The 2ν-SVM: A Cost-Sensitive Extension of the ν-SVM

Mark A. Davenport*
Rice University
Department of Electrical and Computer Engineering
Technical Report TREE 0504

Updated December 2005

Abstract

Standard classification algorithms aim to minimize the probability of making an incorrect classification. In many important applications, however, some kinds of errors are more important than others. In this report we review cost-sensitive extensions of standard support vector machines (SVMs). In particular, we describe cost-sensitive extensions of the C-SVM and the ν-SVM, which we denote the 2C-SVM and 2ν-SVM respectively. The C-SVM and the ν-SVM are known to be closely related, and we prove that the 2C-SVM and 2ν-SVM share a similar relationship. This demonstrates that the 2C-SVM and 2ν-SVM explore the same space of possible classifiers, and gives us a clear understanding of the parameter space for both versions.

1 Introduction

In a standard classification problem the goal is to minimize the probability of making an error. In many important applications, however, some kinds of errors are more important than others. In tumor classification, for example, the impact of mistakenly classifying a benign tumor as malignant is much less than that of the opposite mistake. However, nearly all work on classification to date optimizes a "probability of error" criterion. An exception is a recent body of work known as "cost-sensitive classification" that assigns costs to different errors and attempts to minimize the expected misclassification cost.

Support vector machines (SVMs) can be extended to the cost-sensitive setting by introducing an additional parameter that penalizes the errors asymmetrically. This approach

*Supported by NSF, AFOSR, ONR, and the Texas Instruments Leadership University Program.
Email: md@rice.edu
Web: dsp.rice.edu

has been taken by several authors to adapt the C-SVM to be cost-sensitive [1–4], but this strategy also applies to an alternative SVM formulation, the ν-SVM [5]. We refer to the cost-sensitive extensions as the $2C$-SVM and 2ν-SVM respectively. The primary motivation for these methods is to address the problem described above, but they can also be applied to deal with the difficulties that arise when the class frequencies in the training data do not accurately reflect the true prior probabilities of the classes. Additionally, cost-sensitive classifiers are useful in the Neyman-Pearson classification context [6]. In all of these settings, a critical problem is that of parameter selection: the parameter settings that would result in the "best" performance are not known, and so the user must use the training data to estimate appropriate values for the parameters. Thus, it is vital that we understand how varying the parameters of either the $2C$-SVM or the 2ν-SVM will impact the resulting classifier.

In Section 2 we briefly review SVMs. In Section 3 we then introduce the cost-sensitive extensions of the C-SVM and ν-SVM. The ν-SVM has some properties that make it more attractive than the C-SVM. This is also the case for the 2ν-SVM. We describe these properties in Section 4. Among the contributions of this paper is a proof that the 2ν-SVM is feasible if and only if the parameters lie in a specified range. In Section 5 we show the $2C$-SVM and the 2ν-SVM are closely related. Specifically, we generalize a result of [7] and show that under certain technical conditions, any optimal solution for one of the cost-sensitive SVM formulations is an optimal solution of the other with the right parameter settings. Using these results, we then prove a theorem that precisely relates the parameter spaces and resulting classifiers of the $2C$-SVM and the 2ν-SVM.

2 Review of Support Vector Machines

Support vector machines (SVMs) are among the more effective methods for classification. For a more thorough review see [8–10]. In the following, assume that we have access to training data $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$ where $\mathbf{x}_i \in \mathbb{R}^d$ is a d-dimensional feature vector and $y_i \in \{+1, -1\}$ indicates the class of $\mathbf{x}_i$.

Conceptually, the support vector classifier is constructed in a two step process. In the first step, the $\mathbf{x}_i$ are transformed via a mapping $\Phi : \mathbb{R}^d \to \mathcal{H}$ where $\mathcal{H}$ is a high (possibly infinite) dimensional Hilbert space. The intuition is that the two classes are more easily separated in $\mathcal{H}$ than in $\mathbb{R}^d$. For algorithmic reasons, Φ must be chosen so that the *kernel* operator $k(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle_{\mathcal{H}}$ is positive definite. This allows us to compute inner products in $\mathcal{H}$ without explicitly evaluating Φ.

In the second step, a hyperplane is determined in the induced feature space according to the max-margin principle. In the case where the two classes can be separated by a hyperplane, the SVM finds the hyperplane that maximizes the distance between the decision boundary and the closest point to the boundary, known as the *margin.* When the classes cannot be separated by a hyperplane, the constraints are relaxed through the introduction of slack variables ξ_i. If $\xi_i > 0$, this means that the corresponding $\mathbf{x}_i$ lies inside the margin

and is called a *margin error.* If $\mathbf{w} \in \mathcal{H}$ and $b \in \mathbb{R}$ are the normal vector and affine shift defining the max-margin hyperplane, then the support vector classifier is given by $f_{\mathbf{w},b}(\mathbf{x}) = \operatorname{sgn}(\langle \mathbf{w}, \Phi(\mathbf{x}) \rangle_{\mathcal{H}} + b)$. The offset parameter b is often called the *bias.*

There are two different formulations of the SVM. The original SVM [11], which we shall call the C-SVM, can be formulated as the following quadratic program:

$$(P_C) \qquad \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i$$

$$\begin{aligned} \text{subject to} \qquad & y_i(k(\mathbf{w}, \mathbf{x}_i) + b) \geq 1 - \xi_i && \text{for } i = 1, \ldots, n \\ & \xi_i \geq 0 && \text{for } i = 1, \ldots, n \end{aligned}$$

where $C \geq 0$ is a parameter that controls the tradeoff between minimizing the margin errors and maximizing the margin.

