

## Chapter 2

# LOO Bounds for Support Vector Machines

### 2.1 Introduction

The success of support vector machine depends on the tuning of its several parameters which affect the generalization error. For example, when given a training set, a practitioner will ask how to choose these parameters which will generalize well. An effective approach is to estimate the generalization error and then search for parameters so that this estimator is minimized. This requires that the estimators are both effective and computationally efficient. Devroye *et al.* [57] give an overview of error estimation. While some estimators (e.g., uniform convergence bounds) are powerful theoretical tools, they are of little use in practical applications, since they are too loose. Others (e.g., cross-validation, bootstrapping) give good estimates, but are computationally inefficient.

Leave-one-out (LOO) method is the extreme case of cross-validation, and in this case, a single point is excluded from the training set, and the classifier is trained using the remaining points. It is then determined whether this new classifier correctly labels the point that was excluded. The process is repeated over the entire training set, and the LOO error is computed by taking the average over these trials. LOO error provides an almost unbiased estimate of the generalization error.

However one shortcoming of the LOO method is that it is highly time consuming, thus methods are sought to speed up the process. An effective approach is to approximate the LOO error by its upper bound that is a function of the parameters. Then, we search for parameter so that this upper bound is minimized. This approach has successfully been developed for both support vector classification machine [97, 114, 119, 207] and support vector regression machine [34].

In this chapter we will introduce other LOO bounds for several algorithms of support vector machine [200, 201, 231].

## 2.2 LOO Bounds for $\varepsilon$ -Support Vector Regression

### 2.2.1 Standard $\varepsilon$ -Support Vector Regression

First, we introduce the standard  $\varepsilon$ -support vector regression ( $\varepsilon$ -SVR). Consider a regression problem with a training set

$$T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (R^n \times \mathcal{Y})^l, \quad (2.1)$$

where  $x_i \in R^n$ ,  $y_i \in \mathcal{Y} = R$ ,  $i = 1, \dots, l$ . Suppose that the loss function is selected to be the  $\varepsilon$ -insensitive loss function

$$c(x, y, f(x)) = |y - f(x)|_\varepsilon = \max\{0, |y - f(x)| - \varepsilon\}. \quad (2.2)$$

In support vector regression framework, the input space is first mapped to a higher dimensional space  $\mathcal{H}$  by

$$\mathbf{x} = \Phi(x), \quad (2.3)$$

and the training set  $T$  turns to be

$$\bar{T} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l)\} \in (\mathcal{H} \times \mathcal{Y})^l, \quad (2.4)$$

where  $\mathbf{x}_i = \Phi(x_i) \in \mathcal{H}$ ,  $y_i \in \mathcal{Y} = R$ ,  $i = 1, \dots, l$ . Then in space  $\mathcal{H}$ , the following primal problem is constructed:

$$\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*), \quad (2.5)$$

$$\text{s.t. } (\mathbf{w} \cdot \mathbf{x}_i) + b - y_i \leq \varepsilon + \xi_i, \quad i = 1, \dots, l, \quad (2.6)$$

$$y_i - (\mathbf{w} \cdot \mathbf{x}_i) - b \leq \varepsilon + \xi_i^*, \quad i = 1, \dots, l, \quad (2.7)$$

$$\xi_i, \xi_i^* \geq 0, \quad i = 1, \dots, l. \quad (2.8)$$

And  $\varepsilon$ -SVR solves this problem by introducing its dual problem

$$\begin{aligned} \max_{\alpha_T^{(*)}} J_T(\alpha^{(*)}) = & -\frac{1}{2} \sum_{i,j=1}^l (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j) \\ & - \varepsilon \sum_{i=1}^l (\alpha_i^* + \alpha_i) + \sum_{i=1}^l y_i (\alpha_i^* - \alpha_i), \end{aligned} \quad (2.9)$$

$$\text{s.t. } \sum_{i=1}^l (\alpha_i^* - \alpha_i) = 0, \quad (2.10)$$

$$0 \leq \alpha_i, \alpha_i^* \leq C, \quad i = 1, \dots, l, \quad (2.11)$$

where  $\alpha_T^{(*)} = (\alpha_1, \alpha_1^*, \dots, \alpha_l, \alpha_l^*)^T$ , and  $K(x_i, x_j) = (x_i \cdot x_j) = (\Phi(x_i) \cdot \Phi(x_j))$  is the kernel function. Thus, the algorithm can be established as follows:

**Algorithm 2.1** ( $\varepsilon$ -SVR)

- (1) Given a training set  $T$  defined in (2.1);
- (2) Select a kernel  $K(\cdot, \cdot)$ , and parameters  $C > 0$  and  $\varepsilon > 0$ ;
- (3) Solve problem (2.9)–(2.11) and get its solution  $\bar{\alpha}_T^{(*)} = (\bar{\alpha}_1, \bar{\alpha}_1^*, \dots, \bar{\alpha}_l, \bar{\alpha}_l^*)^T$ ;
- (4) Construct the decision function as

$$f(x) = f_T(x) = (\bar{w} \cdot x) + \bar{b} = \sum_{i=1}^l (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_i, x) + \bar{b}, \quad (2.12)$$

where  $\bar{b}$  is computed as follows: either choose one  $\bar{\alpha}_j \in (0, C)$ , then

$$\bar{b} = y_j - \sum_{i=1}^l (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_i, x_j) + \varepsilon; \quad (2.13)$$

or choose one  $\bar{\alpha}_k^* \in (0, C)$ , then

$$\bar{b} = y_k - \sum_{i=1}^l (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_i, x_k) - \varepsilon. \quad (2.14)$$

The uniqueness of the solution of primal problem (2.5)–(2.8) is shown by the following theorem [29].

**Theorem 2.2** Suppose  $\bar{\alpha}_T^{(*)} = (\bar{\alpha}_1, \bar{\alpha}_1^*, \dots, \bar{\alpha}_l, \bar{\alpha}_l^*)^T$  is an optimal solution of dual problem (2.9)–(2.11), and there exists a subscript  $i$  such that either  $0 < \bar{\alpha}_i < C$  or  $0 < \bar{\alpha}_i^* < C$ . Then the decision function

$$f(x) = f_T(x) = (w \cdot x) + b = \sum_{i=1}^l (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x, x_i) + b$$

obtained by Algorithm 2.1 is unique.

### 2.2.2 The First LOO Bound

The kernel and the parameters in Algorithm 2.1 are reasonably selected by minimizing the LOO error or its bound. In this section, we recall the definition of this error at first, and then estimate its bound.

The definition of LOO error with respect to Algorithm 2.1 is given as follows:

**Definition 2.3** For Algorithm 2.1, consider the  $\varepsilon$ -insensitive loss function (2.2) and the training set

$$T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in \{R^n \times \mathcal{Y}\}^l, \quad (2.15)$$

where  $x_i \in R^n$ ,  $y_i \in \mathcal{Y} = R$ . Let  $f_{T|t}(x)$  be the decision function obtained by Algorithm 2.1 from the training set  $T|t = T \setminus \{(x_t, y_t)\}$ , then the LOO error of Algorithm 2.1 (with respect to the loss function (2.2) and the training set  $T$ ) is defined as

$$R_{\text{LOO}}(T) = \sum_{t=1}^l |f_{T|t}(x_t) - y_t|_\varepsilon. \quad (2.16)$$

Obviously, the computation cost of the LOO error is very expensive if  $l$  is large. In fact, for a training set including  $l$  points, the computing of the LOO error implies  $l$  times of training. So finding a more easily computed approximation of the LOO error is necessary. An interesting approach is to estimate an upper bound of the LOO error, such that this bound can be computed by training only once.

Now we derive an upper bound of the LOO error for Algorithm 2.1. Obviously, its LOO bound is related with the training set  $T|t = T \setminus \{(x_t, y_t)\}$ ,  $t = 1, \dots, l$ . The corresponding primal problem is

$$\min_{w^t, \xi^t, \xi^{*t}} \frac{1}{2} \|w^t\|^2 + C \sum_{i=1}^l (\xi_i^t + \xi_i^{*t}), \quad (2.17)$$

$$\text{s.t. } (w^t \cdot x_i) + b^t - y_i \leq \varepsilon + \xi_i^t, \quad i = 1, \dots, t-1, t+1, \dots, l, \quad (2.18)$$

$$y_i - (w^t \cdot x_i) - b^t \leq \varepsilon + \xi_i^{*t}, \quad i = 1, \dots, t-1, t+1, \dots, l, \quad (2.19)$$

$$\xi_i^t, \xi_i^{*t} \geq 0, \quad i = 1, \dots, t-1, t+1, \dots, l. \quad (2.20)$$

Its dual problem is

$$\begin{aligned} \max_{\alpha_{T|t}^{(*)}} J_{T|t}(\alpha_{T|t}^{(*)}) = & -\frac{1}{2} \sum_{i,j \neq t} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j) \\ & - \varepsilon \sum_{i \neq t} (\alpha_i^* + \alpha_i) + \sum_{i \neq t} y_i (\alpha_i^* - \alpha_i), \end{aligned} \quad (2.21)$$

$$\text{s.t. } \sum_{i \neq t} (\alpha_i^* - \alpha_i) = 0, \quad (2.22)$$

$$0 \leq \alpha_i, \alpha_i^* \leq C, \quad i = 1, \dots, t-1, t+1, \dots, l, \quad (2.23)$$

where  $\alpha_{T|t}^{(*)} = (\alpha_1, \alpha_1^*, \dots, \alpha_{t-1}, \alpha_{t-1}^*, \alpha_{t+1}, \alpha_{t+1}^*, \dots, \alpha_l, \alpha_l^*)^T$ .

Now let us introduce useful lemmas:

**Lemma 2.4** Suppose problem (2.9)–(2.11) has a solution  $\bar{\alpha}_T^{(*)} = (\bar{\alpha}_1, \bar{\alpha}_1^*, \dots, \bar{\alpha}_l, \bar{\alpha}_l^*)^T$  with a subscript  $i$  such that either  $0 < \bar{\alpha}_i < C$  or  $0 < \bar{\alpha}_i^* < C$ . Suppose also that, for all any  $t = 1, \dots, l$ , problem (2.21)–(2.23) has a solution  $\tilde{\alpha}_{T|t}^{(*)} = (\tilde{\alpha}_1, \tilde{\alpha}_1^*, \dots, \tilde{\alpha}_{t-1}, \tilde{\alpha}_{t-1}^*, \tilde{\alpha}_{t+1}, \tilde{\alpha}_{t+1}^*, \dots, \tilde{\alpha}_l, \tilde{\alpha}_l^*)^T$  with a subscript  $j$  such that either

$0 < \tilde{\alpha}_j < C$  or  $0 < \tilde{\alpha}_j^* < C$ . Let  $f_T(x)$  and  $f_{T|t}(x)$  be the decision functions obtained by Algorithm 2.1 respectively from the training set  $T$  and  $T|t = T \setminus \{x_t, y_t\}$ . Then for  $t = 1, \dots, l$ , we have

- (i) If  $\tilde{\alpha}_t = \tilde{\alpha}_t^* = 0$ , then  $|f_{T|t}(x_t) - y_t| = |f(x_t) - y_t|$ ;
- (ii) If  $\tilde{\alpha}_t > 0$ , then  $f_{T|t}(x_t) \geq y_t$ ;
- (iii) If  $\tilde{\alpha}_t^* > 0$ , then  $f_{T|t}(x_t) \leq y_t$ .

*Proof* Prove the case (i) first: Consider the primal problem (2.5)–(2.8) corresponding to the problem (2.9)–(2.11). Denote its solution as  $(\bar{w}, \bar{b}, \bar{\xi}^{(*)})$ . Note that the corresponding Lagrange multiplier vector is just the solution  $\tilde{\alpha}_T^{(*)} = (\tilde{\alpha}_1, \tilde{\alpha}_1^*, \dots, \tilde{\alpha}_l, \tilde{\alpha}_l^*)^T$  of the problem (2.9)–(2.11). Therefore the KKT conditions can be represented as

$$\bar{w} = \sum_{i=1}^l (\tilde{\alpha}_i^* - \tilde{\alpha}_i) x_i, \quad (2.24)$$

$$\sum_{i=1}^l (\tilde{\alpha}_i^* - \tilde{\alpha}_i) = 0, \quad (2.25)$$

$$(\bar{w} \cdot x_i) + \bar{b} - y_i \leq \varepsilon + \bar{\xi}_i, \quad i = 1, \dots, l, \quad (2.26)$$

$$y_i - (\bar{w} \cdot x_i) - \bar{b} \leq \varepsilon + \bar{\xi}_i^*, \quad i = 1, \dots, l, \quad (2.27)$$

$$\bar{\xi}_i, \bar{\xi}_i^* \geq 0, \quad i = 1, \dots, l, \quad (2.28)$$

$$((\bar{w} \cdot x_i) + \bar{b} - y_i - \varepsilon - \bar{\xi}_i) \tilde{\alpha}_i = 0, \quad i = 1, \dots, l, \quad (2.29)$$

$$((\bar{w} \cdot x_i) + \bar{b} - y_i + \varepsilon - \bar{\xi}_i^*) \tilde{\alpha}_i^* = 0, \quad i = 1, \dots, l, \quad (2.30)$$

$$(C - \tilde{\alpha}_i) \bar{\xi}_i = 0, (C - \tilde{\alpha}_i^*) \bar{\xi}_i^* = 0, \quad i = 1, \dots, l, \quad (2.31)$$

$$0 \leq \tilde{\alpha}_i, \tilde{\alpha}_i^* \leq C, \quad i = 1, \dots, l. \quad (2.32)$$

