

Full length article

# Regularization networks with indefinite kernels

Qiang Wu

*Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, TN, 37132, USA*

Received 29 May 2012; received in revised form 10 September 2012; accepted 3 October 2012

Available online 26 October 2012

Communicated by Ding-Xuan Zhou

---

**Abstract**

Learning with indefinite kernels attracted considerable attention in recent years due to their success in various learning scenarios. In this paper we study the asymptotic properties of the regularization kernel networks where the kernels are assumed to be indefinite, without the usual restrictions of symmetry and positive semi-definiteness as in the traditional study of kernel methods. The kernels are characterized in terms of the singular value decomposition of the corresponding kernel integrals. Two reproducing kernel Hilbert spaces are induced to characterize the approximation ability. Capacity independent error bounds are proved. Fast convergence rates are obtained both in reproducing kernel Hilbert spaces and in  $L^2$  sense.

© 2012 Elsevier Inc. All rights reserved.

**Keywords:** Regularization kernel network; Indefinite kernel; Coefficient regularization; Least square regression; Reproducing kernel Hilbert space; Capacity independent error bound; Learning rates

---

**1. Introduction**

Kernel methods are powerful statistical learning techniques due to their good performance in various scenarios. Research in the literature has focused on the positive semi-definite kernels which can be interpreted as generalized inner product in certain reproducing kernel Hilbert spaces. This feature of positive semi-definite kernels has enabled most traditional linear machine learning methods to have their corresponding kernel formulation. Typical examples include the support vector machines [20], kernel principal component analysis [14], kernel regression [3,2]. As the properties of reproducing kernel Hilbert spaces were well explored, theory for learning with positive semi-definite kernels has been extensively studied from various perspectives.

---

*E-mail address:* [qw@mtsu.edu](mailto:qw@mtsu.edu).

Although positive semi-definite kernels have achieved great success in many applications, indefinite kernels (non-positive definite kernels) started to draw attention in recent years. There are mathematical motivations as well as practical needs for these studies. Most commonly used positive semi-definite kernels have parameters and many of them are positive definite only when the parameter is within certain interval while they become non-positive definite for parameter out of the interval. Examples include the sigmoid kernels [20,6] and dot product kernels [17], the former be quite effective in support vector machine classification. This naturally drove research to the question whether these indefinite kernels also work well in machine learning. The answer has been proved to be yes. Indefinite kernels are shown effective in many problem domains and sometimes even slightly outperform definite kernels. These studies are of great mathematical interest and illustrate the positive definiteness is not the key for good statistical performance.

Nevertheless, comparing to mathematical motivations, the practical needs are more important in pushing forward the research of learning with indefinite kernels. In [7] it was found that fractional power polynomials are more powerful in face recognition than usual polynomial kernels while the former is usually not positive semi-definite. It was pointed out in [10] that positive definite kernels are limited in some problem domains due to the non-Euclidean distances used there. Instead, indefinite kernels arise naturally and can handle the problems effectively. In protein similarity analysis the protein sequence similarity measures derived from Smith–Waterman and BLAST score [13] requires learning with a non-positive semi-definite similarity matrix. These works have motivated a lot of algorithms to handle indefinite kernels or matrices. Some researchers choose to regularize the non-positive definite kernels to make them positive semi-definite [4,11,8,25] and some others developed algorithms directly workable with indefinite kernels [17,9,7,10].

As the development of learning algorithms with indefinite kernels and their success in practice, theoretical studies also achieved advances, though not rich yet. In [23,22] the linear programming SVM with indefinite kernels was analyzed within an error decomposition framework. In [5] a feature space interpretation was used to explain the effectiveness of SVM with indefinite kernels. In a series of papers [21,24,15], least square regression with indefinite kernels and  $\ell_1$  coefficient regularization was studied and capacity dependent error analysis was given. The convergence rates highly depend on the smoothness of the kernel function and could be very slow for rough kernels. In [19] capacity independent analysis was studied for least square regression with indefinite kernels and  $\ell_2$  regularization and the consistency was established for arbitrary continuous indefinite kernels. These results have provided elementary mathematical foundations for learning with indefinite kernels.

Kernel networks are special neural networks which include the well known radial basis function network, kernel regression, support vector machines as typical examples. In this paper we will focus specifically on the kernel networks for regression problem.

Let  $X$  be a compact metric space and  $Y = \mathbb{R}$ ,  $\rho$  be a Borel probability distribution on  $Z = X \times Y$  and have a finite second order moment. The regression function  $f_\rho : X \rightarrow Y$  is given by

$$f_\rho(x) = \mathbb{E}(y|x) = \int_Y y d\rho(y|x)$$

where  $\rho(y|x)$  is the conditional distribution of  $y$  for given  $x$ . In the supervised learning framework,  $\rho$  is unknown and the task is to learn a good approximation of  $f_\rho$  from a set of observations  $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in Z^m$  which are drawn independently and identically distributed

according to  $\rho$ . Since  $f_\rho$  is the minimizer of the least square loss functional

$$\mathcal{E}(f) = \mathbb{E}[(y - f(x))^2], \quad \forall f : X \rightarrow Y,$$

the approximation is expected to be obtained by minimizing the empirical least square loss functional

$$\mathcal{E}_Z(f) = \frac{1}{m} \sum_{i=1}^m (y_i - f(x_i))^2$$

over a class of pre-selected functions called hypothesis space. This is a typical ill-posed problem and regularization technique is needed [20]. Tikhonov regularization is commonly used to overcome the ill-posedness which, given the hypothesis space  $\mathcal{H}$ , and a penalty functional  $\Omega : \mathcal{H} \rightarrow \mathbb{R}_+$  called regularizer, searches for an approximation of  $f_\rho$  by the following scheme:

$$f_{Z, \mathcal{H}} = \arg \min_{f \in \mathcal{H}} \left\{ \mathcal{E}_Z(f) + \lambda \Omega(f) \right\}. \quad (1.1)$$

In kernel networks, a kernel function  $K : X \times X \rightarrow \mathbb{R}$  plays the role of generating function and the hypothesis functions are the linear combinations of the kernel function evaluated at certain points,

$$f(x) = \sum_{i=1}^N \alpha_i K(x, c_i).$$

Here  $N$  is the number of neurons and these points  $c_i$  are called centers. Theoretically, these centers are trained from the data as well as the coefficients  $\alpha_i$ . But since the training of the centers are not easy, an alternative simple method is to take  $c_i$  identically the sampling points:  $c_i = x_i$ . In this case the algorithm is termed as interpolation network. It has a sample dependent hypothesis space

$$\mathcal{H}_{K, \mathbf{x}} = \left\{ f_\alpha(x) = \sum_{i=1}^m \alpha_i K(x, x_i) : \alpha = (\alpha_1, \dots, \alpha_m) \in \mathbb{R}^m, m \in \mathbb{N} \right\}.$$

Learning with sample dependent hypothesis space has essential differences from learning with sample independent hypothesis spaces from a statistical analysis point of view due to the difficulty of defining the approximation error; see discussions in [23]. However, it is out of the scope of this paper and will not be discussed in detail here.

For the interpolation kernel network, regularization can be put on the coefficients  $\alpha_i$ . Define  $\Omega(f_\alpha) = \Omega(\alpha)$  by a positive function  $\Omega$  on  $\mathbb{R}^m$ . The interpolation kernel network with coefficient regularization estimates the regression function by

$$f_Z = f_{\alpha_Z} \quad \text{where } \alpha_Z = \arg \min_{\alpha \in \mathbb{R}^m} \left\{ \frac{1}{m} \sum_{i=1}^m (y_i - f_\alpha(x_i))^2 + \lambda \Omega(\alpha) \right\}.$$

The hypothesis space  $\mathcal{H}_{K, \mathbf{x}}$  and the coefficient regularization have some advantages. In learning algorithms using  $\mathcal{H}_{K, \mathbf{x}}$ , one can freely choose the regularizer for different purposes. For instance the sparse representation can be obtained if  $\ell_1$  norm of the coefficients is used as the regularizer. Moreover, it enables the use of both positive definite kernel and indefinite kernels if one has some a priori knowledge and wants to fit the data in certain trend, e.g., the use of

fractional power polynomial for face recognition [7] was driven by the trend that the recognition performance improves when decreasing the power of the polynomial kernels used in the learning process.

