SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

November 2014, Vol. 57 112103:1–112103:10
doi: 10.1007/s11432-014-5090-z

Kernel selection with spectral perturbation stability of kernel matrix

LIU Yong & LIAO ShiZhong*

School of Computer Science and Technology, Tianjin University, Tianjin 300072*, China*

Received August 10, 2013; accepted November 20, 2013; published online March 18, 2014

Abstract Kernel selection is one of the key issues both in recent research and application of kernel methods. This is usually done by minimizing either an estimate of generalization error or some other related performance measure. Use of notions of stability to estimate the generalization error has attracted much attention in recent years. Unfortunately, the existing notions of stability, proposed to derive the theoretical generalization error bounds, are difficult to be used for kernel selection in practice. It is well known that the kernel matrix contains most of the information needed by kernel methods, and the eigenvalues play an important role in the kernel matrix. Therefore, we aim at introducing a new notion of stability, called the spectral perturbation stability, to study the kernel selection problem. This proposed stability quantifies the spectral perturbation of the kernel matrix with respect to the changes in the training set. We establish the connection between the spectral perturbation stability and the generalization error. By minimizing the derived generalization error bound, we propose a new kernel selection criterion that can guarantee good generalization properties. In our criterion, the perturbation of the eigenvalues of the kernel matrix is efficiently computed by solving the derivative of a newly defined generalized kernel matrix. Both theoretical analysis and experimental results demonstrate that our criterion is sound and effective.

Keywords kernel methods, kernel selection, stability, spectral perturbation stability, generalization error bound

Citation Liu Y, Liao S Z. Kernel selection with spectral perturbation stability of kernel matrix. Sci China Inf Sci, 2014, 57: 112103(10), doi: 10.1007/s11432-014-5090-z

1 Introduction

Kernel methods [1,2] have been successfully used in pattern recognition and machine learning. Since the performance of kernel methods greatly depends on the selection of the kernel function, the kernel selection problem becomes an important topic in kernel methods [3–5].

It is common to select the optimal kernel function by choosing the one with the lowest generalization error [3]. Obviously, the generalization error is not directly computable, as the probability distribution generating the data is unknown. The generalization error can be estimated via either a theoretical bound or testing error on some unused data (hold-out testing or cross validation) [3]. To derive the theoretical upper bounds of the generalization error, some measures are introduced: VC dimension [1], Rademacher complexity [6], covering number [7,8], regularized risk [9], radius-margin bound [1], compression coefficient

*Corresponding author (email: szliao@tju.edu.cn)

[10], eigenvalues perturbation [11], etc. Minimizing an empirical estimate of the generalization error is an alternative for kernel selection in practice. Cross-validation (CV) and leave-one-out (LOO) cross-validation [3,12] are two popular empirical estimates. However, CV and LOO require multiple times of training the algorithm under consideration, which are computationally intensive. For the sake of efficiency, some approximate CV and LOO criteria are introduced: span bound [3], influence function [13], Bouligand influence function [14], etc. Nyström methods [15] and multilevel circulant matrices [16] are used to approximate the kernel matrix to expedite the CV process.

Based on the similarity, Cristianini et al. [17] present a new kernel selection criterion called the kernel target alignment (KTA). Similar to KTA, Cortes et al. [18] present a centered kernel target alignment criterion (CKTA) using the centered kernel matrix, which gives better performance in experiments. Nguyen and Ho [19] point out several drawbacks of the KTA, and propose a surrogate measure (called FSM) to evaluate the goodness of a kernel function via the data distribution in the feature space. Although KTA, CKTA and FSM are widely used, the connections between these criteria and the generalization error for specific learning algorithms have not been established; hence so the kernels chosen by these criteria may not guarantee good generalization performance.

In recent years, using the notions of stability to derive the generalization error bounds has attracted much attention. Rogers and Wagner [20] first consider this idea to obtain error bounds. Kearns and Ron [21] investigate it further and introduce several measures of stability formally. Bousquet and Elisseeff [22] obtain exponential bounds under restrictive conditions on the algorithm, using the notion of uniform stability. These conditions are relaxed by Kutin and Niyogi [23]. The link between stability and consistency of the empirical error minimizer is studied by Poggio et al. [24]. Cortes et al. [25] use the notion of algorithmic stability to derive novel generalization error bounds for several families of transductive regression algorithms. The link between learnability, stability and uniform convergence is studied by Shalev-Shwartz et al. [26]. Cortes et al. [27] propose the stability bounds based on the norm of the kernel approximation.

Unfortunately, most of the existing notions of stability, proposed to derive the theoretical generalization error bounds, are very difficult to be used for kernel selection in practice. In this article, we aim at presenting a kernel selection criterion, based on a new notion of stability, called spectral perturbation stability. This proposed stability quantifies the spectral perturbation of the kernel matrix when removing one example from the training set. Different from the existing notions of stability, our spectral perturbation stability is defined on the kernel matrix. Therefore, we can compute the value of the spectral perturbation stability for any given kernel function from empirical data, which makes it usable for kernel selection. Specifically, we first use spectral perturbation stability to derive the generalization error bounds. Then, to guarantee good generalization performance, we propose the new kernel selection criterion by minimizing the derived generalization error bounds. In our proposed criterion, the perturbation of the eigenvalues of the kernel matrix is efficiently computed by solving the derivative of a newly defined generalized kernel matrix. Experimental results show that the kernel selected by our proposed criterion gives better results than those chosen by KTA, CKTA, FSM and CV.

The rest of the article is organized as follows. In Section 2, we introduce some notations and preliminaries. In Section 3, we present a new notion of stability, and use this notion to derive generalization error bounds. In Section 4, we propose a kernel selection criterion by minimizing the derived generalization error bounds. In Section 5 we empirically analyze the performance of our proposed SPS criterion compared with four popular criteria (KTA, CKTA, FSM and 10-CV). Finally, Section 6 concludes this article.

2 Preliminaries

Given a training set $S = \{z_i = (\boldsymbol{x}_i, y_i)\}_{i=1}^m$ of size m drawn identically and independently from an unknown distribution P on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel function, that is, K is symmetric and for any finite set of points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\} \subset \mathcal{X}$, the kernel matrix $\boldsymbol{K} = [K(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^m$ is positive semidefinite. The reproducing kernel Hilbert space $\mathcal{H}_K$ associated with the kernel K is defined to be the

completion of the linear span of the set of functions $\{K_{\boldsymbol{x}} = K(\boldsymbol{x}, \cdot) : \boldsymbol{x} \in \mathcal{X}\}$ with the inner product denoted as $\langle \cdot, \cdot \rangle_K$ satisfying $\langle K_{\boldsymbol{x}}, K_{\boldsymbol{y}} \rangle_K = K(\boldsymbol{x}, \boldsymbol{y})$.