For computational reasons, it is often easier to solve (P_C) by solving the equivalent dual problem:

$$(D_C) \qquad \min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^{n} \alpha_i$$

$$\begin{aligned} \text{subject to} \qquad & 0 \leq \alpha_i \leq C && \text{for } i = 1, \ldots, n \\ & \sum_{i=1}^{n} \alpha_i y_i = 0. \end{aligned}$$

This formulation is derived by forming the Lagrangian ($\boldsymbol{\alpha}$ is a Lagrange multiplier). The primal and the dual are related through $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \Phi(\mathbf{x}_i)$. We will often have that $\alpha_i = 0$ for most $\mathbf{x}_i$. We call the $\mathbf{x}_i$ for which $\alpha_i \neq 0$ the *support vectors.*

An alternative (but equivalent) formulation of the C-SVM is the ν-SVM [12], which replaces C with a different parameter $\nu \in [0, 1]$ that serves as an upper bound on the fraction of margin errors and a lower bound on the fraction of support vectors. The ν-SVM has the primal formulation

$$(P_\nu) \qquad \min_{\mathbf{w},b,\boldsymbol{\xi},\rho} \quad \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{n}\sum_{i=1}^{n} \xi_i$$

$$\begin{aligned} \text{subject to} \qquad & y_i(k(\mathbf{w}, \mathbf{x}_i) + b) \geq \rho - \xi_i && \text{for } i = 1, \ldots, n \\ & \xi_i \geq 0 && \text{for } i = 1, \ldots, n \\ & \rho \geq 0 \end{aligned}$$

and dual formulation

$$(D_\nu) \qquad \min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{subject to} \qquad 0 \le \alpha_i \le \frac{1}{n} \qquad \text{for } i = 1, \ldots, n$$
$$\sum_{i=1}^{n} \alpha_i y_i = 0, \quad \sum_{i=1}^{n} \alpha_i \ge \nu.$$

3 Cost-Sensitive SVMs

The above formulations implicitly penalize errors in both classes equally. However, as described in the Introduction, there may be different costs associated with the two different kinds of errors. To address this issue, cost-sensitive extensions of both the C-SVM and the ν-SVM have been proposed, which we shall denote the $2C$-SVM and the 2ν-SVM respectively.

First we will consider the $2C$-SVM proposed in [1]. Let $I_+ = \{i : y_i = +1\}$ and $I_- = \{i : y_i = -1\}$. The $2C$-SVM has primal

$$(P_{2C}) \qquad \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\gamma \sum_{i \in I_+} \xi_i + C(1-\gamma) \sum_{i \in I_-} \xi_i$$

$$\begin{aligned} \text{subject to} \qquad y_i(k(\mathbf{w}, \mathbf{x}_i) + b) &\ge 1 - \xi_i \qquad \text{for } i = 1, \ldots, n \\ \xi_i &\ge 0 \qquad \text{for } i = 1, \ldots, n \\ \rho &\ge 0 \end{aligned}$$

and dual

$$(D_{2C}) \qquad \min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^{n} \alpha_i$$

$$\begin{aligned} \text{subject to} \qquad 0 \le \alpha_i &\le C\gamma \qquad \text{for } i \in I_+ \\ 0 \le \alpha_i &\le C(1-\gamma) \qquad \text{for } i \in I_- \\ \sum_{i=1}^{n} \alpha_i y_i &= 0 \end{aligned}$$

where $C > 0$ and $\gamma \in [0, 1]$. Similarly, [5] proposed the 2ν-SVM as a cost-sensitive extension of the ν-SVM. The 2ν-SVM has primal

$$(P_{2\nu}) \qquad \min_{\mathbf{w},b,\boldsymbol{\xi},\rho} \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{\gamma}{n}\sum_{i\in I_+}\xi_i + \frac{1-\gamma}{n}\sum_{i\in I_-}\xi_i$$

$$\begin{aligned} \text{subject to} \qquad y_i(k(\mathbf{w},\mathbf{x}_i)+b) &\geq \rho - \xi_i \qquad \text{for } i = 1,\ldots,n \\ \xi_i &\geq 0 \qquad \text{for } i = 1,\ldots,n \\ \rho &\geq 0 \end{aligned}$$

and dual

$$(D_{2\nu}) \qquad \min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j)$$

$$\begin{aligned} \text{subject to} \qquad 0 \leq \alpha_i &\leq \frac{\gamma}{n} \qquad \text{for } i \in I_+ \\ 0 \leq \alpha_i &\leq \frac{1-\gamma}{n} \qquad \text{for } i \in I_- \\ \sum_{i=1}^{n}\alpha_i y_i = 0, \quad \sum_{i=1}^{n}\alpha_i &\geq \nu \end{aligned}$$

where $\nu \in [0, \frac{1}{2}]$ and $\gamma \in [0, 1]$.

4 Properties of the 2ν-SVM

Before illustrating the relationship between the $2C$-SVM and the 2ν-SVM, we establish some of the basic properties of the 2ν-SVM.

Proposition 1. *Fix $\gamma \in [0, 1]$ and let $n_+ = |I_+|$, $n_- = |I_-|$. Then $(D_{2\nu})$ is feasible if and only if $\nu \leq \nu_{\max} \leq \frac{1}{2}$, where*

$$\nu_{\max} = \frac{2\min(\gamma n_+, (1-\gamma)n_-)}{n}.$$

Proof. First, assume that $\nu \leq \nu_{\max}$. Then we can construct an $\boldsymbol{\alpha}$ that satisfies the constraints of $(D_{2\nu})$. Specifically, let

$$\alpha_i = \frac{\nu_{\max}}{2n_+} = \frac{\min(\gamma, (1-\gamma)n_-/n_+)}{n} \leq \frac{\gamma}{n} \; for \; i \; \in I_+$$

and

$$\alpha_i = \frac{\nu_{\max}}{2n_-} = \frac{\min(\gamma n_+/n_-, 1-\gamma)}{n} \leq \frac{1-\gamma}{n} \ for\ i\ \in I_-.$$

Then $\sum_{i\in I_+} \alpha_i + \sum_{i\in I_+} \alpha_i = \nu_{\max} \geq \nu$ and $\sum_{i=1}^n \alpha_i y_i = 0$. Thus we have a feasible solution, and so $(D_{2\nu})$ is feasible.