The aim of this paper is to study the theoretical performance of the least square kernel regression with a particular coefficient regularization:

$$f_{\mathbf{z}} = f_{\alpha_{\mathbf{z}}} \quad \text{where } \alpha_{\mathbf{z}} = \arg \min_{\alpha \in \mathbb{R}^m} \left\{ \frac{1}{m} \sum_{i=1}^m (y_i - f_{\alpha}(x_i))^2 + \lambda m \sum_{i=1}^m \alpha_i^2 \right\}. \quad (1.2)$$

$f_{\mathbf{z}}$  has explicit expressions given by

$$f_{\mathbf{z}} = T \left( \lambda I + ST_*ST \right)^{-1} ST_* \mathbf{y} = \left( \lambda I + TST_*S \right)^{-1} TST_* \mathbf{y}, \quad (1.3)$$

where  $\mathbf{y} = (y_1, \dots, y_m)^{\top} \in \mathbb{R}^m$ ,  $S$  is the sampling operator defined by  $Sf = (f(x_1), \dots, f(x_m))^{\top} \in \mathbb{R}^m$  for any function  $f$ ,  $T$  and  $T_*$  are operators from  $\mathbb{R}^m$  to proper function spaces: for  $\alpha \in \mathbb{R}^m$ ,

$$T\alpha = \frac{1}{m} \sum_{i=1}^m \alpha_i K(\cdot, x_i), \quad T_*\alpha = \frac{1}{m} \sum_{i=1}^m \alpha_i K(x_i, \cdot).$$

Both expressions were obtained and used for the analysis in [19]. Although the consistency and convergence rate has been established there, there are several questions kept open. First, the second expression in (1.3) was only deduced informally. Its strict interpretation requires the invertibility of  $\lambda I + TST_*S$  which is unknown yet. Second, it was proved that the approximation ability can be characterized via a reproducing kernel Hilbert space  $\mathcal{H}_{\tilde{K}}$  associated to the kernel

$$\tilde{K}(x, t) = \mathbb{E}_u [K(x, u)K(t, u)]$$

and the operator  $TST_*S$  weakly converges to the integral operator  $L_{\tilde{K}}$  defined by

$$L_{\tilde{K}} f(x) = \mathbb{E}_t [K(x, t)f(t)].$$

Although weak convergence provided us a useful observation about the population version of  $f_{\mathbf{z}}$  and guided the asymptotic analysis, it itself cannot be used in the analysis. A natural question is whether strong convergence holds true in certain sense and helps improve the convergence analysis. Finally, the estimation error of the regression learning is usually measured by the prediction error, or equivalently, the  $L^2$  distance between  $f_{\mathbf{z}}$  and  $f_{\rho}$ . Note that the convergence in  $L^2$  is the convergence in average sense and does not imply pointwise convergence. At certain points the prediction by  $f_{\mathbf{z}}$  may be very far from the truth. In traditional kernel regression with positive semi-definite kernels, the convergence in  $C(X)$  can be established provided that  $f_{\rho}$  is sufficiently smooth; see e.g. [16]. The last purpose of this paper is to study the possibility of pointwise or stronger convergence for the algorithm (1.2).

We will start with a study of the structure of indefinite kernels via the singular value decomposition of the corresponding integral operator which induced two reproducing kernel Hilbert spaces  $\mathcal{H}_0$  and  $\mathcal{H}_1$ . It turns out  $\mathcal{H}_0$  is more suitable to characterize the properties of the algorithm:  $f_{\mathbf{z}}$  lies in  $\mathcal{H}_0$ ,  $TST_*S$  converges to  $L_{\tilde{K}}$  as operators on  $\mathcal{H}_0$ , and  $\lambda I + TST_*S$  is invertible on  $\mathcal{H}_0$ . These properties will be proved in Sections 2–4. Note the last property provides a strict mathematical interpretation for the second expression of  $f_{\mathbf{z}}$  in (1.3). The first and second properties are the key features that make our analysis superior to that in [19]. They

not only enable the convergence analysis in  $\mathcal{H}_0$  which implies the convergence in  $C(X)$ , but also enable some advanced techniques in the convergence analysis and help improve the learning rate estimates. These will be done in Section 5. Some further discussions will be given in Section 6.

## 2. Structure of indefinite kernels

Let us start with the definition of reproducing kernel Hilbert spaces since they will be used constantly in the sequel. Let  $K : X \times X \rightarrow \mathbb{R}$  be continuous, symmetric, and positive semi-definite, meaning that the kernel matrix  $K_{\mathbf{x}} = [K(x_i, x_j)]_{i,j=1}^m$  evaluated on any subset  $\mathbf{x} = \{x_1, \dots, x_m\}$  of  $X$  is positive semi-definite. Such a function is called a Mercer kernel. The reproducing kernel Hilbert space  $\mathcal{H}_K$  associated to the kernel  $K$  is defined to be the completion of the span of  $\{K_x = K(\cdot, x) : x \in X\}$  with the inner product induced by  $\langle K_x, K_t \rangle_K = K(x, t)$ . The reproducing property

$$f(x) = \langle f, K_x \rangle_K$$

holds for every  $f \in \mathcal{H}_K$ . By Schwartz inequality, this implies the point evaluations are continuous functional and satisfy

$$|f(x)| \leq \sqrt{K(x, x)} \|f\|_K.$$

For more properties of  $\mathcal{H}_K$  we refer to [1].

Indefinite kernels are also functions on  $X \times X$ . But the restrictions of symmetry and positive semi-definiteness are removed and only the continuity condition is kept.

For any kernel function  $K$ , positive semi-definite or indefinite, the notation  $L_K$  will represent integral operator defined by

$$L_K f(x) = \mathbb{E}[K(x, t)f(t)] = \int_X K(x, t)f(t)d\rho_X(t),$$

where  $\rho_X$  is the marginal distribution of  $\rho$  on  $X$ . Throughout this paper we assume  $\rho_X$  is non-degenerate on  $X$ .

For a positive semi-definite kernel  $K$ ,  $L_K$  is a bounded positive operator both on  $L^2_{\rho_X}$ , the space of square integrable functions with respect to the measure  $\rho_X$ , and on  $\mathcal{H}_K$ . Moreover,  $L_K^{\frac{1}{2}}$  is an isomorphism from  $\overline{\mathcal{H}_K}$ , the closure of  $\mathcal{H}_K$  in  $L^2_{\rho_X}$ , to  $\mathcal{H}_K$ , i.e., for each  $f \in \overline{\mathcal{H}_K}$ ,  $L_K^{\frac{1}{2}}f \in \mathcal{H}_K$  and

$$\|f\|_{L^2_{\rho_X}} = \left\| L_K^{\frac{1}{2}}f \right\|_K.$$

For a continuous indefinite kernel, since  $X$  is bounded,  $L_K$  is a compact operator on  $L^2_{\rho_X}$  and has singular value decomposition. We summarize some of its properties in the following lemma. The result is standard in functional analysis (see e.g. [12]). For completeness we give a short proof.

**Lemma 2.1.** *Let  $K$  be a continuous kernel function on  $X \times X$ . There are a set of non-negative numbers  $\{\sigma_\ell\}$  and two sets of orthonormal bases  $\{\phi_\ell\}$  and  $\{\psi_\ell\}$  of  $L^2_{\rho_X}$  such that the kernel  $K$  admits a decomposition*

$$K(x, t) = \sum_{\ell=1}^{\infty} \sigma_\ell \phi_\ell(x) \psi_\ell(t), \quad (2.1)$$

where the series on the right hand side converges in  $L^2(X \times X, \rho_X \otimes \rho_X)$ . The operator  $L_K$  admits the singular value decomposition

$$L_K = \sum_{\ell=1}^{\infty} \sigma_{\ell} \phi_{\ell} \otimes \psi_{\ell}.$$

The two sets of orthonormal bases satisfy

$$L_K \psi_{\ell} = \sigma_{\ell} \phi_{\ell} \quad \text{and} \quad L_K^* \phi_{\ell} = \sigma_{\ell} \psi_{\ell}.$$

**Proof.** Consider the operator  $L_K^* L_K$ . It is symmetric, positive and compact on  $L^2_{\rho_X}$ . Hence it has eigenvalues  $\sigma_{\ell}^2$ , eigenfunctions  $\psi_{\ell}$ , and admits the eigen-decomposition

$$L_K^* L_K = \sum_{\ell=1}^{\infty} \sigma_{\ell}^2 \psi_{\ell} \otimes \psi_{\ell}.$$

Define  $\phi_{\ell} = \sigma_{\ell}^{-1} L_K \psi_{\ell}$ . Then it is easy to check all the conclusions.  $\square$