The learning algorithm we study here is the regularized least squares algorithm:

$$f_S = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^{m} \ell(f, z_i) + \lambda \|f\|_K^2 \right\}, \tag{1}$$

where $\ell(f, z_i) = (f(\boldsymbol{x}_i) - y_i)^2$ is the squared loss function, λ is the regularized parameter, and f_S is the solution of the regularized least squares algorithm with respect to the training set S.

We will consider measuring the performance of the regularized least squares algorithm. The main quantity we are interested in is the *risk* or *generalization error* which is a random variable depending on the training set S and is defined as $R(S) = \mathbb{E}_z[\ell(f_S, z)]$, where $\mathbb{E}_z[\cdot]$ is the expectation when z is sampled according to P. Unfortunately, $R(S)$ cannot be computed since the probability distribution P is unknown. Thus, we consider estimating it using the empirical error $R_{\text{emp}}(S)$ defined as $R_{\text{emp}}(S) = \frac{1}{m} \sum_{i=1}^{m} \ell(f_S, z_i)$.

3 Generalization error bounds

In this section, we first introduce the spectral perturbation stability, and then use this stability to derive the generalization error bounds.

3.1 Spectral perturbation stability

It is well known that the kernel matrix contains most of the information needed by the regularized least squares algorithm [19], and its eigenvalues play a central role in the kernel matrix. Therefore, we introduce a new notion of stability to quantify the perturbation of eigenvalues of the kernel matrix with respect to the changes in the training set for kernel selection.

To this end, we build the ith removed training set $S^i = \{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_m\}$. Let $\boldsymbol{K}^i$ be the $m \times m$ ith removed kernel matrix with

$$\begin{cases} [\boldsymbol{K}^i]_{jk} = K(\boldsymbol{x}_j, \boldsymbol{x}_k), & \text{if } j \neq i \text{ and } k \neq i, \\ [\boldsymbol{K}^i]_{jk} = 0, & \text{if } j = i \text{ or } k = i. \end{cases}$$

Denote the eigenvalues of $\boldsymbol{K}$ and $\boldsymbol{K}^i$ as $\sigma_j(\boldsymbol{K})$ and $\sigma_j(\boldsymbol{K}^i)$, respectively. Note that $\boldsymbol{K}$ and $\boldsymbol{K}^i$ are both positive semidefinite matrices; thus, $\sigma_1(\boldsymbol{K}) \geqslant \ldots \geqslant \sigma_m(\boldsymbol{K}) \geqslant 0$ and $\sigma_1(\boldsymbol{K}^i) \geqslant \ldots \geqslant \sigma_m(\boldsymbol{K}^i) \geqslant 0$.

Definition 1. The kernel function K is of β spectral perturbation stability if the following holds: $\forall S = \{z_i = (\boldsymbol{x}_i, y_i)\}_{i=1}^m \in \mathcal{Z}^m$ and $\forall i, j \in \{1, \ldots, m\}$, $\left|\sigma_j(\boldsymbol{K}) - \sigma_j(\boldsymbol{K}^i)\right| \leqslant \beta$.

According to the above definition, the spectral perturbation stability is used to quantify the spectral perturbation of the kernel matrix when removing an example in the training set. Different from the existing notions of stability, see, e.g., [20–26] and the references therein, our proposed stability is defined on the kernel matrix. Therefore, we can estimate its value from empirical data, which makes this stability usable for kernel selection in practice.

3.2 Generalization error bounds with spectral perturbation stability

We assume $\forall y \in \mathcal{Y}$, $|y| \leqslant M$ and $\sup_{\boldsymbol{x} \in \mathcal{X}} K(\boldsymbol{x}, \boldsymbol{x}) = \kappa$. To obtain the generalization error bound, we first prove the following theorem.

Theorem 1. Assume the training set $S = \{(\boldsymbol{x}_i, y)\}_{i=1}^m$. If the kernel function K is of β spectral perturbation stability, then $\forall i \in \{1, \ldots, m\}$, $\|f_S - f_{S^i}\|_\infty \leqslant \frac{C(\beta + 2\lambda)}{m-1}$, where $C = \frac{\kappa M}{\lambda^2}$, f_S and f_{S^i} are the solutions of the regularized least squares algorithm with respect to S and S^i, respectively.

Proof. The proof of this theorem is given in Appendix A.

This theorem shows that the spectral perturbation stability implies the stability of f_S, and the $\|f_S - f_{S^i}\|_\infty$ is tight when the spectral perturbation stability β is small.

Using the above theorem, we can obtain the generalization error bound.

Theorem 2. Assume the training set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$. If the kernel function K is of β spectral perturbation stability, then for the regularized least squares algorithm, with probability $1-\delta$, we have

$$R(S) \leqslant R_{\text{emp}}(S) + 2F(\beta + 2\lambda) + C^2(\beta + 2\lambda)^2 + \left(F(\beta + 2\lambda) + C^2(\beta + 2\lambda)^2 + Q\right)\sqrt{\frac{\ln 1/\delta}{2m}},$$

where $C = \frac{\kappa M}{\lambda^2(m-1)}$, $Q = \frac{2\kappa^2 M^2}{\lambda^2} + 2M^2$ and $F = \frac{2\kappa M^2}{\lambda^2(m-1)}$.

Proof. The proof of this theorem is given in Appendix B.

To guarantee good generalization performance, we should choose the kernel function with low generalization error $R(S)$. However, the $R(S)$ is not directly computable. The above theorem shows that we can choose the kernel function by minimizing the $R_{\text{emp}}(S) + \beta$ to restrict the value of $R(S)$ to guarantee good generalization performance.

4 Kernel selection criterion

In this section, we will present a kernel selection criterion based on the spectral perturbation stability, and present a strategy for fast calculation of the perturbation of the eigenvalues.

4.1 Spectral perturbation stability criterion

According to the Theorem 2, to guarantee good generalization performance, we should select the kernel with the lowest $R_{\text{emp}}(S) + \beta$. However, by the definition of the β spectral perturbation stability, we need to try all the possibilities of the training set to compute β, which is infeasible in practice. We should estimate it from the available empirical data S. Therefore, we consider using the following spectral perturbation stability criterion:

$$\text{SPS}(K) = R_{\text{emp}}(S) + \frac{1}{m^2}\sum_{j=1}^{m}\sum_{i=1}^{m}\left|\sigma_j(\boldsymbol{K}) - \sigma_j(\boldsymbol{K}^i)\right|, \tag{2}$$

where $R_{\text{emp}}(S)$ is the empirical error, $\sigma_j(\boldsymbol{K})$ and $\sigma_j(\boldsymbol{K}^i)$ are the eigenvalues of the kernel matrix $\boldsymbol{K}$ and the ith removed kernel matrix $\boldsymbol{K}^i$, respectively.

