldsymbol{\alpha}_k)_t - \max_{d_t^* = -1, y_t = 1} \nabla f(\boldsymbol{\alpha}_k)_t < 0. \tag{4.14}$$

Otherwise,

$$\sum_{d_t^* = 1, y_t = 1} \nabla f(\boldsymbol{\alpha}_k)_t - \sum_{d_t^* = -1, y_t = 1} \nabla f(\boldsymbol{\alpha}_k)_t = \sum_{y_t = 1} \nabla f(\boldsymbol{\alpha}_k)_t d_t^* \geq 0$$

contradicts (4.13).

Then (4.14) implies that in Algorithm 1, i and j can be chosen with $d_i = 1$ and $d_j = -1$. This contradicts the assumption that Algorithm 1 outputs a zero vector. $\square$

Theorem 7 *Algorithm 1 solves (4.7).*

Proof. First we note that the set of $\mathbf{d}$ which satisfies $|\{d_t \mid d_t \neq 0, t \in S\}| \leq q$ can be considered as the union of finitely many closed sets of the form $\{\mathbf{d} \mid d_{i_1} = 0, \ldots, d_{i_{l-q}} = 0\}$. Therefore, the feasible region of (4.7) is closed. With the bounded constraints $-1 \leq d_i \leq 1, i = 1, \ldots, l$, the feasible region is compact so there is at least one optimal solution.

As q is an even integer, we assume $q = 2k$. We then finish the proof by induction on k:

$k = 0$: Algorithm 1 correctly finds the solution zero.

$k > 0$: Suppose Algorithm 1 outputs a vector $\mathbf{d}$ with $d_i = 1$ and $d_j = -1$. In this situation the optimal solution of (4.7) cannot be zero. Otherwise, by assigning a vector $\bar{\mathbf{d}}$ with $\bar{d}_i = 1, \bar{d}_j = -1$, and $\bar{d}_t = 0$ for all $t \in S \backslash \{i, j\}$, $obj(\bar{\mathbf{d}}) < 0$ gives a smaller objective value than that of the zero vector. Thus the assumptions of Lemma 7 hold. Then by the fact that (4.7) is solvable and Lemmas 6 and 7, we know that there is an optimal solution $\mathbf{d}^*$ of (4.5) with $d_i^* = 1$ and $d_j^* = -1$.

By induction $\{d_t, t \in S'\}$ is an optimal solution of (4.10). Since $\{d_t^*, t \in S'\}$ is also feasible to (4.10), we have

$$\begin{aligned} obj(\mathbf{d}) &= \nabla f(\boldsymbol{\alpha}_k)_i d_i + \nabla f(\boldsymbol{\alpha}_k)_j d_j + \sum_{t \in S'} \nabla f(\boldsymbol{\alpha}_k)_t d_t \\ &\leq \nabla f(\boldsymbol{\alpha}_k)_i d_i^* + \nabla f(\boldsymbol{\alpha}_k)_j d_j^* + \sum_{t \in S'} \nabla f(\boldsymbol{\alpha}_k)_t d_t^* = obj(\mathbf{d}^*). \end{aligned} \tag{4.15}$$

Thus $\mathbf{d}$, the output of Algorithm 1, is an optimal solution.

Suppose Algorithm 1 does not output a vector $\mathbf{d}$ with $d_i = 1$ and $d_j = -1$. Then $\mathbf{d}$ is actually a zero vector. Immediately from Lemma 8, $\mathbf{d} = 0$ is an optimal solution. $\square$

Since (4.5) is a special case of (4.7), Theorem 7 implies that Algorithm 1 can solve it.