Now assume that $(D_{2\nu})$ is feasible. Then there exists an $\boldsymbol{\alpha}$ such that $\sum_{i=1}^n \alpha_i \geq \nu$ and $\sum_{i\in I_+} \alpha_i = \sum_{i\in I_-} \alpha_i$. Combining this we get $\nu \leq 2\sum_{i\in I_+} \alpha_i$. Since we also have $0 \leq \alpha_i \leq \gamma/n$ for $i \in I_+$, we see that $\nu \leq 2\sum_{i\in I_+} \alpha_i \leq 2\gamma n_+/n$, and therefore, $\nu \leq 2\gamma n_+/n$. Similarly, $\nu \leq 2(1-\gamma)n_-/n$. Thus $\nu \leq \nu_{\max}$.

Finally, note that $\nu_{\max}$ is maximized when $\gamma n_+ = (1-\gamma)n_-$, which occurs when $\gamma = n_-/n$, and thus

$$\nu_{\max} \leq \frac{2n_-n_+}{n^2} \leq \frac{1}{2}$$

□

Remark. *We can use this result to show that $(D_{2\nu})$ is feasible for fixed $\nu \in [0, \nu_{\max}]$ if and only if*

$$\frac{\nu n}{2n_+} \leq \gamma \leq 1 - \frac{\nu n}{2n_-}$$

Proposition 2. *Fix $\gamma \in [0,1]$ and $\nu \in [0, \nu_{\max}]$. There is at least one optimal solution of $(D_{2\nu})$ that satisfies $\sum_{i=1}^n \alpha_i = \nu$. In addition, if the optimal objective value of $(D_{2\nu})$ is not zero, all optimal solutions of $(D_{2\nu})$ satisfy $\sum_{i=1}^n \alpha_i = \nu$.*

Proof. This proposition was proved in [13] for (D_ν). The proof relies only on the form of the objective function of (D_ν), which is identical to that of $(D_{2\nu})$. Thus, we omit it for the sake of brevity and refer the reader to [13]. □

Remark. *The cost-sensitive extension of the 2ν-SVM proposed in [5] is parameterized in a different manner than $(D_{2\nu})$. Specifically, instead of parameters ν and γ, $(D_{2\nu})$ is formulated using ν_+ and ν_-, where*

$$\nu = \frac{2\nu_+\nu_-n_+n_-}{(\nu_+n_+ + \nu_-n_-)n}, \quad \gamma = \frac{\nu_-n_-}{\nu_+n_+ + \nu_-n_-} = \frac{\nu n}{2\nu_+n_+}.$$

or equivalently

$$\nu_+ = \frac{\nu n}{2\gamma n_+}, \quad \nu_- = \frac{\nu n}{2(1-\gamma)n_-}.$$

This parametrization has the benefit that ν_+ and ν_- have a more intuitive meaning illustrated by the following result.

Proposition 3. *Suppose that the optimal objective value of $(D_{2\nu})$ is not zero. Then for the optimal solution of $(D_{2\nu})$:*

1. *ν_+ is an upper bound on the fraction of margin errors from class $+1$.*

2. *ν_- is an upper bound on the fraction of margin errors from class -1.*

3. *ν_+ is a lower bound on the fraction of support vectors from class $+1$.*

4. *ν_- is a lower bound on the fraction of support vectors from class -1.*

Proof. See [5] for the proof. □

Proposition 4. *$(D_{2\nu})$ is feasible if and only if $\nu_+ \leq 1$ and $\nu_- \leq 1$.*

Proof. From Proposition 1 we have that $(D_{2\nu})$ is feasible if and only if

$$\nu \leq \frac{2\min(\gamma n_+, (1-\gamma)n_-)}{n}.$$

Thus, $(D_{2\nu})$ is feasible if and only if

$$\frac{2\nu_+\nu_-n_+n_-}{(\nu_+n_+ + \nu_-n_-)n} \leq \frac{2\min\left(\frac{\nu_-n_+n_-}{\nu_+n_+ + \nu_-n_-}, \frac{\nu_+n_+n_-}{\nu_+n_+ + \nu_-n_-}\right)}{n} \quad \Longleftrightarrow \quad \nu_+\nu_- \leq \min(\nu_-, \nu_+)$$

or equivalently, $\nu_+ \leq 1$ and $\nu_- \leq 1$. □

5 Relationship between the 2ν-SVM and $2C$-SVM

The following theorems illustrate the relationship between (D_{2C}) and $(D_{2\nu})$. The first shows how solutions of (D_{2C}) are related to solutions of $(D_{2\nu})$, and the second shows how solutions of $(D_{2\nu})$ are related to solutions of (D_{2C}). The third theorem, our main result, shows that increasing (decreasing) ν is similar to decreasing (increasing) C. The proofs of these theorems can be found in Section 7.

