Manuscript Number: 2621

Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel

S. Sathiya Keerthi

Department of Mechanical Engineering

National University of Singapore

Singapore 119260, Republic of Singapore

mpessk@guppy.mpe.nus.edu.sg

Chih-Jen Lin

Department of Computer Science and

Information Engineering

National Taiwan University

Taipei 106, Taiwan

cjlin@csie.ntu.edu.tw

Abstract

Support vector machines (SVMs) with the Gaussian (RBF) kernel have been popular for practical use. Model selection in this class of SVMs involves two hyperparameters: the penalty parameter C and the kernel width σ. This paper analyzes the behavior of the SVM classifier when these hyperparameters take very small or very large values. Our results help in a good understanding of the hyperparameter space that leads to an efficient heuristic method of searching for hyperparameter values with small generalization errors. The analysis also indicates that if complete model selection using the Gaussian kernel has been conducted, there is no need to consider linear SVM.

1 Introduction

Given a training set of instance-label pairs $(x_i, y_i), i = 1, \ldots, l$ where $x_i \in R^n$ and $y \in \{1, -1\}^l$, support vector machines (SVMs) (Vapnik 1998) require the solution of the following (primal) optimization problem:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}w^T w + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i(w^T z_i + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, \ldots, l. \end{aligned} \tag{1.1}$$

Here training vectors x_i are mapped into a higher (maybe infinite) dimensional space by the function ϕ as $z_i = \phi(x_i)$. $C > 0$ is the penalty parameter of the error term.

Usually we solve (1.1) by solving the following dual problem:

$$\begin{aligned} \min_{\alpha} \quad & F(\alpha) = \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C, i = 1, \ldots, l, \\ & y^T\alpha = 0, \end{aligned} \tag{1.2}$$

where e is the vector of all ones and Q is an l by l positive semidefinite matrix. The (i,j)-th element of Q is given by $Q_{ij} \equiv y_i y_j K(x_i, x_j)$ where $K(x_i, x_j) \equiv \phi(x_i)^T\phi(x_j)$ is called the kernel function. Then $w = \sum_{i=1}^{l} \alpha_i y_i \phi(x_i)$ and

$$\operatorname{sgn}(w^T\phi(x) + b) = \operatorname{sgn}(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b)$$

is the decision function.

We are particularly interested in the Gaussian kernel:

$$K(\tilde{x}, \bar{x}) = \exp(\frac{-\|\tilde{x} - \bar{x}\|^2}{2\sigma^2}). \tag{1.3}$$

Our aim is to analyze the behaviors of the SVM classifier when C and/or σ^2 take very small or very large values. The motivation is that, such an analysis will help in a good understanding of the hyperparameter space that will lead to efficient heuristic ways of searching for points in that space with small generalization errors. Some of the behaviors that we will discuss are known in the literature (although, details associated with these are usually not written down carefully) but some key behaviors are new results that are not entirely obvious. Here is a quick summary of the asymptotic behaviors of the SVM classifier that are derived in this paper:

- Severe underfitting (the entire data space is assigned to the majority class) occurs in the following cases: (a) σ^2 is fixed and $C \to 0$; (b) $\sigma^2 \to 0$ and C is fixed to a sufficiently small value; and (c) $\sigma^2 \to \infty$ and C is fixed.

- Severe overfitting (small regions around the training examples of the minority class are classified to be that class while the rest of the data space is classified as the majority class) occurs in the case where $\sigma^2 \to 0$ and C is fixed to a sufficiently large value.

- If σ^2 is fixed and $C \to \infty$ the SVM classifier strictly separates the training examples of the two classes; this is a case of overfitting if the problem under consideration has noise.

- If $\sigma^2 \to \infty$ and $C = \tilde{C}\sigma^2$ where $\tilde{C}$ is fixed then the SVM classifier converges to the Linear SVM classifier with penalty parameter $\tilde{C}$.

Figure 1 gives a summary of the asymptotic behaviors.

Asymptotic behaviors of the generalization error associated with the SVM classifier as C and/or σ^2 take extreme values can be understood via a study of corresponding behaviors of the leave-one-out (loo) error. The loo error is computed as follows. For the i-th example, (1.1) and (1.2) are solved after leaving out that example. The resulting classifier is applied to check if the i-th example is misclassified. The procedure is repeated for each i. The fraction of examples that are misclassified is the loo error.

This paper is organized as follows. In Section 2 we analyze the asymptotic behaviors of the SVM classifier using the Gaussian kernel. The results lead to a simple and efficient heuristic model selection strategy which is described in Section 3. Experiments show that the proposed method is competitive with the usual cross validation search strategy in terms of generalization error achieved, while at the same time, it is much more efficient.

2 Asymptotic Behaviors

To establish various asymptotic behaviors of the SVM decision function as well as the loo error, we need the following assumption, which will be assumed throughout

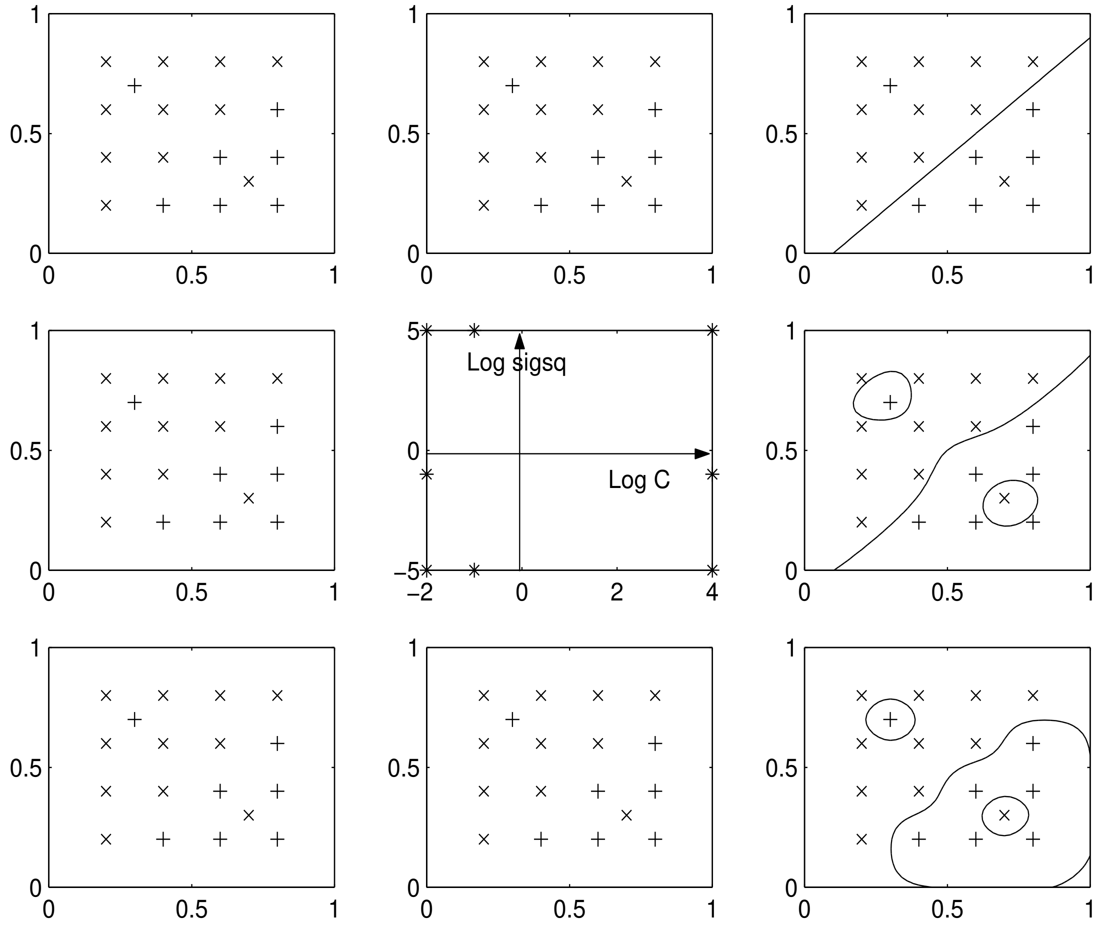

Figure 1: A figurative summary of the asymptotic behaviors. The problem has 11 examples in class 1 (shown by $\times$) and 7 examples in class 2 (shown by $+$). (Thus class 1 is the majority class and class 2 is the minority class.) The plot in the center shows the eight $(\log C, \log \sigma^2)$ pairs tried. The decision curves corresponding to these eight pairs are displayed in the surrounding plots at respective positions. Plots without a decision curve correspond to underfitting classifiers for which the entire input region is classified as class 1.

the paper.