Define

$$\tilde{w} = \bar{w}, \quad \tilde{b} = \bar{b}, \quad (2.33)$$

$$\begin{aligned} \tilde{\xi}^{(*)} &\triangleq (\tilde{\xi}_1, \tilde{\xi}_1^*, \dots, \tilde{\xi}_{t-1}, \tilde{\xi}_{t-1}^*, \tilde{\xi}_{t+1}, \tilde{\xi}_{t+1}^*, \dots, \tilde{\xi}_l, \tilde{\xi}_l^*)^T \\ &= (\tilde{\xi}_1, \tilde{\xi}_1^*, \dots, \tilde{\xi}_{t-1}, \tilde{\xi}_{t-1}^*, \tilde{\xi}_{t+1}, \tilde{\xi}_{t+1}^*, \dots, \tilde{\xi}_l, \tilde{\xi}_l^*)^T, \end{aligned} \quad (2.34)$$

and

$$\begin{aligned} \tilde{\alpha}_{T|t}^{(*)} &\triangleq (\tilde{\alpha}_1, \tilde{\alpha}_1^*, \dots, \tilde{\alpha}_{t-1}, \tilde{\alpha}_{t-1}^*, \tilde{\alpha}_{t+1}, \tilde{\alpha}_{t+1}^*, \dots, \tilde{\alpha}_l, \tilde{\alpha}_l^*)^T \\ &= (\tilde{\alpha}_1, \tilde{\alpha}_1^*, \dots, \tilde{\alpha}_{t-1}, \tilde{\alpha}_{t-1}^*, \tilde{\alpha}_{t+1}, \tilde{\alpha}_{t+1}^*, \dots, \tilde{\alpha}_l, \tilde{\alpha}_l^*)^T. \end{aligned} \quad (2.35)$$

It is not difficult to see, from (2.24)–(2.32), that  $(\tilde{w}, \tilde{b}, \tilde{\xi}^{(*)})$  with the vector  $\tilde{\alpha}_{T|t}^{(*)}$  satisfies the KKT conditions of problem (2.17)–(2.20). Therefore,  $(\tilde{w}, \tilde{b}, \tilde{\xi}^{(*)})$  is

the optimal solution of the problem (2.17)–(2.20). Noticing (2.33) and using Theorem 2.2, we claim that  $f_{T|t}(x) = f(x)$ , so

$$|f_{T|t}(x_t) - y_t| = |f(x_t) - y_t|. \quad (2.36)$$

Next, prove the case (ii): Consider the solution with respect to  $(\bar{w}, \bar{b})$  of problem (2.5)–(2.8) and problem (2.17)–(2.20). There are two possibilities: They have respectively solution  $(\bar{w}, \bar{b})$  and  $(\tilde{w}, \tilde{b})$  with  $(\bar{w}, \bar{b}) = (\tilde{w}, \tilde{b})$ , or have no these solutions. For the former case, it is obvious, from the KKT condition (2.29), that we have

$$f_{T|t}(x_t) = (\tilde{w} \cdot x_t) + \tilde{b} = (\bar{w} \cdot x_t) + \bar{b} = y_t + \varepsilon + \bar{\xi}_t > y_t. \quad (2.37)$$

So we need only to investigate the latter case.

Let  $(\bar{w}, \bar{b}, \bar{\xi}^{(*)})$  and  $(\tilde{w}, \tilde{b}, \tilde{\xi}^{(*)})$  respectively be the solution of primal problem (2.5)–(2.8) and problem (2.17)–(2.20) with

$$(\bar{w}, \bar{b}) \neq (\tilde{w}, \tilde{b}). \quad (2.38)$$

We prove  $f_{T|t}(x_t) \geq y_t$  by a contradiction. Suppose that  $\bar{\alpha}_t > 0$ , and  $f_{T|t}(x_t) < y_t$ . From the KKT condition (2.29), we have

$$(\bar{w} \cdot x_t) + \bar{b} = y_t + \varepsilon + \bar{\xi}_t \geq y_t + \varepsilon > y_t > f_{T|t}(x_t) = (\tilde{w} \cdot x_t) + \tilde{b}. \quad (2.39)$$

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

$$(1 - p)((\bar{w} \cdot x_t) + \bar{b}) + p((\tilde{w} \cdot x_t) + \tilde{b}) = y_t. \quad (2.40)$$

Let

$$(\hat{w}, \hat{b}, \hat{\xi}^{(*)}) = (1 - p)(\bar{w}, \bar{b}, \bar{\xi}^{(*)}) + p(\tilde{w}, \tilde{b}, \tilde{\xi}^{(*)}), \quad (2.41)$$

where  $\tilde{\xi}^{(*)}$  is obtained from  $\tilde{\xi}^{(*)}$  by

$$\tilde{\xi} = (\tilde{\xi}_1, \tilde{\xi}_1^*, \dots, \tilde{\xi}_{t-1}, \tilde{\xi}_{t-1}^*, 0, 0, \tilde{\xi}_{t+1}, \tilde{\xi}_{t+1}^*, \dots, \tilde{\xi}_l, \tilde{\xi}_l^*)^T. \quad (2.42)$$

Thus,  $(\hat{w}, \hat{b}, \hat{\xi}^{(*)})$  with deleting the  $(2t)$ th and  $(2t + 1)$ th components of  $\hat{\xi}^{(*)}$  is a feasible solution of problem (2.17)–(2.20). Therefore, noticing the convexity property, we have

$$\begin{aligned} & \frac{1}{2}(\hat{w} \cdot \hat{w}) + C \sum_{i \neq t} (\hat{\xi}_i + \hat{\xi}_i^*) \\ & \leq (1 - p) \left( \frac{1}{2}(\bar{w} \cdot \bar{w}) + C \sum_{i \neq t} (\bar{\xi}_i + \bar{\xi}_i^*) \right) + p \left( \frac{1}{2}(\tilde{w} \cdot \tilde{w}) + C \sum_{i \neq t} (\tilde{\xi}_i + \tilde{\xi}_i^*) \right) \\ & < \frac{1}{2}(\bar{w} \cdot \bar{w}) + C \sum_{i \neq t} (\bar{\xi}_i + \bar{\xi}_i^*), \end{aligned} \quad (2.43)$$

where the last inequality comes from the fact that  $(\bar{\mathbf{w}}, \bar{b}, \bar{\xi}^{(*)})$  with deleting the  $(2t)$ th and  $(2t + 1)$ th components of  $\bar{\xi}^{(*)}$  is a feasible solution of problem (2.17)–(2.20).

On the other hand, the fact  $\bar{\alpha}_t > 0$ , implies that  $\bar{\xi}_t \geq 0$  and  $\bar{\xi}_t^* = 0$ . Thus, according to (2.42),

$$\hat{\xi}_t = (1 - p)\bar{\xi}_t + p\check{\xi}_t = (1 - p)\bar{\xi}_t \leq \bar{\xi}_t, \quad \hat{\xi}_t^* = (1 - p)\bar{\xi}_t^* + p\check{\xi}_t^* = 0. \quad (2.44)$$

From (2.40), we know that  $(\hat{\mathbf{w}}, \hat{b}, \hat{\xi}, \hat{\xi}^*)$  satisfies the constraint

$$-\varepsilon - \hat{\xi}_t^* \leq (\hat{\mathbf{w}} \cdot \mathbf{x}_t) + \hat{b} - y_t = 0 \leq \varepsilon + \hat{\xi}_t, \quad (2.45)$$

so it is also a feasible solution of problem (2.5)–(2.8). However from (2.43) and (2.44) we have

$$\begin{aligned} \frac{1}{2}(\hat{\mathbf{w}} \cdot \hat{\mathbf{w}}) + C \sum_{i=1}^l (\hat{\xi}_i + \hat{\xi}_i^*) &= \frac{1}{2}(\hat{\mathbf{w}} \cdot \hat{\mathbf{w}}) + C \sum_{i \neq t} (\hat{\xi}_i + \hat{\xi}_i^*) + C(\hat{\xi}_t + \hat{\xi}_t^*) \\ &< \frac{1}{2}(\bar{\mathbf{w}} \cdot \bar{\mathbf{w}}) + C \sum_{i \neq t} (\bar{\xi}_i + \bar{\xi}_i^*) + C(\bar{\xi}_t + \bar{\xi}_t^*) \\ &= \frac{1}{2}(\bar{\mathbf{w}} \cdot \bar{\mathbf{w}}) + C \sum_{i=1}^l (\bar{\xi}_i + \bar{\xi}_i^*), \end{aligned} \quad (2.46)$$

which is contradictive with that  $(\bar{\mathbf{w}}, \bar{\xi}, \bar{\xi}^*)$  is the solution of problem (2.5)–(2.8). Thus if  $\alpha_t > 0$ , there must be  $f_{T|t}(\mathbf{x}_t) \geq y_t$ .

The proof of the case (iii) is similar to case (ii) and is omitted here.  $\square$

**Theorem 2.5** Consider Algorithm 2.1. Suppose  $\bar{\alpha}_T^{(*)} = (\bar{\alpha}_1, \bar{\alpha}_1^*, \dots, \bar{\alpha}_l, \bar{\alpha}_l^*)^T$  is the optimal solution of problem (2.9)–(2.11) and  $f(x)$  is the corresponding decision function. Then the LOO error of this algorithm satisfies

$$\begin{aligned} R_{\text{LOO}}(T) &\leq \sum_{t=1}^l |f(\mathbf{x}_t) - y_t - (\bar{\alpha}_t^* - \bar{\alpha}_t)(R^2 + K(\mathbf{x}_t, \mathbf{x}_t))|_{\varepsilon} \\ &= \sum_{t=1}^l \left| \sum_{i=1}^l (\bar{\alpha}_i^* - \bar{\alpha}_i) K(\mathbf{x}_i, \mathbf{x}_t) + \bar{b} - y_t - (\bar{\alpha}_t^* - \bar{\alpha}_t)(R^2 + K(\mathbf{x}_t, \mathbf{x}_t)) \right|_{\varepsilon}, \end{aligned} \quad (2.47)$$

where

$$R^2 = \max\{K(\mathbf{x}_i, \mathbf{x}_j) \mid i, j = 1, \dots, l\}. \quad (2.48)$$

*Proof* It is sufficient to prove that, for  $t = 1, \dots, l$ ,

$$|f(\mathbf{x}_t) - y_t - (\bar{\alpha}_t^* - \bar{\alpha}_t)(R^2 + K(\mathbf{x}_t, \mathbf{x}_t))|_{\varepsilon} \geq |y_t - f_{T|t}(\mathbf{x}_t)|_{\varepsilon}. \quad (2.49)$$

Its validity will be shown by investigating three cases separately:

(i) The case  $\tilde{\alpha}_t^* = \bar{\alpha}_t = 0$ . In this case, by Lemma 2.4,  $|f_{T|t}(x_t) - y_t| = |f(x_t) - y_t|$ , it is obvious that

$$|f(x_t) - y_t - (\bar{\alpha}_t^* - \bar{\alpha}_t)(R^2 + K(x_t, x_t))| = |f_{T|t}(x_t) - y_t|, \quad (2.50)$$

so the conclusion (2.49) is true.

(ii) The case  $\tilde{\alpha}_t > 0$ . In this case, we have  $\tilde{\alpha}_t^* = 0$ .

First we will construct a feasible solution of problem (2.9)–(2.11) from the solution of problem (2.21)–(2.23) to get some equality.

Suppose problem (2.21)–(2.23) has a solution  $\tilde{\alpha}_{T|t}^{(*)} = (\tilde{\alpha}_1, \tilde{\alpha}_1^*, \dots, \tilde{\alpha}_{t-1}, \tilde{\alpha}_{t-1}^*, \tilde{\alpha}_{t+1}, \tilde{\alpha}_{t+1}^*, \dots, \tilde{\alpha}_l, \tilde{\alpha}_l^*)^T$ , and there exists a subscript  $j$  such that  $0 < \tilde{\alpha}_j < C$  or  $0 < \tilde{\alpha}_j^* < C$ . Now construct  $\gamma^{(*)} = (\gamma_1, \gamma_1^*, \dots, \gamma_l, \gamma_l^*)^T$  as follows:

$$\gamma_i^{(*)} = \begin{cases} \tilde{\alpha}_i^{(*)}, & \text{if } \alpha_i^{(*)} = 0 \text{ or } \alpha_i^{(*)} = C, \\ \tilde{\alpha}_i^{(*)} - v_i^{(*)}, & \text{if } i \in SV^t, \\ \tilde{\alpha}_i^{(*)}, & \text{if } i = t, \end{cases} \quad (2.51)$$

where  $SV^t = \{i \mid 0 < \tilde{\alpha}_i^{(*)} < C, i = 1, \dots, t-1, t+1, l\}$ , and  $v = (v_1, v_1^*, \dots, v_l, v_l^*)^T \geq 0$  satisfies

$$\sum_{i \in SV^t} v_i^* = \bar{\alpha}_t^* = 0, \quad \sum_{i \in SV^t} v_i = \bar{\alpha}_t, \quad v_i = 0 \quad \forall i \notin SV^t. \quad (2.52)$$