After solving (D_ν), we want to calculate ρ and b in (P_ν). The KKT condition (2.5) shows

$$
\begin{aligned}
(\mathbf{Q}\boldsymbol{\alpha})_i - \rho + b y_i &= 0 \text{ if } 0 < \alpha_i < 1/l, \\
&\geq 0 \text{ if } \alpha_i = 0, \\
&\leq 0 \text{ if } \alpha_i = 1/l.
\end{aligned}
$$

Define

$$r_1 \equiv \rho - b,\ r_2 \equiv \rho + b.$$

If $y_i = 1$ the KKT condition becomes

$$
\begin{aligned}
(\mathbf{Q}\boldsymbol{\alpha})_i - r_1 &= 0 \text{ if } 0 < \alpha_i < 1/l, \\
&\geq 0 \text{ if } \alpha_i = 0, \\
&\leq 0 \text{ if } \alpha_i = 1/l.
\end{aligned}
\tag{4.16}
$$

Therefore, if there are α_i which satisfy (4.16), $r_1 = (\mathbf{Q}\boldsymbol{\alpha})_i$. Practically to avoid numerical errors, we can average them:

$$r_1 = \frac{\sum_{0<\alpha_i<1/l, y_i=1} (\mathbf{Q}\boldsymbol{\alpha})_i}{\sum_{0<\alpha_i<1/l, y_i=1} 1}.$$

On the other hand, if there is no such α_i, as r_1 must satisfy

$$\max_{\alpha_i=1/l, y_i=1} (\mathbf{Q}\boldsymbol{\alpha})_i \leq r_1 \leq \min_{\alpha_i=0, y_i=1} (\mathbf{Q}\boldsymbol{\alpha})_i,$$

we take r_1 the midpoint of the range.

For $y_i = -1$, we can calculate r_2 in a similar way.

After r_1 and r_2 are obtained,

$$\rho = \frac{r_1 + r_2}{2} \text{ and } -b = \frac{r_1 - r_2}{2}.$$

Note that the KKT condition can be written as

$$\max_{\alpha_i>0, y_i=1} (\mathbf{Q}\boldsymbol{\alpha})_i \leq \min_{\alpha_i<1/l, y_i=1} (\mathbf{Q}\boldsymbol{\alpha})_i \text{ and } \max_{\alpha_i>0, y_i=-1} (\mathbf{Q}\boldsymbol{\alpha})_i \leq \min_{\alpha_i<1/l, y_i=-1} (\mathbf{Q}\boldsymbol{\alpha})_i.$$

Hence practically we can use the following stopping criterion: The decomposition method stops if the iterate $\boldsymbol{\alpha}$ satisfies the following condition:

$$-(\mathbf{Q}\boldsymbol{\alpha})_i + (\mathbf{Q}\boldsymbol{\alpha})_j < \epsilon, \tag{4.17}$$

where $\epsilon > 0$ is a chosen stopping tolerance, and i and j are the *first* pair obtained from (4.8) or (4.9).

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

5 Numerical Experiments

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 2002)). 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 required time and iterations.

Table 5.1: Solving C-SVM and ν-SVM: $C = 1$ (time in seconds)

Problem	l	ν	C Iter.	ν Iter.	C Time	ν Time	#SV	#FSV	$\lceil \nu l \rceil$
australian	690	0.309619	1040	946	0.34	0.42	244	55	214
diabetes	768	0.574087	395	297	0.4	0.47	447	13	441
german	1000	0.556643	953	909	1.23	1.61	600	88	557
heart	270	0.43103	219	175	0.07	0.08	132	25	117
vehicle	846	0.501182	791	904	0.69	0.91	439	26	424
satimage	4435	0.083544	355	534	8.16	14.05	377	12	371
letter	15000	0.036588	764	897	22.59	35.13	563	26	549
shuttle	43500	0.141534	3267	6982	422.04	1058.0	6159	5	6157
a4a	4781	0.41394	1460	1464	21.14	28.86	2002	53	1980
w7a	24692	0.059718	1896	1721	74.51	102.99	1556	140	1475

Table 5.2: Solving C-SVM and ν-SVM: $C = 1000$ (time in seconds)