Assumption 1

1. *$l_1 > l_2 + 1 > 2$ where l_1 and l_2 are the numbers of training examples in class 1 and class 2, respectively.*[*]

2. *For $i \neq j$, $x_i \neq x_j$. That is, no two examples have identical x vectors.*[†]

The following lemma is useful.

Lemma 1 *For any given (C, σ^2), the solution (α) of (1.2) is unique. Also, for every σ^2, $\{z_i \mid y_i = 1\}$ and $\{z_i \mid y_i = -1\}$ are linearly separable.*

Proof. From (Micchelli 1986), if the Gaussian kernel is used and $x_i \neq x_j \;\; \forall i \neq j$ from Assumption 1, Q is positive definite. By Corollary 1 of (Chang and Lin 2001b), we get linear separability in z-space. Uniqueness of α follows from the fact that (1.2) is a strictly convex quadratic programming problem. □

We now discuss the various asymptotic behaviors. As the results of each case are stated, it is useful to refer to the example shown in Figure 1 given in section 1. Wherever we come across results whose proofs do not shed any insight on the asymptotic behaviors, we only state the results and relegate the proofs to the appendix.

<u>Case 1. σ^2 fixed and $C \to 0$</u>

It can be shown (see the proof of Theorem 5 in (Chang and Lin 2001b) for details) that, if C is smaller than a certain positive value, the following holds:

$$\alpha_i = C \;\; \forall\, i \quad \text{with } y_i = -1. \tag{2.1}$$

Let us take one such C. Using (2.1) together with $\sum_{i=1}^{l} y_i \alpha_i = 0$ and $l_1 > l_2$, it is easy to see that there exists at least one i for which $\alpha_i < C$ and $y_i = 1$. For such

[*] If $l_2 > l_1 + 1 > 2$, then we can always interchange the two classes and apply all the results derived in this paper. Cases where $|l_1 - l_2| \leq 1$ or $\min\{l_1, l_2\} < 2$ correspond to abnormal situations that are not worth discussing in detail since in practice the numbers of examples in the two classes rarely satisfy any of these two conditions.

[†] This is a generic assumption that is easily satisfied if small random perturbations are added to all training examples.

an i we have

$$w^T z_i + b \geq 1. \tag{2.2}$$

For $C \to 0$ we have $\alpha_i \to 0$ and so $w^T z = \sum_{i=1}^{l} \alpha_i y_i K(x_i, x) \to 0$, where $z = \phi(x)$. These imply that, if X is any compact subset of R^n, then for any given $0 < a < 1$ there exists $\bar{C} > 0$ such that for all $C \leq \bar{C}$ we have

$$b \geq a \text{ and } |\sum_{i=1}^{l} \alpha_i y_i K(x_i, x)| \leq \frac{a}{2} \; \forall x \in X. \tag{2.3}$$

Hence, for all $C \leq \bar{C}$

$$f(x) > 0 \; \forall x \in X.$$

In particular, if we take X to be the compact subset of data space that is of interest to the given problem, then for sufficiently small C every point in this subset is classified as class 1.

The first part of Assumption 1 allows us to use similar arguments for the case of (1.2) with one example left out. Then we can also show that, as $C \to 0$, the number of loo errors is l_2. Thus $C \to 0$ corresponds to severe underfitting as expected. Furthermore, we have the following properties as $C \to 0$.

1. $\|w\|^2 = \alpha^T Q \alpha \to 0$
2. $\lim_{C \to 0} \frac{1}{C} \sum_{i=1}^{l} \alpha_i = \lim_{C \to 0} \frac{2}{C} \sum_{i: y_i = -1} \alpha_i = 2 l_2$
3. Using the equality of primal and dual objective function values at optimality and the inequality $\alpha^T Q \alpha \leq l^2 C^2$ we get

$$\lim_{C \to 0} \sum_{i=1}^{l} \xi_i = \lim_{C \to 0} \frac{1}{C} (\sum_{i=1}^{l} \alpha_i - \alpha^T Q \alpha) = 2 l_2.$$

It is useful to interpret the above asymptotic results geometrically; in particular, study the movement of the top, middle and bottom planes defined by $w^T z + b = 1$, $w^T z + b = 0$ and $w^T z + b = -1$ as $C \to 0$. By (2.2) at least one example of class 1 lies on or above the top plane. By property 1 given above, the distance between the top and bottom planes (which equals $2/\|w\|$) goes to infinity. Hence, the middle and bottom planes are forced to move down farther and farther away from the location where the training points are located, causing the half space defined by $w^Tz+b \geq 0$ to entirely cover X, the compact subset of interest to the problem, after C becomes sufficiently small.

Remark 2.1 The results given above for $C \to 0$ are general and apply to non-Gaussian kernels also, assuming, of course, that all hyperparameters associated with the kernel function are kept fixed. The results also apply if Q is a bounded function of C since Theorem 5 of (Chang and Lin 2001b) holds for this case.

Remark 2.2 For kernels whose values are bounded (e.g., the Gaussian kernel), there is $\bar{C}$ such that (2.3) holds for all $x \in R^n$. Thus, for all $C \leq \bar{C}$,

$$f(x) > 0 \; \forall x \in R^n.$$

That is, for all $C \leq \bar{C}$ every point is classified as class 1.

Case 2. σ^2 fixed and $C \to \infty$

By Lemma 1 given at the beginning of this section, $\{z_i \mid y_i = 1\}$ and $\{z_i \mid y_i = -1\}$ are linearly separable. This implies that it is possible to set $\xi_i = 0 \; \forall i$ while still remaining feasible for (1.1). Thus, as $C \to \infty$, the solution of (1.1) approaches the solution of the hard margin problem:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} w^T w \\ \text{subject to} \quad & y_i(w^T z_i + b) \geq 1, i = 1, \ldots, l. \end{aligned} \tag{2.4}$$

A formal treatment of this is in (Lin 2001) which shows that if (2.4) is feasible, then there exists a C^* such that for $C \geq C^*$, the solution set of (1.1) is the same as that of (2.4). An easy way to see this result is to solve (1.2) with $C = \infty$, obtain the $\{\alpha_i\}$ and set $C^* = \max_i \alpha_i$.

The limiting SVM classifier classifies all training examples correctly and so it is an overfitting classifier. In particular, severe overfitting occurs when σ^2 is small since flexibility of the classifier is high when σ^2 is small.

For the case of $C \to \infty$, it is not possible to make any conclusions about the actual value of the loo error. That value depends on the dataset as well as on the value of σ^2. However, after (1.2) is solved using all the examples it is possible to give bounds on the loo error (Joachims 2000; Vapnik and Chapelle 2000) without solving the quadratic programs obtained by leaving out one example at a time.

Case 3. C **is fixed and** $\sigma^2 \to 0$

Let us define $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. Since $e^{-\|x_i - x_j\|^2/(2\sigma^2)} \to \delta_{ij}$ as $\sigma^2 \to 0$, we consider the following problem:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\alpha^T\alpha - e^T\alpha \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C, i = 1, \ldots, l, \\ & y^T\alpha = 0. \end{aligned} \tag{2.5}$$

Using Lemma 2 (proof in Appendix A.1), as $\sigma^2 \to 0$ the solution of (1.2) converges to that of (2.5). Since $l_1 > l_2$, the solution of (1.2) has $0 < \alpha_i < C$ for at least one i.[‡] Thus b is uniquely determined and, as $\sigma^2 \to 0$, it approaches the value of b corresponding to the primal form of (2.5).

