Manuscript Number: 2833

Leave-one-out Bounds for Support Vector Regression Model Selection

Ming-Wei Chang and Chih-Jen Lin
Department of Computer Science and
Information Engineering
National Taiwan University
Taipei 106, Taiwan
cjlin@csie.ntu.edu.tw

Abstract

Minimizing bounds of leave-one-out (loo) errors is an important and efficient approach for support vector machine (SVM) model selection. Past research focuses on their use for classification but not regression. In this article, we derive various loo bounds for support vector regression (SVR) and discuss the difference from those for classification. Experiments demonstrate that the proposed bounds are competitive with Bayesian SVR for parameter selection. We also discuss the differentiability of loo bounds.

1 Introduction

Recently, support vector machines (Boser, Guyon, and Vapnik 1992; Cortes and Vapnik 1995) have been a promising tool for classification and regression. Its success depends on the tuning of several parameters which affect the generalization error. A popular approach is to approximate the error by a bound that is a function of parameters. Then, we search for parameters so that this bound is minimized. Past efforts focus on such bounds for classification, and the aim of this paper is to derive bounds for regression.

We first briefly introduce support vector regression (SVR). Given training vectors $\mathbf{x}_i \in R^n, i = 1, \ldots, l$, and a vector $\mathbf{y} \in R^l$ as their target values, SVR solves

$$\begin{aligned} \min_{\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{2}\sum_{i=1}^{l}\xi_i^2 + \frac{C}{2}\sum_{i=1}^{l}(\xi_i^*)^2 \\ \text{subject to} \quad & -\epsilon - \xi_i^* \le \mathbf{w}^T\phi(\mathbf{x}_i) + b - y_i \le \epsilon + \xi_i, i = 1, \ldots, l. \end{aligned} \tag{1.1}$$

Data are mapped to a higher dimensional space by the function ϕ and an ϵ-insensitive loss function is used. We refer to this form as L2-SVR as a two-norm penalty term

$\xi_i^2 + (\xi_i^*)^2$ is used. As $\mathbf{w}$ may be a huge vector variable after introducing the mapping function ϕ, practically we solve the dual problem:

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \quad & \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T \tilde{K}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \\ & +\epsilon \sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \\ \text{subject to} \quad & \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0, 0 \le \alpha_i, \alpha_i^*, i = 1, \ldots, l, \end{aligned} \tag{1.2}$$

where $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ is the kernel function. $\tilde{K} = K + I/C$ and I is the identity matrix. For optimal $\mathbf{w}$ and $(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)$, the primal-dual relationship shows

$$\mathbf{w} = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)\phi(\mathbf{x}_i),$$

so the approximate function is

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{w}^T\phi(\mathbf{x}) + b \\ &= -\sum_{i=1}^{l} K(\mathbf{x}, \mathbf{x}_i)(\alpha_i - \alpha_i^*) + b. \end{aligned} \tag{1.3}$$

More general information about SVR can be found in the tutorial by Smola and Schölkopf (2004).

One difficulty over classification for parameter selection is that SVR possesses an additional parameter ϵ. Therefore, the search space of parameters is bigger than that for classification. Some work have tried to address SVR parameter selection. In (Momma and Bennett 2002), the authors perform model section by pattern search, so the number of parameters checked is smaller than that by a full grid search. In (Kwok 2001) and (Smola, Murata, Schölkopf, and Müller 1998), the authors analyze the behavior of ϵ and conclude that the optimal ϵ scales linearly with the input noise of the training data. However, this property can be applied only when the noise is known. Gao, Gunn, Harris, and Brown (2002) derive a Bayesian framework for SVR which leads to minimize a function of parameters. However, its performance is not very good compared to a full grid search (Lin and Weng 2004). An improvement is in (Chu, Keerthi, and Ong 2003), which modified the standard SVR formulation. This improved Bayesian SVR will be compared to our approach in this article.

This paper is organized as follows. Section 2 briefly reviews loo bounds for support vector classification. We derive various loo bounds for SVR in Section 3. Implemen-

tation issues are in Section 4 and experiments are in Section 5. Conclusions are in Section 6.

2 Leave-one-out Bounds for Classification: a Review

Given training vectors $\mathbf{x}_i \in R^n, i = 1, \ldots, l$, and a vector $\mathbf{y} \in R^l$ such that $y_i \in \{1, -1\}$, an SVM formulation for two-class classification is:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{2}\sum_{i=1}^{l}\xi_i^2 \\ \text{subject to} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, \ldots, l. \end{aligned} \tag{2.1}$$

Next, we briefly review two loo bounds.

2.1 Radius Margin (RM) Bound for Classification

By defining

$$\tilde{\mathbf{w}} \equiv \begin{bmatrix} \mathbf{w} \\ \sqrt{C}\boldsymbol{\xi} \end{bmatrix}, \tag{2.2}$$

Vapnik and Chapelle (2000) have shown that the following radius margin (RM) bound holds:

$$loo \leq 4\tilde{R}^2\|\tilde{\mathbf{w}}\|^2 = 4\tilde{R}^2\mathbf{e}^T\boldsymbol{\alpha}, \tag{2.3}$$

where loo is the number of loo errors and $\mathbf{e}$ is a vector of all ones. In (2.3), $\boldsymbol{\alpha}$ is the solution of the following dual problem:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \mathbf{e}^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T(Q + \frac{I}{C})\boldsymbol{\alpha} \\ \text{subject to} \quad & 0 \leq \alpha_i, i = 1, \ldots, l, \\ & \mathbf{y}^T\boldsymbol{\alpha} = 0, \end{aligned} \tag{2.4}$$

where $Q_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$. At optimum, $\|\tilde{\mathbf{w}}\|^2 = \mathbf{e}^T\boldsymbol{\alpha}$. Define

$$\tilde{\phi}(\mathbf{x}_i) \equiv \begin{bmatrix} \phi(\mathbf{x}_i) \\ \frac{\mathbf{e}_i}{\sqrt{C}} \end{bmatrix},$$

where $\mathbf{e}_i$ is a zero vector of length l except the ith component is one. Then $\tilde{R}$ in (2.3) is the radius of the smallest sphere containing all $\tilde{\phi}(\mathbf{x}_i), i = 1, \ldots, l$.

The right-hand side of (2.3) is a function of parameters, which will then be minimized for parameter selection.

2.2 Span Bound for Classification

Span bound, another loo bound proposed in (Vapnik and Chapelle 2000), is tighter than the RM bound. Define S_t^2 as the optimal objective value of following problem:

$$\begin{array}{rl} \min\limits_{\boldsymbol{\lambda}} & \|\tilde{\phi}(\mathbf{x}_t) - \sum_{i\in\mathcal{F}\setminus\{t\}} \lambda_i \tilde{\phi}(\mathbf{x}_i)\|^2 \\ \text{subject to} & \sum_{i\in\mathcal{F}\setminus\{t\}} \lambda_i = 1, \end{array} \tag{2.5}$$

where $\mathcal{F} = \{i \mid \alpha_i > 0\}$ is the index set of free components of an optimal $\boldsymbol{\alpha}$ of (2.4). Under the assumption that the set of support vectors remains the same during the leave-one-out procedure, the span bound is:

$$\sum_{t=1}^{l} \alpha_t S_t^2. \tag{2.6}$$

(2.5) indicates that S_t is smaller than $2\tilde{R}$, the diameter of the smallest sphere containing all $\tilde{\phi}(\mathbf{x}_i)$. Thus, (2.6) is tighter than (2.3).

Unfortunately, S_t^2 is not a continuous function (Chapelle, Vapnik, Bousquet, and Mukherjee 2002), so a modified span bound is proposed:

$$\sum_{t=1}^{l} \alpha_t \tilde{S}_t^2, \tag{2.7}$$

where $\tilde{S}_t^2$ is the optimal objective value of

$$\begin{array}{rl} \min\limits_{\boldsymbol{\lambda}} & \|\tilde{\phi}(\mathbf{x}_t) - \sum_{i\in\mathcal{F}\setminus\{t\}} \lambda_i \tilde{\phi}(\mathbf{x}_i)\|^2 + \eta \sum_{i\in\mathcal{F}\setminus\{t\}} \lambda_i^2 \frac{1}{\alpha_i} \\ \text{subject to} & \sum_{i\in\mathcal{F}\setminus\{t\}} \lambda_i = 1. \end{array} \tag{2.8}$$

η is a positive parameter that controls the smoothness of the bound. From (2.5) and (2.8), $S_t^2 \le \tilde{S}_t^2$, so (2.7) is also an loo bound. Define D as an $l \times l$ diagonal matrix where $D_{ii} = \eta/\alpha_i$ and $D_{ij} = 0$ for $i \neq j$. Define a new kernel matrix $\tilde{K}$ with

$$\tilde{K}(\mathbf{x}_i, \mathbf{x}_j) = \tilde{\phi}(\mathbf{x}_i)^T \tilde{\phi}(\mathbf{x}_j).$$

We let

$$\tilde{D} = \begin{bmatrix} D_{\mathcal{F}\mathcal{F}} & \mathbf{0}_{\mathcal{F}} \\ \mathbf{0}_{\mathcal{F}}^T & 0 \end{bmatrix} \text{ and } M = \begin{bmatrix} \tilde{K}_{\mathcal{F}\mathcal{F}} & \mathbf{e}_{\mathcal{F}} \\ \mathbf{e}_{\mathcal{F}}^T & 0 \end{bmatrix}, \tag{2.9}$$

where $\tilde{K}_{\mathcal{F}\mathcal{F}}$ is the sub-matrix of $\tilde{K}$ corresponding to free support vectors and $\mathbf{e}_{\mathcal{F}}$ ($\mathbf{0}_{\mathcal{F}}$) is a vector of $|\mathcal{F}|$ ones (zeros). By defining

$$\tilde{M} = M + \tilde{D}, \tag{2.10}$$

Chapelle, Vapnik, Bousquet, and Mukherjee (2002) showed that

$$\tilde{S}_t^2 = \tilde{K}(\mathbf{x}_t, \mathbf{x}_t) - \mathbf{h}^T(\tilde{M}^t)^{-1}\mathbf{h} = \frac{1}{(\tilde{M}^{-1})_{tt}} - \tilde{D}_{tt}, \tag{2.11}$$

where $\tilde{M}^t$ is the sub-matrix of $\tilde{M}$ with the tth column and row removed and $\mathbf{h}$ is the tth column of $\tilde{M}$ excluding $\tilde{M}_{tt}$.

Note that Chapelle, Vapnik, Bousquet, and Mukherjee (2002) did not give a formal proof on the continuity of (2.7). We address this issue in Section 4.1.

3 Leave-one-out Bounds for Regression

First, the Karash-Kunh-Tucker (KKT) optimality condition of (1.2) is listed here for further analysis: A vector $\boldsymbol{\alpha}$ is optimal for (1.2) if and only if it satisfies constraints of (1.2) and there is a scalar b such that

$$\begin{aligned} -(\tilde{K}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*))_i + b &= y_i + \epsilon, && \text{if } \alpha_i > 0, \\ -(\tilde{K}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*))_i + b &= y_i - \epsilon, && \text{if } \alpha_i^* > 0, \\ y_i - \epsilon \le -(\tilde{K}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*))_i + b &\le y_i + \epsilon, && \text{if } \alpha_i = \alpha_i^* = 0, \end{aligned} \tag{3.1}$$

where $(\tilde{K}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*))_i$ is the ith element of $\tilde{K}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)$. From (3.1), $\alpha_i \alpha_i^* = 0$ when the KKT condition holds. General discussion about KKT conditions can be seen in optimization books (e.g., (Bazaraa, Sherali, and Shetty 1993)).