Problem	l	ν	C Iter.	ν Iter.	C Time	ν Time	#SV	#FSV	$\lceil \nu l \rceil$
australian	690	0.147234	151438	117758	10.98	8.65	222	167	102
diabetes	768	0.421373	216845	137941	18.96	11.79	376	102	324
german	1000	0.069128	79542	81824	11.24	11.37	509	494	70
heart	270	0.033028	11933	11075	0.38	0.35	100	99	9
vehicle	846	0.262569	220973	190324	20.07	17.01	284	111	223
satimage	4435	0.015416	44372	45323	28.3	28.31	136	106	69
letter	15000	0.005789	69052	70604	141.4	134.14	152	100	87
shuttle	43500	0.033965	143273	154558	1215.8	1468.56	1487	17	1478
a4a	4781	0.263506	359618	350818	257.51	244.84	1760	837	1260
w7a	24692	0.023691	187578	187170	1262.15	1112.07	1112	696	585

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

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

The working set selection follows the discussion in Section 4 and here we implement a special case with $q = 2$. Then the working set in each iteration contains only two elements.

For the initial point $\boldsymbol{\alpha}_1$, we assign the first $\lceil \nu l/2 \rceil$ elements with $y_i = 1$ as $[1, \ldots, 1, \nu l/2 - \lfloor \nu l/2 \rfloor]^T$. Similarly, the same numbers are assigned to the first $\lceil \nu l/2 \rceil$ elements with $y_i = -1$. Unlike the decomposition method for (D_C), where the zero vector is usually used as the initial solution so $\nabla f(\boldsymbol{\alpha}_1) = -\mathbf{e}$, now $\boldsymbol{\alpha}_1$ contains $\lceil \nu l \rceil$ nonzero components. In order to obtain $\nabla f(\boldsymbol{\alpha}_1) = \mathbf{Q}\boldsymbol{\alpha}_1$ of (4.5), in the beginning of the decomposition procedure, we must compute $\lceil \nu l \rceil$ columns of $\mathbf{Q}$. This might be a disadvantage of using ν-SVM. Further investigations are needed on this issue.

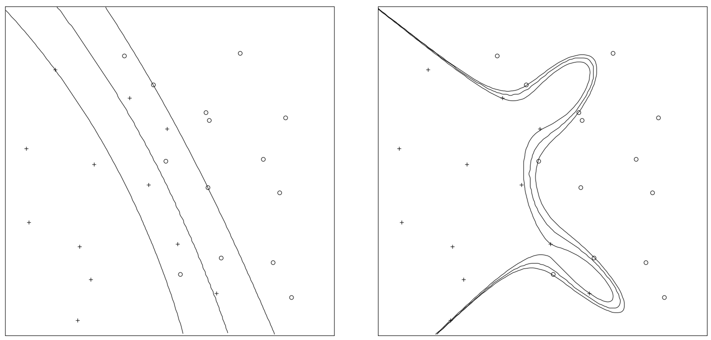
We test the RBF kernel with $Q_{ij} = y_i y_j e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/n}$, where n is the number of attributes of a training data. Our implementation is part of the software LIBSVM* (version 2.03) which is an integrated package for SVM classification and regression.

We test problems from various collections. 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.

As LIBSVM also implements a decomposition method with $q = 2$ for C-SVM (Chang and Lin 2000), we try to conduct some comparisons between C-SVM and ν-SVM. Note that these two codes are nearly the same except different working selections specially for (D_ν) and (D_C). For each problem, we solve its (D_C) form using $C = 1$ and $C = 1000$ first. If $\boldsymbol{\alpha}_C$ is an optimal solution of (D_C), we then

*LIBSVM is available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm` calculate ν by $\mathbf{e}^T \boldsymbol{\alpha}_C/(Cl)$ and solve (D_ν). The stopping tolerance ϵ for solving C-SVM is set to be 10^{-3}. As the $\boldsymbol{\alpha}$ of (4.17) is like the $\boldsymbol{\alpha}$ of (D_C) divided by C and the stopping criterion involves $\mathbf{Q}\boldsymbol{\alpha}$, to have a fair comparison, the tolerance (i.e. ϵ of (4.17)) for (5.1) is set as $10^{-3}/C$.

The computational experiments for this section were done on a Pentium III-500 with 256MB RAM using the gcc compiler. We used 100MB as the cache size of LIBSVM for storing recently used Q_{ij}.