Therefore, let us study the solution of (2.5). In Appendix A.2, we show that its solution is given by: $\alpha_i = \alpha^+$ if $y_i = 1$; $\alpha_i = \alpha^-$ if $y_i = -1$, where

$$\alpha^- = \begin{cases} C_{\lim} & \text{if } C \geq C_{\lim} \\ C & \text{if } C < C_{\lim}, \end{cases} \qquad \alpha^+ = \begin{cases} 2l_2/l & \text{if } C \geq C_{\lim} \\ l_2C/l_1 & \text{if } C < C_{\lim}, \end{cases} \tag{2.6}$$

and $C_{\lim} = 2l_1/l$. The threshold parameter b in the primal form corresponding to (2.5) can be determined using the fact that $0 < \alpha^+ < C$ (and hence all class 1 examples lie on the top plane defined by $w^Tz + b = 1$):

$$b = \begin{cases} (l_1 - l_2)/l & \text{if } C \geq C_{\lim}, \\ 1 - l_2C/l_1 & \text{if } C < C_{\lim}. \end{cases} \tag{2.7}$$

Consider the classifier function $f(x) = w^Tz + b$ corresponding to (2.5). In Appendix A.2 we also show the following.

1. If $C \geq C_{\lim}/2$, f classifies all training examples correctly and classifies the rest of the space as class 1. Thus it overfits the training data.

2. If $C < C_{\lim}/2$, then f classifies the entire space as class 1 and so it underfits the training data.

[‡] As we show below (see (2.6)) the solution of (2.5) is well in the interior of $(0, C)$ for at least one i. Since, for small values of σ^2, the solution of (1.2) approaches that of (2.5), it follows that the solution of (1.2) also has $0 < \alpha_i < C$ for at least one i.

3. The number of loo errors is l_2.

Consider the SVM classifier corresponding to the Gaussian kernel for small values of σ^2. Even though the number of loo errors tends to l_2 for all C, it is important to note that the SVM classifier is qualitatively very different for large C and small C. For large C, there are small regions around each example of class 2 which are classified as class 2 (overfitting) while for small C, there are no such regions (underfitting).

It is interesting to note that, if σ^2 is small and C is greater than a threshold which is around C_{lim}, from (2.6), the SVM classifier does not depend on C. Thus, contour lines of constant generalization error are parallel to the C axis in the region where σ^2 is small and C is large.

Case 4. C is fixed and $\sigma^2 \to \infty$

When $\sigma^2 \to \infty$, we can write

$$\begin{aligned} K(\tilde{x}, \bar{x}) &= \exp(-\|\tilde{x} - \bar{x}\|^2/2\sigma^2) \\ &= 1 - \frac{\|\tilde{x} - \bar{x}\|^2}{2\sigma^2} + o(\|\tilde{x} - \bar{x}\|^2/\sigma^2) \\ &= 1 - \frac{\|\tilde{x}\|^2}{2\sigma^2} - \frac{\|\bar{x}\|^2}{2\sigma^2} + \frac{\tilde{x}^T\bar{x}}{\sigma^2} + o(\|\tilde{x} - \bar{x}\|^2/\sigma^2). \end{aligned} \tag{2.8}$$

Now consider (1.2). Using the simplification given above, we can write the first (quadratic) term of the objective function in (1.2) as

$$\sum_i \sum_j \alpha_i\alpha_j y_i y_j K(x_i, x_j) = T_1 + \frac{T_2 + T_3 + T_4}{2\sigma^2} + \frac{1}{2}\sum_i \sum_j \alpha_i\alpha_j y_i y_j \frac{\Delta_{ij}}{\sigma^2},$$

where

$$T_1 = \sum_i \sum_j \alpha_i\alpha_j y_i y_j\ , \quad T_2 = -\sum_i \sum_j \alpha_i\alpha_j y_i y_j \|x_i\|^2,$$
$$T_3 = -\sum_i \sum_j \alpha_i\alpha_j y_i y_j \|x_j\|^2\ , \qquad T_4 = 2\sum_i \sum_j \alpha_i\alpha_j y_i y_j x_i^T x_j, \ \text{ and}$$
$$\lim_{\sigma^2 \to \infty} \Delta_{ij} = 0. \tag{2.9}$$

By the equality constraint of (1.2), $T_1 = (\sum_i \alpha_i y_i)^2 = 0$. We can also rewrite T_2 as $T_2 = (\sum_i \alpha_i y_i \|x_i\|^2)(\sum_j \alpha_j y_j) = 0$. In a similar way, $T_3 = 0$. By defining

$$\tilde{\alpha}_i = \frac{\alpha_i}{\sigma^2}\ \ \forall\ i, \tag{2.10}$$

(1.2) can be written as [§]

$$
\begin{aligned}
\min_{\tilde{\alpha}} \quad & \frac{F}{\sigma^2} = \frac{1}{2}\sum_i\sum_j \tilde{\alpha}_i\tilde{\alpha}_j y_i y_j \tilde{K}_{ij} - \sum_i \tilde{\alpha}_i \\
\text{subject to} \quad & 0 \le \tilde{\alpha}_i \le \tilde{C}, i = 1, \ldots, l, \\
& y^T\tilde{\alpha} = 0,
\end{aligned}
\tag{2.11}
$$

where $\tilde{K}_{ij} = x_i^T x_j + \Delta_{ij}$ and

$$
\tilde{C} = \frac{C}{\sigma^2}. \tag{2.12}
$$

Remark 2.3 Note that $\tilde{K}_{ij}$ may not correspond to a valid kernel satisfying the Mercer's condition. But that is immaterial since we always operate with the constraint $y^T\tilde{\alpha} = 0$. In the presence of this constraint (1.2) and (2.11) are equivalent.

Remark 2.4 If C is fixed at some value and σ^2 is made large, $\tilde{C}$ of (2.11) goes to zero and so the situation is similar to Case 1 that we discussed at the beginning of this section. By (2.9), $\tilde{K}_{ij}$ is a bounded function for large σ^2 (or, equivalently, for small $\tilde{C}$). By the last sentence of Remark 2.1, results of Case 1 can be applied here. Thus, for C fixed and $\sigma^2 \to \infty$, (2.11) corresponds to a severely underfitting classifier. Since (2.11) and (1.2) correspond to the same problem in different forms, they have the same primal decision function (for full details see (A.8)). Therefore, in this situation we get a severely underfitting classifier.

For a given $\tilde{C}$, as $\sigma^2 \to \infty$ and C varies with σ^2 as given by (2.12), we can see that (2.11) is close to the following linear SVM problem:

$$
\begin{aligned}
\min_{\tilde{\alpha}} \quad & \frac{1}{2}\sum_i\sum_j \tilde{\alpha}_i\tilde{\alpha}_j y_i y_j x_i^T x_j - \sum_i \tilde{\alpha}_i \\
\text{subject to} \quad & 0 \le \tilde{\alpha}_i \le \tilde{C}, i = 1, \ldots, l, \\
& y^T\tilde{\alpha} = 0.
\end{aligned}
\tag{2.13}
$$

[§] To do this, note that we need to divide the objective function by the term σ^2.

We are interested in their corresponding decision functions which can lead us to analyze the performance of (1.1). Now the primal form of (2.13) is

$$\begin{aligned} \min_{\tilde{w},\tilde{b},\tilde{\xi}} \quad & \frac{1}{2}\tilde{w}^T\tilde{w} + \tilde{C}\sum_{i=1}^{l}\tilde{\xi}_i \\ \text{subject to} \quad & y_i(\tilde{w}^T x_i + \tilde{b}) \geq 1 - \tilde{\xi}_i, \\ & \tilde{\xi}_i \geq 0, i = 1, \ldots, l. \end{aligned} \tag{2.14}$$

Let $(w(\sigma^2), b(\sigma^2))$ and $(\tilde{w}, \tilde{b})$ denote primal optimal solutions of (1.1) and (2.14), respectively. We then have the following theorem.