It is easy to verify that

$$\sum_{i=1}^l (\gamma_i^* - \gamma_i) = 0, \quad 0 \leq \gamma_i, \gamma_i^* \leq C, \quad i = 1, \dots, l, \quad (2.53)$$

so  $\gamma^{(*)}$  is a feasible solution of problem (2.9)–(2.11),

$$\begin{aligned} J(\gamma^{(*)}) &= J_{T|t}(\tilde{\alpha}_{T|t}^{(*)}) - \frac{1}{2}(\bar{\alpha}_t^* - \bar{\alpha}_t)^2 K(x_t, x_t) - \varepsilon(\bar{\alpha}_t^* + \bar{\alpha}_t) + y_t(\bar{\alpha}_t^* - \bar{\alpha}_t) \\ &\quad - (\bar{\alpha}_t^* - \bar{\alpha}_t) \sum_{i \neq t} (\tilde{\alpha}_i^* - \tilde{\alpha}_i) K(x_t, x_i) \\ &\quad - \sum_{i \in SV^t} (v_i^* - v_i) \left( y_i + \varepsilon - \sum_{j \neq t} (\tilde{\alpha}_j^* - \tilde{\alpha}_j) K(x_i, x_j) \right) \\ &\quad - \frac{1}{2} \sum_{i, j \in SV^t} (v_i^* - v_i)(v_j^* - v_j) K(x_i, x_j) \\ &\quad + (\bar{\alpha}_t^* - \bar{\alpha}_t) \sum_{i \in SV^t} (v_i^* - v_i) K(x_i, x_t). \end{aligned} \quad (2.54)$$



Because there exist  $0 < \tilde{\alpha}_i < C$  or  $0 < \tilde{\alpha}_i^* < C$ , so the solution with respect to  $b$  of problem (2.21)–(2.23) is unique, and we have

$$y_i + \varepsilon - \sum_{j \neq t} (\tilde{\alpha}_j^* - \tilde{\alpha}_j) K(x_i, x_j) = \tilde{b}; \quad (2.55)$$

furthermore, by

$$\sum_{i \in SV^T} (v_i^* - v_i) = \sum_{i \neq t} (v_i^* - v_i) = (\tilde{\alpha}_t^* - \tilde{\alpha}_t), \quad (2.56)$$

we get

$$\begin{aligned} & (\tilde{\alpha}_t^* - \tilde{\alpha}_t) \left( \sum_{i \neq t} (\tilde{\alpha}_i^* - \tilde{\alpha}_i) K(x_t, x_i) + \tilde{b} \right) \\ &= -J(\gamma^{(*)}) + J_{T|t}(\tilde{\alpha}_{T|t}^{(*)}) - \frac{1}{2} (\tilde{\alpha}_t^* - \tilde{\alpha}_t)^2 K(x_t, x_t) - \varepsilon (\tilde{\alpha}_t^* + \tilde{\alpha}_t) + y_t (\tilde{\alpha}_t^* - \tilde{\alpha}_t) \\ & \quad - \frac{1}{2} \sum_{i, j \in SV^t} (v_i^* - v_i) (v_j^* - v_j) K(x_i, x_j) \\ & \quad + (\tilde{\alpha}_t^* - \tilde{\alpha}_t) \sum_{i \in SV^t} (v_i^* - v_i) K(x_i, x_t). \end{aligned} \quad (2.57)$$

Next, we will construct a feasible solution of problem (2.21)–(2.23) from the solution of problem (2.9)–(2.11) to get another equality.

Similarly, we construct  $\beta_{T|t}^{(*)} = (\beta_1, \beta_1^*, \dots, \beta_{t-1}, \beta_{t-1}^*, \beta_{t+1}, \beta_{t+1}^*, \dots, \beta_l, \beta_l^*)^T$  from the solution  $\tilde{\alpha}^{(*)}$  of problem (2.9)–(2.11) as follows:

$$\beta_i^{(*)} = \begin{cases} \tilde{\alpha}_i^{(*)}, & \text{if } \tilde{\alpha}_i^{(*)} = 0 \text{ or } \tilde{\alpha}_i^{(*)} = C, \\ \tilde{\alpha}_i^{(*)} + \eta_i^{(*)}, & \text{if } i \in SV \setminus \{t\}, \end{cases} \quad (2.58)$$

where  $SV = \{i \mid 0 < \tilde{\alpha}_i^{(*)} < C, i = 1, \dots, l\}$ , and  $\eta_{T|t}^{(*)} = (\eta_1, \eta_1^*, \dots, \eta_{t-1}, \eta_{t-1}^*, \eta_{t+1}, \eta_{t+1}^*, \dots, \eta_l, \eta_l^*)^T \geq 0$  satisfies

$$\sum_{i \in SV \setminus \{t\}} \eta_i^* = \tilde{\alpha}_t^* = 0, \quad \sum_{i \in SV \setminus \{t\}} \eta_i = \tilde{\alpha}_t. \quad (2.59)$$

It is easy to verify that

$$\sum_{i \neq t} (\beta_i^* - \beta_i) = 0, \quad 0 \leq \beta_i, \beta_i^* \leq C, \quad i = 1, \dots, t-1, t+1, \dots, l. \quad (2.60)$$

So  $\beta_{T|t}^{(*)}$  is a feasible solution of problem (2.21)–(2.23), and we have

$$\begin{aligned}
J_{T|t}(\beta_{T|t}^{(*)}) &= J(\bar{\alpha}^{(*)}) + \frac{1}{2}(\bar{\alpha}_t^* - \bar{\alpha}_t)^2 K(x_t, x_t) + \varepsilon(\bar{\alpha}_t^* + \bar{\alpha}_t) - y_t(\bar{\alpha}_t^* - \bar{\alpha}_t) \\
&\quad + (\bar{\alpha}_t^* - \bar{\alpha}_t) \sum_{i \neq t} (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_t, x_i) \\
&\quad + \sum_{i \in SV \setminus \{t\}} (\eta_i^* - \eta_i) \left( y_i + \varepsilon - \sum_{j \neq t} (\bar{\alpha}_j^* - \bar{\alpha}_j) K(x_i, x_j) \right) \\
&\quad - \frac{1}{2} \sum_{i, j \in SV \setminus \{t\}} (\eta_i^* - \eta_i)(\eta_j^* - \eta_j) K(x_i, x_j) \\
&\quad + (\bar{\alpha}_t^* - \bar{\alpha}_t) \sum_{i \in SV \setminus \{t\}} (\eta_i^* - \eta_i) K(x_i, x_t). \tag{2.61}
\end{aligned}$$

Because there exist  $0 < \bar{\alpha}_i < C$  or  $0 < \bar{\alpha}_i^* < C$ , so the solution with respect to  $b$  of problem (2.9)–(2.11) is unique, and we have

$$y_i + \varepsilon - \sum_{j \neq t} (\bar{\alpha}_j^* - \bar{\alpha}_j) K(x_i, x_j) = \bar{b}; \tag{2.62}$$

furthermore,

$$\begin{aligned}
-J(\bar{\alpha}^{(*)}) &= -J_{T|t}(\beta_{T|t}^{(*)}) + \frac{1}{2}(\bar{\alpha}_t^* - \bar{\alpha}_t)^2 K(x_t, x_t) + \varepsilon(\bar{\alpha}_t^* + \bar{\alpha}_t) - y_t(\bar{\alpha}_t^* - \bar{\alpha}_t) \\
&\quad + (\bar{\alpha}_t^* - \bar{\alpha}_t) \left( \sum_{i \neq t} (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_t, x_i) + \bar{b} \right) \\
&\quad - \frac{1}{2} \sum_{i, j \in SV \setminus \{t\}} (\eta_i^* - \eta_i)(\eta_j^* - \eta_j) K(x_i, x_j) \\
&\quad + (\bar{\alpha}_t^* - \bar{\alpha}_t) \sum_{i \in SV \setminus \{t\}} (\eta_i^* - \eta_i) K(x_i, x_t). \tag{2.63}
\end{aligned}$$

Now, by the above two equalities (2.57) and (2.63), and the obvious facts  $J(\gamma^{(*)}) \leq J(\bar{\alpha}^*)$  and  $J_{T|t}(\beta_{T|t}^{(*)}) \leq J_{T|t}(\alpha_{T|t}^{(*)})$ , we get

$$\begin{aligned}
&(\bar{\alpha}_t^* - \bar{\alpha}_t) \left( \sum_{i \neq t} (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_t, x_i) + \bar{b} \right) \\
&\geq (\bar{\alpha}_t^* - \bar{\alpha}_t) \left( \sum_{i \neq t} (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_t, x_i) + \bar{b} \right) \\
&\quad - \frac{1}{2} \sum_{i, j \in SV^t} (v_i^* - v_i)(v_j^* - v_j) K(x_i, x_j)
\end{aligned}$$

$$\begin{aligned}
& -\frac{1}{2} \sum_{i,j \in SV \setminus \{t\}} (\eta_i^* - \eta_i)(\eta_j^* - \eta_j) K(x_i, x_j) \\
& + (\bar{\alpha}_t^* - \bar{\alpha}_t) \sum_{i \in SV^t} (v_i^* - v_i) K(x_i, x_t) \\
& + (\bar{\alpha}_t^* - \bar{\alpha}_t) \sum_{i \in SV \setminus \{t\}} (\eta_i^* - \eta_i) K(x_i, x_t). \tag{2.64}
\end{aligned}$$

By (2.52) and (2.59),

$$(\bar{\alpha}_t^* - \bar{\alpha}_t) \sum_{i \in SV^t} (v_i^* - v_i) K(x_i, x_t) \geq 0, \tag{2.65}$$

$$(\bar{\alpha}_t^* - \bar{\alpha}_t) \sum_{i \in SV \setminus \{t\}} (\eta_i^* - \eta_i) K(x_i, x_t) \geq 0. \tag{2.66}$$

Reminding the definition (2.48), we have

$$\frac{1}{2} \sum_{i,j \in SV^t} (v_i^* - v_i)(v_j^* - v_j) K(x_i, x_j) \leq \frac{1}{2} (\bar{\alpha}_t^* - \bar{\alpha}_t)^2 R^2, \tag{2.67}$$

$$\frac{1}{2} \sum_{i,j \in SV \setminus \{t\}} (\eta_i^* - \eta_i)(\eta_j^* - \eta_j) K(x_i, x_j) \leq \frac{1}{2} (\bar{\alpha}_t^* - \bar{\alpha}_t)^2 R^2, \tag{2.68}$$

therefore

$$\begin{aligned}
& \sum_{i \neq t} (\tilde{\alpha}_i^* - \tilde{\alpha}_i) K(x_t, x_i) + \tilde{b} \\
& \leq \left( \sum_{i \neq t} (\tilde{\alpha}_i^* - \tilde{\alpha}_i) K(x_t, x_i) + \tilde{b} \right) - (\bar{\alpha}_t^* - \bar{\alpha}_t) R^2. \tag{2.69}
\end{aligned}$$

By Lemma 2.4, the fact  $\bar{\alpha}_t > 0$  implies that  $f_{T|t}(x_t) \geq y_t$ . Therefore

$$\begin{aligned}
0 & \leq \left( \sum_{i \neq t} (\tilde{\alpha}_i^* - \tilde{\alpha}_i) K(x_t, x_i) + \tilde{b} \right) - y_t \\
& \leq \left( \sum_{i \neq t} (\tilde{\alpha}_i^* - \tilde{\alpha}_i) K(x_t, x_i) + \tilde{b} \right) - (\bar{\alpha}_t^* - \bar{\alpha}_t) R^2 - y_t \\
& = f(x_t) - y_t - (\bar{\alpha}_t^* - \bar{\alpha}_t)(R^2 + K(x_t, x_t)), \tag{2.70}
\end{aligned}$$

that is,

$$\left| \left( \sum_{i \neq t} (\tilde{\alpha}_i^* - \tilde{\alpha}_i) K(x_t, x_i) + \tilde{b} \right) - y_t \right| \leq |f(x_t) - y_t - (\bar{\alpha}_t^* - \bar{\alpha}_t)(R^2 + K(x_t, x_t))| \quad (2.71)$$

and because the function  $|\cdot|_\varepsilon$  is monotonically increasing, so the conclusion (2.49) is true.

(iii) The case  $\bar{\alpha}_t^* > 0$ . Similar with the discussion of case (ii), the conclusion (2.49) is true.  $\square$

### 2.2.3 A Variation of $\varepsilon$ -Support Vector Regression

If we consider the decision function with the formulation  $f(x) = (w \cdot x)$  in  $\varepsilon$ -SVR, we will get the primal problem

$$\min_{w, \xi, \xi^*} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*), \quad (2.72)$$

$$\text{s.t. } (w \cdot x_i) - y_i \leq \varepsilon + \xi_i, \quad i = 1, \dots, l, \quad (2.73)$$

$$y_i - (w \cdot x_i) \leq \varepsilon + \xi_i^*, \quad i = 1, \dots, l, \quad (2.74)$$

$$\xi_i, \xi_i^* \geq 0, \quad i = 1, \dots, l. \quad (2.75)$$

The corresponding dual problem is

$$\begin{aligned} \max_{\alpha_T^{(*)}} J_T(\alpha_T^{(*)}) = & -\frac{1}{2} \sum_{i,j=1}^l (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j) \\ & - \varepsilon \sum_{i=1}^l (\alpha_i^* + \alpha_i) + \sum_{i=1}^l y_i (\alpha_i^* - \alpha_i), \end{aligned} \quad (2.76)$$

$$\text{s.t. } 0 \leq \alpha_i, \alpha_i^* \leq C, \quad i = 1, \dots, l, \quad (2.77)$$

where  $\alpha_T^{(*)} = (\alpha_1, \alpha_1^*, \dots, \alpha_l, \alpha_l^*)^T$ , and  $K(x_i, x_j) = (x_i \cdot x_j) = (\Phi(x_i) \cdot \Phi(x_j))$  is the kernel function.

