Machine Learning (2021) 110:1145–1173
https://doi.org/10.1007/s10994-021-05955-2

Analysis of regularized least-squares in reproducing kernel Kreĭn spaces

Fanghui Liu[1] · Lei Shi[2] · Xiaolin Huang[3] · Jie Yang[3] · Johan A. K. Suykens[1]

Received: 9 June 2020 / Revised: 3 December 2020 / Accepted: 6 February 2021 / Published online: 19 April 2021

Abstract

In this paper, we study the asymptotic properties of regularized least squares with indefinite kernels in reproducing kernel Kreĭn spaces (RKKS). By introducing a bounded hypersphere constraint to such non-convex regularized risk minimization problem, we theoretically demonstrate that this problem has a globally optimal solution with a closed form on the sphere, which makes approximation analysis feasible in RKKS. Regarding to the original regularizer induced by the indefinite inner product, we modify traditional error decomposition techniques, prove convergence results for the introduced hypothesis error based on matrix perturbation theory, and derive learning rates of such regularized regression problem in RKKS. Under some conditions, the derived learning rates in RKKS are the same as that in reproducing kernel Hilbert spaces (RKHS). To the best of our knowledge, this is the first work on approximation analysis of regularized learning algorithms in RKKS.

Keywords Approximation analysis · Regularized least squares · Indefinite kernel · Matrix perturbation theory

Fanghui Liu and Lei Shi equally contribute to this work.

Editor: Thomas Gärtner.

✉ Fanghui Liu
fanghui.liu@esat.kuleuven.be

Lei Shi
leishi@fudan.edu.cn

Xiaolin Huang
xiaolinhuang@sjtu.edu.cn

Jie Yang
jieyang@sjtu.edu.cn

Johan A. K. Suykens
johan.suykens@esat.kuleuven.be

[1] Department of Electrical Engineering (ESAT-STADIUS), KU Leuven, Leuven, Belgium

[2] Shanghai Key Laboratory for Contemporary Applied Mathematics, and also with School of Mathematical Sciences, Fudan University, Shanghai, People's Republic of China

[3] Institute of Image Processing and Pattern Recognition, and also with Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, People's Republic of China

1 Introduction

Kernel methods (Schölkopf and Smola 2003; Suykens et al. 2002; Liu et al. 2020) have demonstrated success in statistical learning, such as classification (Zhu and Hastie 2002; Shang et al. 2019), regression (Shi et al. 2019; Farooq and Steinwart 2019), and clustering (Dhillon et al. 2004; Terada and Yamamoto 2019; Liu et al. 2020). The key ingredient of kernel methods is a kernel function, that is positive definite (PD) and can be associated with the inner product of two vectors in a reproducing kernel Hilbert space (RKHS). Nevertheless, in real-world applications, the used kernels might be *indefinite* (real, symmetric, but not positive definite) (Ying et al. 2009; Loosli et al. 2016; Oglic and Gärtner 2019) due to *intrinsic* and *extrinsic* factors. Here, *intrinsic* means that we often meet some indefinite kernels by specific domain metrics such as *tanh* kernel (Smola et al. 2001), *TL1* kernel (Huang et al. 2018), *log* kernel (Boughorbel et al. 2005), and hyperbolic kernel (Cho et al. 2019). Meanwhile, *extrinsic* indicates that some positive definite kernels degenerate to indefinite ones in some cases. An intuitive example is that a linear combination of PD kernels (with negative coefficient) (Ong et al. 2005) is an indefinite kernel. Polynomial kernels on the unit sphere are not always PD (Pennington et al. 2015). In manifold learning, the Gaussian kernel with some geodesic distances would lead to be an indefinite one. In neural networks, the sigmoid kernel with various values of hyper-parameters are mostly indefinite (Ong et al. 2004). We refer to a survey (Schleif and Tino 2015) for details.

Efforts on indefinite kernels are often based on a reproducing kernel Kreĭn space (RKKS) (Ong et al. 2004; Loosli et al. 2016; Alabdulmohsin et al. 2016; Saha and Palaniappan 2020) which is endowed by the indefinite inner product. The (reproducing) indefinite kernel associated with RKKS can be decomposed as the difference between two PD kernels, a.k.a, *positive decomposition* (Bognár 1974). The related optimization problem is often non-convex due to the non-positive definiteness of the used indefinite kernel. Since the indefinite inner product in RKKS does not define a norm, most previous works on RKKS (Ong et al. 2004; Loosli et al. 2016; Saha and Palaniappan 2020) focus on *stabilization* instead of risk minimization in RKHS. Here *stabilization* aims to finding a stationary point (more precisely, saddle points) instead of a minimum. Interestingly, *stabilization* in RKKS and minimization in RKHS can be linked together in a projection view (Ando 2009). In this sense, indefinite inner product in RKKS in a projection view can be still served as a valid regularization mechanism (Loosli et al. 2016). It is worth noting that, recently, Oglic and Gärtner (2018, 2019) directly consider empirical risk minimization in RKKS restricted in a hyper-sphere, which is demonstrated to generalize well.

In learning theory, the asymptotic behavior of these regularized indefinite kernel learning based algorithms in RKKS has not been fully investigated in an approximation theory view. Current literature (Wu et al. 2006; Steinwart et al. 2009; Lin et al. 2017; Jun et al. 2019) on approximation analysis often focus on regularized methods in RKHS, but their results could not be directly applied to that in RKKS due to the following two reasons. First, approximation analysis in RKHS often requires a (globally) optimal solution yielded by learning algorithms. While most indefinite kernel based methods via stabilization in RKKS seek for saddle points instead of a minimum. In this case, traditional concentration estimates could be invalid to that in RKKS. Second, in RKKS, the regularizer endowed by the indefinite inner product might be negative, which would fail to quantify complexity of a hypothesis. The classical error decomposition technique (Cucker and Zhou 2007; Lin et al. 2017) might be infeasible to our setting in RKKS.

To overcome the mentioned essential problems, in this paper, we study learning rates of least squares regularized regression in RKKS. Motivated by Oglic and Gärtner (2018), we focus on a typical empirical risk minimization in RKKS, i.e., indefinite kernel ridge regression in a hyper-sphere region endowed by the indefinite inner product. For this purpose, we provide a detailed error analysis and then derive learning rates. To be specific, in algorithm, we demonstrate that, the solution to our considered kernel ridge regression model in RKKS with a spherical constraint can be achieved on the hyper-sphere. Subsequently, albeit non-convex, this model admits a global minimum with a closed form as demonstrated by Oglic and Gärtner (2018).

We start the analysis from the regularized algorithm that has an analytical solution and obtain the first-step to understand the learning behavior in RKKS. In theory, we modify the traditional error decomposition approach, and thus the excess error can be bounded by the sample error, the regularization error, and the additional hypothesis error. We provide estimates for the introduced hypothesis error based on matrix perturbation theory for non-Hermitian and non-diagonalizable matrices and then derive convergence rates of such model. Our analysis is able to bridge the gap between the least squares regularized regression problem in RKHS and RKKS. Under some conditions, the derived learning rates in RKKS is the same as that in RKHS (the best case). To the best of our knowledge, this is the first work to study learning rates of regularized risk minimization in RKKS.

The rest of the paper is organized as follows. In Sect. 2, we briefly introduce the basic concepts of Kreĭn spaces and RKKS. Section 3 presents the problem setting and main results under some fair assumptions. In Sect. 4, we present the least squares regularized regression model in RKKS and give a globally optimal solution to aid the proof. In Sect. 5, we give the framework of convergence analysis for the modified error decomposition technique, detail the estimates for the introduced hypothesis error, and derive the learning rates. In Sect. 6, we report numerical experiments to demonstrate our theoretical results and the conclusion is drawn in Sect. 7.

2 Preliminaries

In this section, we briefly introduce the definitions and basic properties of Kreĭn spaces and the reproducing kernel Kreĭn space (RKKS) that we shall need later. Detailed expositions can be found in the book by Bognár (1974).

We begin with a vector space $\mathcal{H}_\mathcal{K}$ defined on the scalar field $\mathbb{R}$.

Definition 1 (Inner product space) An inner product space is a vector space $\mathcal{H}_\mathcal{K}$ defined on the scalar field $\mathbb{R}$ together with a bilinear form $\langle \cdot, \cdot \rangle_{\mathcal{H}_\mathcal{K}}$ called inner product that satisfies the following conditions

i) symmetry: $\forall f, g \in \mathcal{H}_\mathcal{K}$, we have $\langle f, g \rangle_{\mathcal{H}_\mathcal{K}} = \langle g, f \rangle_{\mathcal{H}_\mathcal{K}}$.

ii) linearity: $\forall f, g, h \in \mathcal{H}_\mathcal{K}$ and two scalars $a, b \in \mathbb{R}$, we have $\langle af + bg, h \rangle_{\mathcal{H}_\mathcal{K}} = a\langle f, h \rangle_{\mathcal{H}_\mathcal{K}} + b\langle g, h \rangle_{\mathcal{H}_\mathcal{K}}$.

iii) non-degenerate: for $f \in \mathcal{H}_\mathcal{K}$, if $\langle f, g \rangle_{\mathcal{H}_\mathcal{K}} = 0$ for all $g \in \mathcal{H}_\mathcal{K}$ implies that $f = 0$.

If $\langle f, f \rangle_{\mathcal{H}_\mathcal{K}} > 0$ holds for any $f \in \mathcal{H}_\mathcal{K}$ with $f \neq 0$, then the inner product on $\mathcal{H}_\mathcal{K}$ is *positive*. If there exists $f, g \in \mathcal{H}_\mathcal{K}$ such that $\langle f, f \rangle_{\mathcal{H}_\mathcal{K}} > 0$ and $\langle g, g \rangle_{\mathcal{H}_\mathcal{K}} < 0$, then the inner product is called *indefinite*, and $\mathcal{H}_\mathcal{K}$ is an indefinite inner product space. Recall that Hilbert spaces satisfy the above conditions and admit the positive inner product. After reviewing the indefinite inner product, we are ready to introduce the definition of Kreĭn space.

Definition 2 (Kreĭn space, Bognár 1974) The vector space $\mathcal{H}_\mathcal{K}$ with the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_\mathcal{K}}$ is a Kreĭn space if there exist two Hilbert spaces $\mathcal{H}_+$ and $\mathcal{H}_-$ such that

i) the vector space $\mathcal{H}_\mathcal{K}$ admits a direct orthogonal sum decomposition $\mathcal{H}_\mathcal{K} = \mathcal{H}_+ \oplus \mathcal{H}_-$.
ii) all $f \in \mathcal{H}_\mathcal{K}$ can be decomposed into $f = f_+ + f_-$, where $f_+ \in \mathcal{H}_+$ and $f_- \in \mathcal{H}_-$, respectively.
iii) $\forall f, g \in \mathcal{H}_\mathcal{K}, \langle f, g\rangle_{\mathcal{H}_\mathcal{K}} = \langle f_+, g_+\rangle_{\mathcal{H}_+} - \langle f_-, g_-\rangle_{\mathcal{H}_-}$.

From the definition, the decomposition $\mathcal{H}_\mathcal{K} = \mathcal{H}_+ \oplus \mathcal{H}_-$ is not necessarily unique. For a fixed decomposition, the inner product $\langle f, g\rangle_{\mathcal{H}_\mathcal{K}}$ is given accordingly (Loosli et al. 2016; Oglic and Gärtner 2018). Kreĭn spaces are indefinite inner product spaces endowed with a Hilbertian topology. The key difference with Hilbert spaces is that the inner products might be negative for Kreĭn spaces, i.e., there exists $f \in \mathcal{H}_\mathcal{K}$ such that $\langle f, f\rangle_{\mathcal{H}_\mathcal{K}} < 0$. If $\mathcal{H}_+$ and $\mathcal{H}_-$ are two RKHSs, the Kreĭn space $\mathcal{H}_\mathcal{K}$ is a RKKS associated with a unique indefinite reproducing kernel k such that the reproducing property holds, i.e., $\forall f \in \mathcal{H}_\mathcal{K}$, $f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}_\mathcal{K}}$.

Proposition 1 (Positive decomposition, Bognár 1974) *An indefinite kernel k associated with a RKKS admits a positive decomposition $k = k_+ - k_-$, with two positive definite kernels k_+ and k_-.*

Typical examples include a wide range of commonly used indefinite kernels, such as a linear combination of PD kernels (with negative coefficients) (Ong et al. 2005; Oglic and Gärtner 2018), and conditionally PD kernels (Schaback 1999; Wendland 2004). It is important to note that, not every indefinite kernel function admits such positive decomposition as a difference between two positive definite kernels. Nevertheless, this can be conducted on finite discrete spaces, e.g., eigenvalue decomposition of indefinite kernel matrices. In fact, for any given an indefinite kernel, whether it can be associated with RKKS still remains a long-lasting open question. For example, the hyperbolic kernel (Cho et al. 2019) is based on the hyperboloid model (Sala et al. 2018), in which the distance between two point is defined as the length of the geodesic path on the hyperboloid that connects the two points. Although the used hyperboloid space stems from a finite-dimensional Kreĭn space, it is unclear whether the derived hyperbolic kernel is associated with RKKS or not. Besides, the existence of positive decomposition for the *TL1* kernel (Huang et al. 2018), defined by the truncated ℓ_1 distance, is also unknown. Our results in this paper are based on RKKS, and can be applied to these kernels if they can be associated with RKKS.

Definition 3 (Associated RKHS of RKKS (Ong et al. 2004)) Let $\mathcal{H}_\mathcal{K}$ be a RKKS with the direct orthogonal sum decomposition into two RKHSs $\mathcal{H}_+$ and $\mathcal{H}_-$. Then the associated RKHS $\mathcal{H}_{\bar{\mathcal{K}}}$ endowed by $\mathcal{H}_\mathcal{K}$ is defined with the positive inner product

$$\langle f, g\rangle_{\mathcal{H}_{\bar{\mathcal{K}}}} = \langle f_+, g_+\rangle_{\mathcal{H}_+} + \langle f_-, g_-\rangle_{\mathcal{H}_-}, \quad \forall f, g \in \mathcal{H}_\mathcal{K}.$$

Note that $\mathcal{H}_{\bar{\mathcal{K}}}$ is the smallest Hilbert space majorizing the RKKS $\mathcal{H}_{\mathcal{K}}$ with $|\langle f,f\rangle_{\mathcal{H}_{\mathcal{K}}}| \leq \|f\|^2_{\mathcal{H}_{\bar{\mathcal{K}}}} = \|f_+\|^2_{\mathcal{H}_+} + \|f_-\|^2_{\mathcal{H}_-}$. Denote $C(X)$ as the space of continuous functions on X with the norm $\|\cdot\|_\infty$, and suppose that $\kappa := \sqrt{2}\sup_{\boldsymbol{x}\in X}\sqrt{k_+(\boldsymbol{x},\boldsymbol{x}) + k_-(\boldsymbol{x},\boldsymbol{x}')} < \infty$. The reproducing property in RKKS indicates that $\forall f \in \mathcal{H}_{\mathcal{K}}$, we have

$$\|f\|_\infty = \sup_{\boldsymbol{x}\in X}\left|\left\langle f, k(\boldsymbol{x},\cdot)\right\rangle\right| \leq \kappa\|f\|_{\mathcal{H}_{\bar{\mathcal{K}}}}. \tag{1}$$

Definition 4 (The empirical covariance operator in RKKS (Pękalska and Haasdonk 2009)) Let k be an indefinite kernel associated with a RKKS $\mathcal{H}_{\mathcal{K}}$, $\psi : X \to \mathcal{H}_{\mathcal{K}}$ be a mapping of the data in $\mathcal{H}_{\mathcal{K}}$ and $\boldsymbol{\Psi} = [\psi(\boldsymbol{x}_1), \psi(\boldsymbol{x}_2), \ldots, \psi(\boldsymbol{x}_m)]$ be a sequence of images of the training data in $\mathcal{H}_{\mathcal{K}}$, then its empirical non-centered covariance operator $T : \mathcal{H}_{\mathcal{K}} \to \mathcal{H}_{\mathcal{K}}$ is defined by

$$T = \frac{1}{m}\boldsymbol{\Psi}\boldsymbol{\Psi}^*, \tag{2}$$

which is not positive definite in the Hilbert sense, but it is in the Kreĭn sense satisfying $\langle\zeta, T\zeta\rangle_{\mathcal{H}_{\mathcal{K}}} \geq 0$ for $\zeta \neq 0$.