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

Figure 2: Training data and separating hyperplanes

Tables 5.1 and 5.2 report results of $C = 1$ and 1000, respectively. In each table, the corresponding ν is listed and the number of iterations and time (in seconds) of both algorithms are compared. Note that for the same problem, fewer iterations do not always lead to less computational time. We think there are two possible reasons: First the computational time for calculating the initial gradient for (D_ν) is more expensive. Second, due to different contents of the cache (or say different numbers of kernel evaluations), the cost of each iteration is different. We also present the number of support vectors (#SV column) as well as free support vectors (#FSV column). It can be clearly seen that the proposed method for (D_ν) performs very well. This comparison has shown the practical viability of using ν-SVM.

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). 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 2(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 smaller. Therefore, here we have given an example that for solving (D_C) and (D_ν), the decomposition method faces a similar difficulty.

6 Discussions and Conclusions

In an earlier version of this paper since we did not know how to design a decomposition method for (D_ν) which has two linear constraints, we tried to remove one of them. For C-SVM, (Mangasarian and Musicant 1999) and (Friess et al. 1998) added $b^2/2$ into the objective function so the dual does not have the linear constraint $\mathbf{y}^T\boldsymbol{\alpha} = 0$. We exploited a similar approach for (P_ν) by considering the following new primal problem:

$$(\bar{P}_\nu) \qquad \begin{aligned} &\min \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{1}{2}b^2 - \nu\rho + \frac{1}{l}\sum_{i=1}^{l}\xi_i \\ &y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq \rho - \xi_i, \\ &\xi_i \geq 0, i = 1, \ldots, l, \rho \geq 0. \end{aligned} \qquad (6.1)$$

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

$$(\bar{D}_\nu) \qquad \begin{aligned} &\min \frac{1}{2}\boldsymbol{\alpha}^T(\mathbf{Q} + \mathbf{y}\mathbf{y}^T)\boldsymbol{\alpha} \\ &\mathbf{e}^T\boldsymbol{\alpha} \geq \nu, \\ &0 \leq \alpha_i \leq 1/l, \qquad i = 1, \ldots, l. \end{aligned} \qquad (6.2)$$

Similar to Theorem 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 and can be solved using existing decomposition methods like SVM^{light}.

Table 6.3: Solving $(\bar{D}_\nu)$: comparing with Table 5.1

Problem	l	ν	ν Iter.	ν Time	#SV	#FSV
australian	690	0.309619	4871	0.64	244	53
diabetes	768	0.574087	1816	0.58	447	13
german	1000	0.556643	1641	1.67	599	87
heart	270	0.43103	527	0.1	130	23
vehicle	846	0.501182	1402	1.04	437	26
satimage	4435	0.083544	3034	15.44	380	16
letter	15000	0.036588	7200	54.6	562	28
shuttle	43500	0.141534	17893	1198.83	6161	8
a4a	4781	0.41394	7500	35.03	2002	54
w7a	24692	0.059718	3109	107.5	1563	149

Table 6.4: Solving $(\bar{D}_\nu)$: comparing with Table 5.2

Problem	l	ν	ν Iter.	ν Time	#SV	#FSV
australian	690	0.147234	597205	36.06	222	167
diabetes	768	0.421373	1811571	132.7	376	102
german	1000	0.069128	504114	56.33	508	493
heart	270	0.033028	48581	1.13	100	99
vehicle	846	0.262569	1626315	125.51	284	112
satimage	4435	0.015416	919695	445.42	136	106
letter	15000	0.005789	1484401	2544.23	150	97
shuttle	43500	0.033965	8364010	59286.83	1487	18
a4a	4781	0.263506	8155518	4905.67	1759	842
w7a	24692	0.023691	28791608	96912.82	1245	830

To be more precise, the working selection becomes:

$$
\begin{aligned}
\min \quad & \nabla f(\boldsymbol{\alpha}_k)^T \mathbf{d} \\
& \mathbf{e}^T\mathbf{d} = 0, \ -1 \le d_i \le 1, \\
& d_i \ge 0, \ \text{if } (\alpha_k)_i = 0, \ d_i \le 0, \ \text{if } (\alpha_k)_i = 1/l, \\
& |\{d_i \mid d_i \ne 0\}| \le q,
\end{aligned}
\tag{6.3}
$$

where $f(\boldsymbol{\alpha})$ is $\frac{1}{2}\boldsymbol{\alpha}^T(\mathbf{Q} + \mathbf{y}\mathbf{y}^T)\boldsymbol{\alpha}$.