Thus the algorithm can be established as follows:

**Algorithm 2.6** (A variation of the standard  $\varepsilon$ -SVR)

- (1) Given a training set  $T$  defined in (2.1);
- (2) Select a kernel  $K(\cdot, \cdot)$ , and parameters  $C$  and  $\varepsilon$ ;
- (3) Solve problem (2.76)–(2.77) and get its solution  $\bar{\alpha}_T^{(*)} = (\bar{\alpha}_1, \bar{\alpha}_1^*, \dots, \bar{\alpha}_l, \bar{\alpha}_l^*)^T$ ;

(4) Construct the decision function

$$f(x) = f_T(x) = (w \cdot x) = \sum_{i=1}^l (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x, x_i). \quad (2.78)$$

Because the objective function of problem (2.72) is strictly convex with respect to  $w$ , so the solution of problem (2.72)–(2.75) with respect to  $w$  is unique. Therefore we have the following theorem.

**Theorem 2.7** *The decision function*

$$f(x) = f_T(x) = (w \cdot x) = \sum_{i=1}^l (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x, x_i)$$

obtained by Algorithm 2.6 is unique.

### 2.2.4 The Second LOO Bound

Now we derive an upper bound of the LOO error for Algorithm 2.6. Obviously, its LOO bound is related with the training set  $T|t = T \setminus \{(x_t, y_t)\}$ ,  $t = 1, \dots, l$ . The corresponding primal problem should be

$$\min_{w^t, \xi^t, \xi^{*t}} \frac{1}{2} \|w^t\|^2 + C \sum_{i=1}^l (\xi_i^t + \xi_i^{*t}), \quad (2.79)$$

$$\text{s.t.} \quad (w^t \cdot x_i) - y_i \leq \varepsilon + \xi_i^t, \quad i = 1, \dots, t-1, t+1, \dots, l, \quad (2.80)$$

$$y_i - (w^t \cdot x_i) \leq \varepsilon + \xi_i^{*t}, \quad i = 1, \dots, t-1, t+1, \dots, l, \quad (2.81)$$

$$\xi_i^t, \xi_i^{*t} \geq 0, \quad i = 1, \dots, t-1, t+1, \dots, l. \quad (2.82)$$

Its dual problem is

$$\begin{aligned} \max_{\alpha_{T|t}^{(*)}} J_{T|t}(\alpha_{T|t}^{(*)}) &\triangleq -\frac{1}{2} \sum_{i,j \neq t} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j) \\ &\quad - \varepsilon \sum_{i \neq t} (\alpha_i^* + \alpha_i) + \sum_{i \neq t} y_i (\alpha_i^* - \alpha_i), \end{aligned} \quad (2.83)$$

$$\text{s.t.} \quad 0 \leq \alpha_i, \alpha_i^* \leq C, \quad i = 1, \dots, t-1, t+1, \dots, l, \quad (2.84)$$

where  $\alpha_{T|t}^{(*)} = (\alpha_1, \alpha_1^*, \dots, \alpha_{t-1}, \alpha_{t-1}^*, \alpha_{t+1}, \alpha_{t+1}^*, \dots, \alpha_l, \alpha_l^*)^T$ .

Now let us introduce two useful lemmas:

**Lemma 2.8** Suppose problem (2.76)–(2.77) has a solution  $\bar{\alpha}_T^{(*)} = (\bar{\alpha}_1, \bar{\alpha}_1^*, \dots, \bar{\alpha}_l, \bar{\alpha}_l^*)^T$ . Let  $f_T(x)$  and  $f_{T|t}(x)$  be the decision functions obtained by Algorithm 2.6 respectively from the training set  $T$  and  $T|t = T \setminus \{x_t, y_t\}$ . Then for  $t = 1, \dots, l$ , we have

- (i) If  $\bar{\alpha}_t = \bar{\alpha}_t^* = 0$ , then  $|f_{T|t}(x_t) - y_t| = |f(x_t) - y_t|$ ;
- (ii) If  $\bar{\alpha}_t > 0$ , then  $f_{T|t}(x_t) \geq y_t$ ;
- (iii) If  $\bar{\alpha}_t^* > 0$ , then  $f_{T|t}(x_t) \leq y_t$ .

*Proof* Its proof is similar with Lemma 2.4, and is omitted here.  $\square$

**Lemma 2.9** Suppose problem (2.76)–(2.77) has a solution  $\bar{\alpha}_T^{(*)} = (\bar{\alpha}_1, \bar{\alpha}_1^*, \dots, \bar{\alpha}_l, \bar{\alpha}_l^*)^T$ . Let  $f_{T|t}(x)$  be the decision function obtained by Algorithm 2.6 from the training set  $T|t = T - \{(x_t, y_t)\}$ . Then for  $t = 1, \dots, l$ , we have

$$-(\bar{\alpha}_t^* - \bar{\alpha}_t) \sum_{i \neq t} (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_i, x_t) \geq -(\bar{\alpha}_t^* - \bar{\alpha}_t) f_{T|t}(x_t). \quad (2.85)$$

*Proof* Obviously problem (2.76)–(2.77) can be expressed as

$$\begin{aligned} \max_{\alpha_T^{(*)}} J_T(\alpha_T^{(*)}) &= J_{T|t}(\alpha_{T|t}^{(*)}) - (\alpha_t^* - \alpha_t) \sum_{i \neq t} (\alpha_i^* - \alpha_i) K(x_i, x_t) \\ &\quad - \frac{1}{2} (\alpha_t^* - \alpha_t)^2 K(x_t, x_t) - \varepsilon (\alpha_t^* + \alpha_t) + y_t (\alpha_t^* - \alpha_t), \end{aligned} \quad (2.86)$$

$$\text{s.t. } 0 \leq \alpha_i, \alpha_i^* \leq C, \quad i = 1, \dots, l, \quad (2.87)$$

where  $J_{T|t}(\alpha_{T|t}^{(*)})$  is given by (2.83). Because  $\bar{\alpha}_T^{(*)} = (\bar{\alpha}_1, \bar{\alpha}_1^*, \dots, \bar{\alpha}_l, \bar{\alpha}_l^*)^T$  is the solution of problem (2.76)–(2.77) or (2.86)–(2.87), then problem (2.86)–(2.87) can be rewritten as

$$\begin{aligned} \max_{\alpha_T^{(*)}} J_{T|t}(\alpha_{T|t}^{(*)}) &= (\alpha_t^* - \alpha_t) \sum_{i \neq t} (\alpha_i^* - \alpha_i) K(x_i, x_t) \\ &\quad - \frac{1}{2} (\alpha_t^* - \alpha_t)^2 K(x_t, x_t) - \varepsilon (\alpha_t^* + \alpha_t) + y_t (\alpha_t^* - \alpha_t), \end{aligned} \quad (2.88)$$

$$\text{s.t. } 0 \leq \alpha_i, \alpha_i^* \leq C, \quad i = 1, \dots, t-1, t+1, \dots, l, \quad (2.89)$$

$$\alpha_t^{(*)} = \bar{\alpha}_t^{(*)}. \quad (2.90)$$

Substitute the equality (2.90) into the objective function directly, the problem (2.88)–(2.90) turns to be

$$\max_{\alpha_{T|t}^{(*)}} \hat{J}_T(\alpha_{T|t}^{(*)}) \triangleq J_{T|t}(\alpha_{T|t}^{(*)}) - (\bar{\alpha}_t^* - \bar{\alpha}_t) \sum_{i \neq t} (\alpha_i^* - \alpha_i) K(x_i, x_t), \quad (2.91)$$

$$\text{s.t. } 0 \leq \alpha_i, \alpha_i^* \leq C, \quad i = 1, \dots, t-1, t+1, \dots, l. \quad (2.92)$$

It is easy to see that

$$\begin{aligned}\hat{\alpha}_{T|t}^{(*)} &\triangleq (\hat{\alpha}_1, \hat{\alpha}_1^*, \dots, \hat{\alpha}_{t-1}, \hat{\alpha}_{t-1}^*, \hat{\alpha}_{t+1}, \hat{\alpha}_{t+1}^*, \dots, \hat{\alpha}_l, \hat{\alpha}_l^*)^T \\ &= (\bar{\alpha}_1, \bar{\alpha}_1^*, \dots, \bar{\alpha}_{t-1}, \bar{\alpha}_{t-1}^*, \bar{\alpha}_{t+1}, \bar{\alpha}_{t+1}^*, \dots, \bar{\alpha}_l, \bar{\alpha}_l^*)^T\end{aligned}\quad (2.93)$$

is an optimal solution of problem (2.91)–(2.92). Because  $\tilde{\alpha}_{T|t}^{(*)}$  is the optimal solution of problem (2.83)–(2.84), and is also a feasible solution of problem (2.91)–(2.92), we have

$$\hat{J}_T(\hat{\alpha}_{T|t}^{(*)}) \geq \hat{J}_T(\tilde{\alpha}_{T|t}^{(*)}). \quad (2.94)$$

Therefore by (2.91) and (2.93),

$$\begin{aligned}J_{T|t}(\hat{\alpha}_{T|t}^{(*)}) - (\bar{\alpha}_t^* - \bar{\alpha}_t) \sum_{i \neq t} (\hat{\alpha}_i^* - \hat{\alpha}_i) K(x_i, x_t) \\ \geq J_{T|t}(\tilde{\alpha}_{T|t}^{(*)}) - (\bar{\alpha}_t^* - \bar{\alpha}_t) \sum_{i \neq t} (\tilde{\alpha}_i^* - \tilde{\alpha}_i) K(x_i, x_t),\end{aligned}\quad (2.95)$$

that is

$$\begin{aligned}-(\bar{\alpha}_t^* - \bar{\alpha}_t) \sum_{i \neq t} (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_i, x_t) \\ \geq J_{T|t}(\tilde{\alpha}_{T|t}^{(*)}) - J_{T|t}(\hat{\alpha}_{T|t}^{(*)}) - (\bar{\alpha}_t^* - \bar{\alpha}_t) \sum_{i \neq t} (\bar{\alpha}_i^* - \tilde{\alpha}_i) K(x_i, x_t).\end{aligned}\quad (2.96)$$

Because  $\tilde{\alpha}_{T|t}^{(*)}$  maximizes the objective function  $J_{T|t}(\alpha_{T|t}^{(*)})$ , we have

$$J_{T|t}(\tilde{\alpha}_{T|t}^{(*)}) - J_{T|t}(\hat{\alpha}_{T|t}^{(*)}) \geq 0. \quad (2.97)$$

So the conclusion comes from (2.96) and (2.97).  $\square$

Now we are in a position to show our main conclusion.

**Theorem 2.10** *Consider Algorithm 2.6. Suppose  $\bar{\alpha}_T^{(*)} = (\bar{\alpha}_1, \bar{\alpha}_1^*, \dots, \bar{\alpha}_l, \bar{\alpha}_l^*)^T$  is the optimal solution of problem (2.76)–(2.77) and  $f(x)$  is the decision function. Then the LOO error of this algorithm satisfies*

$$R_{\text{LOO}}(T) \leq \frac{1}{l} \sum_{t=1}^l |y_t - f(x_t) + (\bar{\alpha}_t^* - \bar{\alpha}_t) K(x_t, x_t)|_{\varepsilon}. \quad (2.98)$$

*Proof* It is sufficient to prove that, for  $t = 1, \dots, l$ ,

$$|y_t - f(x_t) + (\bar{\alpha}_t^* - \bar{\alpha}_t) K(x_t, x_t)|_{\varepsilon} \geq |y_t - f_{T|t}(x_t)|_{\varepsilon}. \quad (2.99)$$

We complete the proof by investigating three cases separately:

(i) The case  $\bar{\alpha}_t^* > 0$ . In this case, we have  $\bar{\alpha}_t = 0$ , so  $(\bar{\alpha}_t^* - \bar{\alpha}_t) > 0$ . Thus by Lemma 2.9,

$$-\sum_{i \neq t} (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_i, x_t) \geq -f_{T|t}(x_t). \quad (2.100)$$

Furthermore, by Lemma 2.8, the fact  $\bar{\alpha}_t^* > 0$  implies that  $f_{T|t}(x_t) \leq y_t$ . Therefore, inequality (2.100) leads to

$$y_t - \sum_{i \neq t} (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_i, x_t) \geq y_t - f_{T|t}(x_t) \geq 0, \quad (2.101)$$

and because the function  $|\cdot|_\varepsilon$  is monotonically increasing, so the conclusion (2.99) is true.

(ii) The case  $\bar{\alpha}_t > 0$ . In this case, we have  $\bar{\alpha}_t^* = 0$ , so  $-(\bar{\alpha}_t^* - \bar{\alpha}_t) > 0$ . Thus, by Lemma 2.9,

$$\sum_{i \neq t} (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_i, x_t) \geq f_{T|t}(x_t). \quad (2.102)$$

Furthermore, by Lemma 2.8, the fact  $\bar{\alpha}_t > 0$  implies that  $f_{T|t}(x_t) \geq y_t$ . Therefore, inequality (2.102) leads to

$$\sum_{i \neq t} (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_i, x_t) - y_t \geq f_{T|t}(x_t) - y_t \geq 0, \quad (2.103)$$

and because the function  $|\cdot|_\varepsilon$  is monotonically increasing, so the conclusion (2.99) is true.

(iii) The validity of the conclusion (2.99) is obvious for the case  $\bar{\alpha}_t^* = \bar{\alpha}_t = 0$  by Lemma 2.9.  $\square$

## 2.2.5 Numerical Experiments

In this section, we will compare the proposed first LOO bound and second LOO bound with the true corresponding LOO errors. Consider the real dataset—"Boston Housing Data", which is a standard regression testing problem. This dataset includes 506 instances, each instance has 13 attributes and a real-valued output.