Since  $\{\phi_{\ell}\}$  and  $\{\psi_{\ell}\}$  are orthonormal bases of  $L^2_{\rho_X}$ , we can use them to define two subspaces of  $L^2_{\rho_X}$  as follows: let  $\Lambda = \{\ell : \sigma_{\ell} > 0\}$  and define

$$\begin{aligned} \mathcal{H}_0 &= \left\{ f \in L^2_{\rho_X} : \|f\|_0^2 = \sum_{\ell \in \Lambda} \frac{\langle f, \phi_{\ell} \rangle_{L^2_{\rho_X}}^2}{\sigma_{\ell}} < \infty \right\}, \\ \mathcal{H}_1 &= \left\{ f \in L^2_{\rho_X} : \|f\|_1^2 = \sum_{\ell \in \Lambda} \frac{\langle f, \psi_{\ell} \rangle_{L^2_{\rho_X}}^2}{\sigma_{\ell}} < \infty \right\}. \end{aligned}$$

Both  $\mathcal{H}_0$  and  $\mathcal{H}_1$  are Hilbert spaces and, as subspaces of  $L^2_{\rho_X}$ , their norms are stronger than  $L^2_{\rho_X}$  norm. Moreover, they are dense in  $L^2_{\rho_X}$  if all  $\sigma_{\ell} > 0$ . The inner products on  $\mathcal{H}_0$  and  $\mathcal{H}_1$  will be denoted by  $\langle \cdot, \cdot \rangle_0$  and  $\langle \cdot, \cdot \rangle_1$  respectively.

Throughout this paper we will make a key assumption on the indefinite kernel function:

**Assumption 1.** Both

$$\kappa_0^2 = \sup_{x \in X} \sum_{\ell \in \Lambda} \sigma_{\ell} \phi_{\ell}^2(x) \quad \text{and} \quad \kappa_1^2 = \sup_{x \in X} \sum_{\ell \in \Lambda} \sigma_{\ell} \psi_{\ell}^2(x)$$

are finite. Denote  $\kappa = \max\{\kappa_0, \kappa_1\}$ .

The following lemma is a quite direct conclusion of this assumption.

**Lemma 2.2.**  $\sum_{\ell \in \Lambda} \sigma_{\ell} \leq \min\{\kappa_0^2, \kappa_1^2\}$ . Consequently,  $\sigma_{\ell} \leq \kappa^2$  for all  $\ell$ .

We can prove the following results.

**Theorem 2.3.** Let  $K$  be a continuous kernel function satisfying [Assumption 1](#). The following statements are true.

(i) Both  $\mathcal{H}_0$  and  $\mathcal{H}_1$  are reproducing kernel Hilbert spaces and their reproducing kernels are

$$K_0(x, t) = \sum_{\ell \in \Lambda} \sigma_{\ell} \phi_{\ell}(x) \phi_{\ell}(t)$$

and

$$K_1(x, t) = \sum_{\ell \in \Lambda} \sigma_\ell \psi_\ell(x) \psi_\ell(t)$$

respectively. These two series converge absolutely and hence both  $K_0$  and  $K_1$  are continuous. Moreover,

$$|K_0(x, t)| \leq \kappa_0^2 \quad \text{and} \quad |K_1(x, t)| \leq \kappa_1^2.$$

- (ii)  $K(\cdot, t) \in \mathcal{H}_0$  and  $\langle K(\cdot, t), K(\cdot, t') \rangle_0 = K_1(t, t')$  for  $t, t' \in X$ . Similarly,  $K(x, \cdot) \in \mathcal{H}_1$  and  $\langle K(x, \cdot), K(x', \cdot) \rangle_1 = K_0(x, x')$  for  $x, x' \in X$ .
- (iii) The integral operator  $L_K$  is a bounded operator from  $\mathcal{H}_1$  to  $\mathcal{H}_0$ . Its adjoint  $L_K^*$  is bounded from  $\mathcal{H}_0$  to  $\mathcal{H}_1$ . Moreover,  $\|L_K\|_{\mathcal{H}_1 \rightarrow \mathcal{H}_0} = \|L_K^*\|_{\mathcal{H}_0 \rightarrow \mathcal{H}_1} \leq \kappa^2$ .
- (iv) The operator  $Jf(t) = \langle f, K(\cdot, t) \rangle_0$  is an isomorphism from  $\mathcal{H}_0$  onto  $\mathcal{H}_1$ . Its adjoint,  $J^*f(x) = \langle f, K(x, \cdot) \rangle_1$ , is an isomorphism from  $\mathcal{H}_1$  to  $\mathcal{H}_0$ . As a consequence,  $\|J\| = \|J^*\| = 1$ . Moreover,  $J^*J = I_{\mathcal{H}_0}$  and  $JJ^* = I_{\mathcal{H}_1}$ . Here  $I$  represents the identity operator on the space specified by the subscript.
- (v)  $JL_K = L_{K_1}$  and  $J^*L_K^* = L_{K_0}$ .
- (vi) For  $f \in L_{\rho_X}^2$ ,  $\|L_K f\|_0 = \langle f, L_{K_1} f \rangle_{L_{\rho_X}^2} = \|L_{K_1}^{\frac{1}{2}} f\|_{L_{\rho_X}^2}$  and  $\|L_K^* f\|_1 = \langle f, L_{K_0} f \rangle_{L_{\rho_X}^2} = \|L_{K_0}^{\frac{1}{2}} f\|_{L_{\rho_X}^2}$ .

**Proof.** (i) It suffices to prove the reproducing properties for  $\mathcal{H}_0$  and  $\mathcal{H}_1$ . They follow by simple calculation. In fact, for  $\mathcal{H}_0$ , given a function  $f = \sum_{\ell \in \Lambda} f_\ell \phi_\ell \in \mathcal{H}_0$  and  $t \in X$ , it is easy to check

$$\langle f, K_0(\cdot, t) \rangle_0 = \sum_{\ell \in \Lambda} f_\ell \phi_\ell(t) = f(t).$$

By

$$|K_0(x, t)| \leq \left( \sum_{\ell \in \Lambda} \sigma_\ell \phi_\ell^2(x) \right)^{1/2} \left( \sum_{\ell \in \Lambda} \sigma_\ell \phi_\ell^2(t) \right)^{1/2}$$

and [Assumption 1](#) we obtain the absolute convergence and uniform bound  $|K_0(x, t)| \leq \kappa_0^2$ . The continuity is a consequence of the absolute convergence.

For assertions about  $\mathcal{H}_1$  and  $K_1$  the proof is analogous.

(ii) By the definition of the norm on  $\mathcal{H}_0$  and [Assumption 1](#),

$$\|K(\cdot, t)\|_0^2 = \sum_{\ell \in \Lambda} \sigma_\ell \psi_\ell^2(t) = K_1(t, t') \leq \kappa_1^2 < \infty$$

implying  $K(\cdot, t) \in \mathcal{H}_0$ . It is direct to obtain

$$\langle K(\cdot, t), K(\cdot, t') \rangle_0 = \sum_{\ell \in \Lambda} \sigma_\ell \psi_\ell(t) \psi_\ell(t') = K_1(t, t').$$

Similarly the second assertion can be proved.

(iii) For any  $f = \sum_{\ell \in \Lambda} f_\ell \psi_\ell \in \mathcal{H}_1$ , we have  $L_K f = \sum_{\ell \in \Lambda} \sigma_\ell f_\ell \phi_\ell$ . Then

$$\|L_K f\|_0 = \left( \sum_{\ell \in \Lambda} \sigma_\ell f_\ell^2 \right)^{\frac{1}{2}} \leq \kappa^2 \left( \sum_{\ell \in \Lambda} \frac{f_\ell^2}{\sigma_\ell} \right)^{\frac{1}{2}} = \kappa^2 \|f\|_1.$$

Similarly, for any  $f \in \mathcal{H}_0$  we can prove  $\|L_K^* f\|_1 \leq \kappa^2 \|f\|_0$ .