The operator T actually depends on the sample set and can be linked to an empirical kernel (Guo and Shi 2019). In our paper, we choose the mapping $\psi(\boldsymbol{x}) := k(\boldsymbol{x},\cdot)$ to obtain the empirical covariance operator T. Since $\langle f, Tf\rangle_{\mathcal{H}_{\mathcal{K}}}$ is nonnegative, we use it as a regularizer to aid our proof.

3 Problem setting and main results

In this section, we introduce our problem setting and present our results under some fair assumptions.

3.1 Problem setting

Let X be a compact metric space and $Y \subseteq \mathbb{R}$, we assume that a sample set $z = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m \in Z^m$ is drawn from a non-degenerate Borel probability measure ρ on $X \times Y$. In the context of statistical learning theory, the *target function* of ρ is defined by $f_\rho(\boldsymbol{x}) = \int_Y y\mathrm{d}\rho(y|\boldsymbol{x}), \boldsymbol{x} \in X$, where $\rho(\cdot|\boldsymbol{x})$ is the conditional distribution of ρ at $\boldsymbol{x} \in X$. The indefinite kernel function $k : X \times X \to \mathbb{R}$ is endowed by the RKKS $\mathcal{H}_{\mathcal{K}}$. The associated indefinite kernel matrix is given by $\boldsymbol{K} = [k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^m$ on the sample set. The goal of a supervised learning task in RKKS endowed by k is to find a hypothesis $f : X \to Y$ such that $f(\boldsymbol{x})$ is a good approximation of the label $y \in Y$ corresponding to a new instance $\boldsymbol{x} \in X$

.

Motivated by Oglic and Gärtner (2018), we consider the least squares regularized regression problem in a bounded region induced by the original regularization mechanism of RKKS

$$f_{z,\lambda} := \operatorname*{argmin}_{f\in\mathcal{B}(r)}\left\{\frac{1}{m}\sum_{i=1}^m \left(f(\boldsymbol{x}_i) - y_i\right)^2 + \lambda\langle f,f\rangle_{\mathcal{H}_{\mathcal{K}}}\right\}, \tag{3}$$

where $\mathcal{B}(r) := \left\{ f \in \operatorname{span}\{k(\boldsymbol{x}_1,\cdot), k(\boldsymbol{x}_2,\cdot),\dots,k(\boldsymbol{x}_m,\cdot)\} : \frac{1}{m}\sum_{i=1}^m \big(f(\boldsymbol{x}_i)\big)^2 \le r^2 \right\}$ is assumed to be spanned by the training data $\{\boldsymbol{x}_i\}_{i=1}^m$ in $\mathcal{H}_\mathcal{K}$ in a bounded hyper-sphere. This setting can be also used in Ong et al. (2004). Here we employ the original regularization mechanism of RKKS, which aims to understand the learning behavior in RKKS and avoid the inconsistency when using various regularizers spanned by different spaces. Our result in fact can be applied to other settings with different regularizers. Following Oglic and Gärtner (2018), we consider a risk minimization problem in a hyper-sphere instead of stabilization. The considered hyper-spherical constraint in RKKS is able to prohibit the objective function value in Problem (3) approaches to infinity, avoiding a meaningless solution. The radius r can be chosen by cross validation or hyper-parameter optimization (Oglic and Gärtner 2018) in practice and is naturally needed and common in classical approximation analysis in RKHS (Wu et al. 2006; Cucker and Zhou 2007; Steinwart et al. 2009). Such risk minimization problem still preserves the specifics of learning in RKKS, i.e., there exists some points $f \in \mathcal{B}(r)$ such that $\langle f, f\rangle_{\mathcal{H}_\mathcal{K}} < 0$, and generalizes well when compared to stabilization, as indicated by Oglic and Gärtner (2018). One main reason why we consider the risk minimization problem is that, stabilization in RKKS does not necessarily have a unique saddle point, which makes approximation analysis infeasible to define the concentration of certain empirical hypotheses around some target hypothesis. Conversely, the studied risk minimization in a hyper-sphere, problem 3, leads to a (globally) optimal solution. This nice result motivates us to obtain the first-step to understand the learning behavior in RKKS.

3.2 Main results

In this section, we state and discuss our main results. To illustrate our analysis, we need the following notations and assumptions.

In the least squares regression problem, the expected (quadratic) risk is defined as $\mathcal{E}(f) = \int_Z (f(\boldsymbol{x}) - y)^2 \mathrm{d}\rho$. The empirical risk functional is defined on the sample z, i.e., $\mathcal{E}_z(f) = \frac{1}{m}\sum_{i=1}^m \big(f(\boldsymbol{x}_i) - y_i\big)^2$. To measure the estimation quality of $f_{z,\lambda}$, one natural way in approximation theory is the *excess risk*: $\mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\rho)$.

Assumption 1 (Existence and boundedness of f_ρ) we assume that the *target function* $f_\rho \in \mathcal{H}_\mathcal{K}$ exists and bounded. There exits a constant $M^* \ge 1$, such that

$$|f_\rho| \le M^* \text{ for almost } \boldsymbol{x} \in X \text{ with respect to } \rho_X.$$

Remark This is a standard assumption in approximation analysis (Cucker and Zhou 2007; Lin et al. 2017; Rudi and Rosasco 2017). Here we remark that existence of f_ρ implies a bounded hyper-sphere region is needed, e.g., the used radius r in problem (3). In fact, the existence of f_ρ is not ensured if we consider a potentially infinite dimensional RKKS $\mathcal{H}_\mathcal{K}$, possibly universal (Steinwart and Andreas 2008). Instead, in approximation analysis, the infinite dimensional RKKS is substituted by a finite one, i.e., $\mathcal{H}_\mathcal{K}^r = \{f \in \mathcal{H}_\mathcal{K} : \|f\| \le r\}$ with r fixed a priori, where the norm $\|f\|$ is defined in some associated Hilbert spaces, e.g., $\mathcal{H}_{\bar{\mathcal{K}}}$ or using the non-negative inner product $\langle f, Tf\rangle_{\mathcal{H}_\mathcal{K}}$ by the empirical covariance operator. In this case, a minimizer of risk $\mathcal{E}$ always exists but r is fixed with a prior and $\mathcal{H}_\mathcal{K}^r$ cannot be universal. As a result, assuming the existence of f_ρ implies that f_ρ belongs to a ball of radius $r_{\rho,\mathcal{H}_\mathcal{K}}$. So this is the reason why the spherical constraint is indeed taken into account in approximation analysis.

For a tighter bound, we need the following *projection operator*.

Definition 5 (Projection operator (Chen et al. 2004)) For $B > 0$, the projection operator $\pi := \pi_B$ is defined on the space of measurable functions $f : X \to \mathbb{R}$ as

$$\pi_B(f)(\boldsymbol{x}) = \begin{cases} B, & \text{if } f(\boldsymbol{x}) > B; \\ -B, & \text{if } f(\boldsymbol{x}) < -B; \\ f(\boldsymbol{x}), & \text{if } -B \le f(\boldsymbol{x}) \le B, \end{cases}$$

and then the projection of *f* is denoted as $\pi_B(f)(\boldsymbol{x}) = \pi_B(f(\boldsymbol{x}))$.

The projection operator is beneficial to the $\|\cdot\|_\infty$-bounds for sharp estimation. Besides, we consider the standard output assumption[1] $|y| \le M$, and then we have $\mathcal{E}_z\big(\pi_B(f_{z,\lambda})\big) \le \mathcal{E}_z\big(f_{z,\lambda}\big)$. So it is more accurate to estimate f_ρ by $\pi_{M^*}(f_{z,\lambda})$ instead of $f_{z,\lambda}$. Therefore, our approximation analysis attempts to bound the error $\|\pi_{M^*}(f_{z,\lambda}) - f_\rho\|^2_{L^{p^*}_{\rho_X}}$ in the space $L^{p^*}_{\rho_X}$ with some $p^* > 0$, where $L^{p^*}_{\rho_X}$ is a weighted L^{p^*}-space with the norm $\|f\|_{L^{p^*}_{\rho_X}} = \Big(\int_X |f(\boldsymbol{x})|^{p^*} d\rho_X(\boldsymbol{x})\Big)^{1/p^*}$. Specifically, in our analysis, the excess error is exactly the distance in $L^2_{\rho_X}$ due to the strong convexity of the squared loss.

To derive the learning rates, we need to consider the approximation ability of $\mathcal{H}_\mathcal{K}$ with respect to its capacity and f_ρ in $L^2_{\rho_X}$. Since the original regularizer $\langle f, f\rangle_{\mathcal{H}_\mathcal{K}}$ in RKKS fails to quantify complexity of a hypothesis, here we use the empirical regularizer $\langle f, Tf\rangle_{\mathcal{H}_\mathcal{K}}$ in Definition 4 as an alternative. Note that, other RKHS regularizers, such as $\langle f, f\rangle_{\mathcal{H}_{\bar{\mathcal{K}}}}$ in Definition 3, is also acceptable, but the used $\langle f, Tf\rangle_{\mathcal{H}_\mathcal{K}}$ will result in elegant and concise theoretical results. Accordingly, the approximation ability of $\mathcal{H}_\mathcal{K}$ can be characterised by the regularization error.

Assumption 2 Rregularity condition) The regularization error of $\mathcal{H}_\mathcal{K}$ is defined as

$$D(\lambda) = \inf_{f \in \mathcal{H}_\mathcal{K}} \Big\{ \mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda \langle f, Tf\rangle_{\mathcal{H}_\mathcal{K}} \Big\}. \tag{4}$$

The target function f_ρ can be approximated by $\mathcal{H}_\mathcal{K}$ with exponent $0 < \beta \le 1$ if there exists a constant C_0 such that

$$D(\lambda) \le C_0 \lambda^\beta, \quad \forall \lambda > 0. \tag{5}$$

Remark This is a natural assumption and approximation theory requires it, e.g., Wu et al. (2006); Wang and Zhou (2011); Steinwart and Andreas (2008). Note that $\beta = 1$ is the best choice as we expect, which is equivalent to $f_\rho \in \mathcal{H}_\mathcal{K}$ when $\mathcal{H}_\mathcal{K}$ is dense.

Furthermore, to quantitatively understand how the complexity of $\mathcal{H}_\mathcal{K}$ affects the learning ability of algorithm (3), we need the capacity (roughly speaking the “size”) of $\mathcal{H}_\mathcal{K}$ measured by covering numbers.

[1] For unbounded outputs, the moment hypothesis (Wang and Zhou 2011) is suitable but the introduced hypothesis error in our analysis depends on the standard output assumption.

Definition 6 (Covering numbers (Cucker and Zhou 2007; Shi et al. 2019)) For a subset Q of $C(X)$ and $\epsilon > 0$, the *covering number* $\mathscr{N}(Q, \epsilon)$ is the minimal integer $l \in \mathbb{N}$ such that there exist l disks with radius ϵ covering Q.

In this paper, the covering numbers of balls are defined by

$$\mathcal{B}_R = \{f \in \mathcal{H}_{\mathcal{K}} : \sqrt{\langle f, Tf \rangle_{\mathcal{H}_{\mathcal{K}}}} \leq R\}, \tag{6}$$

as subsets of $L^\infty(X)$. Note that the used R in Eq. (6) and r in problem (3) admits $R = Cr$ for some positive constant C, as the definition of such non-negative inner product leads to a hyper-sphere with different radius. Hence there is no difference of using R or r in our analysis and thus we directly use R for convenience.

Assumption 3 (Capacity) We assume that for some $s > 0$ and $C_s > 0$ such that

$$\log \mathscr{N}(\mathcal{B}_1, \epsilon) \leq C_s \left(\frac{1}{\epsilon}\right)^s, \quad \forall \epsilon > 0. \tag{7}$$

Remark This is a standard assumption to measure the capacity of $\mathcal{H}_{\mathcal{K}}$ that follows with that of RKHS (Cucker and Zhou 2007; Wang and Zhou 2011; Shi et al. 2019), When X is bounded in $\mathbb{R}^d$ and $k \in C^\tau(X \times X)$, Eq. (7) always holds true with $s = \frac{2d}{\tau}$. In particular, if $k \in C^\infty(X \times X)$, Eq. (7) is still valid for an arbitrary small $s > 0$.

It can be noticed that, the capacity of a RKHS can be also measured by eigenvalue decay of the PSD kernel matrix, which has been has been fully studied, e.g., Steinwart and Andreas (2008), Bach (2013). A small RKHS indicates a fast eigenvalue decay so as to obtain a promising prediction performance. In other words, functions in the RKHS are potentially smoother than what is necessary, which means an arbitrary small s in Assumption 3. Nevertheless, eigenvalue decay of the indefinite kernel matrix has not been studied before due to the extra negative eigenvalues. By virtue of eigenvalue decomposition $\boldsymbol{K} = \boldsymbol{K}_+ - \boldsymbol{K}_-$ with two PSD matrices $\boldsymbol{K}_\pm$, we can easily make the assumption for $\boldsymbol{K}$ based on the eigenvalue decay of $\boldsymbol{K}_\pm$.