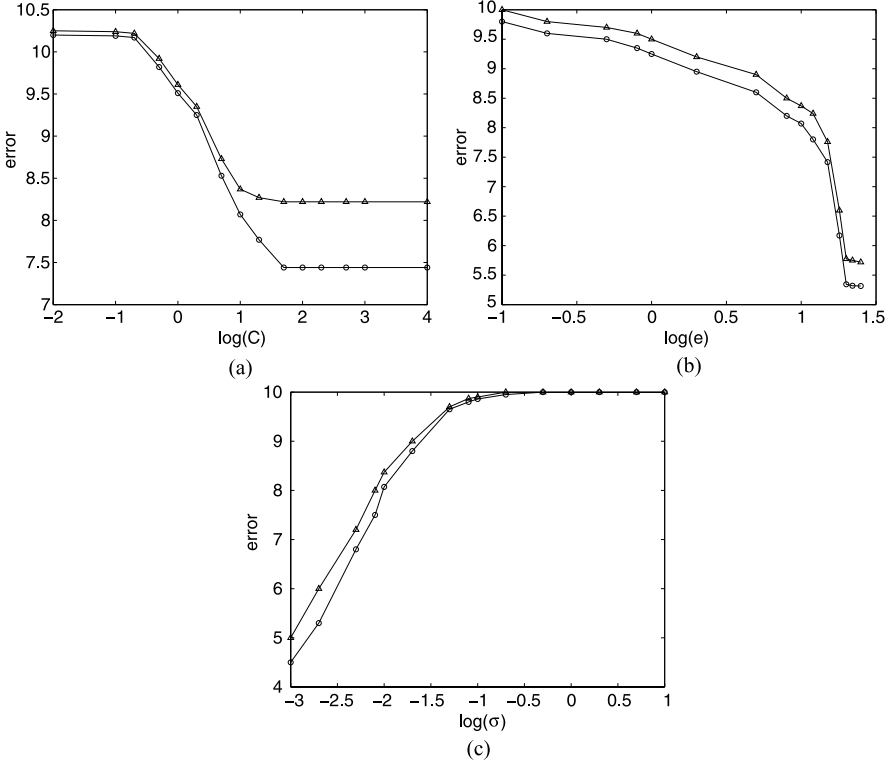
Here we randomly choose 50 instances for training, and the Radial Basis Kernel

$$K(x, x') = \exp\left(\frac{-\|x - x'\|^2}{\sigma^2}\right) \quad (2.104)$$

is applied, where  $\sigma$  is the kernel parameter. So the parameters to be chosen in Algorithm 2.1 and Algorithm 2.6 include  $C$ ,  $\varepsilon$ ,  $\sigma$ , and in our experiments, we choose these three parameters from the following sets:

$$C \in S_C = \{0.01, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 10000\}, \quad (2.105)$$





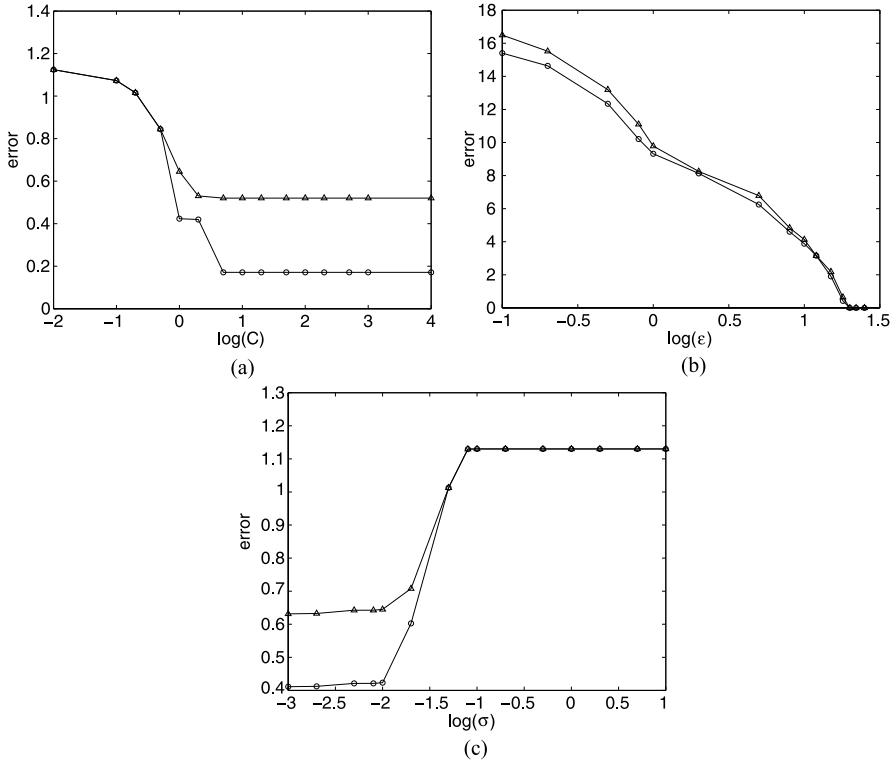
**Fig. 2.1** LOO errors and LOO bounds of Algorithm 2.1

$$\varepsilon \in S_\varepsilon = \{0.1, 0.2, 0.5, 0.8, 1, 2, 5, 8, 10, 12, 15, 18, 20, 22, 25\}, \quad (2.106)$$

$$\sigma \in S_\sigma = \{0.001, 0.002, 0.005, 0.008, 0.01, 0.02, 0.05, 0.08, 0.1, 0.2, 0.5, 1, 2, 5, 10\}. \quad (2.107)$$

However, we do not consider their all combinations. We will only perform three experiments for each algorithm. In the first experiment, we fix  $\varepsilon = 10$ ,  $\sigma = 0.01$ , and choose  $C$  from  $S_C$ . Applying these parameters in Algorithm 2.1 and Algorithm 2.6, and using Definition 2.3, the two LOO errors are computed. On the other hand, according to Theorem 2.5 and Theorem 2.10, the two corresponding LOO error bounds are obtained. Both the LOO errors and the LOO bounds are showed in Fig. 2.1(a) and Fig. 2.2(a), where “o” denotes LOO error and “ $\Delta$ ” denotes the corresponding LOO bound.

Similarly, in the second experiment, let  $C = 10$ ,  $\sigma = 0.01$ , and choose  $\varepsilon$  from  $S_\varepsilon$ , the compared result is showed in Fig. 2.1(b) and Fig. 2.2(b). At last, in the third experiment let  $C = 10$ ,  $\varepsilon = 10$ , and choose  $\sigma$  from  $S_\sigma$ , the compared result is showed in Fig. 2.1(c) and Fig. 2.2(c). Note that in order to be visible clearly, the



**Fig. 2.2** LOO errors and LOO bounds of Algorithm 2.6

values of LOO errors and LOO bounds in the figures are all be divided by 10, and the values of  $[C, \epsilon, \sigma]$  are all changed to  $[\log_{10}(C), \log_{10}(\epsilon), \log_{10}(\sigma)]$ .

From Figs. 2.1 and 2.2, we see that our two LOO bounds are really upper bounds of the corresponding true LOO errors, and more important, they almost have the same trend with the corresponding true LOO errors when the parameters are changing. So in order to choose the optimal parameters in Algorithm 2.1 and Algorithm 2.6, we only need to minimize the proposed LOO bound instead of LOO error itself. Obviously it must cost much less time.

### 2.3 LOO Bounds for Support Vector Ordinal Regression Machine

This section will focus on LOO bounds for support vector ordinal regression machine (SVORM) proposed in [177] which solves ordinal regression problem. Problem of ordinal regression arises in many fields, e.g., in information retrieval [109], in econometric models [194], and in classical statistics [7]. It is complementary to

classification problem and metric regression problem due to its discrete and ordered outcome space. Several methods corresponding with SVM have been proposed to solve this problem, such as in [110] which is based on a mapping from objects to scalar utility values and enforces large margin rank boundaries. SVORM was constructed by applying the large margin principle used in SVM to the ordinal regression problem, and outperforms existing ordinal regression algorithms [177].

Selecting appropriate parameters in SVORM is also an important problem, techniques such as cross-validation and LOO error can also be applied except for their inefficient computation. Therefore, we will present two LOO error bounds for SVORM. The first one corresponds to an upper bound for the C-SVC in [207] by Vapnik and Chapelle, while the second one to an upper bound in [119] by Joachims. Obviously, the derivation of our two bounds are more complicated because multi-class classification, instead of 2-class classification, is solved by SVORM.

### 2.3.1 Support Vector Ordinal Regression Machine

Ordinal regression problem can be described as follows: Suppose a training set is given by

$$T = \{(x_i^j, y_i^j)\}_{i=1, \dots, l^j}^{j=1, \dots, k} \in (R^n \times \mathcal{Y})^l, \quad (2.108)$$

where  $x_i^j \in R^n$  is the input,  $y_i^j = j \in \mathcal{Y} = \{1, \dots, k\}$  is the output or the class label,  $i = 1, \dots, l^j$  is the index with each class and  $l = \sum_{j=1}^k l^j$  is the number of sample points. We need to find  $k - 1$  parallel hyperplanes represented by vector  $w$  and an orderly real sequence  $b_1 \leq \dots \leq b_{k-1}$  defining the hyperplanes  $(w, b_1), \dots, (w, b_{k-1})$  such that the data are separated by dividing the space into equally ranked regions by the decision function

$$f(x) = \min_{r \in \{1, \dots, k\}} \{r : (w \cdot x) - b_r < 0\}, \quad (2.109)$$

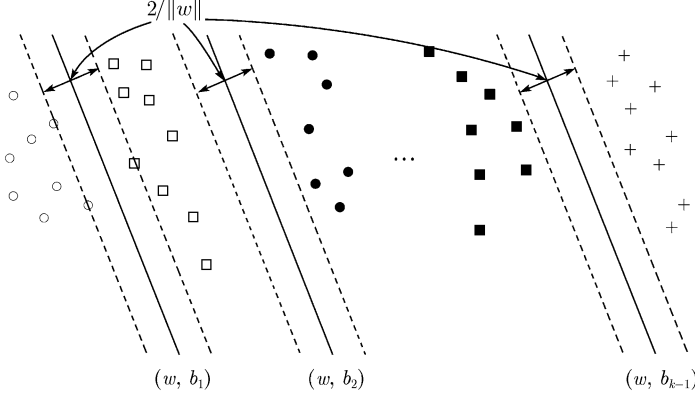
where  $b_k = +\infty$ . In other words, all input points  $x$  satisfying  $b_{r-1} < (w \cdot x) < b_r$  are assigned the rank  $r$ , where  $b_0 = -\infty$ .

Now we briefly introduce the fixed margin version of SVORM as a direct generalization of C-SVM [206]. Figure 2.3 gives out the geometric interpretation of this strategy.

For the training set (2.108), the input is mapped into a Hilbert space by a function  $x = \Phi(x) : x \in R^n \rightarrow x \in \mathcal{H}$ , where  $\mathcal{H}$  is the Hilbert space. Then the primal problem of SVORM is the following optimization problem:

$$\min_{w, b, \xi^{(*)}} \frac{1}{2} \|w\|^2 + C \sum_{j=1}^k \sum_{i=1}^{l^j} (\xi_i^j + \xi_i^{*j}), \quad (2.110)$$

$$\text{s.t. } (w \cdot \Phi(x_i^j)) - b_j \leq -1 + \xi_i^j, \quad j = 1, \dots, k, \quad i = 1, \dots, l^j, \quad (2.111)$$



**Fig. 2.3** Fixed-margin policy for ordinal problem: the margin to be maximized is the one defined by the closest (neighboring) pair of classes. Formally, let  $w, b_j$  be the hyperplane separating the two pairs of classes which are the closest among all the neighboring pairs of classes. Let  $w, b_j$  be scaled such the distance of the boundary points from the hyperplane is 1, i.e., the margin between the classes  $j, j+1$  is  $1/\|w\|$ . Thus, the fixed margin policy for ranking learning is to find the direction  $w$  and the scalars  $b_1, \dots, b_{k-1}$  such that  $\|w\|$  is minimized (i.e., the margin between classes  $j, j+1$  is maximized) subject to the separability constraints (modulo margin errors in the non-separable case)

$$(w \cdot \Phi(x_i^j)) - b_{j-1} \geq 1 - \xi_i^{*j}, \quad j = 1, \dots, k, \quad i = 1, \dots, l^j, \quad (2.112)$$

$$\xi_i^j \geq 0, \quad \xi_i^{*j} \geq 0, \quad j = 1, \dots, k, \quad i = 1, \dots, l^j, \quad (2.113)$$

where  $w \in \mathcal{H}$ ,  $b = (b_1, \dots, b_{k-1})^T$ ,  $b_0 = -\infty$ ,  $b_k = +\infty$ ,  $\xi^{(*)} = (\xi_1^1, \dots, \xi_{l^1}^1, \dots, \xi_1^k, \dots, \xi_{l^k}^k, \xi_1^{*1}, \dots, \xi_{l^1}^{*1}, \dots, \xi_1^{*k}, \dots, \xi_{l^k}^{*k})$  and the penalty parameter  $C > 0$ . The dual of the problem (2.110)–(2.113) can be expressed as [42]:

$$\min_{\alpha^{(*)}} \frac{1}{2} \sum_{j,i} \sum_{j',i'} (\alpha_i^{*j} - \alpha_i^j) (\alpha_{i'}^{*j'} - \alpha_{i'}^{j'}) K(x_i^j, x_{i'}^{j'}) - \sum_{j,i} (\alpha_i^j + \alpha_i^{*j}), \quad (2.114)$$

$$\text{s.t.} \quad \sum_{i=1}^{l^j} \alpha_i^j = \sum_{i=1}^{l^{j+1}} \alpha_i^{*j+1}, \quad j = 1, \dots, k-1, \quad (2.115)$$

$$0 \leq \alpha_i^j, \alpha_i^{*j} \leq C, \quad j = 1, \dots, k, \quad i = 1, \dots, l^j, \quad (2.116)$$

where  $\alpha^{(*)} = (\alpha_1^1, \dots, \alpha_{l^1}^1, \dots, \alpha_1^k, \dots, \alpha_{l^k}^k, \alpha_1^{*1}, \dots, \alpha_{l^1}^{*1}, \dots, \alpha_1^{*k}, \dots, \alpha_{l^k}^{*k})^T$ ,  $\alpha_i^{*1} = 0, i = 1, \dots, l^1, \alpha_i^k = 0, i = 1, \dots, l^k$ .