(iv) First notice that  $J\phi_\ell = \psi_\ell$  and  $J^*\psi_\ell = \phi_\ell$  which directly leads to the conclusion  $J^*J = I_{\mathcal{H}_0}$  and  $JJ^* = I_{\mathcal{H}_1}$ . For  $f = \sum_{\ell \in \Lambda} f_\ell \phi_\ell \in \mathcal{H}_0$  and  $g = \sum_{\ell \in \Lambda} g_\ell \phi_\ell \in \mathcal{H}_0$ , we have

$$\langle Jf, Jg \rangle_1 = \left\langle \sum_{\ell \in \Lambda} f_\ell \psi_\ell, \sum_{\ell \in \Lambda} g_\ell \psi_\ell \right\rangle_1 = \sum_{\ell \in \Lambda} \frac{f_\ell g_\ell}{\sigma_\ell} = \langle f, g \rangle_0.$$

This proves the isomorphism. Similarly for  $f, g \in \mathcal{H}_1$ , we have  $\langle J^*f, J^*g \rangle_0 = \langle f, g \rangle_1$ , proving  $J^*$  is an isomorphism.

(v) For any  $f = \sum_{\ell=1}^\infty f_\ell \psi_\ell \in L_{\rho_X}^2$ , we have  $L_K f = \sum_{\ell \in \Lambda} \sigma_\ell f_\ell \phi_\ell$ . Thus,

$$JL_K f = J \left( \sum_{\ell \in \Lambda} f_\ell \sigma_\ell \phi_\ell \right) = \sum_{\ell \in \Lambda} f_\ell \sigma_\ell \psi_\ell = L_{K_1} f$$

where in the last step we used the decomposition for  $K_1$  in (i). The proof of  $J^*L_K^* = L_{K_0}$  is similar.

(vi) Write  $f = \sum_{\ell=1}^\infty f_\ell \psi_\ell \in L_{\rho_X}^2$ . Then  $L_K f = \sum_{\ell \in \Lambda} f_\ell \sigma_\ell \phi_\ell$  and

$$\begin{aligned} \|L_K f\|_0^2 &= \sum_{\ell \in \Lambda} \sigma_\ell^2 f_\ell^2 = \left\langle \sum_{\ell \in \Lambda} f_\ell \psi_\ell, \sum_{\ell \in \Lambda} \sigma_\ell f_\ell \psi_\ell \right\rangle_{L_{\rho_X}^2} = \langle f, L_{K_1} f \rangle_{L_{\rho_X}^2} \\ &= \left\langle L_{K_1}^{\frac{1}{2}} f, L_{K_1}^{\frac{1}{2}} f \right\rangle_{L_{\rho_X}^2}. \end{aligned}$$

This proves the first identity. The second one follows similarly.  $\square$

From Theorem 2.3(ii),  $f_z \in \mathcal{H}_0$ . This is the first evidence that  $\mathcal{H}_0$  is more suitable for analysis of the algorithm under study. By  $L_{K_0}^2 = L_K L_K^* = L_{\tilde{K}}$ , we see  $\mathcal{H}_{\tilde{K}} = L_{\tilde{K}}^{\frac{1}{2}}(L_{\rho_X}^2) = L_{K_0}(L_{\rho_X}^2) = L_{K_0}^{\frac{1}{2}}(\mathcal{H}_0)$  is a subspace of  $\mathcal{H}_0$ . It is obvious too small for analyzing the properties of  $f_z$ .

Before we finish this section, we remark that we will face many operators, defined on different domains. Moreover, one operator may be defined on different domains and show different properties. For example,  $L_K$  can be regarded as operators on  $L_{\rho_X}^2$ , from  $L_{\rho_X}^2$  to  $\mathcal{H}_0$ , and from  $\mathcal{H}_1$  to  $\mathcal{H}_0$ . In different situations, their norms are different. Subscripts will be used to clarify which sense is considered unless the meaning is quite clear from the context. Also, for simplicity purposes, for the operators on  $\mathcal{H}_0$ ,  $\mathcal{H}_1$  and between them the subscripts will be simplified. For example,  $\|L_K\|_{\mathcal{H}_1 \rightarrow \mathcal{H}_0}$  will be written as  $\|L_K\|_{1,0}$  in the sequel and similar treatments will be adopted for other operators.

### 3. Approximation of integral operators

In this section we study the asymptotic properties of the operator  $TST_*S$ . It is easy to check that  $TST_*S$  converges to  $L_{\tilde{K}}$  weakly, i.e., for each function  $f$ ,  $TST_*Sf$  converges to  $L_{\tilde{K}}f$ . The aim of this section is to show the strong convergence also holds true. Precisely, we will show  $TST_*S$  is a bounded operator on  $\mathcal{H}_0$  and converges to  $L_{\tilde{K}} = L_K L_K^*$  in operator norm.

**Lemma 3.1.** *The operator  $T$  and  $T_*$  are bounded operators from  $\mathbb{R}^m$  to  $\mathcal{H}_0$  and  $\mathcal{H}_1$  respectively. Their operator norms satisfy*

$$\|T\|_{\mathbb{R}^m \rightarrow \mathcal{H}_0} \leq \kappa_1 m^{-\frac{1}{2}} \quad \text{and} \quad \|T_*\|_{\mathbb{R}^m \rightarrow \mathcal{H}_1} \leq \kappa_0 m^{-\frac{1}{2}}.$$



**Proof.** By Theorem 2.3(ii),  $T\alpha \in \mathcal{H}_0$  for any  $\alpha \in \mathbb{R}^m$ . Moreover,

$$\begin{aligned}\|T\alpha\|_0 &= \left( \frac{1}{m^2} \sum_{i=1}^m \alpha_i \alpha_j K_1(x_i, x_j) \right)^{\frac{1}{2}} \leq \frac{\kappa_1}{m} \left( \sum_{i,j=1}^m |\alpha_i| |\alpha_j| \right)^{\frac{1}{2}} = \frac{\kappa_1}{m} \|\alpha\|_{\ell_1} \\ &\leq \kappa_1 m^{-\frac{1}{2}} \|\alpha\|_{\ell_2}.\end{aligned}$$

The assertion for  $T_*$  can be proved analogously.  $\square$

**Lemma 3.2.** We have  $\|S\|_{\mathcal{H}_0 \rightarrow \mathbb{R}^m} \leq \kappa_0 \sqrt{m}$  and  $\|S\|_{\mathcal{H}_1 \rightarrow \mathbb{R}^m} \leq \kappa_1 \sqrt{m}$ .

**Proof.** For  $f \in \mathcal{H}_0$ , we have  $|f(x_i)| \leq \sqrt{K_0(x_i, x_i)} \|f\|_0 \leq \kappa_0 \|f\|_0$ . Thus

$$\|Sf\|_{\ell_2} = \left( \sum_{i=1}^m (f(x_i))^2 \right)^{1/2} \leq \left( \sum_{i=1}^m \kappa_0^2 \|f\|_0^2 \right)^{1/2} = \kappa_0 \|f\|_0 \sqrt{m}.$$

Similarly  $\|Sf\|_{\ell_2} \leq \kappa_1 \sqrt{m} \|f\|_1$  for  $f \in \mathcal{H}_1$ .  $\square$

Throughout this paper we always regard  $T$  and  $T_*$  operators from  $\mathbb{R}^m$  to  $\mathcal{H}_0$  and  $\mathcal{H}_1$  respectively, although they are not the only understanding of these two operators. For their norms we will simply write as  $\|T\|$  and  $\|T_*\|$ . Analogously we will regard  $S$  as the operators on  $\mathcal{H}_0$  and  $\mathcal{H}_1$  and, since the domain is usually clear from the context, the subscripts for the operator norm will be dropped for simplicity.

The following lemma is easy consequence of Lemmas 3.1 and 3.2.

**Lemma 3.3.**  $TS$  is a bounded operator from  $\mathcal{H}_1$  to  $\mathcal{H}_0$  and  $T_*S$  is a bounded operator from  $\mathcal{H}_0$  to  $\mathcal{H}_1$ . Their operator norms satisfy

$$\|TS\|_{1,0} \leq \kappa_0 \kappa_1 \leq \kappa^2 \quad \text{and} \quad \|T_*S\|_{0,1} \leq \kappa_0 \kappa_1 \leq \kappa^2.$$

**Theorem 3.4.** We have  $\mathbb{E}[\|TS - L_K\|_{1,0}^2] \leq \frac{\kappa^4}{m}$  and  $\mathbb{E}[\|T_*S - L_K^*\|_{0,1}^2] \leq \frac{\kappa^4}{m}$ .