(6.3) can be considered as a special problem of (4.2) since $\mathbf{e}$ of $\mathbf{e}^T\mathbf{d} = 0$ is a special case of $\mathbf{y}$. Thus SVM^{light}'s selection procedure can be directly used. An earlier version of LIBSVM implemented this decomposition method for $(\bar{D}_\nu)$. However, later we find that the performance is much worse than that of the method for (D_ν). This can be seen in Tables 6.3 and 6.4 which present the same information as Tables 5.1 and 5.2 for solving $(\bar{D}_\nu)$. As the major difference is on the working set selection, we suspect that the performance gap is similar to the situation happened for C-SVM. In (Hsu and Lin 2002), the authors shown that by directly using SVM^{light}'s strategy, the decomposition method for

$$(\bar{D}_C) \qquad \min \frac{1}{2}\boldsymbol{\alpha}^T(\mathbf{Q}+\mathbf{y}\mathbf{y}^T)\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha}$$
$$0 \le \alpha_i \le C, \qquad i = 1, \ldots, l. \tag{6.4}$$

performs much worse than that for (D_C). Note that the relation between $(\bar{D}_C)$ and $(\bar{D}_\nu)$ is very similar to that of (D_C) and (D_ν) presented earlier. Thus we conjecture that there are some common shortages of using SVM^{light}'s working set selection for $(\bar{D}_C)$ and $(\bar{D}_\nu)$. Further investigations are needed to understand whether explanations in (Hsu and Lin 2002) are true for $(\bar{D}_\nu)$.

In conclusion, this paper discusses the relation between ν-SVM and C-SVM in detail. In particular, we show that solving them is just like solving two different problems with the same optimal solution set. We also have proposed a decomposition method for ν-SVM. Experiments on this method show that it is competitive with methods for C-SVM. Hence we have demonstrated the practical viability of ν-SVM.

Acknowledgments

This work was supported in part by the National Science Council of Taiwan via the grant NSC 89-2213-E-002-013. The second author thanks Craig Saunders for bringing him to the attention of ν-SVM. He also thanks a referee of (Lin 2001a) whose comments lead him to think about the infeasibility of (D_ν). The authors also thank Dr. Bernhard Schölkopf and two anonymous referees 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 Transactions on Neural Networks 11*(4), 1003–1008.

Chang, C.-C. and C.-J. Lin (2000). LIBSVM: Introduction and benchmarks. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.

Crisp, D. J. and C. J. C. Burges (2000). A geometric interpretation of ν-SVM classifiers. In S. Solla, T. Leen, and K.-R. Müller (Eds.), *Advances in Neural Information Processing Systems*, Volume 12, Cambridge, MA. MIT 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 *Proceedings of 15th Intl. Conf. Machine Learning*. Morgan Kaufman Publishers.

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

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

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

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

Lin, C.-J. (2001a). Formulations of support vector machines: a note from an optimization point of view. *Neural Computation 13*(2), 307–317.

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

Mangasarian, O. L. and D. R. Musicant (1999). Successive overrelaxation for support vector machines. *IEEE Transactions on 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, C. C. Taylor, and J. Campbell (Eds.) (1994). *Machine learning, neural and statistical classification*. Upper Saddle River, NJ, USA: Ellis Horwood. Data available at `http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/`.

Murphy, P. M. and D. W. Aha (1994). UCI repository of machine learning databases. Technical report, University of California, Department of Information and Computer Science, Irvine, CA. Data 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*, New York, NY, pp. 130–136. IEEE.

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

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., A. Smola, R. C. Williamson, and P. L. Bartlett (2000). New support vector algorithms. *Neural Computation 12*, 1207–1245.

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: Wiley.