Theorem 2 *For any x, $\lim_{\sigma^2\to\infty} w(\sigma^2)^T z = \tilde{w}^T x$. If the optimal $\tilde{b}$ of (2.14) is unique then $\lim_{\sigma^2\to\infty} b(\sigma^2) = \tilde{b}$ and hence the following also hold:*

1. *For any x,*

$$\lim_{\sigma^2\to\infty} w(\sigma^2)^T z + b(\sigma^2) = \tilde{w}^T x + \tilde{b}; \tag{2.15}$$

2. *If $\tilde{w}^T x + \tilde{b} \neq 0$, then for σ^2 sufficiently large,*

$$sgn(w(\sigma^2)^T z + b(\sigma^2)) = sgn(\tilde{w}^T x + \tilde{b}).$$

The proof is in Appendix A.3. *Thus, for a given $\tilde{C}$, the limiting SVM Gaussian kernel classifier as $\sigma^2 \to \infty$ is same as the SVM linear kernel classifier for $\tilde{C}$.* Hereafter, we will simply refer to the SVM linear kernel classifier as 'Linear SVM'. The above analysis can also be extended to show that, as $\sigma^2 \to \infty$, the loo error corresponding to (1.1) and (2.13) are the same.

The above results also show that, in the part of the hyperparameter space where σ^2 is large, if (C_1, σ_1^2) and (C_2, σ_2^2) are related by $C_1/\sigma_1^2 = C_2/\sigma_2^2 = \tilde{C}$, the classifiers corresponding to the two combinations are nearly the same. Hence, they both will give nearly the same value for generalization error (or an estimate of it such as k-fold cross validation error or loo error). Thus, in this part of the hyperparameter space, contour lines of such functions will be straight lines with slope 1: $\log\sigma^2 = \log C - \log\tilde{C}$. Then all classifiers defined by points on that straight line for large σ^2 are nearly the same as the Linear SVM classifier corresponding to $\tilde{C}$.

Given that for any x, $\lim_{\sigma^2 \to \infty} w(\sigma^2)^T z = \tilde{w}^T x$ holds without any assumption, the assumption on the uniqueness of $\tilde{b}$ in Theorem 2 should only be viewed as a minor technical irritant.¶ For normal situations, the uniqueness assumption is a reasonable one to make. Unless $\tilde{C}$ is very small, typically there will be at least one $\tilde{\alpha}_i$ strictly in between 0 and $\tilde{C}$; when such an $\tilde{\alpha}_i$ exists, Lemma 3 in Appendix A.1 (as applied to (2.14)) implies the uniqueness of $\tilde{b}$. The case $\tilde{C}$ very small corresponds to the upper left part of the plane in which $\log C$ and $\log \sigma^2$ are the horizontal and vertical axes. We can easily see this by considering C fixed and increasing σ^2 to large values (the upper part) or considering σ^2 fixed and decreasing C to small values (the left part). As Remark 2.4 and Case 1 of this section show, each of these asymptotic behaviors corresponds to a severely underfitting SVM decision function.

Finally, Theorem 2 also indicates that if complete model selection on (C, σ^2) using the Gaussian kernel has been conducted, there is no need to consider linear SVM. This helps the selection of kernels.

3 A Method of Model Selection

It is usual to take $\log C$ and $\log \sigma^2$ as the parameters of the hyperparameter space. Putting together the results derived in the previous section, it is easy to see that, in the asymptotic (outer) regions of the $(\log C, \log \sigma^2)$ space there exists a contour of generalization error (or an estimate such as loo error or k-fold cross validation error) that looks like that shown in Figure 2 and which helps separate the hyperparameter space into two regions: an overfitting/underfitting region and a good region (which most likely has the hyperparameter set with the best generalization error). (For loo, recall that, in the underfitting/overfitting region, the number of loo errors is l_2.) The straight line with unit slope in the large σ^2 region ($\log \sigma^2 = \log C - \log \tilde{C}$) corresponds to the choice of $\tilde{C}$ which is small enough to make the Linear SVM an underfitting one. The presence of a

¶ The assumption is needed to state results cleanly. If $\tilde{b}$ is non-unique, SVM classifiers also become non-unique and then it becomes clumsy to talk about convergence of SVM decision functions.

separating contour as outlined in Figure 2 has been observed on a number of real world datasets (Lee 2001).

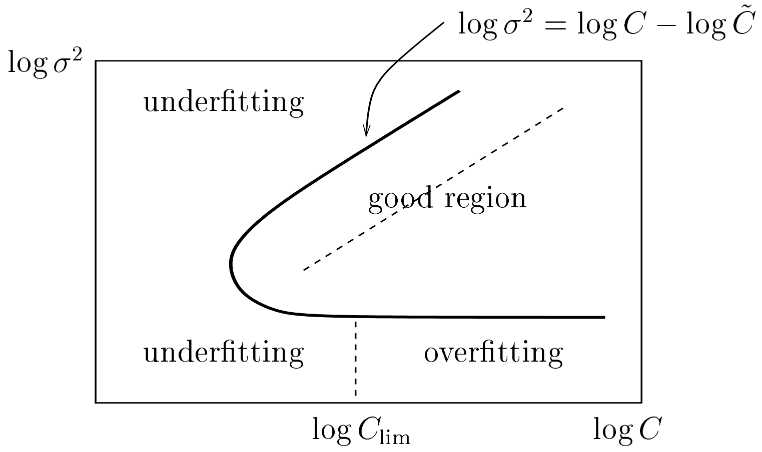

Figure 2: A rough boundary curve separating the underfitting/overfitting region from the "good" region. For each fixed $\tilde{C}$, the equation $\log \sigma^2 = \log C - \log \tilde{C}$ defines a straight line of unit slope. As $\sigma^2 \to \infty$ along this line, the SVM classifier converges to the Linear SVM classifier with penalty parameter $\tilde{C}$. The dotted line corresponds to the choice of $\tilde{C}$ that gives optimal generalization error for the Linear SVM.

When searching for a good set of values for $\log C$ and $\log \sigma^2$, it is usual to form a two dimensional uniform grid (say $r \times r$) of points in this space and find a combination that gives the least value for some estimate of generalization error. This is expensive since it requires the trying of r^2 (C, σ^2) pairs. The earlier discussion relating to Figure 2 suggests a simple and efficient heuristic method for finding a hyperparameter set with small generalization error: form a line of unit slope which cuts through the middle part of the good region (see the dashed line in Figure 2) and search on it for a good set of hyperparameters. The $\tilde{C}$ that defines this line can be set to the optimal value of penalty parameter for the Linear SVM. Thus, we propose the following procedure.

1. Search for the best C of Linear SVM and call it $\tilde{C}$.

2. Fix $\tilde{C}$ from step 1 and search for the best (C, σ^2) satisfying $\log \sigma^2 = \log C - \log \tilde{C}$ using the Gaussian kernel.

The idea is that, as $\sigma^2 \to \infty$, SVM with Gaussian kernel behaves like Linear

SVM and so the best $\tilde{C}$ should happen in the upper part of the "good" region in Figure 2. Then a search on the line defined by $\log\sigma^2 = \log C - \log\tilde{C}$ gives an even better point in the "good" region. In many practical pattern recognition problems, a linear classifier already gives a reasonably good performance and some added nonlinearities help obtain finer improvements in accuracy. Step 2 of our procedure can be thought of as a simple way of injecting the required nonlinearities via the Gaussian kernel. Since the procedure only involves two one dimensional searches, it requires only $2r$ pairs of (C, σ^2) to be tried.

To test the goodness of the proposed method, we compare it with the usual method of using two dimensional grid search. For both, five-fold cross validation was used to obtain estimates of generalization error. For the usual method, we uniformly discretize the $[-10, 10] \times [-10, 10]$ region to $21^2 = 441$ points. At each point, a five-fold cross validation is conducted. The point with the best CV accuracy is chosen and used to predict the test data.

Problem	#inputs	#trg exs	#test exs	Test set error of Usual grid method	Test set error of Proposed method
banana	2	400	4900	0.1235 (6,-0)	0.1178 (-2,-2)
diabetes	8	468	300	0.2433 (4,7)	0.2433 (4,6)
image	18	1300	1010	0.02475 (9,4)	0.02475 (1,0.5)
splice	60	1000	2175	0.09701 (1,4)	0.1011 (0,4)
ringnorm	20	400	7000	0.01429(-2,2)	0.018 (-3,2)
twonorm	20	400	7000	0.031 (1,3)	0.02914 (1,4)
waveform	21	400	4600	0.1078 (0,3)	0.1078 (0,3)
tree	18	700	11692	0.1132 (8,4)	0.1246 (2,2)
adult	123	1605	29589	0.1614 (5,6)	0.1614 (5,6)
web	300	2477	38994	0.02223 (5,5)	0.02223 (5,5)

Table 3.1: Comparison of the model selection methods. For each approach, apart from the test error, the optimal $(\log C, \log\sigma^2)$ pair is also given.