Assumption 4 (Eigenvalue assumption for indefinite kernel matrices) Suppose that the indefinite kernel matrix $\boldsymbol{K} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ has p positive eigenvalues, q negative eigenvalues, and $m - p - q$ zero eigenvalues, i.e., $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_+ + \boldsymbol{\Sigma}_-$, where $\boldsymbol{\Sigma}_+ = \operatorname{diag}(\sigma_1, \sigma_2, \dots \sigma_p, 0, \dots, 0)$, $\boldsymbol{\Sigma}_- = \operatorname{diag}(0, \dots, 0, \sigma_{m-q+1}, \dots, \sigma_m)$ with the decreasing order $\sigma_1 \geq \dots \geq \sigma_p > 0 > \sigma_{m-q+1} \geq \cdots \geq \sigma_m$ and $\sigma_{p+1} = \sigma_{p+2} = \cdots = \sigma_{m-q} = 0$. Here we assume that its (positive) largest eigenvalue satisfies $\sigma_1 \geq c_1 m^{\eta_1}$ with $c_1 > 0$, $\eta_1 > 0$ and its smallest (negative) eigenvalue admits $\sigma_m \leq c_m m^{\eta_2}$ with $c_m < 0$, $\eta_2 > 0$. And we denote $\eta := \min\{\eta_1, \eta_2\}$.

Remark Our assumption only requires the lower bound of the largest (positive) eigenvalue and the upper bound of the smallest (negative) eigenvalue, which is weaker than the common decay of a PSD kernel matrix, e.g., polynomial/exponential decay. In particular, if we take these common eigenvalue decays of $\boldsymbol{K}_\pm$, then our assumption on σ_1 and σ_m is naturally satisfied. To be specific, Bach (2013) considers three eigenvalue decays of a PSD kernel matrix, including i) the exponential decay $\sigma_i \propto me^{-ci}$ with $c > 0$, ii) the polynomial decay $\sigma_i \propto mi^{-2t}$ with $t \geq 1$, and iii) the slowest decay with $\sigma_i \propto m/i$. Hence, under such three eigenvalue decays of $\boldsymbol{K}_{\pm}$, then our assumption on $\sigma_1 \geq c_1 m^{\eta_1}$ and $\sigma_m \leq c_m m^{\eta_2}$ always holds. Specifically, although the number of positive/negative eigenvalues depends on the sample set, our theoretical results will be independent of the unknown p and q.

Formally, our main result about least squares regularized regression in RKKS is stated as follows.

Theorem 1 *Suppose that $|f_\rho(x)| \leq M^*$ with $M^* \geq 1$ in Assumption 1, ρ satisfies the condition in Eq. (5) with $0 < \beta \leq 1$ in Assumption 2, the indefinite kernel matrix $\boldsymbol{K}$ satisfies the eigenvalue assumption in Assumption 4 with $\eta = \min\{\eta_1, \eta_2\} > 0$. Assume that for some $s > 0$ in Assumption 3, take $\lambda := m^{-\gamma}$ with $0 < \gamma \leq 1$. Let*

$$0 < \epsilon < \frac{1}{s} - (\gamma + s\gamma - 1)(2 + s). \tag{8}$$

Then for $0 < \delta < 1$ with confidence $1 - \delta$, when $\gamma + \eta > 1$, we have

$$\|\pi_{M^*}(f_{z,\lambda}) - f_\rho\|^2_{L^2_{\rho_X}} \leq \widetilde{C}\left(\log\frac{2}{\epsilon}\right)^2 \log\frac{2}{\delta} m^{-\Theta},$$

where $\widetilde{C}$ is a constant independent of m or δ and the power index Θ is

$$\Theta = \min\left\{\gamma\beta, \gamma + \eta - 1, \frac{2 - s\gamma(1-\beta)}{2(1+s)}, \frac{2 - s(1-\eta)}{2(1+s)}, \frac{1 - s(\gamma + s\gamma - 1)(2+s) - s\epsilon}{1+s}\right\}, \tag{9}$$

where η is further restricted by $\max\{0, 1 - 2/s\} < \eta < 1$ for a positive Θ, i.e., a valid learning rate.

We hence directly have the following corollary that corresponds to learning rates in RKHS.

Corollary 1 (Link to learning rates in RKHS) *If $\eta := \min\{\eta_1, \eta_2\} \geq 1$ in Assumption 4, the power index Θ in Eq. (9) can be simplified as*

$$\Theta = \min\left\{\gamma\beta, \frac{2 - s\gamma(1-\beta)}{2(1+s)}, \frac{1 - (s\gamma(1+s) - s)(2+s) - s\epsilon}{1+s}\right\}, \tag{10}$$

which is actually the learning rate for least squares regularized regression in RKHS, independent of η.

Remark We provide learning rates in RKKS in Theorem 1 and also demonstrate the relation of the derived learning rates between RKKS and RKHS in Corollary 1. We make the following remarks.

i) In Theorem 1, our results choose $\lambda := m^{-\gamma}$ and the radius R (or r) is implicit in Eq. (9). The estimation for R depends on a bound for $\lambda\langle f_{z,\lambda}, Tf_{z,\lambda}\rangle_{\mathcal{H}_{\mathcal{K}}}$, see Lemma 3 for details. Note that s can be arbitrarily small when the kernel k is $C^\infty(X \times X)$. In this case, Θ in Eq. (9) can be arbitrarily close to $\min(\gamma\beta, \gamma + \eta - 1)$.

Table 1 Comparisons of different least squares regression problems

Learning problem in RKKS	Learning rates
$f_{z,\lambda} := \operatorname*{argmin}_{f \in \mathcal{B}(r)} \left\{ \mathcal{E}_z(f) + \lambda \langle f, f \rangle_{\mathcal{H}_\mathcal{K}} \right\}$	Eq. (9)
$\overline{f_{z,\lambda}} := \operatorname*{argmin}_{f \in \mathcal{B}(r)} \left\{ \mathcal{E}_z(f) + \lambda \langle f, f \rangle_{\mathcal{H}_{\bar{\mathcal{K}}}} \right\}$	Corollary 2 (applied to Oglic and Gärtner (2018))
$\widetilde{f}_{z,\lambda} := \operatorname*{argmin}_{f \in \mathcal{B}(r)} \left\{ \mathcal{E}_z(f) + \lambda \langle f, Tf \rangle_{\mathcal{H}_\mathcal{K}} \right\}$	Corollary 2 (β is different)

ii) Corollary 1 derives the learning rates in RKHS, which recovers the result of Wang and Zhou (2011) for least squares in RKHS. That is, when choosing $\beta = 1$ and s is small enough, the derived learning rate in Corollary 1 can be arbitrary close to 1, and hence is optimal (Wang and Zhou 2011).

iii) Based on Theorem 1 and Corollary 1, we find that if $\eta := \min\{\eta_1, \eta_2\} \geq 1$, our analysis for RKKS is the same as that in RKHS. This is the best case. However, if $\eta < 1$, the derived learning rate in RKKS demonstrated by Eq. (9) is not faster than that in RKHS. This is reasonable since the spanning space of RKKS is larger than that of RKHS.

The proof of Theorem 1 is fairly technical and lengthy, and we briefly sketch some main ideas in the next section.

Furthermore, if problem (3) considers some nonnegative regularizers, such as $\|f\|^2_{\mathcal{H}_{\bar{\mathcal{K}}}}$ in Definition 3 and $\langle f, Tf \rangle_{\mathcal{H}_\mathcal{K}}$ in Definition 4, the analysis would be simplified due to the used nonnegative regularizer. To be specific, denote $\overline{f_{z,\lambda}} := \operatorname{argmin}_{f \in \mathcal{B}(r)} \left\{ \frac{1}{m} \sum_{i=1}^m \left(f(\boldsymbol{x}_i) - y_i \right)^2 + \lambda \|f\|^2_{\mathcal{H}_{\bar{\mathcal{K}}}} \right\}$ as demonstrated by Oglic and Gärtner (2018), its learning rate could be given by the following corollary.

Corollary 2 *Under the same assumption with Theorem* 1 (*without the eigenvalue assumption*), *by defining the regularization error as*

$$D'(\lambda) = \inf_{f \in \mathcal{H}_\mathcal{K}} \left\{ \mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda \|f\|^2_{\mathcal{H}_{\bar{\mathcal{K}}}} \right\},$$

satisfying $D'(\lambda) \leq C_0' \lambda^{\beta'}$ *with a constant* C_0' *and* $\beta' \in (0, 1]$, *we have*

$$\|\pi_{M^*}(\overline{f_{z,\lambda}}) - f_\rho\|^2_{L^2_{\rho_X}} \leq \widetilde{C}' \left(\log \frac{2}{\epsilon} \right)^2 \log \frac{2}{\delta} m^{-\Theta'},$$

where $\widetilde{C}'$ *is a constant independent of m or* δ *and the power index* Θ' *is defined as Eq*. (10) *with* β'.

Remark In fact, Corollary 2 gives the convergence rates of the model in Oglic and Gärtner (2018).

Note that the learning rates would be effected by different regularizers, as indicated by the regularization error in Assumption 2. In Table 1 we summarize the learning rates of problem (3) with different non-negative regularizers. Although the associated Hilbert space norms generated by different decomposition of the Krein space are topologically equivalent (Langer 1962), the derived learning rates cannot be ensured to be the same due to their respective spanning/solving spaces. Besides, Oglic and Gärtner (2019) demonstrate that, stabilization of support vector machine (SVM) in RKKS can be transformed to a risk minimization problem with a PSD kernel matrix by taking the absolute value of negative eigenvalues of the original indefinite one. That means, stabilization of SVM in RKKS could also achieve the same convergence behavior as risk minimization with a PSD kernel matrix in RKHS, e.g., Steinwart and Scovel (2007). Accordingly, the considered problem (3), i.e., risk minimization in RKKS with the original regularizer induced by the indefinite inner product is general. The obtained results in Theorem 1 provide the worst case, and can be improved to the same learning rates as other settings, e.g., least-squares in RKHS, minimization in RKKS but with non-negative regularizers.

4 Solution to regularized least-squares in RKKS

In this section, we study the optimization problem (3), obtain a globally optimal solution to aid our analysis.

By virtue of $f = \sum_{i=1}^{m} \alpha_i k(\boldsymbol{x}_i, \cdot)$, problem (3) can be formulated as

$$\boldsymbol{\alpha}_{z,\lambda} := \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^m:\,\boldsymbol{\alpha}^\top \boldsymbol{K}^2\boldsymbol{\alpha}\leq mr^2} \left\{ \frac{1}{m}\|\boldsymbol{K}\boldsymbol{\alpha}-\boldsymbol{y}\|_2^2 + \lambda\boldsymbol{\alpha}^\top\boldsymbol{K}\boldsymbol{\alpha} \right\}, \tag{11}$$

where the output is $\boldsymbol{y} = [y_1, y_2, \dots, y_m]^\top$. We can see that the above regularized risk minimization problem is in essence non-convex due to the non-positive definiteness of $\boldsymbol{K}$. But more exactly, problem (11) is non-convex when $\frac{1}{m}\boldsymbol{K}^2 + \lambda\boldsymbol{K}$ is indefinite. This condition always holds in practice due to $m \gg \lambda$. Following Oglic and Gärtner (2018), we do not strictly distinguish between the two differences in this paper. This is because, approximation analysis considers the $m \to \infty$ and $\lambda \to 0$ case, so it always holds true when m is large enough. Even if $\frac{1}{m}\boldsymbol{K}^2 + \lambda\boldsymbol{K}$ is PSD, our analysis for problem (11) is still applicable and reduces to a special case (i.e., using a RKHS regularizer), of which the learning rates are demonstrated by Corollary 2.

To obtain a global minimum of problem (11), we need the following proposition.

Proposition 2 *Problem* (11) *is equivalent to*

$$\boldsymbol{\alpha}_{z,\lambda} := \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^m:\,\boldsymbol{\alpha}^\top \boldsymbol{K}^2\boldsymbol{\alpha}= mr^2} \left\{ \frac{1}{m}\|\boldsymbol{K}\boldsymbol{\alpha}-\boldsymbol{y}\|_2^2 + \lambda\boldsymbol{\alpha}^\top\boldsymbol{K}\boldsymbol{\alpha} \right\}. \tag{12}$$

Proof Denote the objective function in problem (11) as $F(\boldsymbol{\alpha}) = \frac{1}{m}\|\boldsymbol{K}\boldsymbol{\alpha}-\boldsymbol{y}\|_2^2 + \lambda\boldsymbol{\alpha}^\top\boldsymbol{K}\boldsymbol{\alpha}$, we aim to prove that the solution $\boldsymbol{\alpha}^* := \operatorname{argmin}_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha})$ of this unconstrained optimization problem would be unbounded. Due to the non-positive definiteness of $\frac{1}{m}\boldsymbol{K}^2 + \lambda\boldsymbol{K}$, there exists an initial solution $\boldsymbol{\alpha}_0$ such that

$$\boldsymbol{\alpha}_0^\top\Big(\frac{1}{m}\boldsymbol{K}^2 + \lambda\boldsymbol{K}\Big)\boldsymbol{\alpha}_0 < 0.$$

By constructing a solving sequence $\{\boldsymbol{\alpha}_i\}_{i=0}^{\infty}$ admitting $\boldsymbol{\alpha}_{i+1} := c\boldsymbol{\alpha}_i$ with $c > 1$, we have

$$F(c\boldsymbol{\alpha}_{i+1}) - cF(\boldsymbol{\alpha}_i) = c(c-1)\boldsymbol{\alpha}_i^\top\Big(\frac{1}{m}\boldsymbol{K}^2 + \lambda\boldsymbol{K}\Big)\boldsymbol{\alpha}_i - \frac{c-1}{m}\|\boldsymbol{y}\|_2^2 < 0,$$

which indicates that, after the t-th iteration, $F(\boldsymbol{\alpha}_t) < c^t F(\boldsymbol{\alpha}_0) < 0$ and $\|\boldsymbol{\alpha}_t\|_2 = c^t \|\boldsymbol{\alpha}_0\|_2$ with $c > 1$. Therefore, the minimum $F(\boldsymbol{\alpha}^*)$ is unbounded, and tends to negative infinity. In this case, $\|\boldsymbol{\alpha}^*\|_2$ would also approach to infinity, i.e., a meaningless solution. Based on the above analyses, for problem $\min_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha})$, by introducing the constraint $\boldsymbol{\alpha}^\top \boldsymbol{K}^2 \boldsymbol{\alpha} \leq mr^2$, its solution is obtained on the hyper-sphere, i.e., $\boldsymbol{\alpha}^\top \boldsymbol{K}^2 \boldsymbol{\alpha} = mr^2$, which concludes the proof.