To compute the value of this criterion, we should compute the eigenvalues perturbation of kernel matrix $\sum_{j=1}^m\sum_{i=1}^m\left|\sigma_j(\boldsymbol{K}) - \sigma_j(\boldsymbol{K}^i)\right|$, which requires the calculation of the eigenvalues of $\boldsymbol{K}$ and $\boldsymbol{K}^i$, $i = 1, \ldots, m$, respectively. The computing cost is too high.

Fortunately, we will show that $\sum_{j=1}^m\sum_{i=1}^m\left|\sigma_j(\boldsymbol{K}) - \sigma_j(\boldsymbol{K}^i)\right|$ can be computed by solving the derivative of a newly defined generalized kernel matrix (see Definition 2), which reduces the times of computation of the eigen-system from $m+1$ to 1.

4.2 A strategy for fast calculation of the spectral perturbation

We will give a strategy for fast calculation of $\sum_{j=1}^m\sum_{i=1}^m\left|\sigma_j(\boldsymbol{K}) - \sigma_j(\boldsymbol{K}^i)\right|$. To this end, we give the definition of the generalized kernel matrix first.

Definition 2 (generalized kernel matrix)**.** Assume that the training set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$. Let $\boldsymbol{D}$ be the $m \times m$ diagonal matrix with $[\boldsymbol{D}]_{ii} = K(\boldsymbol{x}_i, \boldsymbol{x}_i)$. For each $i \in \{1, \ldots, m\}$, let $\boldsymbol{C}^i$ be the $m \times m$ matrix with $[\boldsymbol{C}^i]_{jk} = 0$ if $j \neq i$ and $k \neq i$, $[\boldsymbol{C}^i]_{jk} = K(\boldsymbol{x}_j, \boldsymbol{x}_k)$ if $j = i$ or $k = i$. The generalized kernel matrix $\boldsymbol{K}(\boldsymbol{w})$ is defined as $\boldsymbol{K}(\boldsymbol{w}) = \sum_{i=1}^m w_i\boldsymbol{C}^i + \frac{1}{2}\boldsymbol{D}$, where the parameters $\boldsymbol{w} = (w_1, \ldots, w_m)^{\mathrm{T}} \in \mathbb{R}^m$.

Note that $\boldsymbol{K}(\boldsymbol{w}) = \begin{cases} \boldsymbol{K} & \text{if } \boldsymbol{w} = \boldsymbol{\mathbf{1}}_{\boldsymbol{2}} = \left(\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2}\right)^{\mathrm{T}}, \\ \boldsymbol{K}^i & \text{if } \boldsymbol{w} = \left(\frac{1}{2}, \ldots, \frac{1}{2}, -\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2}\right)^{\mathrm{T}}, \end{cases}$

where $\boldsymbol{w} = (\frac{1}{2}, \ldots, \frac{1}{2}, -\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2})^{\mathrm{T}}$ denotes that the i-th element is $-\frac{1}{2}$, others $\frac{1}{2}$. Therefore, the kernel matrix $\boldsymbol{K}$ and the ith removed kernel matrix $\boldsymbol{K}^i$ can be seen as the special cases of the $\boldsymbol{K}(\boldsymbol{w})$.

Consider the eigen-system of $\boldsymbol{K}(\boldsymbol{w})$:

$$\boldsymbol{K}(\boldsymbol{w})\boldsymbol{q}(\boldsymbol{w})_j = \sigma(\boldsymbol{w})_j \boldsymbol{q}(\boldsymbol{w})_j, \tag{3}$$

where $\boldsymbol{q}(\boldsymbol{w})_j = (q(\boldsymbol{w})_{j1}, \ldots, q(\boldsymbol{w})_{jm})^{\mathrm{T}}$ and $\sigma(\boldsymbol{w})_j$, respectively, denote the jth eigenvector and jth eigenvalue of $\boldsymbol{K}(\boldsymbol{w})$, $\{\boldsymbol{q}(\boldsymbol{w})_j\}_{j=1}^m$ are orthonormal.

Loosely speaking, the derivative of $\sigma(\boldsymbol{w})_j$ with respect to w_i, that is $\frac{\partial \sigma(\boldsymbol{w})_j}{\partial w_i}$, can be thought as how much the jth eigenvalue changed in response to the change of the ith parameter w_i [28]. We consider the differential $d\sigma(\boldsymbol{w})_j$ of $\sigma(\boldsymbol{w})_j$ at w_i, which is expressed as

$$d\sigma(\boldsymbol{w})_j = \left(\frac{\partial \sigma(\boldsymbol{w})_j}{\partial w_i}\right) dw_i. \tag{4}$$

Specifically, when $\boldsymbol{w} = (\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2})^{\mathrm{T}}$ and $dw_i = \frac{1}{2} - (-\frac{1}{2}) = 1$, the corresponding change of the jth eigenvalue can be approximated as follows

$$\Delta\sigma(i)_j \approx \left(\frac{\partial \sigma(\boldsymbol{w})_j}{\partial w_i} | \boldsymbol{w} = (\frac{1}{2}, \ldots, \frac{1}{2})\right). \tag{5}$$

The equation (5) just reveals the change of the jth eigenvalue when the $\boldsymbol{K}$ changes to $\boldsymbol{K}^i$. Therefore, we can employ equation (5) to evaluate $(\sigma_j(\boldsymbol{K}) - \sigma_j(\boldsymbol{K}^i))$. Equation (5) is the first order approximation of Taylor expansion. This approximation error can be bounded by the residue terms. Thus, we employ the following criterion for kernel selection: $\mathrm{SPS}(K) = R_{\mathrm{emp}}(S) + \frac{1}{m^2}\sum_{j=1}^m \sum_{i=1}^m |\Delta\sigma(i)_j|$, where $\Delta\sigma(i)_j = \left(\frac{\partial \sigma(\boldsymbol{w})_j}{\partial w_i} | \boldsymbol{w} = (\frac{1}{2}, \ldots, \frac{1}{2})\right)$.

To compute the $\mathrm{SPS}(K)$, we should calculate the derivative of $\sigma(\boldsymbol{w})_j$ with respect to the parameter w_i. Jiang and Ren [28] present a method to calculate the derivative of eigenvalues of Laplacian matrix with respect to the feature weight coefficient. We extend their method to the generalized kernel matrix for calculating $\frac{\partial \sigma(\boldsymbol{w})_j}{\partial w_i}$:

Theorem 3. The calculation of $\frac{\partial \sigma(\boldsymbol{w})_j}{\partial w_i}$ is formulated as

$$\frac{\partial \sigma(\boldsymbol{w})_j}{\partial w_i} = 2q(\boldsymbol{w})_{ji}\boldsymbol{q}(\boldsymbol{w})_j^{\mathrm{T}}\boldsymbol{k}_i - q(\boldsymbol{w})_{ji}^2 K(\boldsymbol{x}_i, \boldsymbol{x}_i), \tag{6}$$

where $\boldsymbol{k}_i = (K(\boldsymbol{x}_i, \boldsymbol{x}_1), \ldots, K(\boldsymbol{x}_i, \boldsymbol{x}_m))^{\mathrm{T}}$.