For the proposed method, we search for $\tilde{C}$ by five-fold cross validation on Linear SVM using uniformly spaced $\log C$ values in $[-8, 2]$. Then we discretize $[-8, 8]$ as values of $\log\sigma^2$ and check all points satisfying $\log\sigma^2 = \log C - \log\tilde{C}$. Because now fewer points have to be tried, we use the smaller grid spacing of 0.5 for both discretizations. The total number of points tried is 54.

To empirically evaluate the usefulness of the proposed method, we consider several binary problems from (Rätsch 1999). For each problem (Rätsch 1999) gives 100 realizations of the given dataset into (training set, test set) partitions. We consider only the first of those realizations. In addition, the problem adult, from the UCI "adult" data set (Blake and Merz 1998) and the problem web, both as compiled by Platt (1998), are also included. For each of these two datasets also, there are several realizations. For our study here, we only consider the realization with the smallest training set; the full dataset with training data (including duplicated ones) removed is taken as the test set. For all datasets used, Table 3.1 gives the number of input variables, the number of training examples and the number of test examples. All datasets are directly used as given in the mentioned references, without any further normalization or scaling.

The SVM software LIBSVM (Chang and Lin 2001a) which implements a decomposition method is employed for solving (1.2). Table 3.1 presents the test error of the two methods as well as the corresponding chosen values of $\log C$ and $\log \sigma^2$. It can be clearly seen that the new method is very competitive with the usual method in terms of test set accuracy. For large datasets the proposed method has the great advantage that it checks much fewer points on the $(\log C, \log \sigma^2)$ plane and so the savings in computing time can be large.

Note that, in the chosen problems the following quantities have a reasonably wide range: test error (1.5% to 25%), the number of input variables (2 to 300) and the number of training examples (400 to 2477); and so, the empirical evaluation demonstrates the applicability of the proposed approach to different types of datasets.

A remaining issue is how to decide the range of $\log C$ for determining $\tilde{C}$ in step 1. From Table 3.1 we can see that $\log \tilde{C} = \log C - \log \sigma^2$ is usually not a large number. Furthermore, we observe that for all problems, after C is greater than a certain threshold, the cross validation accuracy of the Linear SVM is about the same. Therefore, if we start searching from small C values and go on to large C values, the search can be stopped after the CV accuracy stops varying much. An example of the variation of the five-fold CV accuracy of Linear SVM is given in Figure 3.

For Linear SVMs we can formally establish that there exists a finite limiting

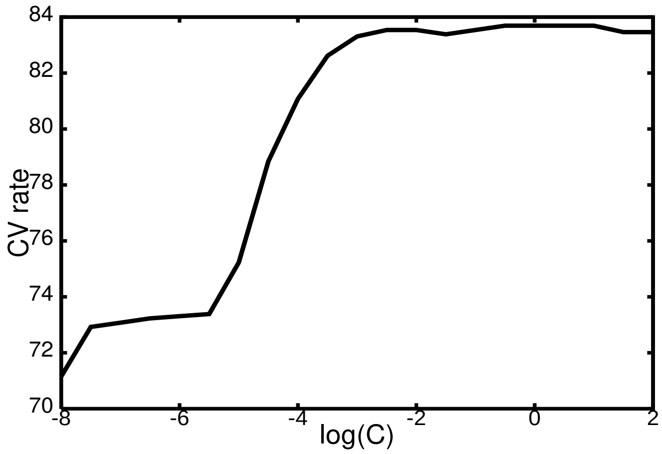

Figure 3: Variation of CV accuracy of Linear SVM with C for the image problem

value C^* such that, for $C \geq C^*$ the solution of the Linear SVM remains unchanged. If $\{x_i : y_i = 1\}$ and $\{x_i : y_i = -1\}$ are linearly separable then the above result is easy to appreciate; the same ideas used in case 2 of section 2 can be applied to show this. However, if $\{x_i : y_i = 1\}$ and $\{x_i : y_i = -1\}$ are not linearly separable (which is typically the case) the result is non-trivial to establish. Here, we prove the following theorem.

Theorem 3 *There exists a finite value C^* and (w^*, b^*) such that $(w, b) = (w^*, b^*)$ solves (1.2) $\forall C \geq C^*$. If this decision function is used, the loo error is same for all $C \geq C^*$. Moreover, this w^* is unique.*

Details of the proof are in Appendix A.4.

A Appendix

A.1 Two useful Lemmas

Lemma 2 *Consider an optimization problem with the form (1.2) and Q is a function of σ^2 (denoted as $Q(\sigma^2)$). Let $\alpha(\sigma^2)$ be its solutions. For a given number a, if*

$$Q^* \equiv \lim_{\sigma^2 \to a} Q(\sigma^2) \text{ exists,}$$

then there exists convergent sequence $\{\alpha(\sigma_k^2)\}$ with $\sigma_k^2 \to a$, and, the limit of any such sequence is an optimal solution of (1.2) with the Hessian matrix Q^*.

Moreover, if Q^* is positive definite, $\lim_{\sigma^2 \to a} \alpha(\sigma^2)$ exists.

Proof. The feasible region of (1.2) is independent of σ^2 so is compact. Then there exists convergent sequence $\{\alpha(\sigma_k^2)\}$ with $\lim_{k\to\infty} \sigma_k^2 = a$. For any one such sequence, we have

$$\begin{aligned} &\frac{1}{2}\alpha(\sigma_k^2)^T Q(\sigma_k^2)\alpha(\sigma_k^2) - e^T\alpha(\sigma_k^2) \le \frac{1}{2}(\alpha^*)^T Q(\sigma_k^2)\alpha^* - e^T\alpha^*, \text{ and} \\ &\frac{1}{2}(\alpha^*)^T Q^*\alpha^* - e^T\alpha^* \le \frac{1}{2}\alpha(\sigma_k^2)^T Q^*(\sigma_k^2)\alpha(\sigma_k^2) - e^T\alpha(\sigma_k^2), \end{aligned} \tag{A.1}$$

where α^* is any optimal solution of (1.2) with the Hessian matrix Q^*. If $\alpha(\sigma_k^2)$ goes to $\bar{\alpha}$, taking the limit of (A.1),

$$\frac{1}{2}(\alpha^*)^T Q^*\alpha^* - e^T\alpha^* = \frac{1}{2}\bar{\alpha}^T Q^*\bar{\alpha} - e^T\bar{\alpha}.$$

Thus, $\bar{\alpha}$ is an optimal solution too.

If Q^* is positive definite, (1.2) is a strictly convex problem with a unique optimal solution. This implies that $\lim_{\sigma^2 \to a} \alpha(\sigma^2)$ exists. □

Lemma 3 *If (1.2) has an optimal solution with at least one free variable (i.e., $0 < \alpha_i < C$ for at least one i), then the optimal b of (1.1) is unique.*

Proof.

The Karush-Kuhn-Tucker (KKT) condition, (i.e. the optimality condition) of (1.2) is: If α is an optimal solution, there is a number b and two nonnegative vectors λ and μ such that

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

where $\nabla F(\alpha) = Q\alpha - e$ is the gradient of $F(\alpha) = 1/2\alpha^T Q\alpha - e^T\alpha$. This can be rewritten as

$$\begin{aligned} \nabla F(\alpha)_i + by_i &\ge 0 \quad && \text{if } \alpha_i = 0, \\ \nabla F(\alpha)_i + by_i &\le 0 \quad && \text{if } \alpha_i = C, \\ \nabla F(\alpha)_i + by_i &= 0 \quad && \text{if } 0 < \alpha_i < C. \end{aligned} \tag{A.2}$$

Note that

$$\nabla F(\alpha)_i = y_i w^T z_i - 1$$

is independent of different optimal solutions α as the primal optimal solution w is unique.