**Proof.** Note each  $f \in \mathcal{H}_1$  can be written as  $f = \sum_{\ell \in A} f_\ell \psi_\ell$  with  $\|f\|_1^2 = \sum_{\ell \in A} \frac{f_\ell^2}{\sigma_\ell}$ . Thus,

$$\begin{aligned}\|TS - L_K\|_{1,0}^2 &= \sup_{\|f\|_1 \leq 1} \|(TS - L_K)f\|_0^2 \\ &= \sup_{\|f\|_1 \leq 1} \left\| \sum_{\ell \in A} f_\ell (TS - L_K)\psi_\ell \right\|_0^2 \\ &\leq \sup_{\|f\|_1 \leq 1} \left( \sum_{\ell \in A} f_\ell \|(TS - L_K)\psi_\ell\|_0 \right)^2 \\ &\leq \sup_{\|f\|_1 \leq 1} \left( \sum_{\ell \in A} \frac{f_\ell^2}{\sigma_\ell} \right) \left( \sum_{\ell \in A} \sigma_\ell \|(TS - L_K)\psi_\ell\|_0^2 \right) \\ &\leq \sum_{\ell \in A} \sigma_\ell \|(TS - L_K)\psi_\ell\|_0^2\end{aligned}$$

and hence

$$\mathbb{E}[\|TS - L_K\|_{1,0}^2] \leq \sum_{\ell \in A} \sigma_\ell \mathbb{E}[\|(TS - L_K)\psi_\ell\|_0^2].$$

Note that  $TS\psi_\ell = \frac{1}{m} \sum_{i=1}^m \psi_\ell(x_i)K(\cdot, x_i)$  and  $L_K\psi_\ell = \sigma_\ell\phi_\ell$ . By Theorem 2.3(ii) and (iv),

$$\|(TS - L_K)\psi_\ell\|_0^2 = \frac{1}{m^2} \sum_{i,j=1}^m \psi_\ell(x_i)\psi_\ell(x_j)K_1(x_i, x_j) - \frac{2}{m} \sum_{i=1}^m \sigma_\ell\psi_\ell^2(x_i) + \sigma_\ell.$$

For  $i \neq j$ ,

$$\mathbb{E}[\psi_\ell(x_i)\psi_\ell(x_j)K_1(x_i, x_j)] = \langle \psi_\ell, L_{K_1}\psi_\ell \rangle_{L_{\rho_X}^2} = \langle \psi_\ell, \sigma_\ell\psi_\ell \rangle_{L_{\rho_X}^2} = \sigma_\ell.$$

For  $i = j$ ,

$$\mathbb{E}[\psi_\ell(x_i)\psi_\ell(x_i)K_1(x_i, x_i)] \leq \kappa_1^2 \mathbb{E}[\psi_\ell^2(x_i)] = \kappa_1^2.$$

Thus,

$$\begin{aligned} \mathbb{E}[\|(TS - L_K)\psi_\ell\|_0^2] &\leq \frac{\kappa_1^2}{m} + \frac{m-1}{m}\sigma_\ell - \frac{2}{m} \sum_{i=1}^n \sigma_\ell \mathbb{E}[\psi_\ell^2(x_i)] + \sigma_\ell \\ &= \frac{\kappa_1^2 - \sigma_\ell}{m} < \frac{\kappa_1^2}{m}. \end{aligned}$$

We therefore obtain

$$\mathbb{E}[\|TS - L_K\|_{1,0}^2] \leq \frac{\kappa_1^2}{m} \sum_{\ell \in \Lambda} \sigma_\ell \leq \frac{\kappa^4}{m}.$$

The proof of  $\mathbb{E}[\|TS - L_K\|_{1,0}^2] \leq \frac{\kappa^4}{m}$  is analogous.  $\square$

**Theorem 3.5.** *The operator  $TST_*S$  is an operator on  $\mathcal{H}_0$  and*

$$\mathbb{E}[\|TST_*S - L_{\tilde{K}}\|^2] \leq \frac{4\kappa^8}{m}.$$

**Proof.** By the fact  $L_{\tilde{K}} = L_K L_K^*$ , Lemma 3.3, and Theorem 2.3(iii), we have

$$\begin{aligned} \|TST_*S - L_{\tilde{K}}\|^2 &\leq \left( \|TS\|_{1,0} \|T_*S - L_K^*\|_{0,1} + \|TS - L_K\|_{1,0} \|L_K^*\| \right)^2 \\ &\leq 2\kappa^4 \left( \|T_*S - L_K^*\|_{0,1}^2 + \|TS - L_K\|_{1,0}^2 \right). \end{aligned}$$

The conclusion then follows from Theorem 3.4.  $\square$

The proof process in fact provides also the stronger convergence in Hilbert–Schmidt norm. But it is not more helpful for our analysis.

The strong convergence of  $TST_*S$  to  $L_{\tilde{K}}$  on  $\mathcal{H}_0$  will play two main roles in our analysis. One is to enable the analysis in  $\mathcal{H}_0$  and hence lead to the pointwise convergence. The second is to enable the application of advanced techniques in  $L_{\rho_X}^2$  convergence analysis. Note in [19]  $\mathcal{H}_{\tilde{K}}$  was used to aid the convergence analysis. But since the image of  $TST_*S$  is not  $\mathcal{H}_{\tilde{K}}$ ,  $TST_*S$  is not a bounded operator on  $\mathcal{H}_{\tilde{K}}$ . It can only be understood as operator from  $\mathcal{H}_{\tilde{K}}$  to  $L_{\rho_X}^2$  or  $C(X)$ . The former only applies to  $L_{\rho_X}^2$  convergence analysis while the latter is hardly applicable due to the lack of Hilbert space structure.

#### 4. Operator expression of $f_z$

In this section we will prove the second expression in (1.3) is mathematically correct. For this purpose we will prove the invertibility of  $\lambda I + TST_*S$ . We will even provide an explicit formulation for inverse operator by which we can estimate its operator norm.

Let  $K_x = [K(x_i, x_j)]_{i,j=1}^m$  be the kernel matrix evaluated on the sampling points. Then

$$ST = \frac{1}{m}K_x \quad \text{and} \quad ST_* = \frac{1}{m}K_x^\top.$$

Thus  $STST_*$  is a symmetric positive semi-definite matrix. Consequently,  $\lambda I + STST_*$  is strictly positive definite and invertible. This fact will play an important role in our proofs below. The explicit formulation of inverse operator  $(\lambda I + TST_*S)^{-1}$  is also given in terms of  $(\lambda I + STST_*)^{-1}$ .

**Theorem 4.1.** *The operator  $\lambda I + TST_*S$  is a bijective linear operator on  $\mathcal{H}_0$ . Consequently  $(\lambda I + TST_*S)^{-1}$  exists and is bijective.*

**Proof.** We first prove that  $\lambda I + TST_*S$  is injective. It suffices to prove  $(\lambda I + TST_*S)f = 0$  implies  $f = 0$ . This can be argued as follows:

$$\begin{aligned} \lambda f + TST_*Sf &= 0 \implies \lambda Sf + STST_*Sf = 0 \\ &\implies (\lambda I + STST_*)Sf = 0 \\ &\implies Sf = 0 \\ &\implies f = -\frac{1}{\lambda}TST_*Sf = 0. \end{aligned}$$

So the injectivity is proved.

To prove the surjectivity, for any  $g \in \mathcal{H}_0$  we should be able to find  $f$  to solve  $\lambda f + TST_*Sf = g$ . It turns out that

$$f = \frac{1}{\lambda} \left( g - TST_*(\lambda I + STST_*)^{-1}Sg \right) \quad (4.1)$$

is the required solution. We can easily check this by direct calculation:

$$\begin{aligned} \lambda f + TST_*Sf &= g - TST_*(\lambda I + STST_*)^{-1}Sg \\ &\quad + \frac{1}{\lambda} \left( TST_*Sg - TST_*STST_*(\lambda I + STST_*)^{-1}Sg \right) \\ &= g + \frac{1}{\lambda}TST_*Sg - TST_* \left( I + \frac{1}{\lambda}STST_* \right) (\lambda I + STST_*)^{-1}Sg \\ &= g + \frac{1}{\lambda}TST_*Sg - \frac{1}{\lambda}TST_*Sg = g. \end{aligned}$$

We proved  $\lambda I + TST_*S$  is bijective on  $\mathcal{H}_0$ . The invertibility is a direct corollary.  $\square$

Next we establish the bound for the norm of the inverse operator which will be used later.