□

As demonstrated by Proposition 2, the inequality constraint in problem (11) can be transformed into an equality constraint, which is also suitable to problem (3). Then, albeit non-convex, problem (12) can be formulated as solving a constrained eigenvalue problem (Gander et al. 1988; Oglic and Gärtner 2018), yielding an optimal solution with closed-form.[2] Accordingly, the optimal solution $\boldsymbol{\alpha}_{z,\lambda}$ is given by

$$\boldsymbol{\alpha}_{z,\lambda} = \frac{1}{m}(\lambda \boldsymbol{I} - \mu \boldsymbol{K})^\dagger \boldsymbol{y}, \tag{13}$$

where the notation $(\cdot)^\dagger$ denotes the pseudo-inverse, $\boldsymbol{I}$ is the identity matrix, and μ is the smallest real eigenvalue of the following non-Hermitian matrix

$$\boldsymbol{G} = \begin{bmatrix} \lambda \boldsymbol{K}^\dagger & -\boldsymbol{I} \\ -\boldsymbol{y}\boldsymbol{y}^\top / m^3 r^2 & \lambda \boldsymbol{K}^\dagger \end{bmatrix}, \tag{14}$$

where $\boldsymbol{K}^\dagger$ is the pseudo-inverse of $\boldsymbol{K}$, i.e. $\boldsymbol{K}^\dagger = \boldsymbol{V} \operatorname{diag}\left(\boldsymbol{\Sigma}_1, \boldsymbol{0}_{m-p-q}, \boldsymbol{\Sigma}_2\right) \boldsymbol{V}^\top$ with two invertible diagonal matrices

$$\boldsymbol{\Sigma}_1 = \operatorname{diag}\left(\frac{\lambda}{\sigma_1}, \ldots, \frac{\lambda}{\sigma_p}\right), \quad \boldsymbol{\Sigma}_2 = \operatorname{diag}\left(\frac{\lambda}{\sigma_{m-q+1}}, \ldots, \frac{\lambda}{\sigma_m}\right). \tag{15}$$

It is clear that we cannot directly calculate μ. However, μ is very important in our analysis and thus we attempt to estimate it based on matrix perturbation theory (Stewart and Sun 1990). We will detail this in Sect. 5.

Besides, to aid our analysis, we introduce another nonnegative regularization scheme in RKKS to problem (3)

$$\widetilde{f_{z,\lambda}} := \operatorname*{argmin}_{f \in \mathcal{B}(r)} \left\{ \frac{1}{m} \sum_{i=1}^m \big(f(\boldsymbol{x}_i) - y_i\big)^2 + \lambda \langle f, Tf \rangle_{\mathcal{H}_\mathcal{K}} \right\}, \tag{16}$$

where the empirical covariance operator T is defined in RKKS but nonnegative, see Definition 4. Based on the above regularized risk minimization problem and Eq. (2), the regularizer can be represented as

$$\langle f, Tf \rangle_{\mathcal{H}_\mathcal{K}} = \frac{1}{m} \sum_{i,i'=1}^m \alpha_i \alpha_{i'} \sum_{j=1}^m k(\boldsymbol{x}_i, \boldsymbol{x}_j) k(\boldsymbol{x}_{i'}, \boldsymbol{x}_j) = \frac{1}{m} \boldsymbol{\alpha}^\top \boldsymbol{K}^2 \boldsymbol{\alpha}.$$

Accordingly, problem (16) can be formulated as

[2] As a generalized trust-region subproblem, problem (12) can be also solved by the S-lemma with equality to yield a globally optimal solution (Adachi and Nakatsukasa 2017; Xia et al. 2016).

$$\widetilde{\boldsymbol{\alpha}_{z,\lambda}} := \underset{\boldsymbol{\alpha}\in\mathbb{R}^m : \boldsymbol{\alpha}^\top \boldsymbol{K}^2\boldsymbol{\alpha}=mr^2}{\operatorname{argmin}} \left\{ \frac{1}{m}\|\boldsymbol{K\alpha}-\boldsymbol{y}\|_2^2 + \frac{\lambda}{m}\boldsymbol{\alpha}^\top\boldsymbol{K}^2\boldsymbol{\alpha} \right\}, \tag{17}$$

with $\widetilde{\boldsymbol{\alpha}_{z,\lambda}} = -\frac{1}{m\widetilde{\mu}}\boldsymbol{K}^\dagger\boldsymbol{y}$, and $\widetilde{\mu}$ is the smallest real eigenvalue of the matrix

$$\widetilde{\boldsymbol{G}} = \begin{bmatrix} \boldsymbol{0}_m & -\boldsymbol{I} \\ -\boldsymbol{y}\boldsymbol{y}^\top/m^3r^2 & \boldsymbol{0}_m \end{bmatrix}. \tag{18}$$

By Sylvester's determinant identity, we directly calculate the largest and smallest real eigenvalues of $\widetilde{\boldsymbol{G}}$ as $\frac{\|\boldsymbol{y}\|_2}{\sqrt{m}mr}$ and $-\frac{\|\boldsymbol{y}\|_2}{\sqrt{m}mr}$, respectively. So we have $\widetilde{\mu} = -\frac{\|\boldsymbol{y}\|_2}{m\sqrt{mr}} < 0$. Note that the regularizer in problem (16) can be also chosen to be other RKHS regularizers, such as $\langle f,f\rangle_{\mathcal{H}_{\widetilde{K}}}$ in Definition 3. But using the empirical kernel regularizer $\langle f, Tf\rangle_{\mathcal{H}_K}$, one obtains elegant and concise theoretical results, i.e., directly compute $\widetilde{\mu}$.

5 Framework of proofs

In this section, we establish the framework of proofs for Theorem 1. By the modified error decomposition technique in Sect. 5.1, the total error can be decomposed into the regularization error, the sample error, and an additional hypothesis error. We detail the estimates for the hypothesis error in Sect. 5.2. These two points are the main elements on novelty in the proof. We briefly introduce estimates for the sample error in Sect. 5.3 and derive the learning rates in Sect. 5.4.

5.1 Error decomposition

In order to estimate error $\|\pi_{M^*}(f_{z,\lambda}) - f_\rho\|$ in the $L^2_{\rho_X}$ space, i.e., to bound $\|\pi_B(f_{z,\lambda}) - f_\rho\|$ for any $B \geq M^*$, we need to estimate the excess error $\mathcal{E}(\pi_B(f_{z,\lambda})) - \mathcal{E}(f_\rho)$ which can be conducted by an error decomposition technique (Cucker and Zhou 2007). However, since $\langle f_{z,\lambda}, f_{z,\lambda}\rangle_{\mathcal{H}_K}$ might be negative, traditional techniques are invalid. Formally, our modified error decomposition technique is given by the following proposition by introducing an additional hypothesis error.

Proposition 3 *Let* $f_\lambda = \operatorname{argmin}_{f\in\mathcal{H}_K}\left\{\mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda\langle f, Tf\rangle_{\mathcal{H}_K}\right\}$, *then* $\mathcal{E}(\pi_B(f_{z,\lambda})) - \mathcal{E}(f_\rho)$ *can be bounded by*

$$\begin{aligned} \mathcal{E}(\pi_B(f_{z,\lambda})) - \mathcal{E}(f_\rho) &\leq \mathcal{E}(\pi_B(f_{z,\lambda})) - \mathcal{E}(f_\rho) + \lambda\langle f_{z,\lambda}, Tf_{z,\lambda}\rangle_{\mathcal{H}_K} \\ &\leq D(\lambda) + S(z,\lambda) + P(z,\lambda), \end{aligned}$$

where $D(\lambda)$ *is the regularization error defined by Eq.* (4). *The sample error* $S(z,\lambda)$ *is given by*

$$S(z,\lambda) = \mathcal{E}(\pi_B(f_{z,\lambda})) - \mathcal{E}_z(\pi_B(f_{z,\lambda})) + \mathcal{E}_z(f_\lambda) - \mathcal{E}(f_\lambda).$$

The introduced hypothesis error $P(z,\lambda)$ *is defined by*

$$P(z,\lambda) = \mathcal{E}_z(f_{z,\lambda}) + \lambda\langle f_{z,\lambda}, Tf_{z,\lambda}\rangle_{\mathcal{H}_K} - \mathcal{E}_z(\widetilde{f_{z,\lambda}}) - \lambda\langle \widetilde{f_{z,\lambda}}, T\widetilde{f_{z,\lambda}}\rangle_{\mathcal{H}_K}, \tag{19}$$

where $f_{z,\lambda}$ *and* $\widetilde{f_{z,\lambda}}$ *are optimal solutions of problem* (3) *and problem* (16), *respectively.*

Proof We write $\mathcal{E}\big(\pi_B(f_{z,\lambda})\big) - \mathcal{E}(f_\rho) + \lambda\langle f_{z,\lambda}, Tf_{z,\lambda}\rangle_{\mathcal{H}_\mathcal{K}}$ as

$$\begin{aligned}
&\mathcal{E}\big(\pi_B(f_{z,\lambda})\big) - \mathcal{E}(f_\rho) + \lambda\langle f_{z,\lambda}, Tf_{z,\lambda}\rangle_{\mathcal{H}_\mathcal{K}} = \Big\{\mathcal{E}\big(\pi_B(f_{z,\lambda})\big) - \mathcal{E}_z\big(\pi_B(f_{z,\lambda})\big)\Big\}\\
&\quad + \Big\{\mathcal{E}_z\big(\pi_B(f_{z,\lambda})\big) + \lambda\langle f_{z,\lambda}, Tf_{z,\lambda}\rangle_{\mathcal{H}_\mathcal{K}}\Big\} - \Big\{\mathcal{E}_z(f_\lambda) + \lambda\langle f_\lambda, Tf_\lambda\rangle_{\mathcal{H}_\mathcal{K}}\Big\} + \Big\{\mathcal{E}_z(f_\lambda) - \mathcal{E}(f_\lambda)\Big\}\\
&\quad + \Big\{\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho) + \lambda\langle f_\lambda, Tf_\lambda\rangle_{\mathcal{H}_\mathcal{K}}\Big\}\\
&\le D(\lambda) + P(z,\lambda) + S(z,\lambda),
\end{aligned}$$

where we use $\mathcal{E}_z\big(\pi_B(f_{\underset{\sim}{z,\lambda}})\big) \le \mathcal{E}_z\big(f_{z,\lambda}\big)$ in the first inequality, and the second inequality holds by the condition that $f_{z,\lambda}$ is a global minimizer of problem (16). □

It can be found that the additional hypothesis error stems from the difference between $\langle f_{z,\lambda}, f_{z,\lambda}\rangle_{\mathcal{H}_\mathcal{K}}$-regularization and $\langle f_{z,\lambda}, Tf_{z,\lambda}\rangle_{\mathcal{H}_\mathcal{K}}$-regularization in essence. Hence, we estimate the introduced hypothesis error in the following descriptions.

5.2 Bound hypothesis error

Since $\widetilde{f_{z,\lambda}}$ is an optimal solution of problem (16), obviously, we have $P(z,\lambda) \ge 0$. To bound the hypothesis error, we need to estimate the objective function value difference of the two learning problems (3) and (17) by the following proposition.

Proposition 4 *Suppose that the spectrum of the indefinite kernel matrix* $\boldsymbol{K}$ *satisfies Assumption 4, denote the condition number of two invertible matrices* $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ *in Eq. (15) as* $C_1, C_2 < \infty$. *When* $\eta + \gamma > 1$ *with* $\eta = \min\{\eta_1, \eta_2\}$, *the hypothesis error defined in Eq. (19) holds with probability* 1 *such that*

$$P(z,\lambda) \le \widetilde{C_1} m^{-\Theta_1},$$

where $\widetilde{C_1} := 2Mr + 2M^2\big(\frac{-c_m}{C_2} + \frac{M^2}{r^2} + \frac{C_1}{c_1}\big)$ *and the power index is* $\Theta_1 = \min\big\{1, \gamma + \eta - 1\big\}$.

Proof The proof can be found in Sect. 5.2.3. □

Remark The condition number of invertible matrices is finite, which is mild as demonstrated by Gao et al. (2015).

In the next, we give the proof of Proposition 4. For better presentation, we divide the proof into three parts: in Sect. 5.2.1, we decompose the hypothesis error $P(z,\lambda)$ into the sum of two terms that would depend on μ, i.e., the smallest real eigenvalue of a non-Hermitian matrix $\boldsymbol{G}$ in Eq. (14). Then we estimate μ in Sect. 5.2.2 so as to bound $P_2(z,\lambda)$ and $P(z,\lambda)$ in Sect. 5.2.3.

5.2.1 Decomposition of hypothesis error

The hypothesis error $P(z,\lambda)$ can be decomposed into the sum of two parts that depend on μ and $\tilde{\mu}$ by the following proposition.

Proposition 5 *Given the hypothesis error $P(z, \lambda)$ defined in Eq. (19), it can be decomposed as*

$$P(z,\lambda) = P_1(z,\lambda) + P_2(z,\lambda) := -\frac{2}{m^2\widetilde{\mu}}\boldsymbol{y}^\top \boldsymbol{K}\boldsymbol{K}^\dagger \boldsymbol{y} - \frac{2}{m^2}\boldsymbol{y}^\top \boldsymbol{K}(\lambda \boldsymbol{I} - \mu \boldsymbol{K})^\dagger \boldsymbol{y},$$

where $P_1(z, \lambda)$ depends on $\widetilde{\mu} := -\frac{\|\boldsymbol{y}\|_2}{m\sqrt{mr}}$ and $P_2(z, \lambda)$ depends on μ, i.e., the smallest real eigenvalue of a non-Hermitian matrix $\boldsymbol{G}$ in Eq. (14).

Proof According to the definition of the hypothesis error $P(z, \lambda)$, we have

$$P(z,\lambda) = \mathcal{E}_z\big(f_{z,\lambda}\big) + \lambda\langle f_{z,\lambda}, Tf_{z,\lambda}\rangle_{\mathcal{H}_K} - \mathcal{E}_z\big(\widetilde{f_{z,\lambda}}\big) - \lambda\langle \widetilde{f_{z,\lambda}}, T\widetilde{f_{z,\lambda}}\rangle_{\mathcal{H}_K},$$

where $f_{z,\lambda}$ and $\widetilde{f_{z,\lambda}}$ are optimal solutions of problem (3) and problem (16), respectively. Therefore, both of them can be obtained on the hyper-sphere. Besides, the regularizer is $\boldsymbol{\alpha}_{z,\lambda}^\top \boldsymbol{K}^2 \boldsymbol{\alpha}_{z,\lambda} = mr^2$ can be canceled out in $P(z, \lambda)$. Based on this, $P(z, \lambda)$ can be further represented as

$$\begin{aligned}
P(z,\lambda) &= \frac{1}{m}\sum_{i=1}^{m}\big(f_{z,\lambda}(\boldsymbol{x}_i) - y_i\big)^2 + \lambda\langle f_{z,\lambda}, Tf_{z,\lambda}\rangle_{\mathcal{H}_K} - \frac{1}{m}\sum_{i=1}^{m}\big(\widetilde{f_{z,\lambda}}(\boldsymbol{x}_i) - y_i\big)^2 - \lambda\langle \widetilde{f_{z,\lambda}}, T\widetilde{f_{z,\lambda}}\rangle_{\mathcal{H}_K} \\
&= \frac{1}{m}\|\boldsymbol{K}\boldsymbol{\alpha}_{z,\lambda} - \boldsymbol{y}\|_2^2 - \frac{1}{m}\|\boldsymbol{K}\widetilde{\boldsymbol{\alpha}_{z,\lambda}} - \boldsymbol{y}\|_2^2 = \frac{2}{m}\boldsymbol{y}^\top \boldsymbol{K}\widetilde{\boldsymbol{\alpha}_{z,\lambda}} - \frac{2}{m}\boldsymbol{y}^\top \boldsymbol{K}\boldsymbol{\alpha}_{z,\lambda} \\
&= \underbrace{-\frac{2}{m^2\widetilde{\mu}}\boldsymbol{y}^\top \boldsymbol{K}\boldsymbol{K}^\dagger \boldsymbol{y}}_{\triangleq P_1(z,\lambda)} \underbrace{-\frac{2}{m^2}\boldsymbol{y}^\top \boldsymbol{K}(\lambda \boldsymbol{I} - \mu \boldsymbol{K})^\dagger \boldsymbol{y}}_{\triangleq P_2(z,\lambda)}.
\end{aligned}$$

□

5.2.2 Estimate μ

To bound $P(z, \lambda)$, we need to bound $P_1(z, \lambda)$ and $P_2(z, \lambda)$ respectively. The estimation for $P_1(z, \lambda)$ is simple (we will illustrate it in the next subsection). However, $P_2(z, \lambda)$ involves with μ, i.e., the smallest real eigenvalue of a non-Hermitian matrix $\boldsymbol{G}$, which makes our estimation for $P_2(z, \lambda)$ quite intractable. Based on this, here we attempt to present an estimation for μ based on matrix perturbation theory (Stewart and Sun 1990).