Let (w, b, ξ) denote a primal solution. As already said, w is unique. By convexity of the solution set, the set of all possible b solutions, B is an interval. Once b is chosen, ξ is uniquely defined. By assumption there exists $b \in B$ and a corresponding Lagrange multiplier vector $\alpha(b)$ with a free alpha, say $0 < \alpha(b)_k < C$. Thus $\alpha(b)$ is an optimal solution of (1.2) and so, by (A.2), $\nabla F(\alpha)_k + by_k = 0$. Denote

$$A_0 = \{i \mid \nabla F(\alpha)_i + by_i > 0\}, \quad A_C = \{i \mid \nabla F(\alpha)_i + by_i < 0\}, \text{ and}$$
$$A_F = \{i \mid i \notin A_0 \cup A_C\}.$$

Let us define $2e$ and $2f$ to be the minimum and maximum of the following set:

$$\{b^{new} \mid \nabla F(\alpha)_i + b^{new} y_i > 0 \text{ if } i \in A_0; \nabla F(\alpha)_i + b^{new} y_i < 0 \text{ if } i \in A_C\}. \quad \text{(A.3)}$$

Clearly $e < b < f$. Suppose B is not a singleton. Now choose $b^{new} \in B \cap [e, f]$ such that $b^{new} \neq b$. Let $\alpha(b^{new})$ be any Lagrange multiplier corresponding to b^{new}. Thus $\alpha(b^{new})$ and b^{new} satisfy (A.2). Suppose $b^{new} > b$ and $y_k = 1$. Then

$$\nabla F(\alpha)_k + b^{new} y_k > 0 \text{ so } \alpha(b^{new})_k = 0 < \alpha(b)_k. \quad \text{(A.4)}$$

If we use (A.2) as applied to $(b, \alpha(b))$ and $(b^{new}, \alpha(b^{new}))$, (A.3) implies the following: $\alpha(b)_i = \alpha(b^{new})_i \ \forall\, i \in A_0 \cup A_C$ with $y_i = y_k$; also, $\alpha(b)_i \geq \alpha(b^{new})_i \ \forall\, i \in A_F$ with $y_i = y_k$. Note that k is an element of this second group. Thus, with (A.4),

$$\sum_{i:y_i=y_k} \alpha(b)_i > \sum_{i:y_i=y_k} \alpha(b^{new})_i. \quad \text{(A.5)}$$

This is a violation of the fact that both $\alpha(b)$ and $\alpha(b^{new})$ are solutions of (1.2) since, for a given dual solution α, dual cost is $(\|w\|^2/2) - 2\sum_{i:y_i=1} \alpha_i$ and the first term is same for $\alpha(b)$ as well as $\alpha(b^{new})$. If $y_k = -1$, the proof is the same but (A.5) becomes $\sum_{i:y_i=y_k} \alpha(b)_i < \sum_{i:y_i=y_k} \alpha(b^{new})_i$. A similar contradiction can be reached if $b^{new} < b$. Thus B is a singleton and b is unique. $\square$

A.2 Optimal solution of (2.5)

Karush-Kuhn-Tucker (KKT) conditions applied to (2.5) correspond to the existence of a scalar b and two nonnegative vectors λ and μ such that

$$\begin{aligned} &\alpha_i - 1 + by_i = \lambda_i - \mu_i, \\ &\alpha_i \lambda_i = 0, (C - \alpha_i)\mu_i = 0, i = 1, \ldots, l. \end{aligned} \tag{A.6}$$

To show that the solution is given by (2.6) and (2.7), all that we need to do is to show the existence of λ and μ so that (A.6) holds. For the solution (2.6), when $C \geq C_{\text{lim}}$, using b defined in (2.7),

$$\alpha_i - 1 + by_i = 0 \ \ \forall\, i$$

so we can simply choose $\lambda = \mu = 0$ so that (A.6) is satisfied.

If $C < C_{\text{lim}}$,

$$\alpha_i - 1 + by_i = \begin{cases} 0 & \text{if } y_i = 1, \\ \frac{Cl}{l_1} - 2 \leq 0 & \text{if } y_i = -1, \end{cases}$$

so (A.6) also holds, Therefore, (2.6) gives an optimal solution for (2.5).

Let us now analyze properties of the classifier function f associated with (2.5). Note using (2.7) that $b > 0$. For $x \neq x_i$, $\exp(-\|x - x_i\|^2/\sigma^2) \to 0$ as $\sigma^2 \to 0$. Therefore, for such x, the classifier function corresponding to (2.5) is given by $f(x) = b$. Since $b > 0$, all points x not in the training set are classified as class 1 irrespective of the value of C. This together with item (2) of Assumption 1 implies that the number of loo errors is equal to l_2.

For a training point x_i we have $f(x_i) \to y_i\alpha_i + b$ as $\sigma^2 \to 0$. Thus after σ^2 is sufficiently small, all class 1 training points are classified correctly by f. For training points x_i in class 2, we can use (2.6) and (2.7) to show that: (i) for $C > C_{\text{lim}}/2$, all of those points are classified correctly by f; and, (ii) for $C \leq C_{\text{lim}}/2$, all of those points are classified incorrectly by f.

A.3 Proof of Theorem 2

To prove Theorem 2, first we write down the primal form of (2.11):

$$\begin{aligned} \min_{\tilde{w},\tilde{b},\tilde{\xi}} \quad & \frac{1}{2}\tilde{w}^T\tilde{w} + \tilde{C}\sum_{i=1}^{l}\tilde{\xi}_i \\ \text{subject to} \quad & y_i(\tilde{w}^T\tilde{\phi}(x_i)+\tilde{b}) \geq 1-\tilde{\xi}_i, \\ & \tilde{\xi}_i \geq 0, i=1,\dots,l, \end{aligned} \tag{A.7}$$

where $\tilde{\phi}(x) \equiv \sigma\phi(x)$.[‖] By defining $w \equiv \sigma\tilde{w}$, multiplying the objective function of (A.7) by σ^2, and using (2.12), (A.7) has exactly the same form as (1.1), so we can say

$$b = \tilde{b}, \xi = \tilde{\xi}, \text{ and } \tilde{w}^T\tilde{\phi}(x) + \tilde{b} = w^T\phi(x) + b. \tag{A.8}$$

A difficulty in proving this theorem is that the solution of (A.7) is an element of a vector space that is different from that of a solution of (2.14). Hence, to build the relation as $\sigma^2 \to \infty$ we will consider their duals using Lemma 2.

Assume $\tilde{\alpha}(\sigma^2)$ is the solution of (2.11) under a given $\tilde{C}$. It is in a bounded region for all $\sigma^2 > 0$ so there is a convergent sequence $\tilde{\alpha}(\sigma_k^2) \to \tilde{\alpha}$ as $\sigma_k^2 \to \infty$. We can apply Lemma 2 as now the ij component of the Hessian of (2.11) is a function of σ^2:

$$y_iy_j\tilde{K}_{ij} = \sigma^2 y_iy_j\big(e^{-\|x_i-x_j\|/(2\sigma^2)} - 1 + \frac{\|x_i\|^2}{2\sigma^2} + \frac{\|x_j\|^2}{2\sigma^2}\big)$$

with the limit $y_iy_jx_i^Tx_j$ as $\sigma^2 \to \infty$. Therefore, $\tilde{\alpha}(\sigma_k^2)$ converges to an optimal solution $\tilde{\alpha}$ of (2.13).

We denote that $\tilde{w}(\sigma^2)$ and $\tilde{w}$ are unique optimal solutions of (A.7) and (2.14), respectively. Then, for any such convergent sequence $\{\tilde{\alpha}(\sigma_k^2)\}_{k=1}^{\infty}$, using $y^T\tilde{\alpha}(\sigma_k^2) =$

[‖] It should be pointed out that (2.11) is not directly the dual of (A.7). The dual of (A.7) reduces to (2.11) when the $y^T\tilde{\alpha} = 0$ constraint is used.