3.1 Radius Margin Bound for Regression

To study the loo error for SVR, we introduce the leave-one-out problem without the tth data:

$$\begin{aligned} \min_{\mathbf{w}^t, b^t, \boldsymbol{\xi}^t, \boldsymbol{\xi}^{t*}} \quad & \frac{1}{2}(\mathbf{w}^t)^T(\mathbf{w}^t) + \frac{C}{2}\sum_{i \ne t}(\xi_i^t)^2 + \frac{C}{2}\sum_{i \ne t}(\xi_i^{t*})^2 \\ \text{subject to} \quad & -\epsilon - \xi_i^{t*} \le (\mathbf{w}^t)^T\phi(\mathbf{x}_i) + b^t - y_i \le \epsilon + \xi_i^t, \\ & i = 1, \ldots, t-1, t+1, \ldots, l. \end{aligned} \tag{3.2}$$

Though $\boldsymbol{\xi}^t$ and $\boldsymbol{\xi}^{t*}$ are vectors with $l-1$ elements, we define $\xi_t^t = \xi_t^{t*} = 0$ to make them have l elements. The approximate function of (3.2) is

$$f^t(\mathbf{x}) = (\mathbf{w}^t)^T\phi(\mathbf{x}) + b^t,$$

so the loo error for SVR is defined as

$$loo \equiv \sum_{t=1}^{l} |f^t(\mathbf{x}_t) - y_t|. \tag{3.3}$$

The loo error is well defined if the approximation function is unique. Note that though $\mathbf{w}$ (and $\mathbf{w}^t$) is unique due to the strictly convex term $\mathbf{w}^T\mathbf{w}$ (and $(\mathbf{w}^t)^T\mathbf{w}^t$), multiple b (or b^t) is possible (see, for example, the discussion in (Lin 2001a)). Therefore, we make the following assumption:

Assumption 1 (1.2) *and the dual problem of* (3.2) *all have free support vectors.*

We say a dual SVR has free support vectors if $(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)$ is optimal and there are some i such that $\alpha_i > 0$ or $\alpha_i^* > 0$ (i.e., $\alpha_i + \alpha_i^* > 0$ as $\alpha_i \alpha_i^* = 0$). Under this assumption, (1.3) and (3.1) imply $\mathbf{w}^T\phi(\mathbf{x}_i) + b = y_i + \epsilon$ if $\alpha_i > 0$ (or $= y_i - \epsilon$ if $\alpha_i^* > 0$). As the optimal $\mathbf{w}$ is unique, so is b. Similarly, b^t is unique as well. We then introduce a useful lemma:

Lemma 1 *Under Assumption 1,*

1. *If* $\alpha_t > 0$, $f^t(\mathbf{x}_t) \geq y_t$.
2. *If* $\alpha_t^* > 0$, $f^t(\mathbf{x}_t) \leq y_t$.

The proof is in Appendix A. If $\alpha_t > 0$, the KKT condition (3.1) implies that $f(\mathbf{x}_t)$ is also larger or equal to y_t. Thus, this lemma reveals the relative position of $f^t(\mathbf{x}_t)$ to $f(\mathbf{x}_t)$ and y_t.

The next lemma gives an error bound on each individual leave-one-out test.

Lemma 2 *Under Assumption 1,*

1. *If* $\alpha_t = \alpha_t^* = 0$, $|f^t(\mathbf{x}_t) - y_t| = |f(\mathbf{x}_t) - y_t| \leq \epsilon$.
2. *If* $\alpha_t > 0$, $f^t(\mathbf{x}_t) - y_t \leq 4\tilde{R}^2\alpha_t + \epsilon$.
3. *If* $\alpha_t^* > 0$, $y_t - f^t(\mathbf{x}_t) \leq 4\tilde{R}^2\alpha_t^* + \epsilon$.

The proof is in Appendix B. Then, when $\alpha_t > 0$, $|f^t(\mathbf{x}_t) - y_t| = f^t(\mathbf{x}_t) - y_t$ from Lemma 1. It follows $|f^t(\mathbf{x}_t) - y_t| \leq 4\tilde{R}^2\alpha_t + \epsilon$ from Lemma 2. After extending this argument to the cases of $\alpha_t^* > 0$ and $\alpha_t = \alpha_t^* = 0$, we have

Theorem 1 *Under Assumption 1, the leave-one-out error (3.3) is bounded by*

$$4\tilde{R}^2\mathbf{e}^T(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*)+l\epsilon. \tag{3.4}$$

We discuss the difference on proving bounds for classification and regression. In classification, the RM bound (2.3) is from the following derivation: If the tth training data is wrongly classified during the loo procedure, then

$$1 \le 4\alpha_t\tilde{R}^2, \tag{3.5}$$

where α_t is the tth element of the optimal solution of (2.4). If the data is correctly classified, the loo error is zero and still smaller than $4\alpha_t\tilde{R}^2$. Therefore, $4\tilde{R}^2\mathbf{e}^T\boldsymbol{\alpha}$ is larger than the number of loo errors. On the other hand, since there is no "wrongly classified" data in regression, we use Lemma 2 instead of (3.5) and Lemma 1 is required.

3.2 Span Bound for L2-SVR

Similar to the above discussion, we can have the span bound for L2-SVR:

Theorem 2 *Under the same assumptions of Theorem 1 and the assumption that the set of support vectors remains the same during the loo procedure, the loo error of L2-SVR is bounded by*

$$\sum_{t=1}^{l}(\alpha_t+\alpha_t^*)S_t^2+l\epsilon, \tag{3.6}$$

where S_t^2 is the optimal objective value of (2.5) with $\mathcal{F}$ replaced by $\{i \mid \alpha_i+\alpha_i^*>0\}$.

The proof is in Appendix C. The same as the case for classification, (3.6) may not be continuous, so we propose a similar modification like (2.7):

$$\sum_{t=1}^{l}(\alpha_t+\alpha_t^*)\tilde{S}_t^2+l\epsilon. \tag{3.7}$$

$\tilde{S}_t^2$ is the optimal objective solution of

$$\begin{array}{rc} \min\limits_{\boldsymbol{\lambda}} & \|\tilde{\phi}(\mathbf{x}_t)-\sum_{i\in\mathcal{F}\setminus\{t\}}\lambda_i\tilde{\phi}(\mathbf{x}_i)\|^2+\eta\sum_{i\in\mathcal{F}\setminus\{t\}}\lambda_i^2\frac{1}{(\alpha_i+\alpha_i^*)} \\ \text{subject to} & \sum_{i\in\mathcal{F}\setminus\{t\}}\lambda_i=1, \end{array} \tag{3.8}$$

where η is a positive parameter and $\mathcal{F}=\{i \mid \alpha_i+\alpha_i^*>0\}$. The calculation of $\tilde{S}_t^2$ is the same as (2.11).

3.3 LOO Bounds for L1-SVR

L1-SVR is another commonly used form for regression. It considers the following objective function

$$\min_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i + C\sum_{i=1}^{l}\xi_i^* \tag{3.9}$$

under the constraints of (1.1) and nonnegative constraints on $\boldsymbol{\xi}, \boldsymbol{\xi}^*$: $\xi_i \geq 0, \xi_i^* \geq 0, i = 1, \ldots, l$. The name "L1" comes from the linear loss function. The dual problem is

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \quad & \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T K(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) + \epsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \\ \text{subject to} \quad & \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0, 0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, \ldots, l, \end{aligned} \tag{3.10}$$

Two main differences between (1.2) and (3.10) are that $\tilde{K}$ is replaced by K and α_i, α_i^* are upper-bounded by C. To derive loo bounds, we still require Assumption 1. With some modifications in the proof (details in Appendix D.1), Lemma 1 still holds. For Lemma 2, results are different as now $\alpha_i, \alpha_t^* \leq C$ and ξ_t plays a role:

Lemma 3

1. *If* $\alpha_t = \alpha_t^* = 0$, $|f^t(\mathbf{x}_t) - y_t| = |f(\mathbf{x}_t) - y_t| \leq \epsilon$.
2. *If* $\alpha_t > 0$, $f^t(\mathbf{x}_t) - y_t \leq 4R^2\alpha_t + \xi_t + \xi_t^* + \epsilon$.
3. *If* $\alpha_t^* > 0$, $y_t - f^t(\mathbf{x}_t) \leq 4R^2\alpha_t^* + \xi_t + \xi_t^* + \epsilon$.

The proof is in Appendix D.2. Note that R is now the radius of the smallest sphere containing all $\phi(\mathbf{x}_i), i = 1, \ldots, l$. Using Lemmas 1 and 3, the bound is

$$4R^2\mathbf{e}^T(\boldsymbol{\alpha} + \boldsymbol{\alpha}^*) + \mathbf{e}^T(\boldsymbol{\xi} + \boldsymbol{\xi}^*) + l\epsilon.$$

Regarding the span bound, the proof for Theorem 2 still holds. However, S_t^2 is redefined as the optimal objective value of the following problem:

$$\begin{aligned} \min_{\boldsymbol{\lambda}} \quad & \|\phi(\mathbf{x}_t) - \textstyle\sum_{i\in\mathcal{F}\setminus\{t\}} \lambda_i\phi(\mathbf{x}_i)\|^2 \\ \text{subject to} \quad & \textstyle\sum_{i\in\mathcal{F}\setminus\{t\}} \lambda_i = 1, \end{aligned} \tag{3.11}$$

where $\mathcal{F} = \{i \mid 0 < \alpha_i + \alpha_i^* < C\}$. Then an loo bound is

$$\sum_{t=1}^{l}(\alpha_t + \alpha_t^*)S_t^2 + \sum_{t=1}^{l}(\xi_t + \xi_t^*) + l\epsilon. \tag{3.12}$$

4 Implementation Issues

In the rest of this paper, we consider only loo bounds using L2-SVR.

4.1 Continuity and Differentiability

To use the bound, $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$ must be well defined functions of parameters. That is, we need the uniqueness of the optimal dual solution. As $\tilde{K}$ contains the term I/C and hence is positive definite, $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$ are unique (Chang and Lin 2002, Lemma 4).

To discuss continuity and differentiability, we make an assumption about the kernel function:

Assumption 2 *The kernel function is differentiable respect to parameters.*

For continuity, we have known that the span bound is not continuous, but others are:

Theorem 3

1. $(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)$ *and* $\tilde{R}^2$ *are continuous, and so is the radius margin bound.*

2. *The modified span bound* (3.7) *is continuous.*

The proof is in Appendix E. To minimize leave-one-out bounds, differentiability is important as we may have to calculate the gradient. Unfortunately, loo bounds for L2-SVR are not differentiable. An example for the radius margin bound is in Appendix F. This situation is different from classification, where the radius margin bound for L2-SVM is differentiable (see more discussion in (Chung, Kao, Sun, Wang, and Lin 2003)). However, we may still use gradient-based methods as gradients exist almost everywhere.

Theorem 4 *Radius margin and modified span bounds are differentiable almost everywhere. If around a given parameter set, zero and non-zero elements of* $(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)$ *are the same, then bounds are differentiable at this parameter set.*

The proof is in Appendix G. The above discussion applies to bounds for classification as well. For differentiable points, we calculate gradients in Section 4.2.

4.2 Gradient Calculation

To have the gradient of loo bounds, we need the gradient of $\boldsymbol{\alpha} + \boldsymbol{\alpha}^*$, $\tilde{R}^2$, and $\tilde{S}_t^2$.

4.2.1 Gradient of $\boldsymbol{\alpha} + \boldsymbol{\alpha}^*$

Define $\hat{\boldsymbol{\alpha}}_{\mathcal{F}} \equiv \boldsymbol{\alpha}^*_{\mathcal{F}} - \boldsymbol{\alpha}_{\mathcal{F}}$ and recall the definition of M in (2.9). For free support vectors, KKT optimality conditions (3.1) imply that

$$M \begin{bmatrix} \hat{\boldsymbol{\alpha}}_{\mathcal{F}} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{p} \\ 0 \end{bmatrix},$$

where

$$p_i = \begin{cases} y_i - \epsilon & \text{if } \hat{\alpha}_i > 0, \\ y_i + \epsilon & \text{if } \hat{\alpha}_i < 0. \end{cases}$$

We have

$$\frac{\partial(\alpha_i + \alpha_i^*)}{\partial\theta} = z_i \frac{\partial \hat{\alpha}_i}{\partial\theta}, \tag{4.1}$$

where

$$z_i = \begin{cases} 1 & \text{if } \hat{\alpha}_i > 0, \\ -1 & \text{if } \hat{\alpha}_i < 0. \end{cases}$$

Except ϵ, all other parameters relate to M but not $\mathbf{p}$, so for any such parameter θ,

$$M \begin{bmatrix} \frac{\partial \hat{\boldsymbol{\alpha}}_{\mathcal{F}}}{\partial\theta} \\ \frac{\partial b}{\partial\theta} \end{bmatrix} + \frac{\partial M}{\partial\theta} \begin{bmatrix} \hat{\boldsymbol{\alpha}}_{\mathcal{F}} \\ b \end{bmatrix} = 0.$$

Thus,

$$\begin{bmatrix} \frac{\partial \hat{\boldsymbol{\alpha}}_{\mathcal{F}}}{\partial\theta} \\ \frac{\partial b}{\partial\theta} \end{bmatrix} = -M^{-1} \begin{bmatrix} \frac{\partial \tilde{K}}{\partial\theta} & \mathbf{0}_{\mathcal{F}} \\ \mathbf{0}_{\mathcal{F}}^T & 0 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\alpha}}_{\mathcal{F}} \\ b \end{bmatrix} = -M^{-1} \begin{bmatrix} \frac{\partial \tilde{K}}{\partial\theta} \hat{\boldsymbol{\alpha}}_{\mathcal{F}} \\ 0 \end{bmatrix}. \tag{4.2}$$