Theorem 1. *Fix $\gamma \in [0,1]$. For each $C > 0$, let $\boldsymbol{\alpha}^*$ be any optimal solution of (D_{2C}) and set $\nu = \sum_{i=1}^n \alpha_i^*/(Cn)$. Then any $\boldsymbol{\alpha}$ is an optimal solution of (D_{2C}) if and only if $\boldsymbol{\alpha}/(Cn)$ is an optimal solution of $(D_{2\nu})$.*

Theorem 2. *Fix $\gamma \in [0,1]$. Assume $(D_{2\nu})$, $0 < \nu \leq \nu_{\max}$, has a nonzero optimal objective value, then $\rho > 0$. Set $C = 1/(\rho n)$. Then any $\boldsymbol{\alpha}$ is an optimal solution of (D_{2C}) if and only if $\boldsymbol{\alpha}/(Cn)$ is an optimal solution of $(D_{2\nu})$.*

Theorem 3. *Fix $\gamma \in [0,1]$ and let $\boldsymbol{\alpha}^*$ be any optimal solution of (D_{2C}). Define*

$$\nu_* = \lim_{C\to\infty} \frac{\sum_{i=1}^n \alpha_i^*}{Cn}$$

and

$$\nu^* = \lim_{C\to 0} \frac{\sum_{i=1}^n \alpha_i^*}{Cn}.$$

Then $0 \leq \nu_* \leq \nu^* = \nu_{\max} \leq \frac{1}{2}$. *Thus, for any* $\nu > \nu^*$, $(D_{2\nu})$ *is infeasible. For any* $\nu \in (\nu_*, \nu^*]$ *the optimal objective value of* $(D_{2\nu})$ *is strictly positive, thus there exists at least one* $C > 0$ *such that any* $\boldsymbol{\alpha}$ *is an optimal solution of* (D_{2C}) *if and only if* $\boldsymbol{\alpha}/(Cn)$ *is an optimal solution of* $(D_{2\nu})$. *For any* $\nu \in [0, \nu_*]$, $(D_{2\nu})$ *is feasible with zero optimal objective value (and a trivial solution).*

Remark. *Consider the case where the data can be perfectly separated by a hyperplane. In this case, if* $C \to \infty$, *margin errors are penalized more heavily, and thus for some sufficiently large* C, *the solution of* (D_{2C}) *will be the* $\boldsymbol{\alpha}^*$ *corresponding to the separating hyperplane. Thus there exists some* C^* *such that* $\boldsymbol{\alpha}^*$ *(corresponding to the separating hyperplane) is an optimal solution of* (D_{2C}) *for all* $C \geq C^*$. *In this case, as* $C \to \infty$, $\sum_{i=1}^{n} \alpha_i^*/Cn \to 0$, *and thus* $\nu_* = 0$.

Remark. *Using the definitions of* ν_+ *and* ν_- *in Section 4, it is easy to see that Theorem 3 implies that if* γ *is fixed and we let* $C \to \infty$, (D_{2C}) *is equivalent to* $(D_{2\nu})$ *(in the sense described above) if we let*

$$\nu_+ \to \frac{\nu_* n}{2\gamma n_+} \geq 0, \quad \nu_- \to \frac{\nu_* n}{2(1-\gamma) n_-} \geq 0.$$

Similarly, if γ *is fixed and we let* $C \to 0$, (D_{2C}) *is equivalent to* $(D_{2\nu})$ *if we let*

$$\nu_+ \to \frac{\nu_{\max} n}{2\gamma n_+} = \min\left(1, \frac{(1-\gamma) n_-}{\gamma n_+}\right), \quad \nu_- \to \frac{\nu_{\max} n}{2(1-\gamma) n_-} = \min\left(1, \frac{\gamma n_+}{(1-\gamma) n_-}\right).$$

6 Conclusion

In this paper we have reviewed extensions of the two main SVM formulations. These extensions address the practical need to penalize errors from the two classes differently in many classification tasks. The $2C$-SVM is commonly used to address this problem, but we have proven that the 2ν-SVM is equivalent to the $2C$-SVM in a certain sense. Additionally, we have shown that the 2ν-SVM has many properties that make it an attractive alternative to the $2C$-SVM. Specifically, as C becomes very large or small, numerical implementations of the $2C$-SVM can become unstable. Thus, when performing parameter estimation, it is typical to restrict C to a range of possible values. However, this range is inevitably arbitrary. The 2ν-SVM replaces C and γ with ν_+ and ν_-. These parameters have a more intuitive meaning, and we have shown that the 2ν-SVM has a feasible solution if and only if $(\nu_+, \nu_-) \in [0, 1]^2$. Thus, the 2ν-SVM offers a much more natural setting for parameter selection, which is a critical issue in practical applications.

7 Proof of Theorems

In order to compare (D_C) and (D_ν), we can rescale (D_C) (by setting $\boldsymbol{\alpha}' = \boldsymbol{\alpha}/Cn$), in which case we obtain:

$$(D'_C) \qquad \min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{Cn}\sum_{i=1}^{n} \alpha_i$$

$$\text{subject to} \qquad 0 \le \alpha_i \le \frac{1}{n} \qquad \text{for } i = 1, \ldots, n$$
$$\sum_{i=1}^{n} \alpha_i y_i = 0.$$

The solutions of (D_C) and (D'_C) have a simple relationship: $\boldsymbol{\alpha}$ is a solution of (D_C) if and only if $\boldsymbol{\alpha}/(Cn)$ is a solution of (D'_C). Thus, in this sense, (D_C) and (D'_C) are equivalent. Furthermore, notice that (D'_C) and (D_ν) differ only in their objective functions and the additional inequality constraint of (D_ν). In [13] this similarity was exploited to establish a detailed relationship between (D_ν) and (D'_C), and hence between (D_ν) and (D_C).