0, we have that for any x,

$$\begin{aligned}
& \lim_{\sigma_k^2\to\infty} \tilde{w}(\sigma_k^2)^T\tilde{\phi}(x) \\
= & \lim_{\sigma_k^2\to\infty} \sum_{i=1}^{l} y_i\tilde{\alpha}(\sigma_k^2)_i\tilde{\phi}(x_i)^T\tilde{\phi}(x) \\
= & \lim_{\sigma_k^2\to\infty} \sum_{i=1}^{l} y_i\tilde{\alpha}(\sigma_k^2)_i\big(\sigma_k^2 - \|x_i\|^2/2 + x_i^Tx - \|x\|^2/2\big) \\
= & \lim_{\sigma_k^2\to\infty} \sum_{i=1}^{l} y_i\tilde{\alpha}(\sigma_k^2)_i\big(-\|x_i\|^2/2 + x_i^Tx\big) && \text{(A.9)} \\
= & \sum_{i=1}^{l} y_i\tilde{\alpha}_i x_i^Tx + d(\tilde{\alpha}) = \tilde{w}^Tx + d(\tilde{\alpha}), && \text{(A.10)}
\end{aligned}$$

where (A.9) follows from (2.8) and $y^T\tilde{\alpha}(\sigma_k^2) = 0$ and $d(\tilde{\alpha}) \equiv -\sum_{i=1}^{l} y_i\tilde{\alpha}\|x_i\|^2/2$. By a similar way we can prove

$$\begin{aligned}
& \lim_{\sigma_k^2\to\infty} \tilde{w}(\sigma_k^2)^T\tilde{w}(\sigma_k^2) \\
= & \lim_{\sigma_k^2\to\infty} \sum_{i=1}^{l}\sum_{j=1}^{l} \tilde{\alpha}(\sigma_k^2)_i\tilde{\alpha}(\sigma_k^2)_j y_i y_j \tilde{\phi}(x_i)^T\tilde{\phi}(x_j) \\
= & \sum_{i=1}^{l}\sum_{j=1}^{l} \tilde{\alpha}(\sigma_k^2)_i\tilde{\alpha}(\sigma_k^2)_j y_i y_j x_i^Tx_j = \tilde{w}^T\tilde{w}. && \text{(A.11)}
\end{aligned}$$

Note that (A.11) follows from the discussion between (2.8) and (2.11). The last equality is via $\tilde{w} = \sum_{i=1}^{l}\tilde{\alpha}_o x_i$ as $\tilde{w}$ is the optimal solution of (2.11).

Next we consider that $(\tilde{w}, \tilde{b})$ is the unique optimal solution of (2.14). The constraints of (A.7) imply that

$$\max_{y_i=1}\{1 - \tilde{w}(\sigma^2)^T\tilde{\phi}(x_i) - \tilde{\xi}(\sigma^2)_i\} \le \tilde{b}(\sigma^2) \le \max_{y_i=-1}\{-1 - \tilde{w}(\sigma^2)^T\tilde{\phi}(x_i) + \tilde{\xi}(\sigma^2)_i\}.$$

Note that the primal-dual optimality condition implies

$$0 \le \tilde{\xi}(\sigma^2)_i \le \sum_{i=1}^{l}\tilde{\xi}(\sigma^2)_i \le \frac{e^T\tilde{\alpha}(\sigma^2)}{\tilde{C}} \le l.$$

With (A.9) and the assumption $l_1 \ge 1$ and $l_2 \ge 1$, after σ^2 is large enough, $\tilde{b}(\sigma^2)$ is in a bounded region. When $(\tilde{w}(\sigma^2), \tilde{b}(\sigma^2))$ is optimal for (A.7), the optimal $\tilde{\xi}(\sigma^2)$ is

$$\tilde{\xi}(\sigma^2)_i \equiv \max(0, 1 - y_i(\tilde{w}(\sigma^2)^T\tilde{\phi}(x_i) + \tilde{b}(\sigma^2))).$$

For any convergent sequence $\tilde{b}(\sigma_k^2) \to b^*$ with $\sigma_k^2 \to \infty$, we can further have a subsequence such that $\{\tilde{\alpha}(\sigma_k^2)\}$ converges. Thus, we can consider any such sequence with both properties. Then, (A.10) implies

$$\tilde{\xi}(\sigma_k^2)_i \to \xi_i^* = \max(0, 1 - y_i(\tilde{w}^T x_i + d(\tilde{\alpha}) + b^*)). \tag{A.12}$$

Hence, $(\tilde{w}, b^* + d(\tilde{\alpha}), \xi^*)$ is feasible for (2.14). By defining

$$\bar{\xi}(\sigma_k^2)_i \equiv \max(0, 1 - y_i(\tilde{w}(\sigma_k^2)^T \tilde{\phi}(x_i) - d(\tilde{\alpha}) + \tilde{b})),$$

$(\tilde{w}(\sigma_k^2), \tilde{b} - d(\tilde{\alpha}), \bar{\xi}(\sigma_k^2))$ is feasible for (A.7). In addition, using (A.10),

$$\tilde{\xi} \equiv \lim_{\sigma_k^2 \to \infty} \bar{\xi}(\sigma_k^2) = \max(0, 1 - y_i(\tilde{w}^T x_i + \tilde{b})), \tag{A.13}$$

so $(\tilde{w}, \tilde{b}, \tilde{\xi})$ is optimal for (2.14). Thus,

$$\frac{1}{2}\tilde{w}^T\tilde{w} + \tilde{C}\sum_{i=1}^{l}\tilde{\xi}_i \leq \frac{1}{2}\tilde{w}^T\tilde{w} + \tilde{C}\sum_{i=1}^{l}\xi_i^*, \text{ and}$$

$$\frac{1}{2}\tilde{w}(\sigma_k^2)^T\tilde{w}(\sigma_k^2) + \tilde{C}\sum_{i=1}^{l}\tilde{\xi}(\sigma_k^2)_i \leq \frac{1}{2}\tilde{w}(\sigma_k^2)^T\tilde{w}(\sigma_k^2) + \tilde{C}\sum_{i=1}^{l}\bar{\xi}(\sigma_k^2)_i. \tag{A.14}$$

With (A.11), (A.12), and (A.13), taking the limit (A.14) becomes

$$\frac{1}{2}\tilde{w}^T\tilde{w} + \tilde{C}\sum_{i=1}^{l}\xi_i^* \leq \frac{1}{2}\tilde{w}^T\tilde{w} + \tilde{C}\sum_{i=1}^{l}\tilde{\xi}_i.$$

Therefore, we have that $(\tilde{w}, b^* + d(\tilde{\alpha}), \xi^*)$ is optimal for (2.14). Since $\tilde{b}$ is unique by assumption,

$$b^* + d(\tilde{\alpha}) = \tilde{b}. \tag{A.15}$$

Now we are ready to prove the main result (2.15). If it is wrong, there is $\epsilon > 0$ and a sequence $\{\tilde{w}(\sigma_k^2)\}$ with $\sigma_k^2 \to \infty$ such that

$$|\tilde{w}(\sigma_k^2)^T\tilde{\phi}(x) + b(\sigma_k^2) - \tilde{w}^T x - \tilde{b}| \geq \epsilon, \forall k. \tag{A.16}$$

Since we can find an infinite subset K such that $\lim_{k \in K, \sigma_k^2 \to \infty} \tilde{b}(\sigma_k^2) = b^*$ and (A.10) holds, with $b(\sigma^2) = \tilde{b}(\sigma^2)$ from (A.8), the above analysis (i.e., (A.10) and (A.15)) shows that

$$\begin{aligned} & \lim_{k \in K, \sigma_k^2 \to \infty} w(\sigma_k^2)^T\phi(x) + b(\sigma_k^2) \\ = \; & \tilde{w}^T x + d(\tilde{\alpha}) + \tilde{b} - d(\tilde{\alpha}) \\ = \; & \tilde{w}^T x + \tilde{b}. \end{aligned}$$

This contradicts (A.16) so (2.15) is valid.