Proof. By differentiating both sides of $\boldsymbol{K}(\boldsymbol{w})\boldsymbol{q}(\boldsymbol{w})_j = \sigma(\boldsymbol{w})_j\boldsymbol{q}(\boldsymbol{w})_j$ with respect to weight coefficient w_i, we have

$$\frac{\partial \boldsymbol{K}(\boldsymbol{w})}{\partial w_i}\boldsymbol{q}(\boldsymbol{w})_j + \boldsymbol{K}(\boldsymbol{w})\frac{\partial \boldsymbol{q}(\boldsymbol{w})_j}{\partial w_i} = \frac{\partial \sigma(\boldsymbol{w})_j}{\partial w_i}\boldsymbol{q}(\boldsymbol{w})_j + \sigma(\boldsymbol{w})_j\frac{\partial \boldsymbol{q}(\boldsymbol{w})_j}{\partial w_i}.$$

Multiply both sides of the above equation by $\boldsymbol{q}(\boldsymbol{w})_j^{\mathrm{T}}$:

$$\boldsymbol{q}(\boldsymbol{w})_j^{\mathrm{T}}\frac{\partial \boldsymbol{K}(\boldsymbol{w})}{\partial w_i}\boldsymbol{q}(\boldsymbol{w})_j + \boldsymbol{q}(\boldsymbol{w})_j^{\mathrm{T}}\boldsymbol{K}(\boldsymbol{w})\frac{\partial \boldsymbol{q}(\boldsymbol{w})_j}{\partial w_i} = \frac{\partial \sigma(\boldsymbol{w})_j}{\partial w_i}\boldsymbol{q}(\boldsymbol{w})_j^{\mathrm{T}}\boldsymbol{q}(\boldsymbol{w})_j + \sigma(\boldsymbol{w})_j\boldsymbol{q}(\boldsymbol{w})_j^{\mathrm{T}}\frac{\partial \boldsymbol{q}(\boldsymbol{w})_j}{\partial w_i}.$$

Since $\boldsymbol{K}(\boldsymbol{w})$ is symmetric and $\boldsymbol{K}(\boldsymbol{w})\boldsymbol{q}(\boldsymbol{w})_j = \sigma(\boldsymbol{w})_j\boldsymbol{q}(\boldsymbol{w})_j$, it is easy to verify that

$$\boldsymbol{q}(\boldsymbol{w})_j^{\mathrm{T}}\boldsymbol{K}(\boldsymbol{w})\frac{\partial \boldsymbol{q}(\boldsymbol{w})_j}{\partial w_i} = \sigma(\boldsymbol{w})_j\boldsymbol{q}(\boldsymbol{w})_j^{\mathrm{T}}\frac{\partial \boldsymbol{q}(\boldsymbol{w})_j}{\partial w_i}.$$

According to the above two equations, we have $\frac{\partial \sigma(\boldsymbol{w})_j}{\partial w_i} = \frac{\boldsymbol{q}(\boldsymbol{w})_j^{\mathrm{T}}\frac{\partial \boldsymbol{K}(\boldsymbol{w})}{\partial w_i}\boldsymbol{q}(\boldsymbol{w})_j}{\boldsymbol{q}(\boldsymbol{w})_j^{\mathrm{T}}\boldsymbol{q}(\boldsymbol{w})_j}$. Hence, $\{\boldsymbol{q}(\boldsymbol{w})_j\}_{j=1}^m$ are orthonormal and $\frac{\partial \sigma(\boldsymbol{w})_j}{\partial w_i} = \boldsymbol{q}(\boldsymbol{w})_j^{\mathrm{T}}\frac{\partial \boldsymbol{K}(\boldsymbol{w})}{\partial w_i}\boldsymbol{q}(\boldsymbol{w})_j$.

Note that $\boldsymbol{K}(\boldsymbol{w}) = \sum_{i=1}^m w_i\boldsymbol{C}^i + \frac{1}{2}\boldsymbol{D}$, so we have $\frac{\partial \boldsymbol{K}(\boldsymbol{w})}{\partial w_i} = \boldsymbol{C}^i$. By the definition of $\boldsymbol{C}^i$, it is easy to verify that $\frac{\partial \sigma(\boldsymbol{w})_j}{\partial w_i} = 2q(\boldsymbol{w})_{ji}\boldsymbol{q}(\boldsymbol{w})_j^{\mathrm{T}}\boldsymbol{k}_i - q(\boldsymbol{w})_{ji}^2 K(\boldsymbol{x}_i, \boldsymbol{x}_i)$. Thus, the theorem is proved.

From the above theorem, we only need to compute the eigen-system of the kernel matrix once to compute the $\mathrm{SPS}(K)$.

4.3 Time complexity analysis

To compute the spectral perturbation stability criterion SPS(K), we need $O(m^3)$ to calculate the empirical error $R_{\text{emp}}(S)$ and to calculate the eigen-system of the kernel matrix, and need $O(m^2)$ to calculate the derivatives of eigenvalues, where m is the size of the training set. Thus, the overall time complexity of SPS(K) is $O(2m^3 + m^2)$.

Remark 1. Instead of choosing a single kernel, some researchers consider combining multiple kernels by some criteria, called multiple kernel learning (MKL), see, e.g., [12,18] and the references therein. Our criterion can be used for MKL. However, in this paper, we mainly want to verify the effectiveness of spectral perturbation stability criterion.

5 Experiments

In this section, we will empirically analyze the performance of our proposed SPS criterion compared with four popular kernel selection criteria: KTA [17], CKTA [18], FSM [29] and 10-fold cross validation (10-CV). The learning algorithm we use here is the regularized least squares algorithm. Experiments are performed on a Dell Vestro PC with 3.4-GHz CPU and 8-GB memory.

The evaluation is made on 10 public available datasets from LIBSVM Data[1)] seen in Table 1. All datasets are normalized to have zero-means and unit-variances on every attribute to avoid numerical problems caused by large-value kernel matrices.

We use Gaussian kernels $K_{\text{Gauss}}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\tau\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2\right)$ and polynomial kernels $K_{\text{Ploy}}(\boldsymbol{x}, \boldsymbol{x}') = (1 + \boldsymbol{x} \cdot \boldsymbol{x}')^d$ as our candidate kernels: 20 Gaussian kernels with $\tau \in \{2^i, i = -10, -9, \ldots, 9\}$ and 20 polynomial kernels with degree $d \in \{1, 2, \ldots, 20\}$. The regularization parameter $\lambda \in \{0.01, 0.1, 1, 10\}$. For each data set, we have run all the methods 20 times with random partition of the datasets (50% of all the examples for training and the other 50% for testing).