If θ is ϵ,

$$\begin{bmatrix} \frac{\partial \hat{\boldsymbol{\alpha}}_{\mathcal{F}}}{\partial\epsilon} \\ \frac{\partial b}{\partial\epsilon} \end{bmatrix} = M^{-1} \begin{bmatrix} \frac{\partial \mathbf{p}}{\partial\epsilon} \\ 0 \end{bmatrix},$$

and

$$\frac{\partial p_i}{\partial\epsilon} = \begin{cases} -1 & \text{if } \hat{\alpha}_i > 0, \\ 1 & \text{if } \hat{\alpha}_i < 0. \end{cases}$$

4.2.2 Gradient of $\tilde{S}_t^2$

Now $\tilde{S}_t^2$ is defined as (2.11) but $D_{tt} = \eta/(\alpha_t + \alpha_t^*)$ for $t \in \mathcal{F}$. Using (2.11),

$$\frac{\partial S_t^2}{\partial\theta} = -\frac{1}{(\tilde{M}^{-1})_{tt}^2} \frac{\partial(\tilde{M}^{-1})_{tt}}{\partial\theta} + \frac{\eta}{(\alpha_t + \alpha_t^*)^2} \frac{\partial(\alpha_t + \alpha_t^*)}{\partial\theta}.$$

Note that

$$\frac{\partial(\tilde{M}^{-1})_{tt}}{\partial\theta} = \left(\frac{\partial \tilde{M}^{-1}}{\partial\theta} \right)_{tt} = \left(\tilde{M}^{-1} \frac{\partial \tilde{M}}{\partial\theta} \tilde{M}^{-1} \right)_{tt}, \tag{4.3}$$

and $\partial(\alpha_t + \alpha_t^*)/\partial\theta$ can be obtained using (4.1) and (4.2). Furthermore, in (4.3),

$$\frac{\partial \tilde{M}}{\partial \theta} = \begin{bmatrix} \frac{\partial \tilde{K}}{\partial \theta} + \frac{\partial \tilde{D}}{\partial \theta} & \mathbf{0}_{\mathcal{F}} \\ \mathbf{0}_{\mathcal{F}}^T & 0 \end{bmatrix},$$

where

$$\left(\frac{\partial \tilde{D}}{\partial \theta}\right)_{ii} = -\frac{\eta}{(\alpha_i + \alpha_i^*)^2} \frac{\partial(\alpha_i + \alpha_i^*)}{\partial \theta}, i \in \mathcal{F}.$$

4.2.3 Gradient of $\tilde{R}^2$

$\tilde{R}^2$ is the optimal object value of

$$\begin{aligned} \max_{\boldsymbol{\beta}} \quad & \sum_{i=1}^{l} \beta_i \tilde{K}(\mathbf{x}_i, \mathbf{x}_i) - \boldsymbol{\beta}^T \tilde{K} \boldsymbol{\beta} \\ \text{subject to} \quad & 0 \le \beta_i, i = 1, \ldots, l, \sum_{i=1}^{l} \beta_i = 1. \end{aligned} \tag{4.4}$$

(see, for example, (Vapnik 1998)). From (Bonnans and Shapiro 1998), it is differentiable and the gradient is

$$\frac{\partial \tilde{R}^2}{\partial \theta} = \sum_{i=1}^{l} \beta_i \frac{\partial \tilde{K}(x_i, x_i)}{\partial \theta} - \boldsymbol{\beta}^T \frac{\partial \tilde{K}}{\partial \theta} \boldsymbol{\beta}.$$

5 Experiments

In this section, different parameter selection methods including the proposed bounds are compared. We consider the same real data used in (Lin and Weng 2004) and some statistics are in Table 1. To have a reliable comparison, for each data set, we randomly produce 30 training/testing splits. Each training set consists of 4/5 of the data and the remaining are for testing. Parameter selection using different methods are applied on each training file. We then report the average and standard deviation of 30 mean squared errors (MSE) on predicting test sets. A method with lower MSE is better.

We compare two proposed bounds with three other parameter selection methods:

1. RM (L2-SVR): the radius margin bound (3.4).
2. MSP (L2-SVR): the modified span bound (3.7).
3. CV (L2-SVR): a grid search of parameters using five-fold cross validation.

Table 1: Data Statistics: n is the number of features and l is the number of data instances.

Problem	n	l
pyrim	27	74
triazines	60	186
mpg	7	392
housing	13	566
add10	10	1,000
cpusmall	12	1,000
spacega	6	1,000
abalone	8	1,000

4. CV (L1-SVR): the same as the previous method but L1-SVR is considered.

5. BSVR: a Bayesian framework which improves the smoothness of the evidence function using a modified SVR (Chu, Keerthi, and Ong 2003).

All methods except BSVR use the Radial Basis Function (RBF) kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(2\sigma^2)}, \tag{5.1}$$

where σ^2 is the kernel parameter. BSVR implements an extension of the RBF kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \kappa_0 e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(2\sigma^2)} + \kappa_b, \tag{5.2}$$

where κ_0 and κ_b are two additional kernel parameters. Both kernels satisfy Assumption 2 on differentiability.

Implementation details and experimental results are in the following subsections.

5.1 Implementations of Various Model Selection Methods

RM and MSP are differentiable almost everywhere, so we implement quasi-Newton, a gradient-based optimization method, to minimize them. The parameter η in the modified span bound (2.7) is set to be 0.1. Section 5.3 will discuss the impact of using different η. Following most earlier work on minimizing loo bounds, we consider parameters in the log scale: $(\ln C, \ln \sigma^2, \ln \epsilon)$. Thus, if f is the function of parameters, the gradient is calculated by

$$\frac{\partial f}{\partial \ln \theta} = \theta \frac{\partial f}{\partial \theta},$$

Table 2: Mean and standard deviation of 30 MSEs (using 30 training/testing splits).

	RM		MSP		L2 CV		L1 CV		BSVR	
Problem	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
pyrim	0.015	0.010	0.007	0.008	0.007	0.007	0.007	0.007	0.007	0.008
triazines	0.042	0.005	0.021	0.006	0.021	0.005	0.023	0.008	0.021	0.007
mpg	8.156	1.598	7.045	1.682	7.122	1.809	7.146	1.924	6.894	1.856
housing	23.14	7.774	9.191	2.733	9.318	2.957	11.26	5.014	10.40	3.950
add10	6.491	1.675	1.945	0.254	1.820	0.182	1.996	0.194	2.298	0.256
cpusmall	34.02	13.33	14.57	4.692	14.73	3.754	15.63	5.344	16.17	5.740
spacega	0.037	0.007	0.013	0.001	0.012	0.001	0.013	0.001	0.014	0.001
abalone	10.69	1.551	5.071	0.678	5.088	0.646	5.247	0.806	5.514	0.912

and formulas in Section 4.2. Suppose $\boldsymbol{\theta}$ is the parameter vector to be determined. The quasi-Newton method is an iterative procedure to minimize $f(\boldsymbol{\theta})$. If k is the index of the loop, the kth iteration for updating $\boldsymbol{\theta}^k$ to $\boldsymbol{\theta}^{k+1}$ is as the following

1. Compute a search direction $\mathbf{p} = -H_k \nabla f(\boldsymbol{\theta}^k)$.
2. Find $\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k + \lambda \mathbf{p}$ using a line search to ensure sufficient decrease.
3. Obtain H_{k+1} by

$$H_{k+1} = \begin{cases} (I - \frac{\mathbf{s}\mathbf{t}^T}{\mathbf{t}^T\mathbf{s}}) H_k (I - \frac{\mathbf{t}\mathbf{s}^T}{\mathbf{t}^T\mathbf{s}}) + \frac{\mathbf{s}\mathbf{s}^T}{\mathbf{t}^T\mathbf{s}} & \text{if } \mathbf{t}^T\mathbf{s} > 0, \\ H_k & \text{otherwise,} \end{cases} \tag{5.3}$$

   where

$$\mathbf{s} = \boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k \text{ and } \mathbf{t} = \nabla f(\boldsymbol{\theta}^{k+1}) - \nabla f(\boldsymbol{\theta}^k).$$

Here, H_k serves as the inverse of an approximate Hessian of f and is set to be the identity matrix in the first iteration. The sufficient decrease by the line search usually means

$$f(\boldsymbol{\theta}^k + \lambda \mathbf{p}) \le f(\boldsymbol{\theta}^k) + \sigma_1 \lambda \nabla f(\boldsymbol{\theta}^k)^T \mathbf{p}, \tag{5.4}$$

where $0 < \sigma_1 < 1$ is a positive constant. We find the largest value λ in a set $\{\gamma^i \mid i = 0, 1, \ldots\}$ such that (5.4) holds ($\gamma = 1/2$ used in this paper). We confine the search in a fixed region, so each parameter θ_i is associated with a lower bound l_i and upper bound u_i. If in the quasi-Newton method, $\theta_i^k + \lambda p_i$ is not in $[l_i, u_i]$, it is projected to the interval. For $\ln C$ and $\ln \sigma^2$, we set $l_i = -8$ and $u_i = 8$. For $\ln \epsilon$,

$l_i = -8$, but $u_i = 1$. We could not use too large u_i as a too large ϵ may cause all data to be in the ϵ-insensitive tube and hence $\boldsymbol{\alpha} = \boldsymbol{\alpha}* = \mathbf{0}$. Then, Assumption 1 does not hold and the loo bound may not be valid. More discussion on the use of quasi-Newton method is in (Chung, Kao, Sun, Wang, and Lin 2003, Section 5)

Table 3: Average $(\ln C, \ln \sigma^2, \ln \epsilon)$ of 30 runs.

	RM			MSP			L2 CV			L1 CV		
Problem	$\ln C$	$\ln \sigma^2$	$\ln \epsilon$	$\ln C$	$\ln \sigma^2$	$\ln \epsilon$	$\ln C$	$\ln \sigma^2$	$\ln \epsilon$	$\ln C$	$\ln \sigma^2$	$\ln \epsilon$
pyrim	4.0	1.0	-2.6	1.0	2.9	-8.0	6.4	3.1	-7.3	2.1	3.4	-6.6
triazines	-1.4	-0.6	-1.8	-0.8	3.0	-8.0	3.5	4.4	-7.2	0.6	4.0	-4.7
mpg	-0.7	0.4	-1.0	4.7	1.1	-7.1	3.6	0.6	-6.6	5.6	0.4	-1.7
housing	-1.5	1.1	-1.0	5.7	1.2	-6.9	6.6	1.6	-6.1	7.5	1.4	-1.7
add10	0.9	4.6	-1.0	8.0	3.3	-8.0	8.0	3.0	-7.8	7.9	2.9	-1.2
cpusmall	-1.1	-0.5	-1.0	7.9	1.6	-5.6	7.9	2.4	-7.0	8.0	1.7	-2.1
spacega	-6.5	-6.2	-1.7	7.4	3.1	-8.0	7.3	0.0	-6.8	6.0	1.1	-4.6
abalone	-8.0	7.0	-1.0	3.7	0.6	-8.0	4.1	0.9	-6.8	6.8	1.2	-2.1

The initial point of the quasi-Newton method is

$$(\ln C, \ln \sigma^2, \ln \epsilon) = (0, 0, -3).$$

The minimization procedure stops when

$$\|\nabla f(\boldsymbol{\theta}^k)\| < (1 + f(\boldsymbol{\theta}^k)) \times 10^{-5}, \text{ or } \frac{f(\boldsymbol{\theta}^{k-1}) - f(\boldsymbol{\theta}^k)}{f(\boldsymbol{\theta}^{k-1})} < 10^{-5} \tag{5.5}$$

happens. Each function (gradient) evaluation involves solving an SVR and is the main computational bottleneck. We use LIBSVM (Chang and Lin 2001) as the underlying solver.

For CV, we try 2,312 parameter sets with $(\ln C, \ln \gamma, \ln \epsilon) = [-8, -7, \ldots, 8] \times [-8, -7, \ldots, 8] \times [-8, -7, \ldots, -1]$. Similar to the case of using loo bounds, here we avoid considering too large ϵ. The one with the lowest five-fold CV accuracy is used to train the model for testing.

For BSVR, we directly use the authors' gradient-based implementation with the same stopping condition (5.5). Note that their evidence function is not differentiable either.

5.2 Experimental Results

Table 2 presents mean and standard deviation of 30 MSEs. CV (L1- and L2-SVR), MSP, and BSVR are similar but RM is worse. In classification, Chapelle, Vapnik,

Table 4: Computational time (in second) for parameter selection.

	RM	SP	L2 CV	L1 CV
pyrim	0.6	0.3	84.3	99.27
triazines	2.5	5.8	310.8	440.0
mpg	10.7	63.0	1536.7	991.59
housing	23.3	159.4	2368.9	1651.5
add10	160.6	957.1	6940.2	5312.7
cpusmall	204.7	931.8	10073.7	6087.0
spacega	53.6	771.9	4246.9	15514.42
abalone	68.0	1030.4	12905.0	7523.3

Table 5: Number of function and gradient evaluations of the quasi-Newton method (average of 30 runs).