Typically, there are three classical and well-known perturbation bounds for matrix eigenvalues, including the Bauer–Fike theorem and the Hoffman-Wielandt theorem for diagonalizable matrices (Hoffman and Wielandt 2003), and Weyl's theorem for Hermitian matrices (Stewart and Sun 1990). However, $\boldsymbol{G}$ is neither Hermitian nor diagonalizable. To aid our proof, we need the following lemma.

Lemma 1 (Henrici theorem (Chu 1986)) *Let $\boldsymbol{A}$ be an $m \times m$ matrix with Schur decomposition $\boldsymbol{Q}^H\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{D} + \boldsymbol{U}$, where $\boldsymbol{Q}$ is unitary, $\boldsymbol{D}$ is a diagonal matrix and $\boldsymbol{U}$ is a strict upper triangular matrix, with $(\cdot)^H$ denoting the Hermitian transpose. For each eigenvalue $\tilde{\sigma}$ of $\boldsymbol{A} + \widetilde{\boldsymbol{\Delta}}$, there exists an eigenvalue $\sigma(\boldsymbol{A})$ of $\boldsymbol{A}$ such that*

$$|\tilde{\sigma} - \sigma(\boldsymbol{A})| \leq \max(\varsigma, \sqrt[b]{\varsigma}), \quad \text{where} \;\; \varsigma := \|\widetilde{\Delta}\|_2 \sum_{i=1}^{b-1} \|\boldsymbol{U}\|_2^i,$$

where $b \leq m$ is the smallest integer satisfying $\boldsymbol{U}^b = 0$, i.e., the nilpotent index of $\boldsymbol{U}$.

Based on the above lemma, μ admits the following representation.

Proposition 6 *Under the assumption of Proposition 4, as the smallest real eigenvalue of a non-Hermitian matrix $\boldsymbol{G}$ in Eq. (14), μ admits the following expression*

$$\mu = \widetilde{c_a}\widetilde{\mu} + \widetilde{c_b}\widetilde{\mu}^2 + \left[\frac{C_2}{c_m} + \widetilde{c_d}\left(\frac{C_1}{c_1} - \frac{C_2}{c_m}\right)\right] m^{-(\gamma+\eta)}, \tag{20}$$

with $\widetilde{c_a} \in [-1, 0) \bigcup (0, 1]$, $\widetilde{c_b} \in [-1, 1]$, $\widetilde{c_d} \in [0, 1]$, and $\widetilde{\mu} := -\frac{\|\boldsymbol{y}\|_2}{m\sqrt{mr}} < 0$.

Proof The non-Hermitian matrix $\boldsymbol{G}$ in Eq. (14) can be reformulated as

$$\boldsymbol{G} = \underbrace{\begin{bmatrix} \lambda \boldsymbol{K}^{\dagger} & -\boldsymbol{I} \\ \boldsymbol{0}_{m\times m} & \lambda \boldsymbol{K}^{\dagger} \end{bmatrix}}_{\triangleq \boldsymbol{G}_1} + \underbrace{\begin{bmatrix} \boldsymbol{0}_{m\times m} & \boldsymbol{0}_{m\times m} \\ -\boldsymbol{y}\boldsymbol{y}^{\top}/m^3 r^2 & \boldsymbol{0}_{m\times m} \end{bmatrix}}_{\triangleq \boldsymbol{G}_2}.$$

As a result, $\boldsymbol{G}$ can be represented as a sum of a block upper triangular matrix $\boldsymbol{G}_1$ with a non-Hermitian perturbation $\boldsymbol{G}_2$.

To estimate $\boldsymbol{G}_1$, by Lemma 1, from the definition of Schur decomposition on $\boldsymbol{G}_1$, it can be easily verified that $\boldsymbol{D}$ and $\boldsymbol{U}$ are

$$\boldsymbol{D} = \operatorname{diag}\left(\frac{\lambda}{\sigma_1}, \ldots, \frac{\lambda}{\sigma_p}, 0, \ldots, 0, \frac{\lambda}{\sigma_{m-q+1}}, \ldots, \frac{\lambda}{\sigma_m}, \frac{\lambda}{\sigma_1}, \ldots, \frac{\lambda}{\sigma_p}, 0, \ldots, 0, \frac{\lambda}{\sigma_{m-p+1}}, \ldots, \frac{\lambda}{\sigma_m}\right),$$
$$\text{and} \quad \boldsymbol{U} = \begin{bmatrix} \boldsymbol{0}_m & -\boldsymbol{I} \\ \boldsymbol{0}_m & \boldsymbol{0}_m \end{bmatrix}.$$

Accordingly, $\boldsymbol{U}$ is a nilpotent matrix with $\boldsymbol{U}^2 = 0$, and thus we have $b = 2$. According to Lemma 1, there exists an eigenvalue of $\boldsymbol{G}_1$ denoting as $\sigma(\boldsymbol{G}_1)$ such that

$$\left|\mu - \sigma(\boldsymbol{G}_1)\right| \leq \max(\varsigma, \sqrt[b]{\varsigma}) \leq \varsigma + \sqrt[b]{\varsigma}, \tag{21}$$

where ς is given by

$$\varsigma := \|\boldsymbol{G}_2\|_2 \sum_{i=1}^{b-1} \left\|\boldsymbol{U}\right\|_2^i = \|\boldsymbol{G}_2\|_2 \|\boldsymbol{U}\|_2 = \|\boldsymbol{G}_2\|_2 = \frac{\|\boldsymbol{y}\|^2}{m^3 r^2}.$$

Then we consider the following three cases based on the sign of $\sigma(\boldsymbol{G}_1)$.

Case 1 $\sigma(\boldsymbol{G}_1) = 0$

The inequality in Eq. (21) can be formulated as

$$-\frac{\|\boldsymbol{y}\|_2}{m\sqrt{mr}} - \frac{\|\boldsymbol{y}\|_2^2}{m^3r^2} \leq \mu \leq \frac{\|\boldsymbol{y}\|_2}{m\sqrt{mr}} + \frac{\|\boldsymbol{y}\|_2^2}{m^3r^2}. \tag{22}$$

Case 2 $\sigma(\boldsymbol{G}_1) > 0$

Without loss of generality, we assume that $\sigma(\boldsymbol{G}_1)$ is λ/σ_l with $l \in \{1, 2, \dots, p\}$. According to the definition of condition number C_1, we have

$$0 < \frac{1}{\sigma_1} \leq \frac{1}{\sigma_l} \leq \frac{C_1}{c_1} m^{-\eta_1} \leq \frac{C_1}{c_1} m^{-\eta}, \quad \eta = \min\{\eta_1, \eta_2\}.$$

Then, the inequality in Eq. (21) can be formulated as

$$-\frac{\|\boldsymbol{y}\|_2}{m\sqrt{mr}} - \frac{\|\boldsymbol{y}\|_2^2}{m^3r^2} \leq \mu \leq \frac{C_1}{c_1} m^{-(\gamma+\eta)} + \frac{\|\boldsymbol{y}\|_2}{m\sqrt{mr}} + \frac{\|\boldsymbol{y}\|_2^2}{m^3r^2}. \tag{23}$$

Case 3 $\sigma(\boldsymbol{G}_1) < 0$

Likewise, we assume that $\sigma(\boldsymbol{G}_1)$ is λ/σ_l with $l \in \{m-q+1, m-q+2, \dots, m\}$. According to the definition of condition number C_2, we have

$$0 > \frac{1}{\sigma_m} \geq \frac{1}{\sigma_l} \geq \frac{C_2}{c_m} m^{-\eta_2} \geq \frac{C_2}{c_m} m^{-\eta}, \quad \eta = \min\{\eta_1, \eta_2\}.$$

Then, the inequality in Eq. (21) can be formulated as

$$\frac{C_2}{c_m} m^{-(\gamma+\eta)} - \frac{\|\boldsymbol{y}\|_2}{m\sqrt{mr}} - \frac{\|\boldsymbol{y}\|_2^2}{m^3r^2} \leq \mu \leq \frac{\|\boldsymbol{y}\|_2}{m\sqrt{mr}} + \frac{\|\boldsymbol{y}\|_2^2}{m^3r^2}. \tag{24}$$

Combining Eq. (22), Eqs. (23) and (24), we have

$$\begin{cases} \mu \geq \dfrac{C_2}{c_m} m^{-(\gamma+\eta)} - \dfrac{\|\boldsymbol{y}\|_2}{m\sqrt{mr}} - \dfrac{\|\boldsymbol{y}\|_2^2}{m^3r^2} \\ \mu \leq \dfrac{C_1}{c_1} m^{-(\gamma+\eta)} + \dfrac{\|\boldsymbol{y}\|_2}{m\sqrt{mr}} + \dfrac{\|\boldsymbol{y}\|_2^2}{m^3r^2}, \end{cases}$$

which can be further written as

$$\frac{C_2}{c_m} m^{-(\gamma+\eta)} + \tilde{\mu} - \tilde{\mu}^2 \leq \mu \leq \frac{C_1}{c_1} m^{-(\gamma+\eta)} - \tilde{\mu} + \tilde{\mu}^2.$$

Therefore, we have $\lim_{m\to\infty} \mu = 0$, and its convergence rate is $\mathcal{O}(1/m)$ due to $\gamma + \eta > 1$. Finally, μ can be represented in Eq. (20) with $\widetilde{c_a} \neq 0$, which concludes the proof. □

5.2.3 Proofs of Proposition 4

Given the expression of μ with the convergence rate $\mathcal{O}(1/m)$ in Proposition 6, we are ready to present the estimates for $P_2(z, \lambda)$ and $P(z, \lambda)$ as demonstrated by Proposition 4.

Proof of Proposition 4 We cast the proof in two steps: firstly prove the *consistency*, i.e., $\lim_{m\to\infty} P(z, \lambda) = 0$, and then derive its convergence rate.

Step 1: Consistency of $P(z, \lambda)$

Based on the decomposition of the hypothesis error $P(z, \lambda)$ in Proposition 5, due to $P(z, \lambda) \geq 0$ for any $m \in \mathbb{N}$, we have $\lim_{m\to\infty} \big(P_1(z, \lambda) + P_2(z, \lambda)\big) \geq 0$ if the limits $\lim_{m\to\infty} P_1(z, \lambda)$ and $\lim_{m\to\infty} P_2(z, \lambda)$ exist. Next we analyse $P_1(z, \lambda)$ and $P_2(z, \lambda)$, respectively.

According to the expression of $P_1(z, \lambda)$, it can be bounded by

$$\begin{aligned} P_1(z, \lambda) &= \frac{2}{m}\boldsymbol{y}^\top \boldsymbol{K}\widetilde{\boldsymbol{\alpha}_{z,\lambda}} = -\frac{2}{m^2\tilde{\mu}}\boldsymbol{y}^\top \boldsymbol{K}\boldsymbol{K}^\dagger \boldsymbol{y} \\ &= -\frac{2}{m^2\tilde{\mu}}\boldsymbol{y}^\top \underbrace{\left(\sum_{i=1}^{p} \boldsymbol{v}_i\boldsymbol{v}_i^\top + \sum_{i=m-q+1}^{m} \boldsymbol{v}_i\boldsymbol{v}_i^\top\right)}_{\triangleq \boldsymbol{\Xi}}\boldsymbol{y} \\ &\leq \frac{2\|\boldsymbol{y}\|_2 r}{\sqrt{m}}, \end{aligned} \tag{25}$$

where $\boldsymbol{v}_i$ is the i-th column of the orthogonal matrix $\boldsymbol{V}$ from the eigenvalue decomposition $\boldsymbol{K} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. The inequality in the above equation holds by $\boldsymbol{y}^\top \boldsymbol{\Xi}\boldsymbol{y} = \boldsymbol{y}^\top(\boldsymbol{I} - \sum_{i=p+1}^{m-q} \boldsymbol{v}_i\boldsymbol{v}_i^\top)\boldsymbol{y} \leq \boldsymbol{y}^\top\boldsymbol{y}$.