**Theorem 4.2.** *We have*

$$\|(\lambda I + TST_*S)^{-1}\| \leq \frac{1}{\lambda} \left( 1 + \frac{\kappa^2}{\sqrt{\lambda}} \right)$$

and

$$\|(\lambda I + TST_*S)^{-1}TS\| \leq \frac{2\kappa^2}{\lambda}.$$

**Proof.** Note that (4.1) implies that

$$(\lambda I + TST_*S)^{-1} = \frac{1}{\lambda} \left( I - TST_*(\lambda I + STST_*)^{-1}S \right).$$

Therefore

$$\|(\lambda I + TST_*S)^{-1}\| \leq \frac{1}{\lambda} \left( 1 + \|T\| \|ST_*(\lambda I + STST_*)^{-1}\| \|S\| \right) \leq \frac{1}{\lambda} \left( 1 + \frac{\kappa^2}{\sqrt{\lambda}} \right),$$

where we used Lemmas 3.1 and 3.2, and the elementary inequality

$$\|ST_*(\lambda I + STST_*)^{-1}\| = \left\| \frac{1}{m} K_{\mathbf{x}}^{\top} \left( \lambda I + \frac{1}{m^2} K_{\mathbf{x}} K_{\mathbf{x}}^{\top} \right)^{-1} \right\| \leq \frac{1}{\sqrt{\lambda}}.$$

Similarly,

$$\|(\lambda I + TST_*S)^{-1}TS\| \leq \frac{1}{\lambda} \left( \|TS\| + \|T\| \|ST_*(\lambda I + STST_*)^{-1}ST\| \|S\| \right) \leq \frac{2\kappa^2}{\lambda},$$

where we used the fact

$$\|ST_*(\lambda I + STST_*)^{-1}ST\| = \left\| \frac{1}{m} K_{\mathbf{x}}^{\top} \left( \lambda I + \frac{1}{m^2} K_{\mathbf{x}} K_{\mathbf{x}}^{\top} \right)^{-1} \frac{1}{m} K_{\mathbf{x}} \right\| \leq 1. \quad \square$$

**Remark.** At the first glance one may expect an operator norm bound  $\frac{1}{\lambda}$  for  $(\lambda I + TST_*S)^{-1}$ . However, since  $TST_*S$  is not a positive operator, such a bound cannot be proved and I conjecture it is probably not true.

## 5. Convergence analysis

After the above preparations we are ready to analyze the asymptotic properties of the solution  $f_{\mathbf{z}}$  of the algorithm (1.2). Two different kinds of convergence will be studied.

The first is the convergence in  $\mathcal{H}_0$ . By Theorem 2.3(ii) we know  $f_{\mathbf{z}} \in \mathcal{H}_0$ . If the target function  $f_{\rho} \in \mathcal{H}_0$  either, we can estimate the convergence rate in  $\mathcal{H}_0$ . This convergence implies the convergence in  $C(X)$  and is stronger than the convergence in  $L_{\rho_X}^2$ .

The second convergence is in the prediction error sense. It is equivalent to the convergence in  $L_{\rho_X}^2$  since  $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\rho}) = \|f_{\mathbf{z}} - f_{\rho}\|_{L_{\rho_X}^2}^2$ . Convergence in  $L_{\rho_X}^2$  has been proved in [19]. Here we will prove a sharper error bound and deduce faster learning rate.

### 5.1. Preliminaries

Let us first discuss our analysis framework and provide some preliminary lemma.

Since  $TST_*S$  converges to  $L_{\tilde{K}}$ , a sample limit of  $f_{\mathbf{z}}$  is given by

$$f_{\lambda} = (\lambda I + L_{\tilde{K}})^{-1} L_{\tilde{K}} f_{\rho}.$$

We write

$$f_{\mathbf{z}} - f_{\rho} = f_{\mathbf{z}} - f_{\lambda} + f_{\lambda} - f_{\rho}$$

where  $f_{\mathbf{z}} - f_{\rho}$  characterizes the variance of the algorithm due to the randomness of the samples and  $f_{\lambda} - f_{\rho}$  characterizes the approximation ability.

The convergence always holds true provided that  $f_{\rho}$  can be approximated, e.g.  $f_{\rho} \in \overline{\mathcal{H}_0}$ . However, by no free lunch principle, to deduce convergence rate we need some further regularity condition on  $f_{\rho}$ . In the sequel we will use the following assumption.

**Assumption 2.**  $f_{\rho} \in L_{K_0}^r(L_{\rho_X}^2)$  for some  $r > 0$ .

Note  $L_{K_0}^r$  characterizes the interpolation space between  $L_{\rho_X}^2$  and  $\mathcal{H}_0$  if  $r \leq \frac{1}{2}$  while it characterizes the subspaces of  $\mathcal{H}_0$  when  $r > \frac{1}{2}$ . This kind of regularity condition has been widely used in learning theory.

Our first result concerns the approximation error estimate.

**Theorem 5.1.** Under Assumption 2, we have

$$\|f_{\lambda} - f_{\rho}\|_{L_{\rho_X}^2} \leq C_1 \lambda^{\min\{1, \frac{r}{2}\}}$$

where  $C_1 = \|L_{K_0}^{-r} f_{\rho}\|_{L_{\rho_X}^2}$  if  $r \leq 2$  and  $C_1 = \kappa_0^{2(r-2)} \|L_{K_0}^{-r} f_{\rho}\|_{L_{\rho_X}^2}$  if  $r \geq 2$ .

If in addition  $r > \frac{1}{2}$ , we have

$$\|f_{\lambda} - f_{\rho}\|_0 \leq C_2 \lambda^{\min\{1, \frac{2r-1}{4}\}}$$

where  $C_2 = \|L_{K_0}^{-r} f_{\rho}\|_{L_{\rho_X}^2}$  if  $\frac{1}{2} < r \leq \frac{5}{2}$  and  $C_2 = \kappa_0^{2r-5} \|L_{K_0}^{-r} f_{\rho}\|_{L_{\rho_X}^2}$  if  $r > \frac{5}{2}$ .

**Proof.** The first conclusion is a restatement of Theorem 2.2 in [19].

To see the second one, write

$$\begin{aligned} \|f_{\mathbf{z}} - f_{\rho}\|_0 &= \|\lambda(\lambda I + L_{K_0}^2)^{-1} f_{\rho}\|_0 = \left\| \lambda(\lambda I + L_{K_0}^2)^{-1} L_{K_0}^{r-\frac{1}{2}} L_{K_0}^{\frac{1}{2}} L_{K_0}^{-r} f_{\rho} \right\|_0 \\ &\leq \lambda \left\| (\lambda I + L_{K_0}^2)^{-1} L_{K_0}^{r-\frac{1}{2}} \right\| \left\| L_{K_0}^{\frac{1}{2}} L_{K_0}^{-r} f_{\rho} \right\|_0 \\ &= \lambda \left\| (\lambda I + L_{K_0}^2)^{-1} L_{K_0}^{r-\frac{1}{2}} \right\| \|L_{K_0}^{-r} f_{\rho}\|_{L_{\rho_X}^2}. \end{aligned}$$

The conclusion follows from

$$\left\| (\lambda I + L_{K_0}^2)^{-1} L_{K_0}^{r-\frac{1}{2}} \right\| \leq \lambda^{-1+\frac{2r-1}{4}}$$

if  $\frac{1}{2} < r \leq \frac{5}{2}$  and

$$\left\| (\lambda I + L_{K_0}^2)^{-1} L_{K_0}^{r-\frac{1}{2}} \right\| \leq \left\| (\lambda I + L_{K_0}^2)^{-1} L_{K_0}^2 \right\| \left\| L_{K_0}^{r-\frac{5}{2}} \right\| \leq \kappa_0^{2r-5}.$$

if  $r > \frac{5}{2}$ .  $\square$

Next we turn to the estimate of the sample error. It is easy to check the fact

$$\lambda f_\lambda = L_{\tilde{K}}(f_\rho - f_\lambda).$$

Therefore we have the following expression

$$\begin{aligned} f_{\mathbf{z}} - f_\lambda &= (\lambda I + TST_*S)^{-1} [TST_*\mathbf{y} - (TST_*S + \lambda I) f_\lambda] \\ &= (\lambda I + TST_*S)^{-1} \left[ TS \left( \frac{1}{m} \sum_{i=1}^m (y_i - f_\lambda(x_i)) K(x_i, \cdot) \right) \right. \\ &\quad \left. - L_K L_K^*(f_\rho - f_\lambda) \right] \\ &= (\lambda I + TST_*S)^{-1} TSU + (\lambda I + TST_*S)^{-1} VW \end{aligned} \quad (5.1)$$

where

$$\begin{aligned} U &= \frac{1}{m} \sum_{i=1}^m (y_i - f_\lambda(x_i)) K(x_i, \cdot) - L_K^*(f_\rho - f_\lambda), \\ V &= (TS - L_K), \quad \text{and} \quad W = L_K^*(f_\rho - f_\lambda). \end{aligned}$$

Let  $\sigma_\rho^2 = \mathbb{E}[(y - f_\rho(x))^2]$  be the variance of the random variable  $y - f_\rho(x)$ , which is the minimal possible least square loss. The following lemma provides an estimate for  $\|U\|_1$ .