We follow a similar course and rescale (D_{2C}) by Cn in order to compare it with $(D_{2\nu})$. This gives us:

$$(D'_{2C}) \qquad \min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{Cn}\sum_{i=1}^{n} \alpha_i$$

$$\begin{aligned} \text{subject to} \qquad & 0 \le \alpha_i \le \frac{\gamma}{n} && \text{for } i \in I_+ \\ & 0 \le \alpha_i \le \frac{1-\gamma}{n} && \text{for } i \in I_- \\ & \sum_{i=1}^{n} \alpha_i y_i = 0. \end{aligned}$$

Rather than proving the theorems in Section 5 directly, we will take advantage of the relationship between (D_{2C}) and (D'_{2C}). We will establish equivalent theorems (which we denote Theorems $1'$, $2'$, and $3'$) relating $(D_{2\nu})$ and (D'_{2C}), which are then trivially extended to the theorems stated in Section 5. We begin by proving the following lemma:

Lemma 1. *Fix $\gamma \in [0,1]$ for both (D'_{2C}) and $(D_{2\nu})$. Assume (D'_{2C}) and $(D_{2\nu})$ share one optimal solution $\boldsymbol{\alpha}^*$ with $\sum_{i=1}^{n} \alpha_i^* = \nu$. Then any $\boldsymbol{\alpha}$ is an optimal solution of (D'_{2C}) if and only if it is an optimal solution of $(D_{2\nu})$.*

Proof. The analogue of this lemma for (D'_C) and (D_ν) is proved in [13]. The proof depends only on the form of the objective functions (specifically not taking the constraints into account) and on the analogue of Proposition 2. Since the objective function of (D_ν) is identical to that of $(D_{2\nu})$ and the objective function of (D'_C) is also identical to that of (D'_{2C}), we refer the reader to [13] and omit the proof. □

For the proofs of Theorems 1′ and 2′, we will need to employ the Karush-Kuhn-Tucker (KKT) conditions. Essentially, the KKT conditions are necessary and sufficient conditions for $\boldsymbol{\alpha}$ to be an optimal solution to our optimization problem. Specifically, $\boldsymbol{\alpha}$ is an optimal solution of (D'_{2C}) if and only if there exist $b \in \mathbb{R}$ and $\boldsymbol{\lambda}, \boldsymbol{\xi} \in \mathbb{R}^n$ satisfying the KKT conditions:

$$\sum_{j=1}^{n} \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{Cn} + b y_i = \lambda_i - \xi_i \qquad \text{for } i = 1, \ldots, n \tag{1}$$

$$\lambda_i \alpha_i = 0, \quad \lambda_i \geq 0, \quad \xi_i \geq 0 \qquad \text{for } i = 1, \ldots, n \tag{2}$$

$$\xi_i \left(\frac{\gamma}{n} - \alpha_i \right) = 0, \;\; 0 \leq \alpha_i \leq \frac{\gamma}{n} \qquad \text{for } i \in I_+ \tag{3}$$

$$\xi_i \left(\frac{1-\gamma}{n} - \alpha_i \right) = 0, \;\; 0 \leq \alpha_i \leq \frac{1-\gamma}{n} \qquad \text{for } i \in I_- \tag{4}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0. \tag{5}$$

Similarly, $\boldsymbol{\alpha}$ is an optimal solution of $(D_{2\nu})$ if and only if there exist $b, \rho \in \mathbb{R}$ and $\boldsymbol{\lambda}, \boldsymbol{\xi} \in \mathbb{R}^n$ satisfying the slightly different KKT conditions:

$$\sum_{j=1}^{n} \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) - \rho + b y_i = \lambda_i - \xi_i \qquad \text{for } i = 1, \ldots, n \tag{6}$$

$$\lambda_i \alpha_i = 0, \quad \lambda_i \geq 0, \quad \xi_i \geq 0 \qquad \text{for } i = 1, \ldots, n \tag{7}$$

$$\xi_i \left(\frac{\gamma}{n} - \alpha_i \right) = 0, \;\; 0 \leq \alpha_i \leq \frac{\gamma}{n} \qquad \text{for } i \in I_+ \tag{8}$$

$$\xi_i \left(\frac{1-\gamma}{n} - \alpha_i \right) = 0, \;\; 0 \leq \alpha_i \leq \frac{1-\gamma}{n} \qquad \text{for } i \in I_- \tag{9}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \quad \sum_{i=1}^{n} \alpha_i \geq \nu, \quad \rho \left(\sum_{i=1}^{n} \alpha_i - \nu \right) = 0. \tag{10}$$

Notice that the two sets of conditions are mostly identical, except for the first and last two of the conditions for $(D_{2\nu})$. Using this observation, we can prove equivalent versions of the first two theorems.

Theorem 1′. *Fix $\gamma \in [0, 1]$. For each $C > 0$, let $\boldsymbol{\alpha}^*$ be any optimal solution of (D'_{2C}) and set $\nu = \sum_{i=1}^{n} \alpha_i^*$. Then any $\boldsymbol{\alpha}$ is an optimal solution of (D'_{2C}) if and only if it is an optimal solution of $(D_{2\nu})$.*

Proof. If $\boldsymbol{\alpha}^*$ is an optimal solution of (D'_{2C}) then it satisfies the KKT conditions for (D'_{2C}). By setting $\nu = \sum_{i=1}^n \alpha_i^*$ and $\rho = 1/(Cn)$, we see that $\boldsymbol{\alpha}^*$ also satisfies the KKT conditions for $(D_{2\nu})$ and thus is an optimal solution of $(D_{2\nu})$. From Lemma 1 we thus have that, for any $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}$ is an optimal solution of (D'_{2C}) if and only if it is an optimal solution of $(D_{2\nu})$. □

Theorem 2′. *Fix $\gamma \in [0,1]$. Assume $(D_{2\nu})$, $0 < \nu \leq \nu_{\max}$, has a nonzero optimal objective value, then $\rho > 0$. Set $C = 1/(\rho n)$. Then any $\boldsymbol{\alpha}$ is an optimal solution of (D'_{2C}) if and only if it is an optimal solution of $(D_{2\nu})$.*