According to the expression of $P_2(z, \lambda)$, it can be rewritten as

$$P_2(z, \lambda) = -\frac{2}{m^2}\boldsymbol{y}^\top \boldsymbol{K}(\lambda\boldsymbol{I} - \mu\boldsymbol{K})^\dagger \boldsymbol{y} = \frac{2}{m^2}\boldsymbol{y}^\top\left(\sum_{i=1}^{p} \frac{-\boldsymbol{v}_i\boldsymbol{v}_i^\top}{\frac{\lambda}{\sigma_i} - \mu} + \sum_{i=m-q+1}^{m} \frac{-\boldsymbol{v}_i\boldsymbol{v}_i^\top}{\frac{\lambda}{\sigma_i} - \mu}\right)\boldsymbol{y}.$$

Since the function $h(\sigma_i) = \frac{-1}{\frac{\lambda}{\sigma_i} - \mu}$ is an increasing function of σ_i, $P_2(z, \lambda)$ can be bounded by

$$-\frac{2}{m^2} \cdot \frac{1}{\frac{\lambda}{\sigma_1} - \mu}\boldsymbol{y}^\top \boldsymbol{\Xi}\boldsymbol{y} \leq P_2(z, \lambda) \leq -\frac{2}{m^2} \cdot \frac{1}{\frac{\lambda}{\sigma_{m-q+1}} - \mu}\boldsymbol{y}^\top \boldsymbol{\Xi}\boldsymbol{y}. \tag{26}$$

By Proposition 6, plugging Eq. (20) into the above inequality, when $\eta + \gamma > 1$, we have

$$\begin{aligned} \lim_{m\to\infty} -\frac{2}{m^2} \cdot \frac{1}{\frac{\lambda}{\sigma_1} - \mu}\boldsymbol{y}^\top \boldsymbol{\Xi}\boldsymbol{y} &= \lim_{m\to\infty} -\frac{2}{m^2} \cdot \frac{1}{\frac{\lambda}{\sigma_{m-q+1}} - \mu}\boldsymbol{y}^\top \boldsymbol{\Xi}\boldsymbol{y} = \lim_{m\to\infty} \frac{2\boldsymbol{y}^\top \boldsymbol{\Xi}\boldsymbol{y}}{\sqrt{m}\|\boldsymbol{y}\|} \cdot \frac{r}{-\widetilde{c_a}} \\ &\leq \lim_{m\to\infty} \frac{2\|\boldsymbol{y}\|_2 r}{-\widetilde{c_a}\sqrt{m}} < \infty, \end{aligned}$$

which holds by $\|\boldsymbol{y}\|_2 = \mathcal{O}(\sqrt{m})$ and $\widetilde{c_a} \neq 0$. According to the squeeze theorem, we conclude that the limit $\lim_{m\to\infty} P_2(z, \lambda)$ exists. Because of $P(z, \lambda) \geq 0$, we have

$$0 \leq \lim_{m\to\infty} \Big(P_1(z, \lambda) + P_2(z, \lambda)\Big) \leq \lim_{m\to\infty} \left[\frac{2\|\boldsymbol{y}\|_2 r}{\sqrt{m}}\left(1 - \frac{1}{\widetilde{c_a}}\right)\right],$$

which indicates that $1 - \frac{1}{\widetilde{c_a}} \geq 0$, i.e., $\widetilde{c_a} \geq 1$. Accordingly, the coefficient in Eq. (20) $\widetilde{c_a} \in [-1, 0) \bigcup (0, 1]$ can be further improved to $\widetilde{c_a} = 1$. In this case, it is obvious that $\lim_{m\to\infty} \Big(P_1(z, \lambda) + P_2(z, \lambda)\Big) = 0$ implies the consistency for $P(z, \lambda)$.

Step 2: Convergence rate of $P(z, \lambda)$

Based on the consistency of $P(z, \lambda)$, we derive its convergence rate as follows. For notational simplicity, we denote $\widetilde{c_e} := \left[\frac{C_2}{c_m} + \widetilde{c_d}\left(\frac{C_1}{c_1} - \frac{C_2}{c_m}\right)\right]$. Accordingly, by virtue of Eqs. (25), (26) and Proposition 6 for μ, we have

$$
\begin{aligned}
P(z, \lambda) &= P_1(z, \lambda) + P_2(z, \lambda) \\
&\leq \frac{2\|\boldsymbol{y}\|_2 r}{\sqrt{m}} + \frac{2\|\boldsymbol{y}\|_2^2}{m^2} \cdot \frac{1}{\frac{\lambda}{\sigma_{m-q+1}} - \mu} \\
&\leq \frac{2\|\boldsymbol{y}\|_2 r}{\sqrt{m}} + \frac{2\|\boldsymbol{y}\|_2^2}{m} \frac{1}{\frac{-1}{\sigma_{m-q+1}} m^{1-\gamma} - \frac{\|\boldsymbol{y}\|_2}{\sqrt{m} r} - \frac{\widetilde{c_b}\|\boldsymbol{y}\|_2^2}{m^2 r^2} - \widetilde{c_e} m^{-\gamma}} \\
&\leq \frac{2\|\boldsymbol{y}\|_2 r}{\sqrt{m}} + \frac{2\|\boldsymbol{y}\|_2^2}{m}\left(- \frac{\sqrt{m} r}{\|\boldsymbol{y}\|} \frac{-c_m}{C_2} m^{1-\gamma-\eta} + \frac{\|\boldsymbol{y}\|_2^2}{m r^2} m^{-1} + |\widetilde{c_e}| m^{-(\gamma+\eta)} \right) \\
&\leq \left(2Mr + 2M^2\left(\frac{-c_m}{C_2} + \frac{M^2}{r^2} + \frac{C_1}{c_1}\right)\right) m^{-\Theta_1} \\
&\triangleq \widetilde{C_1} m^{-\Theta_1},
\end{aligned}
$$

where $\widetilde{C_1} := 2Mr + 2M^2\left(\frac{-c_m}{C_2} + \frac{M^2}{r^2} + \frac{C_1}{c_1}\right)$ and the power index is $\Theta_1 = \min\left\{1, \gamma + \eta - 1\right\}$. Finally, we conclude the proof for Proposition 4. □

5.3 Estimate sample error

The sample error can be decomposed into $S(z, \lambda) = S_1(z, \lambda) + S_2(z, \lambda)$ with

$$
\begin{aligned}
S_1(z, \lambda) &= \mathcal{E}\big(\pi_B(f_{z,\lambda})\big) - \mathcal{E}(f_\rho) - \mathcal{E}_z\big(\pi_B(f_{z,\lambda})\big) + \mathcal{E}_z(f_\rho), \\
S_2(z, \lambda) &= \Big\{\mathcal{E}_z\big(f_\lambda\big) - \mathcal{E}_z(f_\rho)\Big\} - \Big\{\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho)\Big\}.
\end{aligned}
$$

Note that $S_1(z, \lambda)$ involves the samples z. Thus a uniform concentration inequality for a family of functions containing $f_{z,\lambda}$ is needed to estimate $S_1(z, \lambda)$. Since we have $f_{z,\lambda} \in \mathcal{B}_R$ defined by Eq. (6), we shall bound S_1 by the following proposition with a properly chosen R. Considering that the estimates for $S_1(z, \lambda)$ and $S_2(z, \lambda)$ have been extensively investigated in Wu et al. (2006), Cucker and Zhou (2007), Shi et al. (2014), we directly present the corresponding results in Appendix 1 under the existence of f_ρ in Assumption 1, and the regularity condition on ρ in Assumption 3.

5.4 Derive learning rates

Combining the bounds in Propositions 3, 4 and estimates for the sample error, the excess error $\mathcal{E}\big(\pi_B(f_{z,\lambda})\big) - \mathcal{E}(f_\rho)$ can be estimated. Specifically, as aforementioned, algorithmically, the radius r or R in Eq. (6) is determined by cross validation in our experiments. Theoretically, in our analysis, it is estimated by giving a bound for $\lambda\langle f_{z,\lambda}, Tf_{z,\lambda}\rangle_{\mathcal{H}_\mathcal{K}}$. This is conducted by the iteration technique (Wu et al. 2006) to improve learning rates. Under Assumption 1– 4, the proof for learning rates in Theorem 1 can be found in Appendix 2.

6 Numerical experiments

In this section, we validate our theoretical results by numerical experiments in the following three aspects.

6.1 Eigenvalue assumption

Here we verify the justification of our eigenvalue decay assumption in Assumption 4 on four indefinite kernels, including

- the spherical polynomial (SP) kernel (Pennington et al. 2015): $k_p(\boldsymbol{x},\boldsymbol{x}') = (1 + \langle \boldsymbol{x},\boldsymbol{x}' \rangle)^p$ with $p = 10$ on the unit sphere is shift-invariant but indefinite.
- the *TL1* kernel (Huang et al. 2018): $k_{\tau'}(\boldsymbol{x},\boldsymbol{x}') = \max\{\tau' - \|\boldsymbol{x} - \boldsymbol{x}'\|_1, 0\}$ with $\tau' = 0.7d$ as suggested.
- the Delta-Gauss kernel (Oglic and Gärtner 2018): It is formulated as the difference of two Gaussian kernels, i.e., $k\left(\boldsymbol{x},\boldsymbol{x}'\right) = \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/\tau_1\right) - \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/\tau_2\right)$ with $\tau_1 = 1$ and $\tau_2 = 0.1$.
- the *log* kernel (Boughorbel et al. 2005): $k(\boldsymbol{x},\boldsymbol{x}') = -\log(1 + \|\boldsymbol{x} - \boldsymbol{x}'\|)$.

Here the Delta-Gaussian kernel (Oglic and Gärtner 2018) and the log kernel Boughorbel et al. (2005) are associated with RKKS while the SP and `TL1` kernels have not been proved as reproducing kernels in RKKS. It is still an open problem to verify that a kernel admits the decomposition (Liu et al. 2020). The Delta-Gaussian kernel is defined as the difference of two Gaussian kernels, and thus it is clear that σ_1 and σ_m follow with the exponential decay in the same rate, i.e., $\eta_1 = \eta_2$. For the `log` kernel (Boughorbel et al. 2005) is a conditionally positive definite kernel of order one[3] associated with RKKS. According to Theorem 8.5 in Wendland (2004), the kernel matrix induced by this kernel has only one negative eigenvalue. Further, we can conclude that the only one negative eigenvalue admits $\sigma_m = -\sum_{i=1}^{m-1} \sigma_i$ because of $k(\boldsymbol{0}) = \frac{1}{n}\mathrm{tr}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n} \lambda_i = 0$, which implies $\eta_2 > \eta_1$.

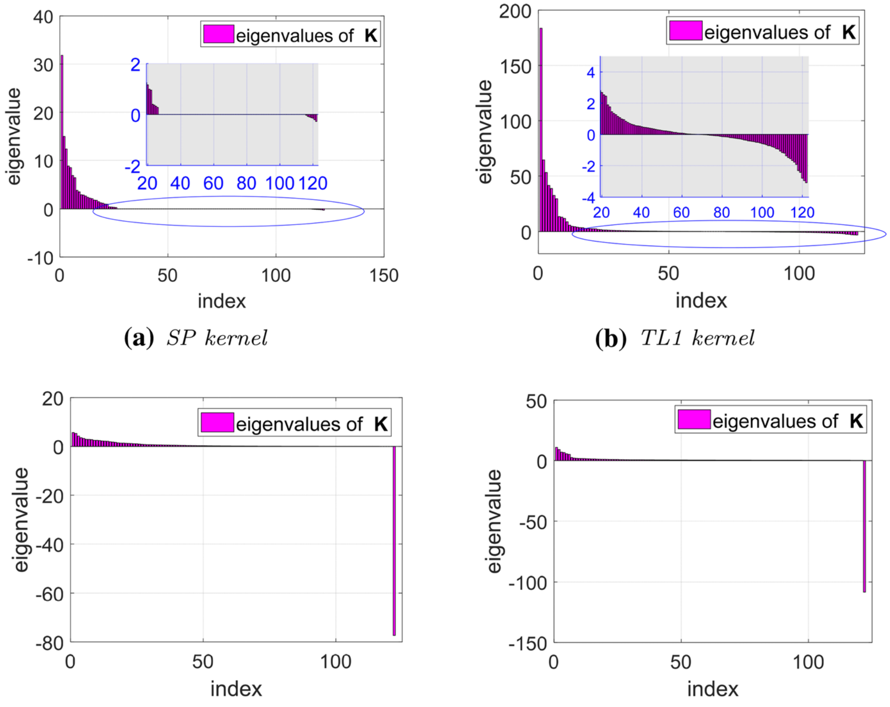
Figure 1 experimentally shows eigenvalue distributions of the above four indefinite kernels on the *monks3* dataset.[4] It can be found that our eigenvalue assumption: $\sigma_1 \geq c_1 m^{\eta_1}$ $(c_1 > 0, \eta_1 > 0)$ and $\sigma_m \leq c_m m^{\eta_2}$ $(c_m < 0, \eta_2 > 0)$ in Definition 4 is reasonable. Specifically, our experiments on the `log` kernel verify that it has only one negative eigenvalue admitting $\sigma_m = -\sum_{i=1}^{m-1} \sigma_i$. Note that although the SP and `TL1` kernels have not been proved as reproducing kernels in RKKS, our eigenvalue assumption still covers them, which demonstrates the feasibility of our assumption.

6.2 Empirical validations of derived learning rates

Here we verify the derived convergence rates on the *monks3* dataset effected by different indefinite kernels. In our experiment, we choose $\lambda := 1/m$ and two indefinite kernels including the Delta-Gauss kernel and the `log` kernel on *monks3* to study in what degree

[3] The order in conditionally positive definite kernels is an important concept, refer to Wendland (2004) for details.

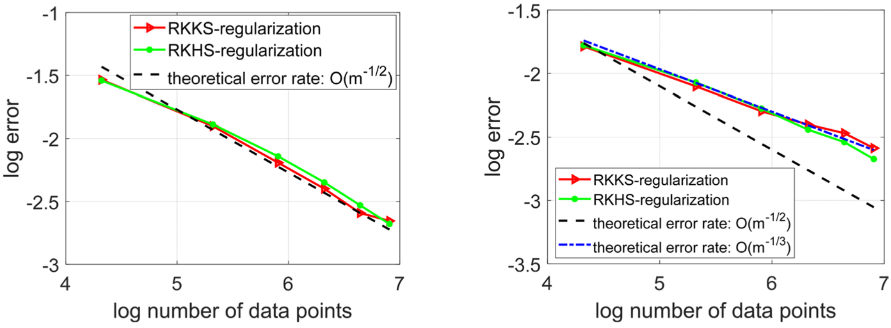
[4] https://archive.ics.uci.edu/ml/datasets.html.



(a) *SP kernel* **(b)** *TL1 kernel*

(c) *Delta-Gauss kernel* **(d)** *log kernel*

Fig. 1 Eigenvalue distribution of kernel matrices generated by various indefinite kernels on the *monks3* dataset



(a) *Delta-Gauss kernel* **(b)** *log kernel*

Fig. 2 The log–log plot of the theoretical and observed risk convergence rates averaged on 100 trials

they would effect the learning rates. Since the selected two kernels are $C^\infty(X \times X)$, s can be arbitrarily small. In this case, by Theorem 1 and Corollary 2, the learning rate of problem (3) with the RKKS regularizer $\langle f, f \rangle_{\mathcal{H}_\mathcal{K}}$ or the RKHS regularizer $\|f\|^2_{\mathcal{H}_{\bar{\mathcal{K}}}}$ is close to $\min\{\beta, \eta\}$. Here the two parameters β and η indicate the approximation ability for f_ρ and the size of RKKS by different indefinite kernels, and thus they will influence the expected risk rate. Figure 2a shows the observed learning rate associated with the Delta-Gauss kernel is $\mathcal{O}(1/\sqrt{m})$, while the excess risk associated with the `log` kernel converges at $\mathcal{O}(m^{-1/3})$ in Fig. 2b. Hence, Fig. 2 demonstrates this difference that the excess risk of problem (3) with the Delta-Gauss kernel converges faster than that with the `log` kernel. This is reasonable and demonstrated by Theorem 1, i.e., different $\mathcal{H}_{\mathcal{K}}$ spanned by various indefinite kernels lead to different convergence rates due to their different approximation ability for f_ρ.

The above experiments validate the rationality of our eigenvalue assumption and the consistency with theoretical results.