5.1 Accuracy

The average test accuracies are reported in Table 1. The elements in this table are obtained as follows. For each training set, each regularized parameter λ, we choose the kernel by each kernel selection criterion on the training set, and evaluate the test accuracies of the chosen parameters on the test set. Then, we compute the means over all runs on the different partitions. The SPS criterion is proposed to choose the kernel function, not the regularization parameter λ; therefore, we do not select this value but report the results under different λ in our experiments. The results in Table 1 can be summarized as follows: (a) SPS is much better than KTA, CKTA and FSM on nearly all datasets. This can be explained by the fact that the connections between these three criteria and generalization error for the regularized least squares algorithm has not been established, such that the kernels chosen by these criteria may not guarantee good generalization performance. (b) SPS is comparable or better than 10-CV on most datasets. (c) The accuracies of the SPS do not depend much on the size of the data sets. Besides experiments on random evenly split datasets, we also have run our method with 70% examples for training and the other 30% for testing, the results turn out to be similar with each other. The above results imply that the influence of the amount of examples is not very large.

Therefore, it implicates that choosing the kernel based on the spectral perturbation stability can guarantee good generalization.

The highest test accuracies for each kernel selection criterion in Table 1 are reported in Figure 1. We can observe that SPS gives the best results on most of the datasets. In particular, SPS outperforms KTA, CKTA and FSM on all the datasets. SPS outperforms 10-CV on 6 (or more) out 10 sets (Breast, Diabetes, Guide, Heart, Sonar and Ionosphere), and also give results close to 10-CV on the remaining datasets.

1) http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Table 1 The test accuracies (%) with standard deviations using different λ

$\lambda = 0.01$						
Method	#Example	SPS	KTA	CKTA	CV	FSM
Australian	690	83.48 ± 1.83	72.61 ± 1.67	75.62 ± 2.09	**83.77** ± 1.30	74.94 ± 1.76
Breast	683	**97.36** ± 0.59	91.60 ± 0.67	91.52 ± 0.86	95.42 ± 1.02	93.51 ± 0.88
Diabetes	768	69.07 ± 1.86	63.75 ± 1.84	64.04 ± 1.92	**70.94** ± 2.02	64.70 ± 2.90
German	1000	**70.18** ± 1.71	66.45 ± 1.78	67.05 ± 1.66	66.19 ± 2.00	65.42 ± 1.85
Guide	1284	**77.65** ± 1.28	71.72 ± 1.38	72.65 ± 1.36	75.62 ± 2.47	65.29 ± 1.22
Heart	270	78.96 ± 3.54	76.80 ± 3.62	77.19 ± 3.58	**79.00** ± 4.23	74.80 ± 3.54
Liver	345	**62.13** ± 3.06	59.87 ± 3.46	61.14 ± 3.92	61.74 ± 3.23	61.74 ± 3.23
Sonar	208	**90.38** ± 3.25	80.38 ± 3.18	82.38 ± 3.25	85.06 ± 9.64	84.36 ± 6.64
Splice	3175	80.84 ± 1.33	71.27 ± 1.52	72.20 ± 3.57	**82.00** ± 1.52	74.90 ± 3.57
Ionosphere	351	**90.86** ± 2.93	86.49 ± 2.80	86.50 ± 2.93	88.78 ± 3.14	86.51 ± 2.92
$\lambda = 0.1$						
Method	#Example	SPS	KTA	CKTA	CV	FSM
Australian	690	84.93 ± 1.66	73.90 ± 2.03	76.23 ± 1.82	**85.71** ± 1.68	78.49 ± 2.03
Breast	683	**96.33** ± 0.84	89.26 ± 0.93	90.13 ± 0.84	95.98 ± 0.63	88.25 ± 0.84
Diabetes	768	**69.32** ± 1.46	62.36 ± 2.31	63.04 ± 1.89	68.55 ± 1.58	57.60 ± 2.46
German	1000	69.96 ± 2.04	63.36 ± 1.42	65.21 ± 1.95	**71.55** ± 1.68	64.15 ± 4.79
Guide	1284	**73.83** ± 1.58	63.44 ± 1.59	66.25 ± 1.58	69.14 ± 1.63	63.62 ± 1.65
Heart	270	**81.58** ± 3.29	75.80 ± 3.62	77.01 ± 2.93	79.05 ± 2.87	75.81 ± 3.03
Liver	345	60.33 ± 3.43	56.47 ± 4.22	57.38 ± 5.44	**61.12** ± 4.24	51.10 ± 4.28
Sonar	208	**85.71** ± 3.20	75.68 ± 4.77	78.10 ± 3.20	83.49 ± 7.57	78.72 ± 4.52
Splice	3175	82.84 ± 1.72	71.27 ± 2.20	72.55 ± 3.45	**83.84** ± 2.52	69.79 ± 2.57
Ionosphere	351	**85.85** ± 2.31	80.57 ± 2.30	80.21 ± 2.52	82.37 ± 2.71	79.18 ± 3.82
$\lambda = 1$						
Method	#Example	SPS	KTA	CKTA	CV	FSM
Australian	690	**81.60** ± 1.34	71.90 ± 2.45	74.32 ± 1.32	80.51 ± 1.27	73.53 ± 2.45
Breast	683	93.66 ± 0.93	91.26 ± 0.94	91.13 ± 0.52	**96.98** ± 0.93	87.40 ± 3.36
Diabetes	768	**71.09** ± 2.06	67.36 ± 2.10	66.56 ± 2.06	69.09 ± 2.28	64.29 ± 2.10
German	1000	**65.20** ± 1.99	61.36 ± 1.58	61.66 ± 2.30	64.12 ± 2.15	58.09 ± 2.13
Guide	1284	70.28 ± 1.38	61.34 ± 1.45	62.31 ± 1.39	**72.89** ± 1.79	63.99 ± 1.45
Heart	270	**75.56** ± 4.20	73.80 ± 4.41	72.32 ± 5.15	**73.41** ± 4.87	69.68 ± 4.41
Liver	345	54.38 ± 3.75	51.32 ± 3.64	53.72 ± 4.28	55.32 ± 3.65	51.32 ± 3.64
Sonar	208	**76.92** ± 3.81	66.68 ± 3.77	68.11 ± 3.81	73.19 ± 5.39	69.81 ± 5.54
Splice	3175	**72.64** ± 2.27	61.03 ± 2.20	63.40 ± 3.08	72.47 ± 2.46	62.22 ± 2.28
Ionosphere	351	**83.71** ± 2.52	72.57 ± 3.47	74.61 ± 2.52	81.37 ± 2.71	73.16 ± 3.47
$\lambda = 10$						
Method	#Example	SPS	KTA	CKTA	CV	FSM
Australian	690	77.97 ± 2.01	65.20 ± 2.51	64.64 ± 3.28	**79.96** ± 2.78	69.68 ± 2.49
Breast	683	92.38 ± 1.39	88.26 ± 1.73	89.07 ± 0.72	**93.68** ± 0.67	88.73 ± 2.27
Diabetes	768	**64.58** ± 2.82	58.36 ± 2.67	58.96 ± 2.93	59.10 ± 2.82	54.32 ± 2.67
German	1000	**61.20** ± 1.59	60.36 ± 1.58	58.22 ± 1.58	58.29 ± 1.61	58.21 ± 1.63
Guide	1284	**68.79** ± 1.34	63.55 ± 1.35	64.66 ± 1.30	67.85 ± 1.34	57.97 ± 1.35
Heart	270	**68.52** ± 3.58	56.61 ± 3.41	57.11 ± 4.08	68.02 ± 4.31	55.65 ± 3.71
Liver	345	**54.38** ± 3.75	52.32 ± 3.64	53.72 ± 4.28	53.32 ± 3.65	51.32 ± 3.64
Sonar	208	**66.73** ± 4.18	56.68 ± 3.87	58.97 ± 4.18	60.25 ± 4.89	59.93 ± 6.24
Splice	3175	64.57 ± 2.24	61.46 ± 2.35	62.40 ± 3.48	**64.64** ± 2.24	59.57 ± 2.96
Ionosphere	351	**72.86** ± 3.34	63.57 ± 3.47	64.84 ± 3.24	68.86 ± 3.27	63.39 ± 3.00