	RM		MSP	
Problem	FEV	GEV	FEV	GEV
pyrim	34.7	13.5	32.2	13.5
triazines	29.8	10.4	30.0	13.2
mpg	38.2	5.0	21.6	11.7
housing	40.5	5.0	33.2	13.2
add10	44.0	4.96	27.8	5.4
cpusmall	43.9	5.0	35.6	6.66
spacega	28.6	9.96	30.1	9.93
abalone	19.0	3.0	19.9	10.3

Bousquet, and Mukherjee (2002) showed that radius margin and span bounds perform similarly. From our experiments on the radius margin bound, parameters are more sensitive in regression than in classification. One possible reason is that the loo error for SVR is a continuous but not a discrete measurement.

The good performance of BSVR indicates that its Bayesian evidence function is accurate, but the use of a more general kernel function may also help. On the other hand, even though MSP uses only the RBF kernel, it is competitive with BSVR. Note that as CV is conducted on a discrete set of parameters, sometimes its MSE is slightly worse than that of MSP or BSVR, which considers parameters in a continuous space.

Table 3 presents the parameters obtained by different approaches. We do not give

those by BSVR as it considers more than three parameters. Clearly these methods obtain quite distinct parameters even though they (except RM) give similar testing errors. This observation indicates that good parameters are in a quite wide region. In other words, SVM is sensitive to parameters but is not too sensitive. Moreover, different regions that RM and MSP lead to also cause the two approaches to have quite different running time. More details are in Table 4 and the discussion below.

To see the performance gain of the bound-based methods, we compare the computational time in Table 4. The experiments were done on a Pentium IV 2.8 GHz computer using the linux operating system. We did not compare the running time of BSVR as the code is available only on MS windows. Clearly, using bounds saves a significant amount of time.

For RM and MSP, the quasi-Newton implementation requires much fewer SVRs than CV. Table 5 lists the average number of function and gradient evaluations of the quasi-Newton method. Note that the number of function evaluations is the same as the number of SVRs solved. From Table 4, MSP is slower than RM though they have similar numbers of function/gradient evaluations. As they do not land at the same parameter region, their respective SVR training time is different. In other words, the individual SVR training time here is related to parameters. Now MSP leads to a good region with smaller testing errors but training SVRs with parameters in this region takes more time.

From Table 4, the computational time of CV using L1-and L2-SVR is not close. As they have different formulas (c.g. C in L1-SVR and $C/2$ in L2-SVR), we do not expect them to be very similar.

5.3 Discussion

The smoothing parameter η of the modified span bound was simply set to be 0.1 for experiments. It is important to check how η affects the performance of the bound. Figure 1 presents the relation between η and the test error. From (2.8), large η causes the modified bound to be away from the original one. Thus, the performance is worse as shown in Figure 1. However, if η is reasonably small, the performance is quite stable. Therefore, the selection of η is not difficult.

It is also interesting to investigate how tight the proposed bounds are in practice. Figure 2 compares different bounds and the loo value. We select the best σ^2 and ϵ from CV and show values of bounds and loo via changing C. Clearly, the span bound is a good approximation of loo, but RM is not when C is large. This situation has

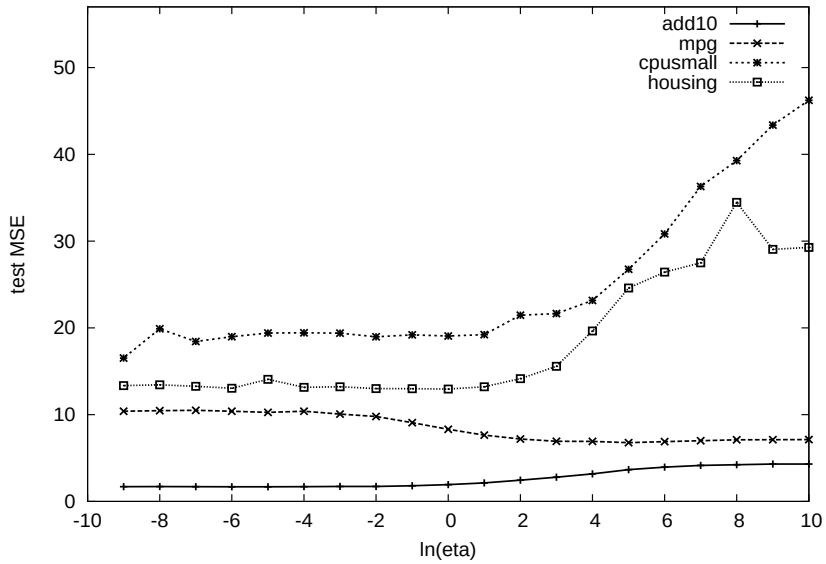

Figure 1: Effect of η on testing errors: using the first training/testing split of the problems add10, mpg, housing, and cpusmall.

happened in classification (Chung, Kao, Sun, Wang, and Lin 2003). The reason is that S_t can be much smaller than $2\tilde{R}$ under some parameters. Recall that $\tilde{R}$ is the radius of the smallest sphere containing $\tilde{\phi}(x_i), i = 1, \ldots, l$, so $\tilde{R}$ is large if there are two far away points. However, the span bound finds a combination of $\mathbf{x}_i, i \notin F \backslash \{t\}$, to be as close to $\mathbf{x}_t$.

6 Conclusions

In this article, we derive loo bounds for SVR and discuss their properties. Experiments demonstrate that the proposed bounds are competitive with Bayesian SVR for parameter selection. A future study is to apply the proposed bounds on feature selection. We also would like to implement non-smooth optimization techniques as bounds here are not really differentiable. The implementation considering L1-SVR is also interesting.

Experiments demonstrate that minimizing the proposed bound is more efficient than cross validation on a discrete set of parameters. For a model with more than two parameters, a grid search is time consuming, so a gradient-based method may be more suitable.

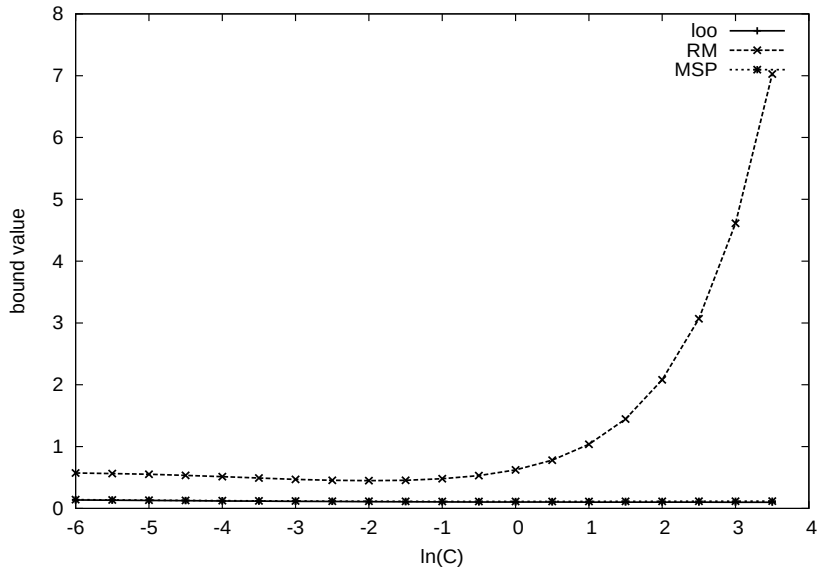

Figure 2: Loo, radius margin bound (RM), and the modified span bound (MSP): σ^2 and ϵ are fixed via using the best parameters from CV (L2-SVR). The training file is spacega.

Acknowledgment

The authors would like to thank Olivier Chapelle for many helpful comments.

References

Bazaraa, M. S., H. D. Sherali, and C. M. Shetty (1993). *Nonlinear programming : theory and algorithms* (Second ed.). Wiley.

Bonnans, J. F. and A. Shapiro (1998). Optimization problems with perturbations: a guided tour. *SIAM Review 40*(2), 228–264.

Boser, B., I. Guyon, and V. Vapnik (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*.

Chang, C.-C. and C.-J. Lin (2001). *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 (2002). Training ν-support vector regression: Theory and algorithms. *Neural Computation 14*(8), 1959–1977.

Chapelle, O., V. Vapnik, O. Bousquet, and S. Mukherjee (2002). Choosing multiple parameters for support vector machines. *Machine Learning 46*, 131–159.

Chu, W., S. Keerthi, and C. J. Ong (2003). Bayesian support vector regression using a unified loss function. *IEEE Transactions on Neural Networks*. To appear.

Chung, K.-M., W.-C. Kao, C.-L. Sun, L.-L. Wang, and C.-J. Lin (2003). Radius margin bounds for support vector machines with the RBF kernel. *Neural Computation 15*, 2643–2681.

Clarke, F. H. (1983). *Optimization and Nonsmooth Analysis*. New York: Wiley.

Cortes, C. and V. Vapnik (1995). Support-vector network. *Machine Learning 20*, 273–297.

Gao, J. B., S. R. Gunn, C. J. Harris, and M. Brown (2002). A probabilistic framework for SVM regression and error bar estimation. *Machine Learning 46*, 71–89.

Golub, G. H. and C. F. Van Loan (1996). *Matrix Computations* (Third ed.). The Johns Hopkins University Press.

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

Kwok, J. T. (2001). Linear dependency between epsilon and the input noise in epsilon-support vector regression. *Proceedings of the International Conference on Artificial Neural Networks ICANN*.

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.

Lin, C.-J. and R. C. Weng (2004). Simple probabilistic predictions for support vector regression. Technical report, Department of Computer Science, National Taiwan University.

Momma, M. and K. P. Bennett (2002). A pattern search method for model selection of support vector regression. *Proceedings of SIAM Conference on Data Mining*.

Smola, A., N. Murata, B. Schölkopf, and K.-R. Müller (1998). Asymptotically optimal choice of epsilon-loss for support vector machines. *Proceeding of the International Conference on Artificial Neural Network*.

Smola, A. J. and B. Schölkopf (2004). A tutorial on support vector regression. *Statistics and Computing 14*(3), 199–222.

Ulbrich, M. (2000). *Nonsmooth Newton-like Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. Ph. D. thesis, Technische Universität München.

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.

A Proof of Lemma 1

We consider the first case, $\alpha_t > 0$. Let $(\mathbf{w}^t, b^t, \boldsymbol{\xi}^t, \boldsymbol{\xi}^{t*})$ and $(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*)$ be the optimal solutions of (3.2) and (1.1), respectively. Though $\boldsymbol{\xi}^t$ and $\boldsymbol{\xi}^{t*}$ are vectors with $l-1$ elements, recall that we define $\xi_t^t = \xi_t^{t*} = 0$ to make $\boldsymbol{\xi}^t$ and $\boldsymbol{\xi}^{t*}$ have l elements.

Note that the only difference between (3.2) and (1.1) is that (1.1) possesses the following constraint:

$$-\epsilon - \xi_t^* \le \mathbf{w}^T\phi(\mathbf{x}_t) + b - y_t \le \epsilon + \xi_t. \tag{A.1}$$

We then prove the lemma by a contradiction. If the result is wrong, there are $\alpha_t > 0$ and $f^t(\mathbf{x}_t) < y_t$. From the KKT condition (3.1), $\alpha_t > 0$ implies $\xi_t > 0$ and $f(\mathbf{x}_t) = y_t + \epsilon + \xi_t > y_t + \epsilon > y_t$. Then, we have

$$f(\mathbf{x}_t) = \mathbf{w}^T\phi(\mathbf{x}_t) + b > y_t > f^t(\mathbf{x}_t) = (\mathbf{w}^t)^T\phi(\mathbf{x}_t) + b^t.$$

Therefore, there is $0 < p < 1$ such that

$$(1-p)(\mathbf{w}^T\phi(\mathbf{x}_t) + b) + p((\mathbf{w}^t)^T\phi(\mathbf{x}_t) + b^t) = y_t. \tag{A.2}$$

Using the feasibility of the two points,

$$(\hat{\mathbf{w}}, \hat{b}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\xi}}^*) = (1-p)(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*) + p(\mathbf{w}^t, b^t, \boldsymbol{\xi}^t, \boldsymbol{\xi}^{t*})$$

is a new feasible solution of (3.2) without considering the tth element of $\hat{\xi}_t$ and $\hat{\xi}_t^*$. Since the objective function is convex and $0 < p < 1$, we have

$$
\begin{aligned}
& \frac{1}{2}\hat{\mathbf{w}}^T\hat{\mathbf{w}} + \frac{C}{2}\sum_{i\neq t}(\hat{\xi}_i)^2 + \frac{C}{2}\sum_{i\neq t}(\hat{\xi}_i^*)^2 \\
\leq\ & (1-p)\left(\frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{2}\sum_{i\neq t}(\xi_i)^2 + \frac{C}{2}\sum_{i\neq t}(\xi_i^*)^2\right) + \\
& p\left(\frac{1}{2}(\mathbf{w}^t)^T(\mathbf{w}^t) + \frac{C}{2}\sum_{i\neq t}(\xi_i^t)^2 + \frac{C}{2}\sum_{i\neq t}(\xi_i^{t*})^2\right) \\
\leq\ & \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{2}\sum_{i\neq t}(\xi_i)^2 + \frac{C}{2}\sum_{i\neq t}(\xi_i^*)^2. \qquad \text{(A.3)}
\end{aligned}
$$

The last inequality comes from the fact that $(\mathbf{w}^t, b^t, \boldsymbol{\xi}^t, \boldsymbol{\xi}^{t*})$ is optimal for (3.2) and if the constraint (A.1) of (1.1) is not considered, $(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*)$ is feasible for (3.2).