Proof. If $\boldsymbol{\alpha}^*$ is an optimal solution of $(D_{2\nu})$ then it satisfies the KKT conditions for $(D_{2\nu})$. From the KKT conditions we have

$$\sum_{i=1}^n \left(\sum_{j=1}^n \alpha_j^* y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) - \rho + b y_i \right) \alpha_i^* = \sum_{i=1}^n (\lambda_i - \xi_i) \alpha_i^*$$

which, by applying the remaining KKT conditions, reduces to

$$\sum_{i,j=1}^n \alpha_i^* \alpha_j^* y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) - \rho \sum_{i=1}^n \alpha_i^* = -\frac{\gamma}{n} \sum_{i=1}^n \xi_i.$$

By assumption, $(D_{2\nu})$ has a nonzero optimal objective value. Thus from Proposition 2 we have that $\sum_{i=1}^n \alpha_i^* = \nu$, and we have

$$\rho = \frac{1}{\nu} \left(\sum_{i,j=1}^n \alpha_i^* \alpha_j^* y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) + \frac{\gamma}{n} \sum_{i=1}^n \xi_i \right) > 0.$$

Thus we can choose $C > 0$ such that $C = 1/(\rho n)$ and $\boldsymbol{\alpha}^*$ is a KKT point of (D'_{2C}). Thus from Lemma 1 any $\boldsymbol{\alpha}$ is an optimal solution of (D'_{2C}) if and only if it is an optimal solution of $(D_{2\nu})$. □

We will need the following lemmas to prove Theorem 3′.

Lemma 2. *Fix $\gamma \in [0,1]$ and let $\boldsymbol{\alpha}^*$ be an optimal solution of (D'_{2C}). Define $\nu = \sum_{i=1}^n \alpha_i^*$. If the optimal objective value of $(D_{2\nu})$ is zero, then $\nu = \nu_{\max}$ and any $\boldsymbol{\alpha}$ is an optimal solution of $(D_{2\nu})$ if and only if it is an optimal solution for all (D'_{2C}), $C > 0$.*

Proof. By setting $\rho = 1/Cn$, $\boldsymbol{\alpha}^*$ is a KKT point of $(D_{2\nu})$. Therefore, if the objective function of $(D_{2\nu})$ is zero, $\sum_{i=1}^n \sum_{j=1}^n \alpha_i^* \alpha_j^* y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) = 0$. Since k is a positive definite kernel, we also have $\sum_{j=1}^n \alpha_j^* y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) = 0$. In this case, (1) of (D'_{2C})'s KKT conditions becomes

$$-\frac{1}{Cn} + b y_i = \lambda_i - \xi_i \quad \text{for} \quad i = 1, \ldots, n$$

or

$$-\frac{1}{Cn} + b = \lambda_i - \xi_i \quad \text{for} \quad i \in I_+$$
$$-\frac{1}{Cn} - b = \lambda_i - \xi_i \quad \text{for} \quad i \in I_-.$$

Assume that $b \geq 0$, then

$$\lambda_i - \xi_i < 0 \quad \text{for} \quad i \in I_-.$$

This implies that $\xi_i > 0$ for all $i \in I_-$ since both λ_i and ξ_i are nonnegative. Therefore, in order for the KKT condition $\xi_i((1-\gamma)/n - \alpha_i^*) = 0$ to hold, we must have $\alpha_i^* = (1-\gamma)/n$ for all $i \in I_-$. From condition (5) we have that $\sum_{i \in I_+} \alpha_i^* = \sum_{i \in I_-} \alpha_i^*$, thus we need $\sum_{i \in I_+} \alpha_i^* = (1-\gamma)n_-/n \leq \gamma n_+/n$. Therefore, if $(1-\gamma)n_- > \gamma n_+$ then we have a contradiction, and it must be that $b < 0$.

However, assume without loss of generality that $(1-\gamma)n_- \leq \gamma n_+$, in which case $b \geq 0$ and $\alpha_i^* = (1-\gamma)/n$ for all $i \in I_-$. There are three possibilities for $i \in I_+$:

1. $\lambda_i - \xi_i < 0$
2. $\lambda_i - \xi_i > 0$
3. $\lambda_i - \xi_i = 0$.

In case 1, where $\lambda_i - \xi_i < 0$, we have that $\xi_i > 0$ for all $i \in I_+$. For the KKT condition $\xi_i(\gamma/n - \alpha_i) = 0$ to hold, we need $\alpha_i^* = \gamma/n$ for all $i \in I_+$. The requirement that $\sum_{i \in I_+} \alpha_i^* = \sum_{i \in I_-} \alpha_i^*$ and the fact that $\alpha_i^* = (1-\gamma)/n)$ for all $i \in I_-$ imply that $\sum_{i=1}^n \alpha_i^* = 2n_+\gamma/n = 2n_-(1-\gamma)/n = \nu_{\max}$. Furthermore, the objective function for (D'_{2C}) in this case becomes

$$\min_{\boldsymbol{\alpha}} \quad -\frac{1}{Cn} \sum_{i=1}^n \alpha_i$$

which is clearly minimized by $\boldsymbol{\alpha}^*$ (in which case $\sum_{i=1}^n \alpha_i^* = \nu_{\max}$) for all $C > 0$, thus $\boldsymbol{\alpha}^*$ is an optimal solution of (D'_{2C}) for all $C > 0$. By Lemma 1, any $\boldsymbol{\alpha}$ is an optimal solution of $(D_{2\nu})$ if and only if it is an optimal solution for all (D'_{2C}), $C > 0$.