7 Conclusion

In this paper, we provide approximation analysis of the least squares problem associated with the $\langle f,f\rangle_{\mathcal{H}_{\mathcal{K}}}$ regularization scheme in RKKS. For this non-convex problem with the bounded hyper-sphere constraint, we can get an attainable optimal solution, which makes it possible to conduct approximation analysis in RKKS. Accordingly, we start the analysis from the learning problem that has an analytical solution, and thus obtain the first-step to understand the learning behavior in RKKS. Our analysis and experimental validation bridge the gap between the regularized risk minimization problem in RKHS and RKKS.

Appendix 1: Proof for the sample error

The asymptotic behaviors of $S_1(z,\lambda)$ and $S_2(z,\lambda)$ are usually illustrated by the convergence of the empirical mean $\frac{1}{m}\sum_{i=1}^m \xi_i$ to its expectation $\mathbb{E}\xi$, where $\left\{\xi_i\right\}_{i=1}^m$ are independent random variables on (Z,ρ) defined as

$$\xi(\boldsymbol{x},y) := \left(y - f_\lambda(\boldsymbol{x})\right)^2 - \left(y - f_\rho(\boldsymbol{x})\right)^2.$$

For $R \geq 1$, denote

$$\mathscr{W}(R) = \left\{ z \in Z^m \,:\, \sqrt{\langle f_{z,\lambda}, Tf_{z,\lambda}\rangle_{\mathcal{H}_{\mathcal{K}}}} \leq R \right\}.$$

Lemma 2 *If ξ is a symmetric real-valued function on $X \times Y$ with mean $\mathbb{E}(\xi)$. Assume that $\mathbb{E}(\xi) \geq 0$, $|\xi - \mathbb{E}\xi| \leq T$ almost surely and $\mathbb{E}\xi^2 \leq c_1'(\mathbb{E}\xi)^\theta$ for some $0 \leq \theta \leq 1$ and $c_1' \geq 0$, $T \geq 0$. Then for every $\epsilon > 0$ there holds*

$$\mathrm{Prob}\left\{ \frac{\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbb{E}\xi}{\sqrt{(\mathbb{E}\xi)^\theta + \epsilon^\theta}} \geq \epsilon^{1-\frac{\theta}{2}} \right\} \leq \exp\left\{ \frac{-m\epsilon^{2-\theta}}{2c_1' + \frac{2}{3}T\epsilon^{1-\theta}} \right\}.$$

Now we can bound $S_2(z,\lambda)$ by the following proposition.

Proposition 7 *Suppose that $|f_\rho(\boldsymbol{x})| \leq M^*$ with $M^* \geq 1$, for any $0 < \delta < 1$, there exists a subset of Z_1 of Z^m with confidence at least $1 - \delta/2$, such that for any $\forall z \in Z_1$*

$$S_2(z,\lambda) \le \frac{1}{2}D(\lambda) + \frac{1}{m}\left(\kappa\sqrt{\frac{D(\lambda)}{\lambda}} + M^* + 12\right)\log\frac{2}{\delta}.$$

Proof From the definition of f_λ in Proposition 3, combining Eqs. (1) and (4), we have

$$\|f_\lambda\|_\infty \le \kappa\sqrt{\langle f_\lambda, Tf_\lambda\rangle_{\mathcal{H}_{\bar{K}}}} \le \kappa\sqrt{\frac{D(\lambda)}{\lambda}} \le \kappa\sqrt{C_0}\lambda^{\frac{\beta-1}{2}}, \tag{27}$$

which leads to $\|f_\lambda\|_\infty \le \kappa\sqrt{\frac{D(\lambda)}{\lambda}}$. The first equality holds because the reproducing kernel $k_+ + k_-$ associated with $\mathcal{H}_{\bar{K}}$ is the square root of the limiting kernel in Guo and Shi (2019) associated with the empirical covariance operator T. Due to $f_\rho(\boldsymbol{x})$ contained in $[-M^*, M^*]$, we can get

$$\left|\xi - \mathbb{E}(\xi)\right| \le \kappa\sqrt{\frac{D(\lambda)}{\lambda}} + M^*.$$

For least squared loss, $\mathbb{E}(\xi^2) \le 4\mathbb{E}(\xi)$ indicates $c_1' = 4$ and $\theta = 1$. Applying Lemma 2, there exists a subset Z_1 of Z^m with confidence $1 - \delta/2$, we have

$$\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mathbb{E}\xi \le \sqrt{(\mathbb{E}\xi)^\theta + \epsilon^\theta}\epsilon^{1-\frac{\theta}{2}} \le \frac{1}{2}\mathbb{E}\xi + \frac{3}{2}\epsilon,$$

Then, we obtain

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mathbb{E}\xi &\le \frac{\theta}{2}\Big\{\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho)\Big\} + \frac{T + 3c_1'}{m}\log\frac{2}{\delta}\\ &\le \frac{1}{2}D(\lambda) + \frac{\kappa\sqrt{\frac{D(\lambda)}{\lambda}} + M^* + 12}{m}\log\frac{2}{\delta},\end{aligned}$$

which concludes the proof. □

In the next, we attempt to bound $S_1(z,\lambda)$ with respect to the samples z. Thus a uniform concentration inequality for a family of functions containing $f_{z,\lambda}$ is needed to estimate S_1. Since we have $f_{z,\lambda} \in \mathcal{B}_R$, which is defined by Eq. (6), we shall bound S_1 by the following proposition with a properly chosen R.

Proposition 8 *Suppose that* $|f_\rho(\boldsymbol{x})| \le M^*$ *with* $M^* \ge 1$ *in Assumption* 1, *and* ρ *satisfies the regularity condition in Assumption* 3, *for any* $0 < \delta < 1$, $R \ge 1$, $B > 0$, *there exists a subset* Z_2 *of* Z^m *with confidence at least* $1 - \delta/2$, *such that for any* $z \in \mathscr{W}(R) \cap Z_2$,

$$S_1(z,\lambda) \le \frac{136(M^* + B)}{m}\log\frac{2}{\delta} + \frac{1}{2}\Big\{\mathcal{E}\big(\pi_B(f_{z,\lambda})\big) - \mathcal{E}(f_\rho)\Big\} + 144C_s(M^* + B)m^{-\frac{1}{1+s}}R^{\frac{s}{1+s}}.$$

Proof Consider the function set $\mathcal{F}_R$ with $R > 0$ by

$$\mathcal{F}_R := \Big\{\big(y - \pi_B(f)(\boldsymbol{x})\big)^2 - \big(y - f_\rho(\boldsymbol{x})\big)^2 : f \in \mathcal{B}_R\Big\}.$$

We can easily see that each function $g \in \mathcal{F}_R$ satisfies $\|g\|_\infty \le B + M^*$, and thus we have $|g - \mathbb{E}g| \le B + M^*$.

So using $\mathscr{N}(\mathcal{F}_R, \epsilon) \le \mathscr{N}(\mathcal{B}_1, \epsilon)$ and applying Lemma 2 to the function set $\mathcal{F}_R$ with the covering number condition in Eq. (7) in Assumption 3, we have

$$\operatorname*{Prob}_{z\in Z^m}\left\{ \sup_{f\in\mathcal{F}_R} \frac{\mathbb{E}g - \frac{1}{m}\sum_{i=1}^m g(\boldsymbol{x}_i, y_i)}{\sqrt{(\mathbb{E}g)^\theta + \epsilon^\theta}} \ge 4\epsilon^{1-\frac{\theta}{2}} \right\} \le \exp\left\{ C_s\left(\frac{R}{\epsilon}\right)^s - \frac{m\epsilon^{2-\theta}}{2c_1' + \frac{2}{3}(B+M^*)\epsilon^{1-\theta}} \right\},$$

with $\mathbb{E}g = \mathcal{E}(\pi_B(f)) - \mathcal{E}(f_\rho)$. Hence there holds a subset Z_2 of Z^m with confidence at least $1-\delta/2$ such that $\forall z \in Z_2 \cap \mathscr{W}(R)$

$$\sup_{f\in\mathcal{F}_R} \frac{\mathbb{E}g - \frac{1}{m}\sum_{i=1}^m g(\boldsymbol{x}_i, y_i)}{\sqrt{(\mathbb{E}g)^\theta + \left(\epsilon^*(m,R,\frac{\delta}{2})\right)^\theta}} \le 4\left(\epsilon^*(m,R,\frac{\delta}{2})\right)^{1-\frac{\theta}{2}},$$

where $\epsilon^*(m,R,\frac{\delta}{2})$ is the smallest positive number ϵ satisfying

$$C_s\left(\frac{R}{\epsilon}\right)^s - \frac{m\epsilon^{2-\theta}}{2c_1' + \frac{2}{3}(M^*+B)\epsilon^{1-\theta}} = \log\frac{\delta}{2},$$

using Lemma 7.2 in Cucker and Zhou (2007), we have

$$\begin{aligned}\epsilon^* &\le \max\left\{ \frac{48+2(M^*+B)}{3m}\log\frac{2}{\delta}, \left(\frac{48+4(B+M^*)}{3m}C_sR^s\right)^{\frac{1}{1+s}} \right\} \\ &\le \frac{17(M^*+B)}{m}\log\frac{2}{\delta} + 18C_s(M^*+B)m^{-\frac{1}{1+s}}R^{\frac{s}{1+s}},\end{aligned}$$

where we use $M^* \ge 1$. For $z \in \mathcal{B}(R) \cap Z_2$, we have

$$S_1(z,\lambda) \le 8\epsilon^*\left(m,R,\frac{\delta}{2}\right) + \frac{1}{2}\left\{\mathcal{E}\left(\pi_B(f_{z,\lambda})\right) - \mathcal{E}(f_\rho)\right\}.$$

□

Appendix 2: Proof for learning rates

Combining the bounds in Propositions 3, 4, 7, 8, and Eq. (27), let Eq. (7) with $s > 0$, Eq. (5) with $0 < \beta \le 1$, take $\lambda = m^{-\gamma}$ with $0 < \gamma < 1$, the excess error $\mathcal{E}(\pi_B(f_{z,\lambda})) - \mathcal{E}(f_\rho)$ can be bounded by

$$\begin{aligned}\mathcal{E}\left(\pi_B(f_{z,\lambda})\right) - \mathcal{E}(f_\rho) + \lambda\langle f_{z,\lambda}, Tf_{z,\lambda}\rangle_{\mathcal{H}_K} &\le 3C_0 m^{-\gamma\beta} + \widetilde{C_1}m^{-\Theta_1} + \widetilde{C_2}\log\frac{2}{\delta}m^{-1} \\ &\quad + \widetilde{C_3}m^{-\frac{1}{1+s}}R^{\frac{s}{1+s}}\log\frac{2}{\delta} + 2\kappa\sqrt{C_0}m^{-\left(\frac{\gamma(\beta-1)}{2}+1\right)}\log\frac{2}{\delta},\end{aligned} \tag{28}$$

where $\widetilde{C_1}$ is given in Proposition 4. Two constants $\widetilde{C_2}$ and $\widetilde{C_3}$ are given by

$$\widetilde{C_2} = 274M^* + 272B + 24, \quad \widetilde{C_3} = 288(M^*+B)C_s.$$

In the next, we attempt to find a $R > 0$ by giving a bound for $\lambda\langle f_{z,\lambda}, Tf_{z,\lambda}\rangle_{\mathcal{H}_K}$.

Lemma 3 *Suppose that* ρ *satisfies the condition in Eq. (5) with* $0 < \beta \le 1$ *in Assumption 2. For some* $s > 0$ *in Assumption 3, take* $\lambda = m^{-\gamma}$ *with* $0 < \gamma \le 1$. *Then for* $0 < \epsilon < 1$ *and* $0 < \delta < 1$ *with confidence* $1 - \delta$, *we have*

$$\sqrt{\langle f_{z,\lambda}, Tf_{z,\lambda}\rangle_{\mathcal{H}_{\mathcal{K}}}} \le 4\widetilde{C_3}\widetilde{C_X}\left(\log\frac{2}{\epsilon}\right)^2\sqrt{\log\frac{2}{\delta}}m^{\theta_\epsilon}, \tag{29}$$

where $\widetilde{C_X}$ *is given by*

$$\widetilde{C_X} = \left(1 + \sqrt{\widetilde{C_2}} + \sqrt{2\kappa\sqrt{C_0}} + \sqrt{3C_0} + \sqrt{\widetilde{C_1}}\right),$$

and θ_ϵ *is*

$$\theta_\epsilon = \max\left\{\frac{\gamma(1-\beta)}{2}, \frac{1-\eta}{2}, (\gamma(1+s)-1)(2+s)+\epsilon\right\}. \tag{30}$$

Proof According to Eq. (28), we know that for any $R \ge 1$ there exists a subset V_R of Z_m with measure at most δ such that

$$\sqrt{\langle f_{z,\lambda}, Tf_{z,\lambda}\rangle_{\mathcal{H}_{\mathcal{K}}}} \le a_m R^{\frac{s}{2+2s}} + b_m, \quad \forall z \in \mathscr{W}(R)\backslash V_R,$$

where $a_m = \sqrt{\widetilde{C_3}}m^{\frac{\gamma}{2} - \frac{1}{2(1+s)}}$, and b_m is defined as

$$b_m = \left(\sqrt{\widetilde{C_2}\log\frac{2}{\delta}} + \sqrt{2\kappa\sqrt{C_0}\log\frac{2}{\delta}} + \sqrt{3C_0} + \sqrt{\widetilde{C_1}}\right)m^\zeta,$$

where the power index ζ is

$$\begin{aligned}\zeta &= \max\left\{\frac{\gamma(1-\beta)}{2}, \frac{\gamma-1}{2}, \frac{\gamma}{2} - \frac{\gamma(\beta-1)+2}{4}, \frac{1-\eta}{2}\right\}\\ &= \max\left\{\frac{\gamma(1-\beta)}{2}, \frac{1-\eta}{2}\right\}.\end{aligned}$$

It tells us that $\mathscr{W}(R) \subseteq \mathscr{W}\left(a_m R^{\frac{s}{2+2s}} + b_m\right) \bigcup V_R$. Define a sequence $\{R^{(j)}\}_{j=0}^J$ with $R^{(j)} = a_m(R^{(j-1)})^{s/(2+2s)} + b_m$ with $J \in \mathbb{N}$, we have $Z^m = \mathscr{W}(R^{(0)})$ satisfying

$$\mathscr{W}(R^{(0)}) \subseteq \mathscr{W}(R^{(1)}) \bigcup V_{R^{(0)}} \subseteq \cdots \subseteq \mathscr{W}(R^{(J)}) \bigcup \left(\bigcup_{j=0}^{J-1} V_{R^{(j)}}\right)$$

Since each set $V_{R^{(j)}}$ is at most δ, the set $\mathscr{W}(R^{(J)})$ has measure at least $1 - J\delta$.