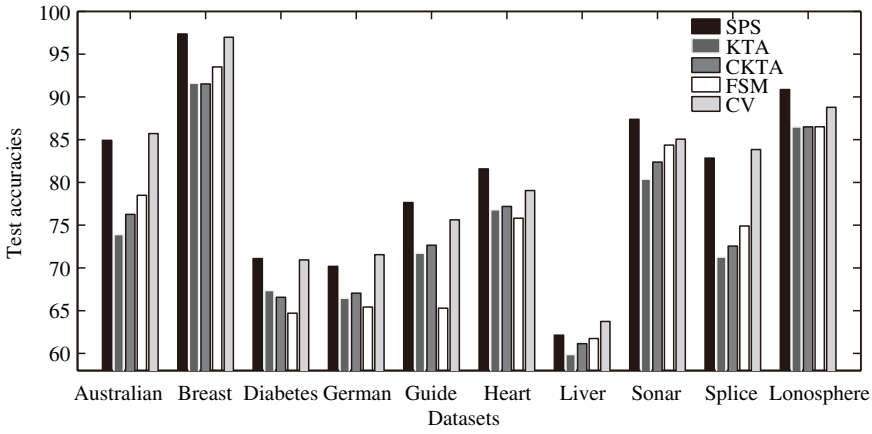

Figure 1 Comparison among SPS, KTA, CKTA, FSM and 10-CV criteria. The highest test accuracies for each kernel selection criterion in Table 1.

6 Conclusion

In this paper, we propose a new kernel selection criterion based on the spectral perturbation stability, which quantifies the spectral perturbation of the kernel matrix with respect to the changes in the training set. This criterion is theoretically justified and obtain good results in practice. We believe that our analysis opens new perspectives on the application of the stability to practical problem.

We can extend the results of Theorem 2 to SVM via the similar proof of the regularized least squares algorithm. We have obtained the generalization error bound for SVM: $R(S) \leqslant R_{\mathrm{emp}}(S) + \sqrt{\frac{Q^2+6Qm\beta^{\frac{1}{4}}(1+(\beta/(2\kappa))^{\frac{1}{4}})}{2m\delta}}$, where C and Q are some constants. Our criterion can also be applied to MKL: $\max_{\boldsymbol{\mu}=(\mu_1,\ldots,\mu_k)} \mathrm{SPS}(K_{\boldsymbol{\mu}}), \mathrm{s.t.} \|\boldsymbol{\mu}\|_p = 1, \boldsymbol{\mu} \geqslant 0$, where $K_{\boldsymbol{\mu}} = \sum_{i=1}^{k} \mu_i K_i$. This optimization can be efficiently solved by the projected gradient algorithm similar to our previous work [12].

Future work will use the Nyström methods to speed up our proposed criterion, and extend this criterion to other kernel based method (such as SVM, LSSVM), and apply this criterion for multiple kernel learning.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61170019), the Natural Science Foundation of Tianjin (Grant No. 11JCYBJC00700), and the Tianjin Key Laboratory of Cognitive Computing and Application.

References

1 Vapnik V. The Nature of Statistical Learning Theory. New York: Springer, 2000
2 Xu C, Peng Z M, Jing W F. Sparse kernel logistic regression based on $\ell_{1/2}$ regularization. Sci China Inf Sci, 2013, 56: 042308
3 Chapelle O, Vapnik V, Bousquet O, et al. Choosing multiple parameters for support vector machines. Mach Learn, 2002, 46: 131–159
4 Xu Z B, Dai M, Meng D Y. Fast and efficient strategies for model selection of Gaussian support vector machine. IEEE Trans Syst Man Cybern B Cybern, 2009, 39: 1292–1307
5 Li G Z, Zhao R W, Qu H N, et al. Model selection for partial least squares based dimension reduction. Pattern Recognit Lett, 2012, 33: 524–529
6 Bartlett P, Mendelson S. Rademacher and Gaussian complexities: Risk bounds and structural results. J Mach Learn Res, 2002, 3: 463–482
7 Zhang T. Covering number bounds of certain regularized linear function classes. J Mach Learn Res, 2002, 2: 527–550
8 Zou B, Peng Z M, Xu Z B. The learning performance of support vector machine classification based on markov sampling. Sci China Inf Sci, 2013, 56: 032110