In this case, $\alpha_t > 0$, which implies $\xi_t > 0$ and $\xi_t^* = 0$ from KKT conditions (3.1). Since $\xi_t \neq 0$, $\hat{\xi}_t \neq 0$ as well. As $\hat{\xi}_t$ and $\hat{\xi}_t^*$ are not considered for deriving (A.3), we can redefine $\hat{\xi}_t = \hat{\xi}_t^* = 0$ such that $(\hat{\mathbf{w}}, \hat{b}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\xi}}^*)$ satisfies (A.1) and thus is also a feasible solution of (1.1). Then, from (A.3), $\hat{\xi}_t = 0 < \xi_t$, and $\hat{\xi}_t^* = \xi_t^* = 0$, we have

$$
\begin{aligned}
& \frac{1}{2}\hat{\mathbf{w}}^T\hat{\mathbf{w}} + \frac{C}{2}\sum_{i=1}^{l}(\hat{\xi}_i)^2 + \frac{C}{2}\sum_{i=1}^{l}(\hat{\xi}_i^*)^2 \\
<\ & \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{2}\sum_{i=1}^{l}(\xi_i)^2 + \frac{C}{2}\sum_{i=1}^{l}(\xi_i^*)^2. \qquad \text{(A.4)}
\end{aligned}
$$

Therefore, $(\hat{\mathbf{w}}, \hat{b}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\xi}}^*)$ is a better solution than $(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*)$ of (1.1), a contradiction. The proof of the other case is similar. □

B Proof of Lemma 2

For easier description, in this proof we introduce a different representation of SVR dual:

$$
\begin{aligned}
\min_{\bar{\boldsymbol{\alpha}}} \quad & \frac{1}{2}\bar{\boldsymbol{\alpha}}^T\bar{Q}\bar{\boldsymbol{\alpha}} + \mathbf{p}^T\bar{\boldsymbol{\alpha}} \qquad \text{(B.1)} \\
\text{subject to} \quad & \mathbf{z}^T\bar{\boldsymbol{\alpha}} = 0, \\
& \bar{\alpha}_i \geq 0, i = 1, \ldots, 2l,
\end{aligned}
$$

where

$$
\bar{\boldsymbol{\alpha}} = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\alpha}^* \end{bmatrix}, \mathbf{p} = \begin{bmatrix} \epsilon\mathbf{e} + \mathbf{y} \\ \epsilon\mathbf{e} - \mathbf{y} \end{bmatrix}, \mathbf{z} = \begin{bmatrix} -\mathbf{e} \\ \mathbf{e} \end{bmatrix},
$$

and

$$\bar{Q} = \begin{bmatrix} K + I/C & -K - I/C \\ -K - I/C & K + I/C \end{bmatrix}. \tag{B.2}$$

We proceed the proof by considering three cases:

1) $\alpha_t = \alpha_t^* = 0$

If $\alpha_t = \alpha_t^* = 0$, then

$$\alpha_1, \ldots, \alpha_{t-1}, \alpha_{t+1}, \ldots, \alpha_l, \tag{B.3}$$

and

$$\alpha_1^*, \ldots, \alpha_{t-1}^*, \alpha_{t+1}^*, \ldots, \alpha_l^*, \tag{B.4}$$

is a feasible and optimal solution of the dual problem of (3.2) because it satisfies the KKT condition. Since there are free support vectors, b can be uniquely determined from (3.1) using some nonnegative α_i or α_i^*. It follows that $b = b^t$ and $f(\mathbf{x}_t) = f^t(\mathbf{x}_t)$. From (3.1),

$$|f^t(\mathbf{x}_t) - y_t| = |f(\mathbf{x}_t) - y_t| \leq \epsilon.$$

2) $\alpha_t > 0$

In this case, we mainly use the formulation (B.1) to represent L2-SVR with the optimal solution $\bar{\boldsymbol{\alpha}}$. We also represent the dual of (3.2) by a form similar to (B.1) and denote $\bar{\boldsymbol{\alpha}}^t$ as any its unique optimal solution. Note that (3.2) has $l - 1$ constraints, so $\bar{\boldsymbol{\alpha}}^t$ has $2(l-1)$ elements. Here, we define $\bar{\boldsymbol{\alpha}}_t^t = \bar{\boldsymbol{\alpha}}_{t+l}^t = 0$ to make $\bar{\boldsymbol{\alpha}}^t$ a vector with $2l$ elements. The KKT condition of (B.1) can be rewritten as:

$$\begin{aligned} (\bar{Q}\bar{\boldsymbol{\alpha}})_i + bz_i + p_i = 0, \text{ if } \bar{\alpha}_i > 0, \\ (\bar{Q}\bar{\boldsymbol{\alpha}})_i + bz_i + p_i > 0, \text{ if } \bar{\alpha}_i = 0. \end{aligned} \tag{B.5}$$

Next, we follow the procedure of (Vapnik and Chapelle 2000) and (Joachims 2000). In (B.1), the approximate value of the training vector $\mathbf{x}_t$ can be written as:

$$f(\mathbf{x}_t) = -\sum_{i=1}^{2l} \bar{\alpha}_i \bar{Q}_{it} + b,$$

which equals $f(\mathbf{x}_t)$ defined in (1.3). Here, we intend to consider

$$f'(\mathbf{x}_t) = -\sum_{i \neq t, t+l} \bar{\alpha}_i \bar{Q}_{it} + b \tag{B.6}$$

as an approximation of $f^t(\mathbf{x}_t)$ since both f' and f^t do not consider the tth training vector. However, (B.6) is not applicable since (B.3) and (B.4) is not a feasible solution of dual of (3.2) when $\alpha_t > 0$. Therefore, we construct $\boldsymbol{\gamma}$ from $\bar{\boldsymbol{\alpha}}$ where $\boldsymbol{\gamma}$ is feasible for the problem with the tth data removed from (B.1).

Define a set $\mathcal{F}^t = \{i \mid \bar{\alpha}_i > 0, i \neq t, t+l\}$. In order to make $\boldsymbol{\gamma}$ a feasible solution, let $\boldsymbol{\gamma} = \bar{\boldsymbol{\alpha}} - \boldsymbol{\eta}$, where $\boldsymbol{\eta}$ satisfies

$$\begin{aligned} \eta_i \leq \bar{\alpha}_i, \quad & i \in \mathcal{F}^t, \\ \eta_i = 0, \quad & i \notin \mathcal{F}^t, i \neq t, i \neq t+l, \\ \eta_i = \bar{\alpha}_i, \quad & i = t, i = t+l, \end{aligned} \tag{B.7}$$

and

$$\mathbf{z}^T \boldsymbol{\eta} = 0. \tag{B.8}$$

For example, we can set all $\eta_1, \ldots, \eta_l$ to zero except $\eta_t = \alpha_t$. Then, using $\sum_{i=l+1}^{2l} \bar{\alpha}_i \geq \alpha_t$, we can find $\boldsymbol{\eta}$ which satisfies (B.7) and

$$\sum_{i=l+1}^{2l} \eta_i = \alpha_t. \tag{B.9}$$

Denote F as the objective function of (B.1). Then,

$$\begin{aligned} F(\boldsymbol{\gamma}) - F(\bar{\boldsymbol{\alpha}}) =& \frac{1}{2}(\bar{\boldsymbol{\alpha}} - \boldsymbol{\eta})^T \bar{Q}(\bar{\boldsymbol{\alpha}} - \boldsymbol{\eta}) + \mathbf{p}^T(\bar{\boldsymbol{\alpha}} - \boldsymbol{\eta}) - \\ & \frac{1}{2}(\bar{\boldsymbol{\alpha}})^T \bar{Q}(\bar{\boldsymbol{\alpha}}) - \mathbf{p}^T \bar{\boldsymbol{\alpha}} \\ =& \frac{1}{2}\boldsymbol{\eta}^T \bar{Q} \boldsymbol{\eta} - \boldsymbol{\eta}^T(\bar{Q}\bar{\boldsymbol{\alpha}} + \mathbf{p}). \end{aligned} \tag{B.10}$$

From (B.5) and (B.7), for any $i \in \mathcal{F}^t \cup \{t\}$, $(\bar{Q}\bar{\boldsymbol{\alpha}} + \mathbf{p})_i = -z_i b$ and for any $i \notin \mathcal{F}^t \cup \{t\}$, $\eta_i = 0$. Using (B.8), (B.10) is reduced to

$$F(\boldsymbol{\gamma}) - F(\bar{\boldsymbol{\alpha}}) = \frac{1}{2}\boldsymbol{\eta}^T \bar{Q} \boldsymbol{\eta}. \tag{B.11}$$

Similarly, from $\bar{\boldsymbol{\alpha}}^t$, we construct a vector $\boldsymbol{\delta}$ that is a feasible solution of (B.1). Let $\bar{\mathcal{F}}^t = \{i \mid \bar{\alpha}_i^t > 0, i \neq t, t+l\}$. In order to make $\boldsymbol{\delta}$ feasible, define $\boldsymbol{\delta} = \bar{\boldsymbol{\alpha}}^t - \boldsymbol{\mu}$, where $\boldsymbol{\mu}$ satisfies

$$\begin{aligned} \mu_i \leq \bar{\alpha}_i^t, \quad & i \in \bar{\mathcal{F}}^t, \\ \mu_i = 0, \quad & i \notin \bar{\mathcal{F}}^t, i \neq t, i \neq t+l, \\ \mu_i = -\bar{\alpha}_i, \quad & i = t, i = t+l, \end{aligned} \tag{B.12}$$

and

$$\mathbf{z}^T \boldsymbol{\mu} = 0. \tag{B.13}$$

The existence of $\boldsymbol{\mu}$ easily follows from Assumption 1. With the condition $\mathbf{z}^T\bar{\boldsymbol{\alpha}}^t = 0$, this assumption implies that at least one of $\bar{\alpha}^t_{l+1}, \ldots, \bar{\alpha}^t_{2l}$ is positive. Thus, while $\bar{\alpha}^t_t$ is increased from zero to $\bar{\alpha}_t$, we can increase some positive $\bar{\alpha}^t_{l+1}, \ldots, \bar{\alpha}^t_{2l}$ so that $\mathbf{z}^T\bar{\boldsymbol{\alpha}}^t = 0$ still holds.