For optimal solutions  $w$  and  $\alpha^{(*)}$ , the primal–dual relationship shows

$$w = \sum_{j=1}^k \sum_{i=1}^{l^j} (\alpha_i^{*j} - \alpha_i^j) \Phi(x_i^j). \quad (2.117)$$

So in the decision function (2.109) the real value function  $g(x)$  is given by

$$g(x) = (w \cdot x) = \sum_{j=1}^k \sum_{i=1}^{l^j} (\alpha_i^{*j} - \alpha_i^j) K(x_i^j, x), \quad (2.118)$$

where  $K(x_i^j, x) = (\Phi(x_i^j) \cdot \Phi(x))$  is the kernel function. The scalars  $b_1, \dots, b_{k-1}$  can be obtained from the KKT conditions of primal problem (2.110)–(2.113).

This leads to the following algorithm:

**Algorithm 2.11** (SVORM)

- (1) Given a training set (2.108);
- (2) Select a scalar  $C > 0$  and a kernel function  $K(x, x')$ . Solve the dual problem (2.114)–(2.116), and get its optimal solution  $\alpha^{(*)} = (\alpha_1^1, \dots, \alpha_{l^1}^1, \dots, \alpha_1^k, \dots, \alpha_{l^k}^k, \alpha_1^{*1}, \dots, \alpha_{l^1}^{*1}, \dots, \alpha_1^{*k}, \dots, \alpha_{l^k}^{*k})^T$ ;
- (3) Compute

$$g(x) = (w \cdot x) = \sum_{j=1}^k \sum_{i=1}^{l^j} (\alpha_i^{*j} - \alpha_i^j) K(x_i^j, x); \quad (2.119)$$

- (4) For  $j = 1, \dots, k-1$ , execute the following steps:
  - (4.1) Choose a component  $\alpha_i^{*j} \in (0, C)$  in  $\alpha^{(*)}$ . If we get such subscript  $i$ , set

$$b_j = 1 + \sum_{j'=1}^k \sum_{i'=1}^{l^{j'}} (\alpha_{i'}^{*j'} - \alpha_{i'}^{j'}) K(x_{i'}^{j'}, x_i^j);$$

otherwise go to step (4.2);

- (4.2) Choose a component  $\alpha_i^{*j+1} \in (0, C)$  in  $\alpha^{(*)}$ . If we get such subscript  $i$ , set

$$b_j = \sum_{j'=1}^k \sum_{i'=1}^{l^{j'}} (\bar{\alpha}_{i'}^{*j'} - \bar{\alpha}_{i'}^{j'}) K(x_{i'}^{j'}, x_i^{j+1}) - 1;$$

otherwise go to step (4.3);

- (4.3) Set

$$b_j = \frac{1}{2}(b_j^{\text{dn}} + b_j^{\text{up}}),$$

where

$$b_j^{\text{dn}} = \max \left\{ \max_{i \in I_1^j} (g(x_i^j) + 1), \max_{i \in I_4^j} (g(x_i^{j+1}) - 1) \right\},$$

$$b_j^{\text{up}} = \min \left\{ \min_{i \in I_3^j} (g(x_i^j) + 1), \min_{i \in I_2^j} (g(x_i^{j+1}) - 1) \right\},$$

and

$$\begin{aligned} I_1^j &= \{i \in \{1, \dots, l^j\} \mid \alpha_i^j = 0\}, & I_2^j &= \{i \in \{1, \dots, l^{j+1}\} \mid \alpha_i^{*j+1} = 0\}, \\ I_3^j &= \{i \in \{1, \dots, l^j\} \mid \alpha_i^j = C\}, \\ I_4^j &= \{i \in \{1, \dots, l^{j+1}\} \mid \alpha_i^{*j+1} = C\}; \end{aligned}$$

- (5) If there exists  $j \in \{1, \dots, k\}$  such that  $b_j \leq b_{j-1}$ , stop or go to step (2);  
 (6) Define  $b_k = +\infty$ , construct the decision function

$$f(x) = \min_{r \in \{1, \dots, k\}} \{r : g(x) - b_r < 0\}. \quad (2.120)$$

In addition, in order to derive the LOO error bounds for SVORM, we firstly give the definitions of its support vector and its LOO error in the LOO procedure.

**Definition 2.12** (Support vector) Suppose that  $\alpha^{(*)}$  is the optimal solution of the dual problem (2.114)–(2.116) for the training set  $T$  (2.108). Then

- (i) The input  $x_i^j$  is called *non-margin support vector about  $\alpha = (\alpha_1^1, \dots, \alpha_{l^1}^1, \dots, \alpha_1^k, \dots, \alpha_{l^k}^k)^T$* , if the corresponding component  $\alpha_i^j$  of  $\alpha^{(*)}$  is equal to  $C$ . For  $j = 1, \dots, k$  define the index set

$$N(\alpha, j) = \{(j, i) \mid i = 1, \dots, l^j, \alpha_i^j = C\}. \quad (2.121)$$

The input  $x_i^j$  is called *non-margin support vector about  $\alpha^* = (\alpha_1^{*1}, \dots, \alpha_{l^1}^{*1}, \dots, \alpha_1^{*k}, \dots, \alpha_{l^k}^{*k})^T$* , if the corresponding component  $\alpha_i^{*j}$  of  $\alpha^{(*)}$  is equal to  $C$ . For  $j = 1, \dots, k$  define the index set

$$N(\alpha^*, j) = \{(j, i) \mid i = 1, \dots, l^j, \alpha_i^{*j} = C\}. \quad (2.122)$$

The input  $x_i^j$  is called *non-margin support vector about  $\alpha^{(*)}$* , if  $x_i^j$  is either non-margin support vector about  $\alpha$  or non-margin support vector about  $\alpha^*$ . For  $j = 1, \dots, k$  define the index set

$$N(\alpha^{(*)}, j) = N(\alpha, j) \cup N(\alpha^*, j). \quad (2.123)$$

- (ii) The input  $x_i^j$  is called *margin support vector about  $\alpha = (\alpha_1^1, \dots, \alpha_{l^1}^1, \dots, \alpha_1^k, \dots, \alpha_{l^k}^k)^T$* , if the corresponding component  $\alpha_i^j$  of  $\alpha^{(*)}$  is in the interval  $(0, C)$  and the component  $\alpha_i^{*j}$  of  $\alpha^{(*)}$  is not equal to  $C$ . For  $j = 1, \dots, l$  define the index set

$$M(\alpha, j) = \{(j, i) \mid i = 1, \dots, l^j, \alpha_i^j \in (0, C)\} \setminus N(\alpha^*, j). \quad (2.124)$$

The input  $x_i^j$  is called *margin support vector about  $\alpha^* = (\alpha_1^{*1}, \dots, \alpha_{l^1}^{*1}, \dots, \alpha_1^{*k}, \dots, \alpha_{l^k}^{*k})^T$* , if the corresponding component  $\alpha_i^{*j}$  of  $\alpha^{(*)}$  is in the interval

$(0, C)$  and the component  $\alpha_i^j$  of  $\alpha^{(*)}$  is not equal to  $C$ . For  $j = 1, \dots, l$  define the index set

$$M(\alpha^*, j) = \{(j, i) \mid i = 1, \dots, l^j, \alpha_i^{*j} \in (0, C)\} \setminus N(\alpha, j). \quad (2.125)$$

The input  $x_i^j$  is called *margin support vector about  $\alpha^{(*)}$* , if  $x_i^j$  is either margin support vector about  $\alpha$  or margin support vector about  $\alpha^*$ . For  $j = 1, \dots, k$  define the index set

$$M(\alpha^{(*)}, j) = M(\alpha, j) \cup M(\alpha^*, j). \quad (2.126)$$

(iii) The input  $x_i^j$  is called *support vector about  $\alpha = (\alpha_1^1, \dots, \alpha_{l^1}^1, \dots, \alpha_1^k, \dots, \alpha_{l^k}^k)^T$* , if  $x_i^j$  is either non-margin support vector about  $\alpha$  or margin support vector about  $\alpha$ . For  $j = 1, \dots, k$  define the index set

$$V(\alpha, j) = M(\alpha, j) \cup N(\alpha, j). \quad (2.127)$$

The input  $x_i^j$  is called *support vector about  $\alpha^* = (\alpha_1^{*1}, \dots, \alpha_{l^1}^{*1}, \dots, \alpha_1^{*k}, \dots, \alpha_{l^k}^{*k})^T$* , if  $x_i^j$  is either non-margin support vector about  $\alpha^*$  or margin support vector about  $\alpha^*$ . For  $j = 1, \dots, k$  define the index set

$$V(\alpha^*, j) = M(\alpha^*, j) \cup N(\alpha^*, j). \quad (2.128)$$

The input  $x_i^j$  is called *support vector about  $\alpha^{(*)}$* , if  $x_i^j$  is either non-margin support vector about  $\alpha^{(*)}$  or margin support vector about  $\alpha^{(*)}$ . For  $j = 1, \dots, k$  define the index set

$$V(\alpha^{(*)}, j) = V(\alpha, j) \cup V(\alpha^*, j). \quad (2.129)$$

**Definition 2.13** (LOO error) Consider the training set  $T_p^q = T \setminus \{(x_p^q, y_p^q)\}$ ,  $q = 1, \dots, k$ ,  $p = 1, \dots, l^q$ , where  $T$  is given by (2.108). Suppose that  $f_{T_p^q}^q(x)$  is the decision function obtained by executing Algorithm 2.11 for  $T_p^q$ , then the leave-one-out error, or LOO error for short, is defined as

$$R_{\text{LOO}}(T) = \frac{1}{l} \sum_{q=1}^k \sum_{p=1}^{l^q} c(x_p^q, y_p^q, f_{T_p^q}^q(x_p^q)), \quad (2.130)$$

where  $c$  is the 0–1 loss function

$$c(x, y, f(x)) = \hat{c}(y - f(x)),$$

with

$$\hat{c}(\xi) = \begin{cases} 0, & \text{if } \xi = 0, \\ 1, & \text{otherwise.} \end{cases}$$

From Definition 2.13 we can see that the computation of LOO error for SVORM is time-consuming and inefficient. So researching for LOO error bounds for SVORM will be necessary.

### 2.3.2 The First LOO Bound

In this section, we study the derivation of our first LOO bound for Algorithm 2.11 by the concept of a span.

#### Definition and Existence of Span

We now define an S-span of a margin support vector about  $\alpha$  and  $\alpha^*$  respectively.

**Definition 2.14** (S-span about  $\alpha$ ) For any margin support vector  $x_p^q$  about  $\alpha$ , define its S-span by

$$S^2(q, p) := \min\{\|x_p^q - \tilde{x}_p^q\|^2 | \tilde{x}_p^q \in \Lambda_p^q\}, \quad (2.131)$$

where  $\Lambda_p^q$  is

$$\Lambda_p^q := \left\{ \sum_{i \in M_p^q(\alpha, q)} \lambda_i^q x_i^q + \sum_{i \in M_p^q(\alpha^*, q+1)} \lambda_i^{q+1} x_i^{q+1} \right\}, \quad (2.132)$$

with the following conditions:

$$0 \leq \alpha_i^q + \lambda_i^q \alpha_p^q \leq C, \quad 0 \leq \alpha_i^{*q} + \lambda_i^q \alpha_p^{*q} \leq C, \quad (2.133)$$

$$0 \leq \alpha_i^{q+1} - \lambda_i^{q+1} \alpha_p^{*q} \leq C, \quad 0 \leq \alpha_i^{*q+1} - \lambda_i^{q+1} \alpha_p^q \leq C, \quad (2.134)$$

$$\sum_{i \in M_p^q(\alpha, q)} \lambda_i^q + \sum_{i \in M_p^q(\alpha^*, q+1)} \lambda_i^{q+1} = 1, \quad \lambda_p^q = -1, \quad (2.135)$$

and

$$M_p^q(\alpha, j) = M(\alpha, j) \setminus \{(q, p)\}, \quad M_p^q(\alpha^*, j) = M(\alpha^*, j) \setminus \{(q, p)\}. \quad (2.136)$$

**Definition 2.15** (S-span about  $\alpha^*$ ) For any margin support vector  $x_p^q$  about  $\alpha^*$ , define its S-span by

$$S^{*2}(q, p) := \min\{\|x_p^q - \hat{x}_p^q\|^2 | \hat{x}_p^q \in \Lambda_p^{*q}\}, \quad (2.137)$$

where  $\Lambda_p^{*q}$  is

$$\Lambda_p^{*q} := \left\{ \sum_{i \in M_p^q(\alpha, q-1)} \lambda_i^{q-1} x_i^{q-1} + \sum_{i \in M_p^q(\alpha^*, q)} \lambda_i^q x_i^q \right\}, \quad (2.138)$$



with the following conditions:

$$0 \leq \alpha_i^{q-1} - \lambda_i^{q-1} \alpha_p^{*q} \leq C, \quad 0 \leq \alpha_i^{*q-1} - \lambda_i^{q-1} \alpha_p^q \leq C, \quad (2.139)$$

$$0 \leq \alpha_i^q + \lambda_i^q \alpha_p^q \leq C, \quad 0 \leq \alpha_i^{*q} + \lambda_i^q \alpha_p^{*q} \leq C, \quad (2.140)$$

$$\sum_{i \in M_p^q(\alpha, q-1)} \lambda_i^{q-1} + \sum_{i \in M_p^q(\alpha^*, q)} \lambda_i^q = 1, \quad \lambda_p^q = -1, \quad (2.141)$$

and

$$M_p^q(\alpha, j) = M(\alpha, j) \setminus \{(q, p)\}, \quad M_p^q(\alpha^*, j) = M(\alpha^*, j) \setminus \{(q, p)\}. \quad (2.142)$$

For the S-span  $S^2(q, p)$  and  $S^{*2}(q, p)$  defined above, it is necessary to show that the set  $\Lambda_p^q$  and  $\Lambda_p^{*q}$  are non-empty. To this end, we make use of the following lemma.