Denote $\Delta = s/(2+2s) < 1/2$, the definition of the sequence $\{R^{(j)}\}_{j=0}^J$ indicates that

$$R^{(J)} = \underbrace{a_m^{1+\Delta+\cdots+\Delta^{J-1}}(R^{(0)})^{\Delta^J}}_{R_1^{(J)}} + \underbrace{\sum_{j=1}^{J-1} a_m^{1+\Delta+\cdots+\Delta^{j-1}} b_m^{\Delta^j} + b_m}_{R_2^{(J)}}.$$

The first term $R_1^{(J)}$ can be bounded by

$$R_1^{(J)} \leq \widetilde{C_3} m^{(\gamma(1+s)-1)(2+s)} m^{\frac{1}{1+s}2^{-J}},$$

where J is chosen to be the smallest integer satisfying $J \geq \frac{\log(1/\epsilon)}{\log 2}$. Besides, $R_2^{(J)}$ can be bounded by

$$R_2^{(J)} \leq m^{(\gamma(1+s)-1)(2+s)} \widetilde{C_3} b_1 \sum_{j=0}^{J-1} m^{\left(\zeta-(\gamma(1+s)-1)(2+s)\right)\frac{s^j}{(2+2s)^j}},$$

with $b_1 := \sqrt{\widetilde{C_2} \log \frac{2}{\delta}} + \sqrt{2\kappa\sqrt{C_0} \log \frac{2}{\delta}} + \sqrt{3C_0} + \sqrt{\widetilde{C_1}}$. When $\zeta \leq (\gamma(1+s)-1)(2+s)$, $R_2^{(J)}$ can be bounded by $\widetilde{C_3} b_1 J m^{(\gamma(1+s)-1)(2+s)}$. When $\zeta > (\gamma(1+s)-1)(2+s)$, $R_2^{(J)^2}$ can be bounded by $\widetilde{C_3} b_1 J m^{\zeta}$. Based on the above discussion, we have

$$R^{(J)} \leq (\widetilde{C_3} + \widetilde{C_3} b_1 J) m^{\theta_\epsilon},$$

with $\theta_\epsilon = \max\{\zeta, (\gamma(1+s)-1)(2+s)+\epsilon\}$. So with confidence $1 - J\delta$, there holds

$$\sqrt{\langle f_{z,\lambda}, T f_{z,\lambda} \rangle_{\mathcal{H}_K}} \leq R^{(J)} \leq \widetilde{C_3} \widetilde{C_X} J \sqrt{\log \frac{2}{\delta}} m^{\theta_\epsilon},$$

which follows by replacing δ by δ/J and noting $J \leq 2\log(2/\epsilon)$. Finally, we conclude the proof. □

Now, by Lemma 3 and Eq. (28), we are able to prove our main result in Theorem 1.

Proof Take R to be the right hand side of Eq. (29) by Lemma 3, there exists a subset V_R' of Z_m with measure at most δ such that $Z^m / V_R' \subseteq \mathscr{W}(R)$. Therefore, there exists another subset V_R of Z^m with measure at most δ such that for any $z \in \mathscr{W}(R)/V_R$, Eq. (28) can be formulated as

$$\mathcal{E}\left(\pi_B(f_{z,\lambda})\right) - \mathcal{E}(f_\rho) \leq 3C_0 m^{-\gamma\beta} + \widetilde{C_1} m^{-\Theta_1} + \widetilde{C_2} \log \frac{2}{\delta} m^{-1} + 2\kappa\sqrt{C_0} m^{-\left(\frac{\gamma(\beta-1)}{2}+1\right)} \log \frac{2}{\delta}$$
$$+ \widetilde{C_4} \left(\log \frac{2}{\epsilon}\right)^2 \sqrt{\log \frac{2}{\delta}} m^{\frac{s\theta_\epsilon - 1}{1+s}},$$

where $\widetilde{C_4} = \widetilde{C_X}(4\widetilde{C_3})^{\frac{s}{1+s}}$. Accordingly, by setting the constant $\widetilde{C}$ with

$$\widetilde{C} = 3C_0 + \widetilde{C_1} + \widetilde{C_2} + 2\kappa\sqrt{C_0} + \widetilde{C_4},$$

we have the following error bound

$$\|\pi_{M^*}(f_{z,\lambda}) - f_\rho\|_{L^2_{\rho_X}}^2 \leq \widetilde{C} \left(\log \frac{2}{\epsilon}\right)^2 \log \frac{2}{\delta} m^{-\Theta},$$

with confidence $1 - \delta$ and the power index Θ is

$$\Theta = \min\left\{\gamma\beta, \gamma + \eta - 1, \frac{1 - s\theta_\epsilon}{1+s}\right\}, \tag{31}$$

provided that $\theta_\epsilon < 1/s$. Combining Eqs. (30) and (31), when $0 < \eta < 1$, we have

$$\Theta = \min\left\{ \gamma\beta, \gamma + \eta - 1, \frac{2 - s\gamma(1-\beta)}{2(1+s)}, \frac{2 - s(1-\eta)}{2(1+s)}, \frac{1 - s(\gamma(1+s) - 1)(2+s) - s\epsilon}{1+s} \right\},$$

where ϵ is given by Eq. (8) and η needs to be further restricted by $\max\{0, 1 - 2/s\} < \eta < 1$. These two restrictions ensure that Θ is positive for a valid learning rate. Specifically, when $\eta \geq 1$, the power index Θ can be simplified as

$$\Theta = \min\left\{ \gamma\beta, \frac{2 - s\gamma(1-\beta)}{2(1+s)}, \frac{1 - s(\gamma(1+s) - 1)(2+s) - s\epsilon}{1+s} \right\},$$

which concludes the proof. □

Acknowledgements The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program/ERC Advanced Grant E-DUALITY (787960). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information. This work was supported in part by Research Council KU Leuven: Optimization frameworks for deep kernel machines C14/18/068; Flemish Government: FWO projects: GOA4917N (Deep Restricted Kernel Machines: Methods and Foundations), PhD/Postdoc grant. This research received funding from the Flemish Government (AI Research Program). This work was supported in part by Ford KU Leuven Research Alliance Project KUL0076 (Stability analysis and performance improvement of deep reinforcement learning algorithms), EU H2020 ICT-48 Network TAILOR (Foundations of Trustworthy AI - Integrating Reasoning, Learning and Optimization), Leuven.AI Institute; and in part by the National Natural Science Foundation of China (Grants Nos. 61876107, 61977046, U1803261) and NSFC/RGC Joint Research Scheme (Nos. 1201101029 and N_CityU102/20), in part by the National Key Research and Development Project (No. 2018AAA0100702), in part by National Key R&D Program of China (No. 2019YFB1311503), and in part by Shanghai Science and Technology Research Program (20JC1412700 and 19JC1420101).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Adachi, S., & Nakatsukasa, Y. (2017). Eigenvalue-based algorithm and analysis for nonconvex QCQP with one constraint. *Mathematical Programming, 1,* 1–38.

Alabdulmohsin, I., Cisse, M., Gao, X., & Zhang, X. (2016). Large margin classification with indefinite similarities. *Machine Learning, 103*(2), 215–237.

Ando, T. (2009). Projections in krein spaces. *Linear Algebra and Its Applications, 431*(12), 2346–2358.

Bach, F. (2013). Sharp analysis of low-rank kernel matrix approximations. In *Proceedings of conference on learning theory* (pp. 185–209).

Bognár, J. (1974). *Indefinite inner product spaces*. Berlin: Springer.

Boughorbel, S., Tarel, J. P., & Boujemaa, N. (2005). Conditionally positive definite kernels for SVM based image recognition. In *Proceedings of IEEE international conference on multimedia and expo* (pp. 113–116).

Chen, D., Wu, Q., Ying, Y., & Zhou, D. (2004). Support vector machine soft margin classifiers: Error analysis. *Journal of Machine Learning Research, 5*(3), 1143–1175.

Cho, H., DeMeo, B., Peng, J., & Berger, B. (2019). Large-margin classification in hyperbolic space. In *Proceedings of international conference on artificial intelligence and statistics* (pp. 1832–1840). PMLR.

Chu, K. W. E. (1986). Generalization of the Bauer-Fike theorem. *Numerische Mathematik, 49*(6), 685–691.

Cucker, F., & Zhou, D. (2007). *Learning theory: An approximation theory viewpoint* (Vol. 24). Cambridge University Press.

Dhillon, I. S., Guan, Y., & Kulis, B. (2004). Kernel k-means: Spectral clustering and normalized cuts. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 551–556). ACM.

Farooq, M., & Steinwart, I. (2019). Learning rates for kernel-based expectile regression. *Machine Learning, 108*(2), 203–227.

Gander, W., Golub, G. H., & Matt, U. V. (1988). A constrained eigenvalue problem. *Linear Algebra and Its Applications, 114–115,* 815–839.

Gao, C., Ma, Z., Ren, Z., & Zhou, H. H. (2015). Minimax estimation in sparse canonical correlation analysis. *Annals of Statistics, 43*(5), 2168–2197.

Guo, Z. C., & Shi, L. (2019). Optimal rates for coefficient-based regularized regression. *Applied and Computational Harmonic Analysis, 47*(3), 662–701.

Hoffman, A. J., & Wielandt, H. W. (2003). The variation of the spectrum of a normal matrix. In *Selected papers of Alan J Hoffman: With commentary* (pp. 118–120). World Scientific.

Huang, X., Suykens, J. A. K., Wang, S., Hornegger, J., & Maier, A. (2018). Classification with truncated ℓ_1 distance kernel. *IEEE Transactions on Neural Networks and Learning Systems, 29*(5), 2025–2030.

Jun, K. S., Cutkosky, A., & Orabona, F. (2019). Kernel truncated randomized ridge regression: Optimal rates and low noise acceleration. In *Proceedings of advances in neural information processing systems* (pp. 15358–15367).

Langer, H. (1962). Zur spektraltheoriej-selbstadjungierter operatoren. *Mathematische Annalen, 146*(1), 60–85.

Lin, S. B., Guo, X., & Zhou, D. X. (2017). Distributed learning with regularized least squares. *Journal of Machine Learning Research, 18*(1), 3202–3232.

Liu, F., Huang, X., Chen, Y., & Suykens, J. A. K. (2021). Fast learning in reproducing kernel Kreĭn spaces via generalized measures. In *Proceedings of the international conference on artificial intelligence and statistics* (pp. 388–396).

Liu, F., Huang, X., Gong, C., Yang, J., & Li, L. (2020). Learning data-adaptive non-parametric kernels. *Journal of Machine Learning Research, 21*(208), 1–39.

Liu, X., Zhu, E., & Liu, J. (2020). SimpleMKKM: Simple multiple kernel k-means. arXiv preprint arXiv:2005.04975.

Loosli, G., Canu, S., & Cheng, S. O. (2016). Learning SVM in Kreĭn spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(6), 1204–1216.

Oglic, D., & Gärtner, T. (2018). Learning in reproducing kernel Kreĭn spaces. In *Proceedings of the international conference on machine learning* (pp. 3859–3867).

Oglic, D., & Gärtner, T. (2019). Scalable learning in reproducing kernel Kreĭn spaces. In *Proceedings of international conference on machine learning* (pp. 4912–4921).

Ong, C. S., Mary, X., & Smola, A. J. (2004). Learning with non-positive kernels. In *Proceedings of the international conference on machine learning* (pp. 81–89).

Ong, C. S., Smola, A. J., & Williamson, R. C. (2005). Learning the kernel with hyperkernels. *Journal of Machine Learning Research, 6,* 1043–1071.

Pękalska, E., & Haasdonk, B. (2009). Kernel discriminant analysis for positive definite and indefinite kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(6), 1017–1032.

Pennington, J., Yu, F. X. X., & Kumar, S. (2015). Spherical random features for polynomial kernels. In *Proceedings of advances in neural information processing systems* (pp. 1846–1854).

Rudi, A., & Rosasco, L. (2017). Generalization properties of learning with random features. In *Proceedings of advances in neural information processing systems* (pp. 3215–3225.

Saha, A., & Palaniappan, B. (2020). Learning with operator-valued kernels in reproducing kernel Kreĭn spaces. In *Proceedings of advances in neural information processing systems* (pp. 1–11).

Sala, F., De Sa, C., Gu, A., & Re, C. (2018). Representation tradeoffs for hyperbolic embeddings. In *Proceedings of international conference on machine learning* (pp. 4460–4469).

Schaback, R. (1999). Native Hilbert spaces for radial basis functions. I. In *New developments in approximation theory* (pp. 255–282). Springer.

Schleif, F. M., & Tino, P. (2015). Indefinite proximity learning: A review. *Neural Computation, 27*(10), 2039–2096.

Schölkopf, B., & Smola, A. J. (2003). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press.

Shang, R., Meng, Y., Liu, C., Jiao, L., Esfahani, A. M. G., & Stolkin, R. (2019). Unsupervised feature selection based on kernel fisher discriminant analysis and regression learning. *Machine Learning, 108*(4), 659–686.

Shi, L., Huang, X., Feng, Y., & Suykens, J. A. K. (2019). Sparse kernel regression with coefficient-based ℓ_q- regularization. *Journal of Machine Learning Research, 20*(161), 1–44.

Shi, L., Huang, X., Tian, Z., & Suykens, J. A. K. (2014). Quantile regression with ℓ_1-regularization and Gaussian kernels. *Advances in Computational Mathematics, 40*(2), 517–551.

Smola, A. J., Ovari, Z. L., & Williamson, R. C. (2001). Regularization with dot-product kernels. In *Proceedings of advances in neural information processing systems* (pp. 308–314).

Steinwart, I., & Andreas, C. (2008). *Support vector machines*. Springer.

Steinwart, I., Hush, D. R., & Scovel, C. (2009). Optimal rates for regularized least squares regression. In *Proceedings of conference on learning theory* (pp. 1–10).

Steinwart, I., & Scovel, C. (2007). Fast rates for support vector machines using Gaussian kernels. *Annals of Statistics, 35*(2), 575–607.

Stewart, G. W., & Sun, J. (1990). *Matrix perturbation theory*. Harcourt Brace Jovanoich.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least squares support vector machines*. World Scientific.

Terada, Y., & Yamamoto, M. (2019). Kernel normalized cut: A theoretical revisit. In *Proceedings of international conference on machine learning* (pp. 6206–6214).

Wang, C., & Zhou, D. X. (2011). Optimal learning rates for least squares regularized regression with unbounded sampling. *Journal of Complexity, 27*(1), 55–67.

Wendland, H. (2004). *Scattered data approximation* (Vol. 17). Cambridge University Press.

Wu, Q., Ying, Y., & Zhou, D. (2006). Learning rates of least-square regularized regression. *Foundations of Computational Mathematics, 6*(2), 171–192.

Xia, Y., Wang, S., & Sheu, R. L. (2016). S-lemma with equality and its applications. *Mathematical Programming, 156*(1–2), 513–547.

Ying, Y., Campbell, C., & Girolami, M. (2009). Analysis of SVM with indefinite kernels. In *Proceedings of advances in neural information processing systems* (pp. 2205–2213).

Zhu, J., & Hastie, T. (2002). Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics, 14*(1), 185–205.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.