Next, define $\boldsymbol{\delta} = \bar{\boldsymbol{\alpha}}^t - \boldsymbol{\mu}$ and note that $\boldsymbol{\delta}$ is a feasible solution of (1.2). It follows:

$$\begin{aligned} F(\boldsymbol{\delta}) - F(\bar{\boldsymbol{\alpha}}^t) =& \frac{1}{2}(\bar{\boldsymbol{\alpha}}^t - \boldsymbol{\mu})^T\bar{Q}(\bar{\boldsymbol{\alpha}}^t - \boldsymbol{\mu}) + \mathbf{p}^T(\bar{\boldsymbol{\alpha}}^t - \boldsymbol{\mu}) - \\ & \frac{1}{2}(\bar{\boldsymbol{\alpha}}^t)^T\bar{Q}(\bar{\boldsymbol{\alpha}}^t) - \mathbf{p}^T\bar{\boldsymbol{\alpha}}^t \\ =& \frac{1}{2}\boldsymbol{\mu}^T\bar{Q}\boldsymbol{\mu} - \boldsymbol{\mu}^T(\bar{Q}\bar{\boldsymbol{\alpha}}^t + \mathbf{p}). \end{aligned} \tag{B.14}$$

From (B.5),

$$(\bar{Q}\bar{\boldsymbol{\alpha}}^t + \mathbf{p})_i = -z_i b^t, \forall i \in \bar{\mathcal{F}}^t, \tag{B.15}$$

where $\bar{\boldsymbol{\alpha}}^t$ and b^t are optimal for (3.2) and its dual. By (B.12), (B.13), and (B.15),

$$\begin{aligned} \boldsymbol{\mu}^T(\bar{Q}\bar{\boldsymbol{\alpha}}^t + \mathbf{p}) =& - b^t \sum_{i \neq t, t+l}^{2l} \mu_i z_i + \mu_t(\bar{Q}\bar{\boldsymbol{\alpha}}^t + \mathbf{p})_t + \mu_{t+l}(\bar{Q}\bar{\boldsymbol{\alpha}}^t + \mathbf{p})_{t+l} \\ =& - \bar{\alpha}_t(\bar{Q}\bar{\boldsymbol{\alpha}}^t + \mathbf{p} + b^t\mathbf{z})_t - \bar{\alpha}_{t+l}(\bar{Q}\bar{\boldsymbol{\alpha}}^t + \mathbf{p} + b^t\mathbf{z})_{t+l} \\ =& \bar{\alpha}_t(f^t(\mathbf{x}_t) - y_t - \epsilon) + \bar{\alpha}_{t+l}(y_t - f^t(\mathbf{x}_t) - \epsilon). \end{aligned}$$

Thus, (B.14) is simplified to

$$\begin{aligned} & \bar{\alpha}_t(f^t(\mathbf{x}_t) - y_t - \epsilon) + \bar{\alpha}_{t+l}(y_t - f^t(\mathbf{x}_t) - \epsilon) \\ =& F(\bar{\boldsymbol{\alpha}}^t) - F(\boldsymbol{\delta}) + \frac{1}{2}\boldsymbol{\mu}^T\bar{Q}\boldsymbol{\mu}. \end{aligned} \tag{B.16}$$

Here we claim that $\bar{\alpha}_{t+l} = 0$ when $\bar{\alpha}_t > 0$ since from (B.5),

$$\alpha_t\alpha^*_t = \bar{\alpha}_t\bar{\alpha}_{t+l} = 0. \tag{B.17}$$

Note that $F(\boldsymbol{\delta}) \geq F(\bar{\boldsymbol{\alpha}})$ as $\bar{\boldsymbol{\alpha}}$ is the optimal solution of (B.1). Similarly, $F(\boldsymbol{\gamma}) \geq F(\bar{\boldsymbol{\alpha}}^t)$. Combining (B.17), (B.16), and (B.11),

$$\begin{aligned} & \bar{\alpha}_t(f^t(\mathbf{x}_t) - y_t - \epsilon) \\ =& F(\bar{\boldsymbol{\alpha}}^t) - F(\boldsymbol{\delta}) + \frac{1}{2}\boldsymbol{\mu}^T\bar{Q}\boldsymbol{\mu} \\ \leq& F(\boldsymbol{\gamma}) - F(\bar{\boldsymbol{\alpha}}) + \frac{1}{2}\boldsymbol{\mu}^T\bar{Q}\boldsymbol{\mu} \\ =& \frac{1}{2}\boldsymbol{\eta}^T\bar{Q}\boldsymbol{\eta} + \frac{1}{2}\boldsymbol{\mu}^T\bar{Q}\boldsymbol{\mu}. \end{aligned} \tag{B.18}$$

Let $B_{\boldsymbol{\eta}}$ be the set containing all feasible $\boldsymbol{\eta}$. That is, all $\boldsymbol{\eta}$ satisfy (B.7) and (B.8). Similarly, $B_{\boldsymbol{\mu}}$ is the set containing all feasible $\boldsymbol{\mu}$. Then,

$$\bar{\alpha}_t(f^t(\mathbf{x}_t) - y_t - \epsilon) \leq \min_{\boldsymbol{\eta}\in B_{\boldsymbol{\eta}}} \frac{1}{2}\boldsymbol{\eta}^T\bar{Q}\boldsymbol{\eta} + \min_{\boldsymbol{\mu}\in B_{\boldsymbol{\mu}}} \frac{1}{2}\boldsymbol{\mu}^T\bar{Q}\boldsymbol{\mu}, \tag{B.19}$$

since (B.18) is valid for all feasible $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$.

Recall $\tilde{K} = K + I/C$ and define $g_i = \eta_i - \eta_{i+l}$ for any $i = 1, \ldots, l$. From the definition of (B.1) and (B.2),

$$\boldsymbol{\eta}^T\bar{Q}\boldsymbol{\eta} = \sum_{i=1}^{l}\sum_{j=1}^{l} g_i g_j \tilde{K}_{ij} = g_t^2\tilde{K}_{tt} + 2g_t\sum_{i\neq t} g_i\tilde{K}_{it} + \sum_{i\neq t}\sum_{j\neq t} g_i g_j \tilde{K}_{ij}.$$

Moreover, we can rewrite

$$\boldsymbol{\eta}^T\bar{Q}\boldsymbol{\eta} = g_t^2(\tilde{K}_{tt} - 2\sum_{i\neq t}\lambda_i\tilde{K}_{it} + \sum_{i\neq t}\sum_{j\neq t}\lambda_i\lambda_j\tilde{K}_{ij}), \tag{B.20}$$

where

$$\lambda_i = -\frac{g_i}{g_t}, \quad i = 1, \ldots, l. \tag{B.21}$$

When $\alpha_t > 0$, $\alpha_{t+l} = 0$ from (B.17). Therefore, g_t is not zero since $\eta_{t+l} = 0$ and $g_t = \bar{\alpha}_t$. From (B.8),

$$\sum_{i=1, i\neq t}^{l} \lambda_i = 1. \tag{B.22}$$

Note that

$$\eta_i\eta_{i+l} = 0. \tag{B.23}$$

from (B.7) and (B.17). Therefore, from $\tilde{K}_{ij} = \tilde{\phi}(\mathbf{x}_i)^T\tilde{\phi}(\mathbf{x}_j)$, and (B.23), so (B.20), (B.21), and (B.22) imply

$$\min_{\boldsymbol{\eta}\in B_{\boldsymbol{\eta}}} \boldsymbol{\eta}^T\bar{Q}\boldsymbol{\eta} = g_t^2 d^2(\tilde{\phi}(\mathbf{x}_t), \Lambda_t) = \bar{\alpha}_t^2 d^2(\tilde{\phi}(\mathbf{x}_t), \Lambda_t), \tag{B.24}$$

where

$$\Lambda_t = \{\sum_{i=1, i\neq t}^{l} \lambda_i\tilde{\phi}(\mathbf{x}_i) \mid \sum_{i\neq t}\lambda_i = 1,\ \lambda_i \geq -\frac{\bar{\alpha}_i}{g_t} \text{ if } \bar{\alpha}_i \geq 0, \lambda_i \leq \frac{\bar{\alpha}_{i+l}}{g_t} \text{ if } \bar{\alpha}_{i+l} \geq 0\},$$

and $d(\tilde{\phi}(\mathbf{x}_t), \Lambda_t)$ is the distance between $\tilde{\phi}(\mathbf{x}_t)$ and the set Λ_t in the feature space.

Define a subset of Λ_t as:

$$\Lambda_t^+ = \{\sum_{i=1, i\neq t}^{l} \lambda_i\tilde{\phi}(\mathbf{x}_i) \mid \sum_{i\neq t}\lambda_i = 1, \lambda_i \geq 0,\ \lambda_i \geq -\frac{\bar{\alpha}_i}{g_t} \text{ if } \bar{\alpha}_i \geq 0, \lambda_i \leq \frac{\bar{\alpha}_{i+l}}{g_t} \text{ if } \bar{\alpha}_{i+l} \geq 0\}.$$

Recall that in (B.9), one way to find feasible $\bar{\boldsymbol{\alpha}}-\boldsymbol{\eta}$ is to decrease some free $\bar{\alpha}_{l+1}, \ldots, \bar{\alpha}_{2l}$. With $g_t = \eta_t = \bar{\alpha}_t > 0$, this is achieved by using positive λ_i: $\bar{\alpha}_{i+l} - \lambda_i g_t$. Thus, Λ_t^+ is nonempty. As Λ_t^+ is a subset of the convex hull by $\tilde{\phi}(\mathbf{x}_i), i = 1, \ldots, l, i \neq t$,

$$d^2(\tilde{\phi}(\mathbf{x}_t), \Lambda_t) \leq d^2(\tilde{\phi}(\mathbf{x}_t), \Lambda_t^+) \leq \max_{i \neq t} \|\tilde{\phi}(\mathbf{x}_t) - \tilde{\phi}(\mathbf{x}_i)\|^2 \leq 4\tilde{R}^2. \tag{B.25}$$

(B.24) then implies

$$\min_{\boldsymbol{\eta} \in B_{\boldsymbol{\eta}}} \boldsymbol{\eta}^T \bar{Q} \boldsymbol{\eta} \leq 4\bar{\alpha}_t^2 \tilde{R}^2. \tag{B.26}$$

Similarly,

$$\min_{\boldsymbol{\mu} \in B_{\boldsymbol{\mu}}} \boldsymbol{\mu}^T \bar{Q} \boldsymbol{\mu} \leq 4\bar{\alpha}_t^2 \tilde{R}^2. \tag{B.27}$$

Combining the above inequalities and (B.19), and canceling out $\bar{\alpha}_t$, we have

$$f^t(\mathbf{x}_t) - y_t \leq 4\tilde{R}^2 \bar{\alpha}_t + \epsilon = 4\tilde{R}^2 \alpha_t + \epsilon. \tag{B.28}$$

3) $\alpha_t^* > 0$

The result can be proved through a similar procedure for the case of $\alpha_t > 0$. □

C Proof of Theorem 2

Define $\bar{\boldsymbol{\eta}} = -\bar{\boldsymbol{\mu}} = \bar{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}^t$. Under the assumption the set of support vectors remains the same during the loo procedure, $\bar{\boldsymbol{\eta}} \in B_{\boldsymbol{\eta}}$ and $\bar{\boldsymbol{\mu}} \in B_{\boldsymbol{\mu}}$. Then, (B.18) becomes an equality:

$$\bar{\alpha}_t(f^t(\mathbf{x}_t) - y_t - \epsilon) = \bar{\boldsymbol{\eta}}^T \bar{Q} \bar{\boldsymbol{\eta}} = \min_{\boldsymbol{\eta} \in B_{\boldsymbol{\eta}}} \boldsymbol{\eta}^T \bar{Q} \boldsymbol{\eta} = \bar{\alpha}_t^2 d^2(\tilde{\phi}(\mathbf{x}_t), \Lambda_t).$$

Since $\bar{\boldsymbol{\eta}} = \bar{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}^t$ and the sets of support vectors of $\bar{\boldsymbol{\alpha}}$ and $\bar{\boldsymbol{\alpha}}^t$ are the same,

$$d^2(\tilde{\phi}(\mathbf{x}_t), \Lambda_t) = d^2(\tilde{\phi}(\mathbf{x}_t), \Lambda_t'),$$

where

$$\Lambda_t' = \{ \sum_{i \neq t, \alpha_i + \alpha_i^* > 0} \lambda_i \tilde{\phi}(\mathbf{x}_i) \mid \sum_{i \neq t} \lambda_i = 1 \}.$$

The reason is that, using the assumption, we do not need to consider the constraints associated with free support vectors in the definition of Λ_t. Therefore, it follows that

$$loo \leq \sum_{t=1}^{l} (\alpha_t + \alpha_t^*) S_t^2 + l\epsilon,$$

where S_t^2 is the optimal objective value of (2.5) with $\mathcal{F}$ replaced by $\{i \mid i \neq t, \alpha_i + \alpha_i^* > 0\}$.

D LOO Bounds for L1-SVR

D.1 Modifications in the Proof of Lemma 1

In Lemma 1, $\alpha_t > 0$ implies $\xi_t > 0$, which is used for the strict inequality in (A.4). However, for L1-SVR, ξ_t may be zero even if $\alpha_t > 0$. To prove the inequality, we consider (A.3) in which the equality actives only if

$$\frac{1}{2}(\mathbf{w}^t)^T(\mathbf{w}^t) + C\sum_{i\neq t}\xi_i^t + C\sum_{i\neq t}\xi_i^{t*} = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i\neq t}\xi_i + C\sum_{i\neq t}\xi_i^*.$$

Therefore, $(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*)$ is optimal for (3.2) as well. Using Assumption 1, $(\mathbf{w}, b) = (\mathbf{w}^t, b^t)$, so

$$f^t(\mathbf{x}_t) = f(\mathbf{x}_t) \geq y_t$$

contradicts the assumption that $f^t(\mathbf{x}_t) < y_t$.

D.2 Modifications in the Proof of Lemma 2

The proof for the case of $\alpha_t = \alpha_t^* = 0$ is exactly the same, so we focus on the case of $\alpha_t > 0$. Similar to the L2 case, we consider a form like (B.1) and $\bar{\boldsymbol{\alpha}}$ becomes the dual variable.