**Lemma 2.16** *The both sets  $\Lambda_p^q$  and  $\Lambda_p^{*q}$  defined by (2.132) and (2.138) are non-empty.*

The proof is omitted here, which can be referred to [230] and [231].

According to the above Definition 2.14 and Definition 2.15, we have the following two lemmas.

**Lemma 2.17** *Suppose that  $\alpha^{(*)}$  is an optimal solution of the dual problem (2.114)–(2.116) for the training set  $T$  (2.108) and  $x_p^q$  is a margin support vector about  $\alpha$ . Then we can construct a feasible solution  $\tilde{\alpha}^{(*)}$  of the dual problem (2.114)–(2.116) for the training set  $T_p^q = T \setminus \{(x_p^q, y_p^q)\}$  by*

$$\tilde{\alpha}_i^q = \alpha_i^q + \lambda_i^q \alpha_p^q, \quad \tilde{\alpha}_i^{*q} = \alpha_i^{*q} + \lambda_i^q \alpha_p^{*q}, \quad i \in M_p^q(\alpha, q), \quad (2.143)$$

$$\tilde{\alpha}_i^{q+1} = \alpha_i^{q+1} - \lambda_i^{q+1} \alpha_p^{*q}, \quad \tilde{\alpha}_i^{*q+1} = \alpha_i^{*q+1} - \lambda_i^{q+1} \alpha_p^q, \quad (2.144)$$

$$\tilde{\alpha}_i^q = \alpha_i^q, \quad \tilde{\alpha}_i^{*q} = \alpha_i^{*q}, \quad i \notin M_p^q(\alpha, q), \quad (2.145)$$

$$\tilde{\alpha}_i^{q+1} = \alpha_i^{q+1}, \quad \tilde{\alpha}_i^{*q+1} = \alpha_i^{*q+1}, \quad i \notin M_p^q(\alpha^*, q+1), \quad (2.146)$$

$$\tilde{\alpha}_i^j = \alpha_i^j, \quad \tilde{\alpha}_i^{*j} = \alpha_i^{*j}, \quad j = 1, \dots, q-1, q+2, \dots, k, \quad i = 1, \dots, l^j, \quad (2.147)$$

and

$$\sum_{i \in M_p^q(\alpha, q)} \lambda_i^q x_i^q + \sum_{i \in M_p^q(\alpha^*, q+1)} \lambda_i^{q+1} x_i^{q+1} \in \Lambda_p^q.$$

The proof is omitted here, which can be referred to [230] and [231].

**Lemma 2.18** Suppose that  $\alpha^{(*)}$  is an optimal solution of the dual problem (2.114)–(2.116) for the training set  $T$  (2.108) and  $x_p^q$  is a margin support vector about  $\alpha^*$ . Then we can construct a feasible solution  $\tilde{\alpha}^{(*)}$  of the dual problem (2.114)–(2.116) for the training set  $T_p^q = T \setminus \{(x_p^q, y_p^q)\}$  by

$$\hat{\alpha}_i^{q-1} = \alpha_i^{q-1} - \lambda_i^{q-1} \alpha_p^{*q}, \quad \hat{\alpha}_i^{*q-1} = \alpha_i^{*q-1} - \lambda_i^{q-1} \alpha_p^q, \\ i \in M_p^q(\alpha, q-1), \quad (2.148)$$

$$\hat{\alpha}_i^q = \alpha_i^q + \lambda_i^q \alpha_p^q, \quad \hat{\alpha}_i^{*q} = \alpha_i^{*q} + \lambda_i^q \alpha_p^{*q}, \quad i \in M_p^q(\alpha^*, q), \quad (2.149)$$

$$\hat{\alpha}_i^{q-1} = \alpha_i^{q-1}, \quad \hat{\alpha}_i^{*q-1} = \alpha_i^{*q-1}, \quad i \notin M_p^q(\alpha, q-1), \quad (2.150)$$

$$\hat{\alpha}_i^q = \alpha_i^q, \quad \hat{\alpha}_i^{*q} = \alpha_i^{*q}, \quad i \notin M_p^q(\alpha^*, q), \quad (2.151)$$

$$\hat{\alpha}_i^j = \alpha_i^j, \quad \hat{\alpha}_i^{*j} = \alpha_i^{*j}, \quad j = 1, \dots, q-2, q+1, \dots, k, \quad i = 1, \dots, l^j, \quad (2.152)$$

and

$$\sum_{i \in M_p^q(\alpha, q-1)} \lambda_i^{q-1} x_i^{q-1} + \sum_{i \in M_p^q(\alpha^*, q)} \lambda_i^q x_i^q \in \Lambda_p^{*q}.$$

The proof is omitted here, which can be referred to [230] and [231].

## The Bound

Now we are in a position to introduce our first LOO error bound:

**Lemma 2.19** Suppose that  $\alpha^{(*)}$  is the optimal solution the dual problem (2.114)–(2.116) for the training set  $T$  (2.108) and  $f_{T_p^q}$  is the decision function obtained by Algorithm 2.11 for the training set  $T_p^q = T \setminus \{(x_p^q, y_p^q)\}$ . For a margin support vector  $x_p^q$  about  $\alpha^{(*)}$ , we have

- (1) If  $x_p^q$  is a margin support vector about  $\alpha$  and is recognized incorrectly by the decision function  $f_p^q$ , then the following inequality holds

$$(\alpha_p^{*q} - \alpha_p^q)^2 S^2(p, q) \geq \min \left( C, \frac{1}{D_{q,q+1}^2} \right), \quad (2.153)$$

where  $D_{q,q+1}$  is diameter of the smallest sphere containing the  $q$ th class training points and the  $(q+1)$ th class training points in the training set  $T$  (2.108), and we have the following expression

$$D_{q,q+1} = \min_{D_{q,q+1}, c} \left\{ D_{q,q+1} \|x_i^j - c\|^2 \leq \frac{D_{q,q+1}^2}{4}, \quad j = q, q+1, \quad i = 1, \dots, l^j \right\}; \quad (2.154)$$

- (2) If  $x_p^q$  is a margin support vector about  $\alpha^*$  and is recognized incorrectly by the decision function  $f_p^q$ , then the following inequality holds

$$(\alpha_p^{*q} - \alpha_p^q)^2 S^{*2}(p, q) \geq \min\left(C, \frac{1}{D_{q-1,q}^2}\right), \quad (2.155)$$

where  $D_{q-1,q}$  is diameter of the smallest sphere containing the  $(q-1)$ th class training points and the  $q$ th class training points in the training set  $T$  (2.108), and we have the following expression

$$D_{q-1,q} = \min_{D_{q-1,q}, c} \left\{ D_{q-1,q} \|x_i^j - c\|^2 \leq \frac{D_{q-1,q}^2}{4}, \right. \\ \left. j = q-1, q, i = 1, \dots, l^j \right\}. \quad (2.156)$$

The proof is omitted here, which can be referred to [230] and [231].

The above lemma leads to the following theorem:

**Theorem 2.20** For Algorithm 2.11, the bound of LOO error is estimated by

$$R_{\text{LOO}}(T) \leq \frac{1}{l} \sum_{q=1}^k \sum_{p=1}^{l^q} \left[ \left| \left\{ (q, p) : (\alpha_p^{*q} - \alpha_p^q)^2 S^2(q, p) \geq \min\left(C, \frac{1}{D_{q,q+1}^2}\right) \text{ or } \right. \right. \right. \\ \left. \left. \left. (\alpha_p^{*q} - \alpha_p^q)^2 S^{*2}(q, p) \geq \min\left(C, \frac{1}{D_{q-1,q}^2}\right) \right\} \right| + |N(\alpha^{(*)}, q)| \right], \quad (2.157)$$

where  $D_{q,q+1}$  and  $D_{q-1,q}$  are given by (2.154) and (2.156) respectively,  $N(\alpha^{(*)}, q)$  is defined by (2.123) and  $|\cdot|$  is the number of elements in the set.

*Proof* Considering the Definition 2.13 of LOO error. Denote the number of error made by the LOO procedure as  $\mathcal{L}(T)$

$$\mathcal{L}(T) = \sum_{q=1}^k \sum_{p=1}^{l^q} c(x_p^q, y_p^q, f_{T_p^q}(x_p^q)) = \sum_{q=1}^k \sum_{p=1}^{l^q} \mathbb{I}_{(w_p^q \cdot x_p^q) - b_q > 0 \text{ or } (w_p^q \cdot x_p^q) - b_{q-1} < 0}, \quad (2.158)$$

where

$$\mathbb{I}_P = \begin{cases} 1, & P \text{ is true;} \\ 0, & P \text{ is false.} \end{cases}$$

In order to estimate  $\mathcal{L}(T)$ , define two index sets

$$I_{\text{LOO}}^q \triangleq \{(q, p) : (w_p^q \cdot x_p^q) - b_q > 0 \text{ or } (w_p^q \cdot x_p^q) - b_{q-1} < 0\} \cup N(\alpha^{(*)}, q), \\ \hat{I}_{\text{LOO}}^q \triangleq \{(q, p) : (w_p^q \cdot x_p^q) - b_q > -1 \text{ or } (w_p^q \cdot x_p^q) - b_{q-1} < 1\} \cup N(\alpha^{(*)}, q).$$

It is easy to see that

$$\mathcal{L}(T) = |I_{\text{LOO}}^q| \leq |\hat{I}_{\text{LOO}}^q|. \quad (2.159)$$

By the Lemma 2.19, we have

$$\begin{aligned} |\hat{I}_{\text{LOO}}^q| = & \left| \left\{ (q, p) : (\alpha_p^{*q} - \alpha_p^q)^2 S^2(q, p) \geq \min \left( C, \frac{1}{D_{q,q+1}^2} \right) \right. \right. \\ & \left. \left. \text{or } (\alpha_p^{*q} - \alpha_p^q)^2 S^2(q, p) \geq \min \left( C, \frac{1}{D_{q-1,q}^2} \right) \right\} \right| + |N(\alpha^{(*)}, q)|. \end{aligned} \quad (2.160)$$

So the LOO error bound (2.157) is obtained from (2.159) and (2.160).  $\square$

### 2.3.3 The Second LOO Bound

In this section, we study the derivation our second LOO error bound.

Remind that the dual problem for the training set  $T$  (2.108) is presented in (2.114)–(2.116). For the training set  $T_p^q = T / \{(x_p^q, y_p^q)\}$ , the dual problem is

$$\begin{aligned} \max_{\alpha_p^{(*)q}} & W_p^q(\alpha^{(*)}) \\ = & \sum_{(j,i) \in I \setminus \{(q,p)\}} (\alpha_i^j + \alpha_i^{*j}) \\ & - \frac{1}{2} \sum_{(j,i) \in I \setminus \{(q,p)\}} \sum_{(j',i') \in I \setminus \{(q,p)\}} (\alpha_i^{*j} - \alpha_i^j)(\alpha_{i'}^{*j'} - \alpha_{i'}^{j'}) K(x_i^j, x_{i'}^{j'}) \\ \text{s.t. } & \sum_{i=1}^{l^j} \alpha_i^j = \sum_{i=1}^{l^{j+1}} \alpha_i^{*j+1}, \quad j = 1, \dots, k-1, \quad (j, i) \neq (q, p), \end{aligned} \quad (2.161)$$

$$0 \leq \alpha_i^j, \alpha_i^{*j} \leq C, \quad (j, i) \in I \setminus \{(q, p)\}, \quad (2.162)$$

where  $\alpha_p^{(*)q} = (\alpha_p^{qT}, \alpha_p^{*qT})^T$ ,  $\alpha_p^q = (\alpha_1^1, \dots, \alpha_{l^1}^1, \dots, \alpha_1^q, \dots, \alpha_{p-1}^q, \alpha_{p+1}^q, \dots, \alpha_{l^q}^q, \dots, \alpha_1^k, \dots, \alpha_{l^k}^k)^T$ ,  $\alpha_p^{*q} = (\alpha_1^{*1}, \dots, \alpha_{l^1}^{*1}, \dots, \alpha_1^{*q}, \dots, \alpha_{p-1}^{*q}, \alpha_{p+1}^{*q}, \dots, \alpha_{l^q}^{*q}, \dots, \alpha_1^{*k}, \dots, \alpha_{l^k}^{*k})^T$ ;  $\alpha_i^{*1} = 0, i = 1, 2, \dots, l^1$ ,  $\alpha_i^k = 0, i = 1, 2, \dots, l^k$ ,  $I = \{(j, i) \mid j = 1, \dots, k, i = 1, \dots, l^j\}$ .

In order to derive the second LOO error bound, we give the following lemma firstly.