In case 2, where $\lambda_i - \xi_i > 0$, we have that $\lambda_i > 0$ for all $i \in I_-$. For the KKT condition $\lambda_i \alpha_i^* = 0$ to hold, we need $\alpha_i^* = 0$ for all $i \in I_+$. However, the requirement that $\sum_{i \in I_+} \alpha_i^* = \sum_{i \in I_-} \alpha_i^*$ and the fact that $\alpha_i^* = (1-\gamma)/n$ for all $i \in I_-$ lead to a contradiction if I_- is nonempty. Hence all the training vectors are in the same class, and $\alpha_i^* = 0$ for all i. Thus, $\sum_{i=1}^n \alpha_i^* = 0 = \nu_{\max}$. Furthermore, if all the data are from the same class then $\boldsymbol{\alpha}^* = \mathbf{0}$ is an optimal solution of (D'_{2C}) for all $C > 0$. Thus, by Lemma 1, any $\boldsymbol{\alpha}$ is an optimal solution of $(D_{2\nu})$ if and only if it is an optimal solution for all (D'_{2C}), $C > 0$.

In case 3, where $\lambda_i - \xi_i = 0$, we have that either $\lambda_i = \xi_i \neq 0$ or $\lambda_i = \xi_i = 0$ for each $i \in I_+$. However, $\lambda_i = \xi_i \neq 0$ leads to a contradiction because the KKT conditions would require both $\alpha_i^* = 0$ and $\alpha_i^* = \gamma/n$. Thus, $\lambda_i = \xi_i = 0$ and the KKT conditions involving

λ_i and ξ_i impose no conditions on α_i^* for $i \in I_+$. Since $\alpha_i^* = (1-\gamma)/n$ for all $i \in I_-$, and $(1-\gamma)n_- \leq \gamma n_+$, we have $\sum_{i\in I_+} \alpha_i^* = \sum_{i\in I_-} \alpha_i^* = (1-\gamma)n_+/n$. Thus, $\sum_{i=1}^n \alpha_i^* = \nu_{\max}$. Furthermore, by setting $b = 1/(Cn)$, $\boldsymbol{\alpha}^*$ is an optimal solution of (D'_{2C}) for all $C > 0$. Thus, by Lemma 1, any $\boldsymbol{\alpha}$ is an optimal solution of $(D_{2\nu})$ if and only if it is an optimal solution for all (D'_{2C}), $C > 0$. □

Lemma 3. *Assume $\boldsymbol{\alpha}^*$ is any optimal solution of (D'_{2C}), then $\sum_{i=1}^n \alpha_i^*$ is a continuous decreasing function of C on $(0, \infty)$.*

Proof. Again, the analogue of this lemma for (D'_C) is proved in [13]. Since the proof depends only on the form of the objective function and the analogues of Theorems 1′ and 2′, we omit the proof and refer the reader to [13]. □

Using these lemmas, we are now ready to prove the equivalent of the main theorem:

Theorem 3′. *Fix $\gamma \in [0, 1]$ and let $\boldsymbol{\alpha}^*$ be any optimal solution of (D'_{2C}). Define*

$$\nu_* = \lim_{C\to\infty} \sum_{i=1}^n \alpha_i^*$$

and

$$\nu^* = \lim_{C\to 0} \sum_{i=1}^n \alpha_i^*.$$

Then $0 \leq \nu_* \leq \nu^* = \nu_{\max} \leq \frac{1}{2}$. Thus, for any $\nu > \nu^*$, $(D_{2\nu})$ is infeasible. For any $\nu \in (\nu_*, \nu^*]$, the optimal objective value of $(D_{2\nu})$ is strictly positive, thus there exists at least one $C > 0$ such that any $\boldsymbol{\alpha}$ is an optimal solution of (D'_{2C}) if and only if it is an optimal solution of $(D_{2\nu})$. For any $\nu \in [0, \nu_*]$, $(D_{2\nu})$ is feasible with zero optimal objective value (and a trivial solution).

Proof. From Lemma 3 and the fact that $0 \leq \sum_{i=1}^n \alpha_i^* \leq \nu_{\max}$ we know that the above limits exist and can be defined without any problems.

For the any optimal solution of (D'_{2C}), we have that the KKT condition (1) holds:

$$\sum_{j=1}^n \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{Cn} + b = \lambda_i - \xi_i \quad \text{for} \quad i \in I_+$$

$$\sum_{j=1}^n \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{Cn} - b = \lambda_i - \xi_i \quad \text{for} \quad i \in I_-.$$

Assume that $b \geq 0$. Since $\boldsymbol{\alpha}^*$ is bounded, when C is sufficiently small, we will have $\lambda_i - \xi_i < 0$ for $i \in I_+$, thus $\xi_i > 0$ and from the KKT conditions, $\alpha_i^* = \gamma/n$ for all $i \in I_+$. If $\gamma n_+/n \geq (1-\gamma)n_-/n$, then this $\boldsymbol{\alpha}^*$ is feasible and $\sum_{i=1}^n \alpha_i^* = \nu_{\max}$. However, if $\gamma n_+/n < (1-\gamma)n_-/n$

then we have a contradtiction, and thus $b < 0$. In this case, for C sufficiently small, $\lambda_i - \xi_i < 0$ for $i \in I_i$. As above, this implies that $\alpha_i^* = (1-\gamma)/n$ for all $i \in I_-$, and thus $\sum_{i=1}^n \alpha_i^* = \nu_{\max}$. Hence, $\nu^* = \sum_{i=1}^n \alpha_i^* = \nu_{\max}$. In this case, from Proposition 1 we immediately know that $(D_{2\nu})$ is infeasible if $\nu > \nu^*$.