9 Schölkopf B, Smola A. Learning with Kernels. London: MIT Press, 2002
10 Luxburg U, Bousquet O, Schölkopf B. A compression approach to support vector model selection. J Mach Learn Res, 2004, 5: 293–323
11 Liu Y, Jiang S L, Liao S Z. Eigenvalues perturbation of integral operator for kernel selection. In: Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, San Francisco, 2013. 2189–2198
12 Liu Y, Liao S Z, Hou Y X. Learning kernels with upper bounds of leave-one-out error. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, Glasgow, 2011. 2205–2208
13 Debruyne M, Hubert M, Suykens J. Model selection in kernel based regression using the influence function. J Mach Learn Res, 2008, 9: 2377–2400
14 Liu Y, Jiang S L, Liao S Z. Efficient approximation of cross-validation for kernel methods using Bouligand influence function. In: Proceedings of the 31st International Conference on Machine Learning, Beijing, 2014. 324–332
15 Ding L Z, Liao S Z. Nyström approximate model selection for LSSVM. In: Proceedings of the 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Kuala Lumpur, 2012. 282–293
16 Ding L Z, Liao S Z. Approximate model selection for large scale LSSVM. In: Proceedings of the 3rd Asian Conference on Machine Learning, Taoyuan, 2011. 165–180
17 Cristianini N, Shawe-Taylor J, Elisseeff A, et al. On kernel-target alignment. In: Proceedings of 2001 Neural Information Processing Systems Conference, Vancouver, 2001. 367–373
18 Cortes C, Mohri M, Rostamizadeh A. Two-stage learning kernel algorithms. In: Proceedings of the 27th International Conference on Machine Learning, Haifa, 2010. 239–246
19 Nguyen C H, Ho T B. An efficient kernel matrix evaluation measure. Pattern Recognit, 2008, 41: 3366–3372
20 Rogers W, Wagner T. A finite sample distribution-free performance bound for local discrimination rules. Ann Stat, 1978, 6: 506–514
21 Kearns M, Ron D. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. Neural Comput, 1999, 11: 1472–1453
22 Bousquet O, Elisseeff A. Stability and generalization. J Mach Learn Res, 2002, 2: 499–526
23 Kutin S, Niyogi P. Almost-everywhere algorithmic stability and generalization error. In: Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence, Alberta, 2002. 275–282
24 Poggio T, Rifkin R, Mukherjee S, et al. General conditions for predictivity in learning theory. Nature, 2004, 428: 419–422
25 Cortes C, Mohri M, Pechyony D, et al. Stability of transductive regression algorithms. In: Proceedings of the 25th International Conference on Machine Learning, Helsinki, 2008. 176–183
26 Shalev-Shwartz S, Shamir O, Srebro N, et al. Learnability, stability and uniform convergence. J Mach Learn Res, 2010, 11: 2635–2670
27 Cortes C, Mohri M, Talwalkar A. On the impact of kernel approximation on learning accuracy. In: Proceeding of the International Conference on Artificial Intelligence and Statistics, Sardinia, 2010. 113–120
28 Jiang Y, Ren J T. Eigenvalue sensitive feature selection. In: Proceedings of the 28th International Conference on Machine Learning, Washington, 2011. 89–96
29 Nguyen C H, Ho T B. Kernel matrix evaluation. In: Proceedings of the 20th International Joint Conference on Artifficial Intelligence, Hyderabad, 2007. 987–992

Appendix A Proof of Theorem 1

Proof. Denote the vectors $\boldsymbol{k}$, $\boldsymbol{k}_i$, $\boldsymbol{y}$ and $\boldsymbol{y}_i$ as $\boldsymbol{k} = (K(\boldsymbol{x}, \boldsymbol{x}_1), K(\boldsymbol{x}, \boldsymbol{x}_2) \ldots, K(\boldsymbol{x}, \boldsymbol{x}_n))^{\mathrm{T}}, \boldsymbol{k}_i = (K(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_{i-1}), K(\boldsymbol{x}, \boldsymbol{x}_{i+1}), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_m))^{\mathrm{T}}, \boldsymbol{y} = (y_1, y_2, \ldots, y_m)^{\mathrm{T}}, \boldsymbol{y}_i = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_m)^{\mathrm{T}}$, respectively. Let $\boldsymbol{K}_i$ be the $(m-1) \times (m-1)$ kernel matrix with respect to the ith removed training set S^i with $[\boldsymbol{K}_i]_{j,k} = K(\boldsymbol{x}_j, \boldsymbol{x}_k), \boldsymbol{x}_j, \boldsymbol{x}_k \in S^i$.

The solutions of the regularized least squares algorithm with respect to the training sets S and S^i can be, respectively, written as $f_S(\boldsymbol{x}) = \boldsymbol{k}^{\mathrm{T}}(\boldsymbol{K} + m\lambda \boldsymbol{I})^{-1}\boldsymbol{y}$, $f_{S^i}(\boldsymbol{x}) = \boldsymbol{k}_i^{\mathrm{T}}(\boldsymbol{K}_i + (m-1)\lambda \boldsymbol{I}_i)^{-1}\boldsymbol{y}_i$, where $\boldsymbol{I}$ and $\boldsymbol{I}_i$ are the $m \times m$ and $(m-1) \times (m-1)$ identity matrices, respectively.

According to the definition of the ith removed kernel matrix $\boldsymbol{K}^i$, it is easy to verify that the $f_{S^i}(\boldsymbol{x})$ can be written as

$$f_{S^i}(\boldsymbol{x}) = \boldsymbol{k}^{\mathrm{T}} \left((\boldsymbol{K}^i + (m-1)\lambda \boldsymbol{I})^{-1} - \boldsymbol{A}_i \right) \boldsymbol{y},$$

where $\boldsymbol{A}_i = \mathrm{diag}(0, \ldots, 0, 1/\left((m-1)\lambda\right), 0, \ldots, 0)$ is a diagonal matrix, with the ith diagonal element $1/((m-1)\lambda)$, others 0.

Let $\boldsymbol{G} = \boldsymbol{K} + m\lambda \boldsymbol{I}$ and $\boldsymbol{G}_i = \boldsymbol{K}^i + (m-1)\lambda \boldsymbol{I}$, we can obtain that

$$f_S(\boldsymbol{x}) - f_{S^i}(\boldsymbol{x}) = \boldsymbol{k}^{\mathrm{T}} \boldsymbol{A}_i \boldsymbol{y} + \boldsymbol{k}^{\mathrm{T}} (\boldsymbol{G}^{-1} - \boldsymbol{G}_i^{-1}) \boldsymbol{y}.$$

Note that $\boldsymbol{M}'^{-1} - \boldsymbol{M}^{-1} = -\boldsymbol{M}'^{-1}(\boldsymbol{M}' - \boldsymbol{M})\boldsymbol{M}^{-1}$ is valid for any invertible matrices $\boldsymbol{M}$ and $\boldsymbol{M}'$. Therefore,

$$\boldsymbol{G}^{-1} - \boldsymbol{G}_i^{-1} = -\boldsymbol{G}^{-1}\left(\boldsymbol{K} - \boldsymbol{K}^i + \lambda \boldsymbol{I}\right)\boldsymbol{G}_i^{-1},$$