**Lemma 2.21** Suppose that  $\alpha^{(*)}$  is the optimal solution of the dual problem (2.114)–(2.116) for the training set  $T$  (2.108), and  $f_{T_p^q}$  is the decision function obtained by

Algorithm 2.11 for the training set  $T_p^q = T \setminus \{(x_p^q, y_p^q)\}$ . For the components of optimal solution  $\alpha^{(*)}$ :  $\alpha_i^q, \alpha_i^{*q}, i = 1, \dots, l^q$ ,

- (1) If in  $\{\alpha_i^q \mid i = 1, \dots, l^q\}$ , there exists some  $\alpha_i^q \in (0, C)$  and  $x_p^q$  is recognized incorrectly by the decision function  $f_p^q$ , then the following inequality holds

$$\left[ \sum_{j,i} (\alpha_i^{*j} - \alpha_i^j) K(x_p^q, x_i^j) - b_q \right] - (\alpha_p^{*q} - \alpha_p^q) (K(x_p^q, x_p^q) + R^2) \geq 0, \quad (2.163)$$

where  $R^2 = \max\{K(x_i^j, x_i^j) \mid j = 1, \dots, k, i = 1, \dots, l^j\}$ .

- (2) If in  $\{\alpha_i^{*q} \mid i = 1, \dots, l^q\}$ , there exists  $\alpha_i^{*q} \in (0, C)$  and  $x_p^q$  is recognized incorrectly by the decision function  $f_p^q$ , then the following inequality holds

$$-\left[ \sum_{j,i} (\alpha_i^{*j} - \alpha_i^j) K(x_p^q, x_i^j) - b_{q-1} \right] + (\alpha_p^{*q} - \alpha_p^q) (K(x_p^q, x_p^q) + R^2) \geq 0, \quad (2.164)$$

where  $R^2 = \max\{K(x_i^j, x_i^j) \mid j = 1, \dots, k, i = 1, \dots, l^j\}$ .

The proof is omitted here, which can be referred to [230] and [231].

The above lemma leads to the following theorem for Algorithm 2.11:

**Theorem 2.22** For Algorithm 2.11 the bound of LOO error is estimated by

$$\begin{aligned} R_{\text{LOO}}(T) \leq \frac{1}{l} \left\{ \sum_{q \in I_1} \sum_{p=1}^{l^q} \left| \left[ \sum_{j,i} (\alpha_i^{*j} - \alpha_i^j) K(x_p^q, x_i^j) - b_q \right] \right. \right. \\ \left. \left. - (\alpha_p^{*q} - \alpha_p^q) (K(x_p^q, x_p^q) + R^2) \geq 0 \right. \right. \\ \text{or } - \left[ \sum_{j,i} (\alpha_i^{*j} - \alpha_i^j) K(x_p^q, x_i^j) - b_{q-1} \right] \\ \left. \left. + (\alpha_p^{*q} - \alpha_p^q) (K(x_p^q, x_p^q) + R^2) \geq 0 \right| + \sum_{q \in I_2} \sum_{p=1}^{l^q} |N(\alpha^{(*)}, q)| \right\}, \quad (2.165) \end{aligned}$$

where

$$I_1 = \{q \mid \text{in } (\alpha_1^q, \dots, \alpha_{l_q}^q), (\alpha_1^{*q+1}, \dots, \alpha_{l_{q+1}}^{*q+1}) \text{ there exists } \alpha_i^q \in (0, C) \\ \text{or } \alpha_i^{*q+1} \in (0, C)\},$$

$$I_2 = \{q \mid \text{in } (\alpha_1^q, \dots, \alpha_{l_q}^q), (\alpha_1^{*q+1}, \dots, \alpha_{l_{q+1}}^{*q+1}) \text{ there exist not } \alpha_i^q \in (0, C) \\ \text{and } \alpha_i^{*q+1} \in (0, C)\},$$

$$R^2 = \max\{K(x_i^j, x_i^j) \mid j = 1, \dots, k, i = 1, \dots, l^j\},$$

$N(\alpha^{(*)}, q)$  is defined by (2.123) and  $|\cdot|$  is the number of elements in the set.

*Proof* Assume that the point  $x_p^q$  belongs to the class  $q \in I_1 = \{q \mid \text{in } (\alpha_1^q, \dots, \alpha_{l_q}^q), (\alpha_1^{*q+1}, \dots, \alpha_{l_{q+1}}^{*q+1}) \text{ there exists } \alpha_i^q \in (0, C) \text{ or } \alpha_i^{*q+1} \in (0, C)\}$ . Then according to Lemma 2.21, when the LOO error procedure commits an error at the point  $x_p^q$ , the following one of two inequalities holds

$$\left[ \sum_{j,i} (\alpha_i^{*j} - \alpha_i^j) K(x_p^q, x_i^j) - b_q \right] - (\alpha_p^{*q} - \alpha_p^q) (K(x_p^q, x_p^q) + R^2) \geq 0, \quad (2.166)$$

$$- \left[ \sum_{j,i} (\alpha_i^{*j} - \alpha_i^j) K(x_p^q, x_i^j) - b_{q-1} \right] + (\alpha_p^{*q} - \alpha_p^q) (K(x_p^q, x_p^q) + R^2) \geq 0, \quad (2.167)$$

where  $R^2 = \max\{K(x_i^j, x_i^j) \mid j = 1, \dots, k, i = 1, \dots, l^j\}$ .

If being left out the point  $x_p^q$  belongs to the class  $q \in I_2 = \{q \mid \text{in } (\alpha_1^q, \dots, \alpha_{l_q}^q), (\alpha_1^{*q+1}, \dots, \alpha_{l_{q+1}}^{*q+1}) \text{ there does not exist } \alpha_i^q \in (0, C) \text{ and } \alpha_i^{*q+1} \in (0, C)\}$ , then the number of error made by the LOO error procedure is  $|N(\alpha^{(*)}, q)|$ , where  $N(\alpha^{(*)}, q)$  is defined by (2.123) and  $|\cdot|$  is the number of elements in the set.

So we get the LOO error bound (2.165) for Algorithm 2.11.  $\square$

### 2.3.4 Numerical Experiments

In this section, we describe the performance of the two LOO error bounds with four ordinal regression datasets [10]. The datasets are (1) “Employee Rejection\Acceptance” (ERA), (2) “Employee Selection” (ESL), (3) “Lecturers Evaluation” (LEV), (4) “Social Workers Decisions” (SWD). A summary of the characteristics of these datasets is presented in Table 2.1.

In our experiment, because the cost of computing LOO error is very high, we select randomly only 60 training points from each dataset and merge these 4 multi-class problems into 3-class problems. For each problem, we choose randomly 20 points from each class and get training set expressed as

$$T = \{(x_1^1, y_1^1), \dots, (x_{20}^1, y_{20}^1), (x_1^2, y_1^2), \dots, (x_{20}^2, y_{20}^2), (x_1^3, y_1^3), \dots, (x_{20}^3, y_{20}^3)\}, \quad (2.168)$$

where  $x_i^j$  is the input,  $y_i^j = j$  is the output. In this way, corresponding to ERA, ESL, LEV and SWD, we obtain the following training sets

$$T_{\text{ERA}}, \quad T_{\text{ESL}}, \quad T_{\text{LEV}} \quad \text{and} \quad T_{\text{SWD}}, \quad (2.169)$$

which are tested in our experiments.

Gaussian kernel function

$$K(x, x') = \exp\left(\frac{-\|x - x'\|^2}{\sigma^2}\right) \quad (2.170)$$

**Table 2.1** Characteristics of the selected datasets from the ordinal datasets

Dataset	Features	Classes	Patterns
ERA	4	9	1000
ESL	4	9	488
LEV	4	5	1000
SWD	10	4	1000

is selected in our experiment, while the parameters  $C$  and  $\sigma$  are selected respectively from the following two sequences

$$C = \text{logspace}(-2, 4, 12), \quad (2.171)$$

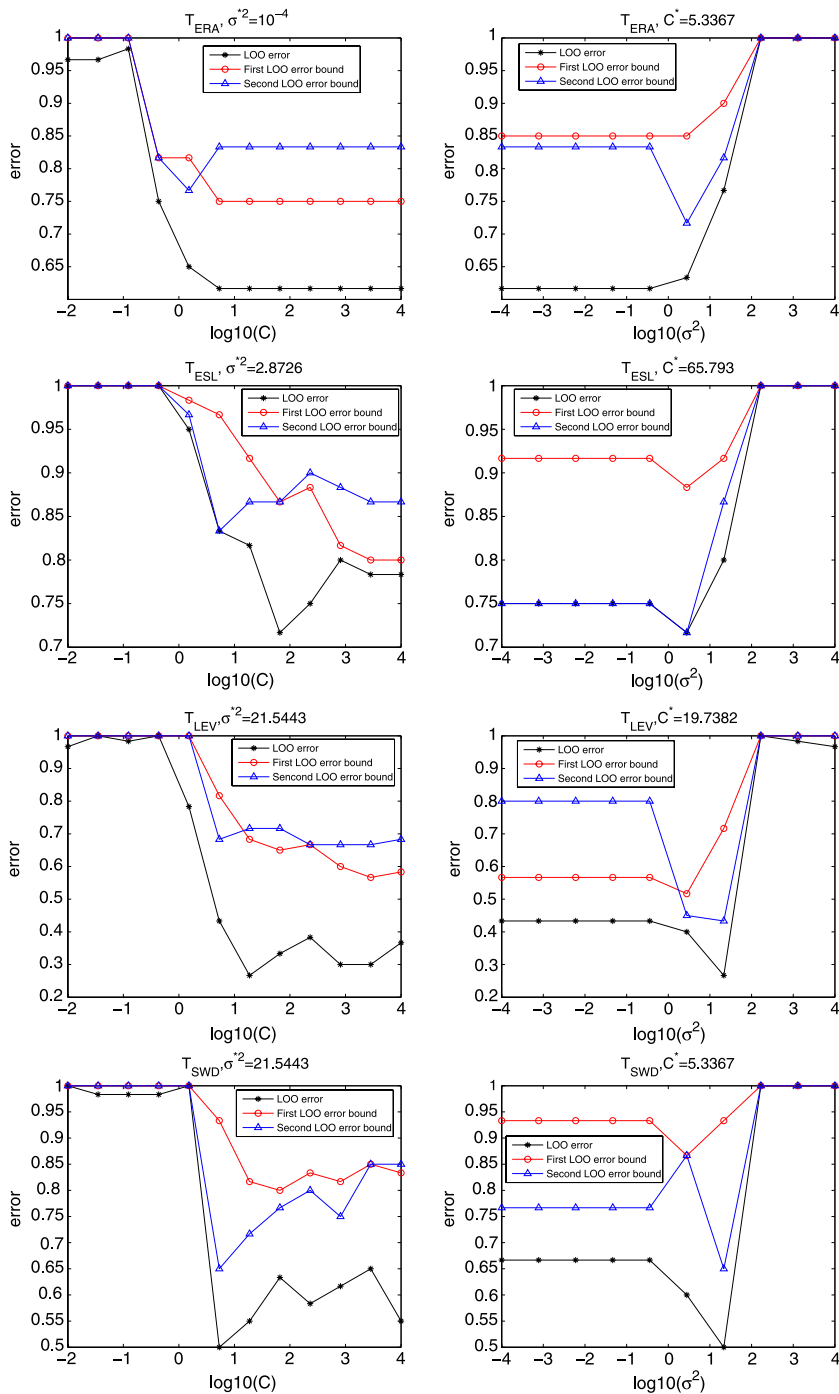
and

$$\sigma = \text{logspace}(-4, 4, 10), \quad (2.172)$$

where  $\text{logspace}$  is a logarithmically spaced vector in MATLAB. More precisely, firstly, we find the optimal parameters  $C^*, \sigma^{*2}$  in (2.171) and (2.172) to our four training sets  $T_{\text{ERA}}, T_{\text{ESL}}, T_{\text{LEV}}$  and  $T_{\text{SWD}}$  by minimizing the LOO error respectively. Secondly, either  $C$  or  $\sigma$  is fixed to be its optimal value obtained, while the other one takes the values in (2.171) or (2.172). Figure 2.4 shows the performance of two LOO error bounds and LOO error itself. For example, the top-left figure corresponds to the training set  $T_{\text{ERA}}$  with  $\sigma = \sigma^* = 10^{-4}$ , and  $C$  take the value in (2.171). “LOO error” stands for the actual LOO error, “First LOO error bound” is the bound given by Theorem 2.20 and “Second LOO error bound” by Theorem 2.22.

By and large, it can be observed from Fig. 2.4 that changing trend of both LOO error bounds is almost consistent with that of LOO error itself. Concretely, when the penalty parameter  $C$  is fixed and the kernel parameter  $\sigma^2$  is changed, our proposed both LOO error bounds are good performance. In other words, the lowest points of both LOO error bounds are close to those of LOO error, some difference of only one step length. So it is reasonable that the optimal parameters can be selected by minimizing these LOO error bounds instead of LOO error itself. Obviously, this strategy is highly efficient.

In this section, we derive two LOO error bounds for SVORM. The second LOO error bound is more effective than the first LOO error bound, because if we will compute the first LOO error bound, we must solve some quadratic programming problems, whereas the second LOO error bound doesn’t need. Experiments demonstrate that these bounds are valid and it is hopeful to get the optimal parameter by minimizing the proposed bounds instead of the LOO error itself. In the further, we improve our proposed both LOO error bounds by smart way handling non-margin support vectors, due to the assumption that all non-margin support vectors are leave-one-out errors. In addition, an interesting study is to apply the proposed bounds on feature selection.



**Fig. 2.4** Results of two LOO error bounds and LOO error



Optimization Based Data Mining: Theory and  
Applications

Shi, Y.; Tian, Y.; Kou, G.; Peng, Y.; Li, J.

2011, XVI, 316 p., Hardcover

ISBN: 978-0-85729-503-3