**Lemma 5.2.** *There holds*

$$\mathbb{E}\|U\|_1^2 \leq \frac{\kappa}{m} (\sigma_\rho^2 + \|f_\lambda - f_\rho\|_{L_{\tilde{\rho}_X}^2}).$$

**Proof.** Consider the  $\mathcal{H}_1$  valued random variable  $\xi(z) = (y - f_\lambda(x))K(x, \cdot)$  on  $Z$ . Since  $\mathbb{E}\xi = L_K^*(f_\rho - f_\lambda)$ , a direct computation gives

$$\begin{aligned} \mathbb{E} \left[ \left\| \frac{1}{m} \sum_{i=1}^m \xi(z_i) - L_K^*(f_\rho - f_\lambda) \right\|_1^2 \right] &= \mathbb{E} \left[ \frac{1}{m^2} \sum_{i=1}^m \|\xi(z_i)\|_1^2 \right] - \frac{1}{m} \|L_K^*(f_\rho - f_\lambda)\|_1^2 \\ &\leq \frac{1}{m} \mathbb{E} \left[ \|(y - f_\lambda(x))K(x, \cdot)\|_1^2 \right] \\ &= \frac{1}{m} \mathbb{E} \left[ (y - f_\lambda(x))^2 K_0(x, x) \right] \\ &\leq \frac{\kappa_0^2}{m} \int_Z (y - f_\lambda(x))^2 d\rho \\ &= \frac{\kappa^2}{m} \left( \int_Z (y - f_\rho(x))^2 d\rho \right. \\ &\quad \left. + \int_X (f_\rho(x) - f_\lambda(x))^2 d\rho_X \right) \\ &\leq \frac{\kappa^2}{m} \left( \sigma_\rho^2 + \|f_\rho - f_\lambda\|_{L_{\tilde{\rho}_X}^2}^2 \right). \end{aligned}$$

This proves the lemma.  $\square$

The following lemma estimates  $W$ .

**Lemma 5.3.** Under Assumption 2, we have

$$\|W\|_1 \leq C_3 \lambda^{\min\left(\frac{1+2r}{4}, 1\right)}$$

with  $C_3 = \|L_{K_0}^{-r} f_\rho\|_{L_{\rho_X}^2}$  if  $0 < r \leq \frac{3}{2}$  and  $C_3 = \kappa_0^{2r-3} \|L_{K_0}^{-r} f_\rho\|_{L_{\rho_X}^2}$  if  $r > \frac{3}{2}$ .

**Proof.** By the fact  $f_\lambda - f_\rho = -\lambda(\lambda I + L_K^2)^{-1} f_\rho$ , Theorem 2.3(iv), and Assumption 2,

$$\begin{aligned} \|L_K^*(f_\lambda - f_\rho)\|_1 &= \left\| L_{K_0}^{\frac{1}{2}} (f_\lambda - f_\rho) \right\|_{L_{\rho_X}^2} = \lambda \left\| (\lambda I + L_{K_0}^2)^{-1} L_{K_0}^{\frac{1}{2}+r} L_K^{-r} f_\rho \right\|_{L_{\rho_X}^2} \\ &\leq C_3 \lambda^{\min\left(\frac{1+2r}{4}, 1\right)} \end{aligned}$$

where we used the estimates

$$\left\| (\lambda I + L_{K_0}^2)^{-1} L_{K_0}^{\frac{1}{2}+r} \right\| \leq \lambda^{\frac{1+2r}{4}}$$

when  $r \leq \frac{3}{2}$  and

$$\left\| (\lambda I + L_{K_0}^2)^{-1} L_{K_0}^{\frac{1}{2}+r} \right\| \leq \|(\lambda I + L_{K_0}^2)^{-1} L_{K_0}^2\| \left\| L_{K_0}^{\frac{1}{2}+r-2} \right\| \leq \kappa_0^{2r-3}$$

when  $r > \frac{3}{2}$ .  $\square$

Next we can prove our convergence results.

## 5.2. Convergence in $\mathcal{H}_0$

Under our assumptions we can obtain the following convergence rates in  $\mathcal{H}_0$ .

**Theorem 5.4.** Under Assumptions 1 and 2, we have

(i) if  $\frac{1}{2} \leq r \leq \frac{5}{2}$ , choosing  $\lambda \sim m^{-\frac{2}{3+2r}}$ , then

$$\mathbb{E}[\|f_{\mathbf{Z}} - f_\rho\|_0] = O\left(m^{-\frac{r-\frac{1}{2}}{3+2r}}\right);$$

(ii) if  $r > \frac{5}{2}$ , choosing  $\lambda \sim m^{-\frac{1}{2}}$ , then

$$\mathbb{E}[\|f_{\mathbf{Z}} - f_\rho\|_0] = O\left(m^{-\frac{1}{4}}\right).$$

This theorem follows from the combination of the approximation error estimate in Theorem 5.1 and the following sample error bound.

**Theorem 5.5.** If  $\lambda \leq 1$  and  $r \geq \frac{1}{2}$  we have

$$\mathbb{E}[\|f_{\mathbf{Z}} - f_\lambda\|_0] \leq C_4 \lambda^{-1} m^{-\frac{1}{2}}$$

where  $C_4 = 2\kappa^3(\sigma + C_1) + (1 + \kappa^2)C_3$ .

**Proof.** By (5.1),

$$\|f_{\mathbf{z}} - f_{\lambda}\|_0 \leq \|(\lambda I + TST_*S)^{-1}TS\| \|U\|_1 + \|(\lambda I + TST_*S)^{-1}\| \|V\|_{1,0} \|W\|_1.$$

Then the bound follows from Theorem 4.2, Lemma 5.2, Theorem 3.4 and Lemma 5.3.  $\square$

### 5.3. Convergence in $L^2_{\rho_X}$

For the convergence rate in  $L^2_{\rho_X}$  we have the following result.

**Theorem 5.6.** Under Assumptions 1 and 2,

(i) if  $0 < r \leq \frac{1}{2}$ , choosing  $\lambda \sim m^{-\frac{1}{2}}$ , then

$$\mathbb{E}[\|f_{\mathbf{z}} - f_{\rho}\|_{L^2_{\rho_X}}] = O\left(m^{-\frac{r}{4}}\right);$$

(ii) if  $\frac{1}{2} < r \leq 2$ , choosing  $\lambda \sim m^{-\frac{2}{3+2r}}$ , then

$$\mathbb{E}[\|f_{\mathbf{z}} - f_{\rho}\|_{L^2_{\rho_X}}] = O\left(m^{-\frac{r}{3+2r}}\right);$$

(iii) if  $r > 2$ , choosing  $\lambda \sim m^{-\frac{2}{7}}$ , then

$$\mathbb{E}[\|f_{\mathbf{z}} - f_{\rho}\|_{L^2_{\rho_X}}] = O\left(m^{-\frac{2}{7}}\right).$$

Under the same assumption on  $f_{\rho}$ , the learning rate obtained in [19] is  $O(m^{-\frac{r}{6+2r}})$  for  $r \leq 2$  and  $O(m^{-\frac{1}{5}})$  for  $r \geq 2$ . The rates in Theorem 5.6 are clearly much faster in all cases.

Theorem 5.6 can be obtained by optimizing the total error bound which is the combination of the approximation error estimates in Theorem 5.1 and the following sample error bound.

**Theorem 5.7.** We have

$$\mathbb{E}[\|f_{\mathbf{z}} - f_{\lambda}\|_{L^2_{\rho_X}}] \leq C_5 \left( \lambda^{-\frac{7}{4}} m^{-1} + \lambda^{-\max(2-\frac{r}{2}, \frac{5}{4})} m^{-1} + \lambda^{-\frac{3}{4}} m^{-\frac{1}{2}} \right)$$

where  $C_5 = \max(2\kappa^5(\sigma_{\rho} + C_1), 2(1 + \kappa^2)\kappa^6 C_3, \kappa^4 + \kappa^2 C_3)$ .