Thus, we can obtain that

$$\begin{aligned}\|(\boldsymbol{G}^{-1} - \boldsymbol{G}_i^{-1})\boldsymbol{y}\| &\leqslant \|\boldsymbol{K} - \boldsymbol{K}^i + \lambda \boldsymbol{I}\|\|\boldsymbol{y}\|\|\boldsymbol{G}^{-1}\|\|\boldsymbol{G}_i^{-1}\| \\ &\leqslant \frac{\|\boldsymbol{K} - \boldsymbol{K}^i + \lambda \boldsymbol{I}\|\|\boldsymbol{y}\|}{\lambda_{\min}(\boldsymbol{G})\lambda_{\min}(\boldsymbol{G}_i)} \leqslant \frac{\|\boldsymbol{K} - \boldsymbol{K}^i\|\|\boldsymbol{y}\| + \|\lambda \boldsymbol{I}\|\|\boldsymbol{y}\|}{\lambda_{\min}(\boldsymbol{G})\lambda_{\min}(\boldsymbol{G}_i)},\end{aligned}$$

where $\lambda_{\min}(\boldsymbol{G})$ and $\lambda_{\min}(\boldsymbol{G}_i)$ are the smallest eigenvalue of $\boldsymbol{G}$ and $\boldsymbol{G}_i$, respectively. Thus, we have

$$\begin{aligned}|f_S(\boldsymbol{x}) - f_{S^i}(\boldsymbol{x})| &= |\boldsymbol{k}^{\mathrm{T}}\boldsymbol{A}_i\boldsymbol{y} + \boldsymbol{k}^{\mathrm{T}}(\boldsymbol{G}^{-1} - \boldsymbol{G}_i^{-1})\boldsymbol{y}| \leqslant \|\boldsymbol{k}\|\|(\boldsymbol{G}^{-1} - \boldsymbol{G}_i^{-1})\boldsymbol{y}\| + |\boldsymbol{k}^{\mathrm{T}}\boldsymbol{A}_i\boldsymbol{y}| \\ &\leqslant \frac{\|\boldsymbol{k}\|\|\boldsymbol{K} - \boldsymbol{K}^i\|\|\boldsymbol{y}\| + \|\boldsymbol{k}\|\|\lambda \boldsymbol{I}\|\|\boldsymbol{y}\|}{\lambda_{\min}(\boldsymbol{G})\lambda_{\min}(\boldsymbol{G}_i)} + \left|\frac{K(\boldsymbol{x}, \boldsymbol{x}_i)y_i}{(m-1)\lambda}\right|.\end{aligned}$$

Hence the fact that $\lambda_{\min}(\boldsymbol{G})$ and $\lambda_{\min}(\boldsymbol{G}_i)$ are larger than or equal to $m\lambda$ and $(m-1)\lambda$. Note that $\|\boldsymbol{y}\| \leqslant \sqrt{m}M$ and $\|\boldsymbol{k}\| \leqslant \sqrt{m}\kappa$; therefore, we have

$$|f_S(\boldsymbol{x}) - f_{S^i}(\boldsymbol{x})| \leqslant \frac{\kappa M\|\boldsymbol{K} - \boldsymbol{K}^i\|}{\lambda^2(m-1)} + \frac{\kappa M}{(m-1)\lambda} + \frac{\kappa M}{(m-1)\lambda}.$$

Denote the diagonal matrix $\boldsymbol{\Lambda}$ as $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1(\boldsymbol{K}), \ldots, \lambda_m(\boldsymbol{K}))$. Note the fact that

$$\|\boldsymbol{K} - \boldsymbol{K}^i\| = \|\boldsymbol{K} - \boldsymbol{\Lambda} + \boldsymbol{\Lambda} - \boldsymbol{K}^i\| \leqslant \|\boldsymbol{K} - \boldsymbol{\Lambda}\| + \|\boldsymbol{\Lambda} - \boldsymbol{K}^i\| = 0 + \sup_{i\in\{1,\ldots,m\}}\left|\lambda_i(\boldsymbol{K}) - \lambda_i(\boldsymbol{K}^i)\right|.$$

According to the definition of β spectral perturbation stability, we have $\sup_{i\in\{1,\ldots,m\}}\left|\lambda_i(\boldsymbol{K}) - \lambda_i(\boldsymbol{K}^i)\right| \leqslant \beta$. Thus, $|f_S(\boldsymbol{x}) - f_{S^i}(\boldsymbol{x})| \leqslant \frac{\beta\kappa M}{\lambda^2(m-1)} + \frac{2\kappa M}{(m-1)\lambda} = \frac{\kappa M}{\lambda^2(m-1)}(\beta + 2\lambda)$.

Appendix B Proof of Theorem 2

Definition B1. An algorithm A has uniform stability γ with respect to the loss function ℓ for the following:

$$\forall S = \{z_i\}_{i=1}^m \in \mathcal{Z}^m, \forall i \in \{1, \ldots, m\}, \|\ell(f_S, \cdot) - \ell(f_{S^i}, \cdot)\|_\infty \leqslant \gamma.$$

Theorem B1 (Theorem 12 in [22])**.** Let A be an algorithm with uniform stability γ with respect to a loss function ℓ such that $0 \leqslant \ell(f_S, z) \leqslant L$, for all $z \in \mathcal{Z}$ and all sets S. Then, for any $m \geqslant 1$, and any $\delta \in (0, 1)$, the following bounds hold (separately) with probability at least $1 - \delta$ over the random draw of the sample S,

$$R(S) \leqslant R_{\mathrm{emp}}(S) + 2\gamma + (4m\gamma + L)\sqrt{\frac{\ln 1/\delta}{2m}}.$$

Proof of Theorem 2. According to Theorem 1, we know that $\|f_S - f_{S^i}\|_\infty \leqslant C(\beta + 2\lambda)$. Thus, $\forall z \in \mathcal{Z}$

$$\begin{aligned}|\ell(f_S, z) - \ell(f_{S^i}, z)| &= |(y - f_S(\boldsymbol{x}))^2 - (y - f_{S^i}(\boldsymbol{x}))^2| = |f_S(\boldsymbol{x}) - f_{S^i}(\boldsymbol{x})| \cdot |2y - f_S(\boldsymbol{x}) + f_{S^i}(\boldsymbol{x})| \\ &\leqslant C(\beta + 2\lambda)(2M + C(\beta + 2\lambda)) = 2MC(\beta + 2\lambda) + C^2(\beta + 2\lambda)^2.\end{aligned}$$

From the definition of uniform stability (see in Definition B1, we know that the regularized least squares algorithm is also $2MC(\beta + 2\lambda) + C^2(\beta + 2\lambda)^2$ uniform stability.

Note that $|y| \leqslant M$ and $f_S(\boldsymbol{x}) = \boldsymbol{k}^{\mathrm{T}}(\boldsymbol{K} + m\lambda \boldsymbol{I})^{-1}\boldsymbol{y}$; therefore, we have $|f(\boldsymbol{x})| \leqslant \frac{\|\boldsymbol{k}\|\|\boldsymbol{y}\|}{\lambda_{\min}(\boldsymbol{K} + m\lambda \boldsymbol{I})} \leqslant \frac{\kappa M}{\lambda}$. Therefore, $\ell(f_S, z) = (f_S(\boldsymbol{x}) - y)^2 \leqslant 2f_S^2(\boldsymbol{x}) + 2|y|^2 \leqslant \frac{2\kappa^2 M^2}{\lambda^2} + 2M^2$. According to the Theorem B1, the assertion is proved.