From Proposition 1, we know that for all $\nu \leq \nu^*$ $(D_{2\nu})$ is feasible. From Lemma 3 we know that $\sum_{i=1}^n \alpha_i^*$ is a continuous decreasing function. Thus for any $\nu \in (\nu_*, \nu^*]$, there is a $C > 0$ such that $\sum_{i=1}^n \alpha_i^* = \nu$, and any $\boldsymbol{\alpha}$ is an optimal solution of $(D_{2\nu})$ if and only if it is an optimal solution for (D'_{2C}).

If $\nu < \nu_*$, $(D_{2\nu})$ must have an optimal objective value of zero because of Theorem 2 and the definition of ν_*. If $\nu = \nu_* = 0$, the optimal objective value of $(D_{2\nu})$ is zero, as $\boldsymbol{\alpha}^* = \mathbf{0}$ is a feasible solution. If $\nu = \nu_* > 0$, the fact that feasible regions of $(D_{2\nu})$ are bounded by $0 \leq \alpha_i \leq \gamma/n$ for $i \in I_+$ and $0 \leq \alpha_i \leq (1-\gamma)/n$ for $i \in I_-$, and Proposition 2 imply that there exists a sequence $\{\boldsymbol{\alpha}^{\nu_j}\}$, $\nu_1 \leq \nu_2 \leq \cdots \leq \nu_*$ such that $\boldsymbol{\alpha}^{\nu_j}$ is an optimal solution of $(D_{2\nu})$ with $\nu = \nu_j$, $\sum_{i=1}^n \alpha_i^{\nu_j} = \nu_j$, and $\hat{\boldsymbol{\alpha}} = \lim_{\nu_j \to \nu_*} \boldsymbol{\alpha}^{\nu_j}$ exists. Since $\sum_{i=1}^n \alpha_i^{\nu_j} = \nu_j$, $\sum_{i=1}^n \hat{\alpha}_i = \lim_{\nu_j \to \nu_*} \sum_{i=1}^n \alpha_i^{\nu_j} = \nu_*$. We also have that $0 \leq \hat{\alpha}_i \leq \gamma/n$ for $i \in I_+$, $0 \leq \hat{\alpha}_i \leq (1-\gamma)/n$ for $i \in I_-$, and $\sum_{i=1}^n y_i \hat{\alpha}_i = \lim_{\nu_j \to \nu_*} y_i \sum_{i=1}^n \alpha_i^{\nu_j} = 0$ so $\hat{\boldsymbol{\alpha}}$ is feasible to $(D_{2\nu})$ for $\nu = \nu_*$. However, $\sum_{l,m=1}^n \hat{\alpha}_l \hat{\alpha}_m y_l y_m k(\mathbf{x}_l, \mathbf{x}_m) = \lim_{\nu_j \to \nu_*} \sum_{l,m=1}^n \alpha_l^{\nu_j} \alpha_m^{\nu_j} y_l y_m k(\mathbf{x}_l, \mathbf{x}_m) = 0$ as $\sum_{l,m=1}^n \alpha_l^{\nu_j} \alpha_m^{\nu_j} y_l y_m k(\mathbf{x}_l, \mathbf{x}_m) = 0$ for all ν_j. Therefore the optimal objective value of $(D_{2\nu})$ is zero if $\nu = \nu_*$.

Finally, from the above discussion, if $\nu \leq \nu_*$, the objective value of $(D_{2\nu})$ is zero. If the objective value of $(D_{2\nu})$ is zero but $\nu > \nu_*$, then by Lemma 3 there is a $C > 0$ such that, if $\boldsymbol{\alpha}^*$ is an optimal solution of (D'_{2C}), then $\sum_{i=1}^n \alpha_i^* = \nu$. Thus, from Lemma 2, we have that $\nu = \nu_{\max} = \nu^* < \nu_*$, a contradiction. Thus the objective value of $(D_{2\nu})$ is zero if and only if $\nu \leq \nu_*$. In this case, $\mathbf{w} = 0$ and we say that the solution is *trivial*. □

References

[1] E. Osuna, R. Freund, and F. Girosi, "Support vector machines: Training and applications," MIT Artificial Intelligence Laboratory, Tech. Rep. A.I. Memo No. 1602, March 1997.

[2] K. Veropoulos, N. Cristianini, and C. Campbell, "Controlling the sensitivity of support vector machines," in *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 1999.

[3] Y. Lin, Y. Lee, and G. Wahba, "Support vector machines for classification in non-standard situations," University of Wisconsin, Dept. of Statistics, Tech. Rep. Technical Report No. 1016, March, 2000.

[4] H. G. Chew, R. E. Bogner, and C. C. Lim, "Target detection in radar imagery using support vector machines with training size biasing," in *Proc. Int. Conf. on Control, Automation, Robotics, and Vision (ICARCV)*, 2000.

[5] ——, "Dual-ν support vector machine with error rate and training size biasing," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, 2001, pp. 1269–1272.

[6] C. Scott and R. Nowak, "A Neyman-Pearson approach to statistical learning," *IEEE Trans. on Information Theory*, November, 2005.

[7] H. G. Chew, C. C. Lim, and R. E. Bogner, "Dual-nu support vector machines and applications in multi-class image recognition," in *Proc. Int. Conf. on Optimization: Techniques and Applications (ICOTA)*, 2004.

[8] B. Schölkopf and A. J. Smola, *Learning with Kernels*. Cambridge, MA: MIT Press, 2002.

[9] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[10] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1999.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[12] B. Schölkopf, A. J. Smola, R. Williams, and P. Bartlett, "New support vector algorithms," *Neural Computation*, vol. 12, pp. 1083–1121, 2000.

[13] C. C. Chang and C. J. Lin, "Training ν-support vector classifiers: Theory and algorithms," *Neural Computation*, vol. 13, pp. 2119–2147, 2001.