Therefore, if $\tilde{w}^T x + \tilde{b} \neq 0$, after σ^2 is sufficiently large,

$$\operatorname{sgn}(w(\sigma^2)^T \phi(x) + b(\sigma^2)) = \operatorname{sgn}(\tilde{w}^T x + \tilde{b}).$$

A.4 Proof of Theorem 3

Let α^1 be a feasible vector of (1.2) for $C = C_1$ and α^2 be a feasible vector of (1.2) for $C = C_2$. We say that α^1 and α^2 are on the same face if the following hold: (i) $\{i \mid 0 < \alpha_i^1 < C_1\} = \{i \mid 0 < \alpha_i^2 < C_2\}$; (ii) $\{i \mid \alpha_i^1 = C_1\} = \{i \mid \alpha_i^2 = C_2\}$; and, (iii) $\{i \mid \alpha_i^1 = 0\} = \{i \mid \alpha_i^2 = 0\}$. To prove Theorem 3 we need the following result.

Lemma 4 *If $C_1 < C_2$ and their corresponding duals have optimal solutions at the same face, then for any $C_1 \leq C \leq C_2$, there is at least one optimal solution at the same face. Furthermore, there are optimal solutions α and b of (1.2) which form linear functions of C in $[C_1, C_2]$.*

Proof of Lemma 4:

If α^1 and α^2 are optimal solutions at the same face corresponding to C_1 and C_2, then they satisfy the following KKT conditions, respectively:

$$Q\alpha^1 - e + b^1 y = \lambda^1 - \mu^1, \lambda_i^1 \alpha_i^1 = 0, (C^1 - \alpha_i^1)\mu_i^1 = 0,$$
$$Q\alpha^2 - e + b^2 y = \lambda^2 - \mu^2, \lambda_i^2 \alpha_i^2 = 0, (C^2 - \alpha_i^2)\mu_i^2 = 0.$$

Since they are at the same face,

$$\begin{aligned} &\lambda_i^2 \alpha_i^1 = 0, \lambda_i^1 \alpha_i^2 = 0, \\ &(C^1 - \alpha_i^1)\mu_i^2 = 0, (C^2 - \alpha_i^2)\mu_i^1 = 0. \end{aligned} \tag{A.17}$$

As $C_1 \leq C \leq C_2$, we can have $0 \leq \tau \leq 1$ such that

$$C = \tau C_1 + (1 - \tau) C_2. \tag{A.18}$$

Let

$$\begin{aligned} &\alpha \equiv \tau\alpha^1 + (1-\tau)\alpha^2, \lambda \equiv \tau\lambda^1 + (1-\tau)\lambda^2, \\ &\mu \equiv \tau\mu^1 + (1-\tau)\mu^2, b \equiv \tau b^1 + (1-\tau) b^2. \end{aligned} \tag{A.19}$$

Then α, λ, μ, b satisfy the KKT condition at C:

$$Q\alpha - e + by = \lambda - \mu, \lambda_i \alpha_i = 0, (C - \alpha_i)\mu_i = 0,$$
$$0 \le \alpha_i \le C, \lambda_i \ge 0, \mu_i \ge 0, y^T \alpha = 0.$$

Using (A.18),

$$\tau = \frac{C - C_2}{C_1 - C_2}.$$

Putting it into (A.19), α and b are linear functions of C where $C \in [C_1, C_2]$. This proves the lemma. □

Let us now prove Theorem 3. As we already mentioned, if the points of the two classes are linearly separable in x space then the proof of the result is straightforward. So let us only give a proof for the case of linearly non-separable points. Since the number of faces is finite, by Lemma 4 there exists a C^* such that for $C \ge C^*$, there are optimal solutions at the same face. For the rest of the proof let us only consider optimal solutions on a single face.

For any $C_1 > C^*$, Lemma 4 implies that there are optimal solutions α and b which form linear functions of C in the interval $[C^*, C_1]$. Since

$$\sum_{i=1}^{l} \xi_i = \sum_{i:\alpha_i = C} -[(Q\alpha)_i - 1 + by_i], \tag{A.20}$$

$\sum_{i=1}^{l} \xi_i$ is a linear function of C in this interval and can be represented as

$$\sum_{i=1}^{l} \xi_i = AC + B, \tag{A.21}$$

where A and B are constants. If we consider another $C_2 > C_1$, $\sum_{i=1}^{l} \xi_i$ is also a linear function of C in $[C^*, C_2]$. For each C, the optimal $\frac{1}{2} w^T w$ as well as $\sum_{i=1}^{l} \xi_i$ are unique. Thus, the two linear functions have the same values at more than two points, so they are indeed identical. Therefore, (A.21) holds for any $C \ge C^*$.

Since $\sum_{i=1}^{l} \xi_i$ is a decreasing function of C (e.g., using techniques similar to (Chang and Lin 2001b, Lemma 4)), $A \le 0$. However, A cannot be negative as otherwise $\sum_{i=1}^{l} \xi_i$ goes to $-\infty$ as C increases. Hence, $A = 0$ and so $\sum_{i=1}^{l} \xi_i$ is a constant for $C \ge C^*$.

If (w^1, b^1, ξ^1) and (w^2, b^2, ξ^2) are optimal solutions at $C = C_1$ and C_2, respectively, then

$$\frac{1}{2}(w^1)^T w^1 + C_1 \sum_{i=1}^{l} \xi_i^1 \leq \frac{1}{2}(w^2)^T w^2 + C_1 \sum_{i=1}^{l} \xi_i^2$$

and

$$\frac{1}{2}(w^2)^T w^2 + C_2 \sum_{i=1}^{l} \xi_i^2 \leq \frac{1}{2}(w^1)^T w^1 + C_2 \sum_{i=1}^{l} \xi_i^1$$

imply that $(w^1)^T w^1 = (w^2)^T w^2$. That is, $\alpha^T Q \alpha$ is a constant for $C \geq C^*$.

Therefore (w^2, b^2, ξ^2) is also feasible and optimal when $C = C_1$. Since the solution of w is unique (e.g., (Lin 2001, Lemma 1)), $w^1 = w^2$.

If $F = \{i \mid 0 < \alpha_i < C\} \neq \emptyset$, there is x_i such that $w^T x_i + b = y_i$. Hence $b^1 = b^2$ and so the decision functions as well as the loo rate are the same for $C \geq C^*$.

On the other hand, if $F = \emptyset$ and we denote $\alpha(C)$ the solution of (1.2) at a given C, then $\alpha(C) = (C/C^*)\alpha(C^*)$ for all $C \geq C^*$. As $w^T w = \alpha^T Q \alpha$ becomes a constant, we have $w = 0$ after $C \geq C^*$. However, since $F = \emptyset$, the optimal b might not be unique under the same C. For any one of such b, (w, b) is optimal for (1.2) for all $C \geq C^*$.

Finally, since $w^T w$ becomes a constant, for $C \geq C^*$, the solution of (1.2) is also a solution of

$$\begin{aligned} \min_{w,b,\xi} \quad & \sum_{i=1}^{l} \xi_i \\ \text{subject to} \quad & y_i(w^T x_i + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, \ldots, l. \end{aligned}$$

This completes the proof of Theorem 3. □

Acknowledgments

The second author was partially supported by the National Science Council of Taiwan via the grant NSC 90-2213-E-002-111.

References

Blake, C. L. and C. J. Merz (1998). UCI repository of machine learning databases. Technical report, University of California, Department of Information and Computer Science, Irvine, CA. Available at `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

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

Chang, C.-C. and C.-J. Lin (2001b). Training ν-support vector classifiers: Theory and algorithms. *Neural Computation 13*(9), 2119–2147.

Joachims, T. (2000). Estimating the generalization performance of a SVM efficiently. In *Proceedings of the International Conference on Machine Learning*, San Francisco. Morgan Kaufman.

Lee, J.-H. (2001). Model selection of the bounded SVM formulation using the RBF kernel. Master's thesis, Department of Computer Science and Information Engineering, National Taiwan University.

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

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

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.

Rätsch, G. (1999). Benchmark data sets. Available at `http://ida.first.gmd.de/~raetsch/data/benchmarks.htm`.

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

Vapnik, V. and O. Chapelle (2000). Bounds on error expectation for support vector machines. *Neural Computation 12*(9), 2013–2036.