Now $\mathcal{F}^t$ is redefined as $\{i \mid 0 < \bar{\alpha}_i < C, i \neq t, t+l\}$. We claim that $\boldsymbol{\eta}$ still exists so that

$$\begin{aligned} 0 \leq \bar{\alpha}_i - \eta_i \leq C, \quad & i \in \mathcal{F}^t, \\ \eta_i = 0, \quad & i \notin \mathcal{F}^t, i \neq t, i \neq t+l, \\ \eta_i = \bar{\alpha}_i, \quad & i = t, i = t+l, \end{aligned} \tag{D.1}$$

and

$$\mathbf{z}^T\boldsymbol{\eta} = 0 \tag{D.2}$$

are satisfied. In order to decrease $\bar{\alpha}_t$ to zero, one may decrease some free $\bar{\alpha}_{l+1}, \ldots, \bar{\alpha}_{2l}$ so that (B.9) is satisfied. However, for L1-SVR, it is possible that after all free $\bar{\alpha}_{l+1}, \ldots, \bar{\alpha}_{2l}$ are decreased to 0, $\bar{\alpha}_t$ is not zero yet. At this point we must increase some free $\bar{\alpha}_1, \ldots, \bar{\alpha}_l$. Since we keep $\mathbf{e}^T\bar{\boldsymbol{\alpha}} = 0$ and all $\bar{\alpha}_{l+1}, \ldots, \bar{\alpha}_{2l}$ have been updated to zero or remain at C,

$$\bar{\alpha}_t + \sum_{\substack{i=1,\ldots,l, i\neq t, \\ 0<\bar{\alpha}_i<C}} \bar{\alpha}_i = \Delta C, \tag{D.3}$$

where $\Delta \geq 1$ is an integer. (D.3) implies that one can reduce $\bar{\alpha}_t$ to zero and increase free $\bar{\alpha}_i, i = 1, \ldots, l$ without exceeding the upper bound.

In Appendix B, from (B.10) to (B.11), we use the property that for any $i \in \mathcal{F}^t \cup \{t\}$, $(\bar{Q}\bar{\boldsymbol{\alpha}} + \mathbf{p})_i = -z_i b_i$. Now this equality may not hold when $i = t$. If $\bar{\alpha}_t = C$,

$$(\bar{Q}\bar{\boldsymbol{\alpha}} + \mathbf{p})_t = -z_t b_t - \xi_t,$$

and $\xi_t^* = 0$, so (B.11) becomes

$$F(\boldsymbol{\gamma}) - F(\bar{\boldsymbol{\alpha}}) = \frac{1}{2}\boldsymbol{\eta}^T \bar{Q}\boldsymbol{\eta} + \bar{\alpha}_t(\xi_t + \xi_t^*). \tag{D.4}$$

For $\bar{\boldsymbol{\alpha}}^t$, now $\mathcal{F}^t = \{i \mid 0 < \bar{\alpha}_i^t < C, i \neq t, t + l\}$. Using Assumption 1, $\mathcal{F}^t \neq \emptyset$. By a similar argument on the existence of $\boldsymbol{\eta}$, there is $\boldsymbol{\mu}$ such that

$$\begin{aligned} 0 \leq \bar{\alpha}_i^t - \mu_i \leq C, \quad & i \in \bar{\mathcal{F}}^t, \\ \mu_i = 0, \quad & i \notin \bar{\mathcal{F}}^t, i \neq t, i \neq t + l, \\ \mu_i = -\bar{\alpha}_i, \quad & i = t, i = t + l, \end{aligned} \tag{D.5}$$

and

$$\mathbf{z}^T \boldsymbol{\mu} = 0. \tag{D.6}$$

Thus, (B.16) holds. With (D.4), an inequality similar to (B.19) is

$$\bar{\alpha}_t(f^t(\mathbf{x}_t) - y_t - \epsilon) \leq \min_{\boldsymbol{\eta} \in B_{\boldsymbol{\eta}}} \frac{1}{2}\boldsymbol{\eta}^T \bar{Q}\boldsymbol{\eta} + \min_{\boldsymbol{\mu} \in B_{\boldsymbol{\mu}}} \frac{1}{2}\boldsymbol{\mu}^T \bar{Q}\boldsymbol{\mu} + \bar{\alpha}_t(\xi_t + \xi_t^*).$$

We then use the same derivation from (B.20) to (B.27), but in the definition of Λ_t and Λ_t^+, λ_i is confined by

$$0 \leq \bar{\alpha}_i + \lambda_i g_t \leq C \text{ and } 0 \leq \bar{\alpha}_{i+l} - \lambda_i g_t \leq C, i = 1, \ldots, l. \tag{D.7}$$

In the discussion near (D.3), a feasible $\bar{\boldsymbol{\alpha}} - \boldsymbol{\eta}$ can be obtained by decreasing some free $\bar{\alpha}_{l+1}, \ldots, \bar{\alpha}_{2l}$, an operation which uses positive λ_i in $\bar{\alpha}_{i+l} - \lambda_i g_t$. We may have to increase some free $\bar{\alpha}_1, \ldots, \bar{\alpha}_l$ as well. This also requires positive λ_i in $\bar{\alpha}_i + \lambda_i g_t$. Therefore, $\Lambda_t^+ \neq \emptyset$, so (B.26) (and similarly (B.27)) follows.

Finally, (B.28) becomes

$$f^t(\mathbf{x}_t) - y_t \leq 4R^2\alpha_t + \xi_t + \xi_t^* + \epsilon.$$

E Proof of Theorem 3

We consider the formula (B.1). $\bar{\boldsymbol{\alpha}}$ is continuous if for any θ', $\lim_{\theta\to\theta'} \bar{\boldsymbol{\alpha}}(\theta) = \bar{\boldsymbol{\alpha}}(\theta')$. If this is wrong, there is a convergent sequence $\{\bar{\boldsymbol{\alpha}}(\theta_i)\}$ such that $\lim_{i\to\infty}\{\theta_i\} = \theta'$ but $\lim_{i\to\infty} \bar{\boldsymbol{\alpha}}(\theta_i) = \bar{\boldsymbol{\alpha}}' \neq \bar{\boldsymbol{\alpha}}(\theta')$. Note that the existence of the convergent sequence requires that $\{\bar{\boldsymbol{\alpha}}(\theta_i)\}$ are in a compact set. This property has been discussed in (Lin 2001b) for L2-SVM. Since $\bar{\boldsymbol{\alpha}}(\theta_i)$ and $\bar{\boldsymbol{\alpha}}(\theta)$ are both optimal solutions at θ_i and θ, respectively,

$$\begin{aligned}
&\frac{1}{2}\bar{\boldsymbol{\alpha}}(\theta_i)^T\bar{Q}(\theta_i)\bar{\boldsymbol{\alpha}}(\theta_i) + \mathbf{p}^T\bar{\boldsymbol{\alpha}}(\theta_i) \le \frac{1}{2}\bar{\boldsymbol{\alpha}}(\theta')^T\bar{Q}(\theta_i)\bar{\boldsymbol{\alpha}}(\theta') + \mathbf{p}^T\bar{\boldsymbol{\alpha}}(\theta'), \text{ and} \\
&\frac{1}{2}\bar{\boldsymbol{\alpha}}(\theta')^T\bar{Q}(\theta')\bar{\boldsymbol{\alpha}}(\theta') + \mathbf{p}^T\bar{\boldsymbol{\alpha}}(\theta') \le \frac{1}{2}\bar{\boldsymbol{\alpha}}(\theta_i)^T\bar{Q}(\theta_i)\bar{\boldsymbol{\alpha}}(\theta_i) + \mathbf{p}^T\bar{\boldsymbol{\alpha}}(\theta_i).
\end{aligned} \tag{E.1}$$

With Assumption 2 that all kernel elements are continuous, $\lim_{i\to\infty}\bar{Q}(\theta_i) = \bar{Q}(\theta')$. Taking the limit of (E.1),

$$\frac{1}{2}\bar{\boldsymbol{\alpha}}(\theta')^T\bar{Q}(\theta')\bar{\boldsymbol{\alpha}}(\theta') + \mathbf{p}^T\bar{\boldsymbol{\alpha}}(\theta') = \frac{1}{2}(\bar{\boldsymbol{\alpha}}')^T\bar{Q}(\theta')\bar{\boldsymbol{\alpha}}' + \mathbf{p}^T\bar{\boldsymbol{\alpha}}'.$$

Thus, $\bar{\boldsymbol{\alpha}}'$ is an optimal solution, too. Since the optimal solution is unique under θ', $\bar{\boldsymbol{\alpha}} = \boldsymbol{\alpha}(\theta')$, a contradiction. Therefore, $\boldsymbol{\alpha}$ is continuous.

About $\tilde{R}^2$, it is the optimal objective value of (4.4). By the same procedure, $\boldsymbol{\beta}$ is continuous, and so is $\tilde{R}^2$. Therefore, the radius margin bound $4\tilde{R}^2\mathbf{e}^T(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*) + l\epsilon$ is continuous.

Next, we prove the continuity of the modified span bound. As we have proved that $\bar{\boldsymbol{\alpha}}$ is continuous, it is sufficient to consider $\tilde{S}_t^2$ only. Define any sequence that converges to θ' as $\{\theta_i\}$. There are corresponding sequences $\{\bar{\boldsymbol{\alpha}}(\theta_i)\}$ and $\{\tilde{S}_t^2(\theta_i)\}$. If for any convergent $\{\theta_i\}$, $\{\tilde{S}_t^2(\theta_i)\}$ converges to $\tilde{S}_t^2(\theta')$, then $\tilde{S}_t^2$ is continuous at θ'. Thus, the convergence of $\{\tilde{S}_t^2(\theta_i)\}$ to $\tilde{S}_t^2(\theta')$ is what we are going to show next.

Note that for any $\bar{\boldsymbol{\alpha}}$, we can define two index sets:

$$\mathcal{F} = \{i \mid 0 < \bar{\alpha}_i\} \text{ and } \mathcal{L} = \{i \mid \bar{\alpha}_i = 0\}. \tag{E.2}$$

They include the indices of free and lower-bounded elements of $\bar{\boldsymbol{\alpha}}$. Thus, for any $\bar{\boldsymbol{\alpha}}(\theta)$, there are associated $\mathcal{F}_\theta$ and $\mathcal{L}_\theta$. Usually we call these sets the face of $\bar{\boldsymbol{\alpha}}(\theta)$. Later if we state that the faces of $\bar{\boldsymbol{\alpha}}(\theta_1)$ and $\bar{\boldsymbol{\alpha}}(\theta_2)$ are identical, it means that $\mathcal{F}_{\theta_1} = \mathcal{F}_{\theta_2}$ and $\mathcal{L}_{\theta_1} = \mathcal{L}_{\theta_2}$.

Because there is only a finite number of possible faces of $\bar{\boldsymbol{\alpha}}$, we can separate $\{\bar{\boldsymbol{\alpha}}(\theta_i)\}$ into a finite number of subsequences such that all elements of each subsequence have

the same face. As it suffices to prove that for any such subsequence, $\{\tilde{S}_t^2(\theta_i)\}$ converges to $\tilde{S}_t^2(\theta')$, without loss of generality, we assume that $\{\bar{\boldsymbol{\alpha}}(\theta_i)\}$ are all at the same face. Since it is a convergent sequence and $\bar{\boldsymbol{\alpha}}(\theta)$ is a continuous function, there is a fixed (maybe empty) set $J \subset \{1, \ldots, l\}$ such that $\alpha_j(\theta) + \alpha_j^*(\theta) > 0$ for any $\theta \in \{\theta_i\}, j \in J$ and

$$\lim_{i \to \infty} \alpha_j(\theta_i) + \alpha_j^*(\theta_i) = 0. \tag{E.3}$$

Now, we calculate the limit of $(\tilde{M}^t)^{-1}$ which was defined in (2.11). We decompose $\tilde{M}^t$ to four blocks

$$\tilde{M}^t = \begin{bmatrix} A_1 & A_2 \\ A_2^T & A_3 \end{bmatrix}.$$

Here, we rearrange $\tilde{M}^t$ such that

$$A_1 = \tilde{K}_{JJ} + \tilde{D}_{JJ} = \tilde{K}_{JJ} + D_{JJ},$$

Hence, the first $|J|$ columns and rows of $\tilde{M}^t$ correspond to indices satisfying (E.3). A_3 is the sub-matrix of $\tilde{M}^t$ without the first $|J|$ columns and rows. We have

$$\begin{bmatrix} A_1 & A_2 \\ A_2^T & A_3 \end{bmatrix}^{-1} = \begin{bmatrix} A_1^{-1}(I + A_2 B^{-1} A_2^T A_1^{-1}) & -A_1^{-1} A_2 B^{-1} \\ -B^{-1} A_2^T A_1^{-1} & B^{-1} \end{bmatrix}, \tag{E.4}$$

where $B = A_3 - A_2^T A_1^{-1} A_2$. From (E.3), it follows every diagonal element of A_1 converges to infinity when $\{\theta_i\}$ approaches θ'. Therefore, from Lemma 2.3.3 of (Golub and Van Loan 1996),

$$\lim_{i \to \infty} A_1(\theta_i)^{-1} = A_1(\theta')^{-1} = O, \tag{E.5}$$

where O is a $|J| \times |J|$ zero matrix. According to (E.4) and (E.5),

$$\lim_{i \to \infty} \tilde{M}^t(\theta_i)^{-1} = \begin{bmatrix} O & O_1 \\ O_1^T & A_3(\theta')^{-1} \end{bmatrix},$$

and

$$\lim_{i \to \infty} \mathbf{h}(\theta_i) = \begin{bmatrix} \mathbf{h}_J(\theta') \\ \mathbf{h}'(\theta') \end{bmatrix},$$

where O_1 is a $|J| \times q$ zero matrix if q is the number of columns of A_3. $\mathbf{h}_j(\theta')$ and $\mathbf{h}'(\theta')$ are sub-vectors of $\mathbf{h}(\theta')$ with the first $|J|$ and the remaining elements, respectively. Then,

$$\lim_{i \to \infty} \tilde{S}_t^2(\theta_i) = \lim_{i \to \infty} \mathbf{h}(\theta_i)^T \tilde{M}^t(\theta_i)^{-1} \mathbf{h}(\theta_i) = \mathbf{h}'(\theta')^T A_3(\theta')^{-1} \mathbf{h}'(\theta') = \tilde{S}_t^2(\theta').$$

Therefore, for any $\{\theta_i\}$ which converges to θ', $\{\tilde{S}_t^2(\theta_i)\}$ converges to $\tilde{S}_t^2(\theta')$ so $\tilde{S}_t^2$ is continuous at θ'.