**Proof.** We use the fact  $\|f_{\mathbf{z}} - f_{\lambda}\|_{L^2_{\rho_X}} = \|L_{K_0}^{\frac{1}{2}}(f_{\mathbf{z}} - f_{\lambda})\|_0$ .

Write

$$\begin{aligned} L_{K_0}^{\frac{1}{2}}(\lambda I + TST_*S)^{-1} &= L_{K_0}^{\frac{1}{2}} \left[ (\lambda I + TST_*S)^{-1} - (\lambda I + L_{\tilde{K}})^{-1} + (\lambda I + L_{\tilde{K}})^{-1} \right] \\ &= L_{K_0}^{\frac{1}{2}} (\lambda I + L_{\tilde{K}})^{-1} \left[ L_{\tilde{K}} - TST_*S \right] (\lambda I + TST_*S)^{-1} \\ &\quad + L_{K_0}^{\frac{1}{2}} (\lambda I + L_{\tilde{K}})^{-1} \\ &= Q_1 + Q_2. \end{aligned}$$

By  $\|L_{K_0}(\lambda I + L_{\tilde{K}})^{-1}\| \leq \lambda^{-\frac{3}{4}}$  and Theorem 4.2 we obtain

$$\|Q_1\| \leq (1 + \kappa^2) \lambda^{-\frac{9}{4}} \|TST_*S - L_{\tilde{K}}\|,$$



$$\|Q_1 T S\| \leq 2\kappa^2 \lambda^{-\frac{7}{4}} \|T S T_* S - L_{\tilde{K}}\|,$$

$$\|Q_2\| \leq \lambda^{-\frac{3}{4}},$$

$$\|Q_2 T S\| \leq \kappa^2 \lambda^{-\frac{3}{4}}.$$

By (5.1) we get

$$\begin{aligned} \|f_{\mathbf{z}} - f_{\lambda}\|_{L^2_{\rho_X}} &\leq 2\kappa^2 \lambda^{-\frac{7}{4}} \|T S T_* S - L_{\tilde{K}}\| \|U\|_1 + (1 + \kappa^2) \lambda^{-\frac{9}{4}} \\ &\quad \times \|T S T_* S - L_{\tilde{K}}\| \|V\|_{1,0} \|W\|_1 \\ &\quad + \kappa^2 \lambda^{-\frac{3}{4}} \|U\|_1 + \lambda^{-\frac{3}{4}} \|V\|_{1,0} \|W\|_1. \end{aligned}$$

Taking expectation, using Theorems 3.5 and 3.4, Lemmas 5.2 and 5.3, and applying Schwartz inequality, we obtain the desired bound.  $\square$

## 6. Discussions

In this paper a novel approach is presented to study the asymptotic properties of least square kernel networks with indefinite kernels and coefficient regularization. The first step is to use the singular value decomposition of the kernel integral to characterize the indefinite kernel function. Two associated reproducing kernel Hilbert spaces  $\mathcal{H}_0$  and  $\mathcal{H}_1$  are then defined. The solution to the learning algorithm turns out to belong to  $\mathcal{H}_0$ . An operator expression of the solution is proved to be mathematically strict. By the aid of these properties the capacity independent error bounds and convergence rates are obtained both in  $\mathcal{H}_0$  and  $L^2_{\rho_X}$ . The results show that learning with indefinite kernels are consistent and could converge very fast. The consistency and rate analysis provide us confidence to apply indefinite kernels in situations where the indefinite kernels are inevitable or no good positive kernel is available.

One may notice that, although the approach in this paper greatly improved the error analysis, the rate is still worse than that of learning with definite kernel which, with the kernel  $K_0$  and under the same assumptions, is  $O(m^{-\frac{r}{1+2r}})$  for  $r \leq 1$  and  $O(m^{-\frac{1}{3}})$  for  $r \geq 1$ . This shows that indefiniteness does result in difficulties. A typical one is the non-positivity of the operator  $T S T_* S$  which prevents the sharper bound  $\frac{1}{\lambda}$  for the inverse operator  $(\lambda I + T S T_* S)^{-1}$  and the use of operator monotone inequality [18]. However, we should point out that theoretical analysis only gives results for the worst situations and does not provide any useful guidance on the comparison of empirical effectiveness. In practice indefinite kernels do not show worse performance than positive definite kernels.

Nevertheless, it seems still reasonable to prefer positive definite kernels because of their good learning performance as well as optimization and geometrical advantages. Our analysis on learning with indefinite kernels also seems to support this preference since, if an indefinite kernel  $K$  works well, learning with the positive definite kernel  $K_0$  will be even better. This, however, we think is only partially true. A good positive kernel exists theoretically does not mean it is constructible. Facing a real problem, the positive definite kernels are preferable only when they are practically constructible. We should feel free to turn our eyes to indefinite kernels when positive definite kernels fail to provide good learning performance.

## Acknowledgments

The author thanks the referees for their valuable comments. This work is partially supported by NNSF of China (No. 11101403).

## References

- [1] N. Aronszajn, Theory of reproducing kernels, *Transactions of the American Mathematical Society* 68 (1950) 337–404.
- [2] F. Cucker, S. Smale, On the mathematical foundations of learning, *Bulletin of the American Mathematical Society* 39 (2001) 1–49.
- [3] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, *Advances in Computational Mathematics* 13 (2000) 1–50.
- [4] T. Graepel, R. Herbrich, P. Bollmann-Sdorra, K. Obermayer, Classification on pairwise proximity data, in: *Advances in Neural Information Processing Systems 11: Proceedings of the 1998 Conference*, The MIT Press, 1999, p. 438.
- [5] B. Haasdonk, Feature space interpretation of svms with indefinite kernels, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (4) (2005) 482–492.
- [6] H. Lin, C. Lin, A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods, Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, 2003.
- [7] C. Liu, Gabor-based kernel PCA with fractional power polynomial models for face recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26 (2004) 572–581.
- [8] R. Luss, A. d'Aspremont, Support vector machine classification with indefinite kernels, *Mathematical Programming Computation* 1 (2009) 97–118.
- [9] C. Ong, X. Mary, S. Canu, A.J. Smola, Learning with non-positive kernels, in: *Proceedings of the 21st International Conference on Machine Learning*, 2004.
- [10] E. Pekalska, B. Haasdonk, Kernel discriminant analysis for positive definite and indefinite kernels, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2008) 1017–1032.
- [11] E. Pekalska, P. Paclik, R. Duin, A generalized kernel approach to dissimilarity-based classification, *The Journal of Machine Learning Research* 2 (2002) 175–211.
- [12] W. Rudin, *Functional Analysis*, McGraw-Hill, Inc., 1991.
- [13] H. Saigo, J. Vert, N. Ueda, T. Akutsu, Protein homology detection using string alignment kernels, *Bioinformatics* 20 (2004) 1682–1689.
- [14] B. Scholkopf, A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*, MIT Press, 2002.
- [15] L. Shi, Y. Feng, D. Zhou, Concentration estimates for learning with  $\ell_1$ -regularizer and data dependent hypothesis spaces, *Applied and Computational Harmonic Analysis* 31 (2011) 286–302.
- [16] S. Smale, D.X. Zhou, Learning theory estimates via integral operators and their approximations, *Constructive Approximation* 26 (2007) 153–172.
- [17] A. Smola, Z. Ovari, R. Williamson, Regularization with dot-product kernels, in: *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*, The MIT Press, 2001, p. 308.
- [18] S. Sun, Q. Wu, A note on application of integral operator in learning theory, *Applied and Computational Harmonic Analysis* 26 (2009) 416–421.
- [19] H. Sun, Q. Wu, Least square regression with indefinite kernels and coefficient regularization, *Applied and Computational Harmonic Analysis* 30 (2011) 96–109.
- [20] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.
- [21] H. Wang, Q. Xiao, D. Zhou, Learning with  $\ell_1$ -regularization for regression, Preprint.
- [22] Q. Wu, Classification and Regularization in Learning Theory, VDM Verlag, 2009.
- [23] Q. Wu, D.-X. Zhou, Learning with sample dependent hypothesis spaces, *Computers & Mathematics with Applications* 56 (2008) 2896–2907.
- [24] Q. Xiao, D. Zhou, Learning by nonsymmetric kernel with data dependent spaces and  $\ell_1$ -regularizer, *Taiwanese Journal of Mathematics* 14 (2010) 1821–1836.
- [25] Y. Ying, C. Campbell, M. Girolami, Analysis of SVM with indefinite kernels, in: *Advances in Neural Information Processing Systems, NIPS*, 2009.