F An Example that $\mathbf{e}^T(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*)$ of L2-SVR Is Not Differentiable

Consider an L2-SVR problem with $\epsilon = 0.1$,

$$\tilde{K} = \begin{bmatrix} 1+\frac{1}{C} & 0.3 & 0.6 \\ 0.3 & 1+\frac{1}{C} & 0.3 \\ 0.6 & 0.3 & 1+\frac{1}{C} \end{bmatrix}, \tag{F.1}$$

and

$$\mathbf{y} = \begin{bmatrix} 0.4 & -0.1 & 0.9 \end{bmatrix}^T.$$

(F.1) can be the kernel matrix when using the RBF kernel. For example, if $\sigma = 1/\sqrt{2}$, then

$$\begin{aligned} \|\mathbf{x}_1 - \mathbf{x}_2\| = \sqrt{-\log 0.3} &\approx 1.0973, \\ \|\mathbf{x}_1 - \mathbf{x}_3\| = \sqrt{-\log 0.6} &\approx 0.7147, \\ \|\mathbf{x}_2 - \mathbf{x}_3\| = \sqrt{-\log 0.3} &\approx 1.0973, \end{aligned}$$

There are $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ which form a triangle satisfying the above.

Assume δ is a small positive number. If $C = 2+\delta$, then the optimal solution is

$$\boldsymbol{\alpha} = \begin{bmatrix} \frac{5C(C-2)}{\Delta} & \frac{13C(2C+5)}{\Delta} & 0 \end{bmatrix}^T \text{ and } \boldsymbol{\alpha}^* = \begin{bmatrix} 0 & 0 & \frac{31C^2+55C}{\Delta} \end{bmatrix}^T,$$

where

$$\Delta = 6(4C+5)(2C+5).$$

It follows:

$$\lim_{C\to 2^+} \frac{\partial \mathbf{e}^T(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*)}{\partial C} = \frac{55}{351}.$$

If $C = 2-\delta$, then the optimal solution is

$$\boldsymbol{\alpha} = \begin{bmatrix} 0 & \frac{4C}{7C+10} & 0 \end{bmatrix}^T \text{ and } \boldsymbol{\alpha}^* = \begin{bmatrix} 0 & 0 & \frac{4C}{7C+10} \end{bmatrix}^T.$$

Then,

$$\lim_{C\to 2^-} \frac{\partial \mathbf{e}^T(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*)}{\partial C} = \frac{5}{36}.$$

Therefore, $\mathbf{e}^T(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*)$ of L2-SVR may not be a differentiable function of C.

G Proof of Theorem 4

We prove these bounds are piecewise differentiable. Such a property implies that the function is locally Lipschitz continuous and hence differentiable almost everywhere (Clarke 1983). First we define piece-wise differentiable functions:

Definition 1
1) C^k is a function class in which every function is differentiable k times.
2) A function $\mathbf{f} : R^n \to R^m$ is called a PC^k function (kth piecewise differentiable), $1 \leq k \leq \infty$, if $\mathbf{f}$ is continuous and for every point $\mathbf{x} \in R^n$, there exists a neighborhood W of $\mathbf{x}$ and a finite collection of C^k-function $\mathbf{f}^i : W \to R^m$, $i = 1, \ldots, N$, such that

$$\mathbf{f}(\mathbf{x}) \in \{\mathbf{f}^1(\mathbf{x}), \ldots, \mathbf{f}^N(\mathbf{x})\}, \forall \mathbf{x} \in W.$$

PC^1 functions are also called piecewise differentiable functions. There are some useful properties about PC^k functions.

Theorem 5

1. *A function $\mathbf{f}(\mathbf{x}) : V \to R^m$ defined on the open set $V \subset R^n$ is piecewise differentiable if and only if its component functions $\mathbf{f}_i(\mathbf{x}) : V \to R, i = 1, \ldots, m$ are also piecewise differentiable.*

2. *(Ulbrich 2000, Proposition 2.20) The class of PC^k-functions is closed under composition, finite summation, and multiplication.*

We then have a lemma to prove that $\bar{\boldsymbol{\alpha}}(\theta)$ is a piece-wise differentiable function.

Lemma 4 *If in Assumption 2, the kernel function is k-times differentiable, $\bar{\boldsymbol{\alpha}}(\theta)$, the optimal solution of* (B.1)*, is a PC^k function.*

Proof. From Definition 1, a function is PC^k if, at any point, there is a neighborhood such that the function in this neighborhood consists of a finite number of k-times differentiable functions. Below we construct finite k-times differentiable functions, so $\bar{\boldsymbol{\alpha}}(\theta)$ is composed of them for any θ.

As $\bar{\boldsymbol{\alpha}}$ is an optimal solution of (B.1), its KKT optimality condition (3.1) can be rewritten as

$$\begin{aligned} (\bar{Q}\bar{\boldsymbol{\alpha}})_i + p_i + z_i b \quad &\geq 0, \text{ if } \bar{\alpha}_i = 0, \\ &= 0, \text{ if } 0 < \bar{\alpha}_i. \end{aligned} \tag{G.1}$$

Assume $\bar{\boldsymbol{\alpha}}$ is obtained under a parameter set θ_1 and is at the face $\mathcal{F}$ and $\mathcal{L}$ defined in (E.2). We first consider the case where $\mathcal{F} \neq \emptyset$. The KKT condition of free support vectors can be written as:

$$\bar{Q}_{\mathcal{F}\mathcal{F}}\bar{\boldsymbol{\alpha}}_{\mathcal{F}} + \bar{Q}_{\mathcal{F}\mathcal{L}}\bar{\boldsymbol{\alpha}}_{\mathcal{L}} + b\mathbf{z}_{\mathcal{F}} = -\mathbf{p}_{\mathcal{F}}. \tag{G.2}$$

If we combine (G.2) and the linear constraint, $\mathbf{z}^T\bar{\boldsymbol{\alpha}} = 0$, $\bar{\boldsymbol{\alpha}}_{\mathcal{F}}$ and b are the solution of the following linear system:

$$\begin{bmatrix} \bar{Q}_{\mathcal{F}\mathcal{F}} & \mathbf{z}_{\mathcal{F}} \\ \mathbf{z}_{\mathcal{F}}^T & 0 \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{\alpha}}_{\mathcal{F}} \\ b \end{bmatrix} = \begin{bmatrix} -\mathbf{p}_{\mathcal{F}} \\ 0 \end{bmatrix}. \tag{G.3}$$

As $\mathcal{F}$ and $\mathcal{L}$ are the face of $\bar{\boldsymbol{\alpha}}$, the above equation uses the fact that α_i equals to 0 for every $i \in \mathcal{L}$.

From (G.3),

$$\begin{bmatrix} \bar{\boldsymbol{\alpha}}_{\mathcal{F}} \\ b \end{bmatrix} = M^{-1}\mathbf{h}, \tag{G.4}$$

where

$$M = \begin{bmatrix} \bar{Q}_{\mathcal{F}\mathcal{F}} & \mathbf{z}_{\mathcal{F}} \\ \mathbf{z}_{\mathcal{F}}^T & 0 \end{bmatrix} \text{ and } \mathbf{h} = \begin{bmatrix} -\mathbf{p}_{\mathcal{F}} \\ 0 \end{bmatrix}.$$

Next, we build a function $\boldsymbol{\gamma}_{\mathcal{F}}(\theta) = (M^{-1}\mathbf{h})_{\mathcal{F}}$ which is made by removing the last component of (G.4). We claim that $\boldsymbol{\gamma}_{\mathcal{F}}(\theta)$ is a k-times differentiable function since the matrix M is invertible and both M and $\mathbf{h}$ are k-times differentiable functions of θ. Furthermore, we can construct a k-times differentiable function $\boldsymbol{\gamma}(\theta)$ as follows:

$$\boldsymbol{\gamma}(\theta) = \begin{bmatrix} \boldsymbol{\gamma}_{\mathcal{F}}(\theta) \\ \mathbf{0}_{\mathcal{L}} \end{bmatrix}, \tag{G.5}$$

where $\mathbf{0}_{\mathcal{L}}$ is the vector containing $|\mathcal{L}|$ zeros. For the other case where $\mathcal{F} = \emptyset$, we can construct $\boldsymbol{\gamma}(\theta) = \mathbf{0}_{\mathcal{L}}$, which is also a k-times differentiable function.

Notice that when $\theta = \theta_1$, $\boldsymbol{\gamma}(\theta_1) = \bar{\boldsymbol{\alpha}}(\theta_1)$. Moreover, for all parameters whose corresponding optimal solutions are at the same face, $\boldsymbol{\alpha}$'s are the same as values of a k-times differentiable function. That is, for any parameter θ_2 where $\bar{\boldsymbol{\alpha}}(\theta_2)$ is at the same face as $\bar{\boldsymbol{\alpha}}(\theta_1)$, $\boldsymbol{\gamma}(\theta_2) = \bar{\boldsymbol{\alpha}}(\theta_2)$.

Next, we collect all possible functions like $\boldsymbol{\gamma}(\theta)$, which can cover $\bar{\boldsymbol{\alpha}}(\theta)$ at any value of θ. As l is the number of training data, there is a finite number of possible faces:

$$\mathcal{F}^i, \mathcal{L}^i, i = 1, \ldots, N, \tag{G.6}$$

where $N \leq 2^{2l}$. For each face, we construct a function $\boldsymbol{\gamma}^i(\theta)$, which, following the explanation earlier, is a k-times differentiable function.

Therefore, for any θ' we have

$$\bar{\boldsymbol{\alpha}}(\theta') \in \{\boldsymbol{\gamma}^1(\theta'), \boldsymbol{\gamma}^2(\theta'), \ldots, \boldsymbol{\gamma}^N(\theta')\}.$$

Since $\bar{\boldsymbol{\alpha}}(\theta)$ is continuous by Theorem 3, $\bar{\boldsymbol{\alpha}}(\theta) \in PC^k$ following from Definition 1. $\square$

Thus, using Theorem 5, the radius margin bound is a PC^k function. Next, we discuss $\tilde{S}_t^2$. Since $\bar{\boldsymbol{\alpha}} \in PC^k$ for every θ', there exists a neighborhood W of θ' and a finite collection of C^k-functions $\bar{\boldsymbol{\alpha}}^i$, $i = 1, \ldots, N$, such that

$$\bar{\boldsymbol{\alpha}}(\theta) \in \{\bar{\boldsymbol{\alpha}}^1(\theta), \ldots, \bar{\boldsymbol{\alpha}}^N(\theta)\}, \forall \theta \in W.$$

For any $\bar{\boldsymbol{\alpha}}^i$ function, we construct a C^k function $\tilde{S}^i$ by (2.11). Note that the first item is a C^k function from Assumption 2, and only the second term involves with $\bar{\boldsymbol{\alpha}}$. Then,

$$\tilde{S}_t^2(\theta) \in \{\tilde{S}^1(\theta), \ldots, \tilde{S}^N(\theta)\}, \forall \theta \in W.$$

Since we have shown that $\tilde{S}_t^2$ is continuous, $\tilde{S}_t^2 \in PC^k$. Furthermore, from Theorem 5, $\sum_{t=1}^{l} \alpha_t \tilde{S}_t^2$ is a PC^k function.

In the above analysis, if around a given parameter set, all $(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)$ share the same face, then the bound is the same as a differentiable function in a neighborhood of this parameter set. Thus, the bound is differentiable there.