Journal of Machine Learning Research 17 (2016) 1-34 Submitted 7/11; Revised 5/15; Published 6/16

Sparsity and Error Analysis of Empirical Feature-Based Regularization Schemes

Xin Guo X.GUO@POLYU.EDU.HK
Department of Applied Mathematics
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong, China
Jun Fan JUNFAN@STAT.WISC.EDU
Department of Statistics
University of Wisconsin-Madison
1300 University Avenue, Madison, WI53706, USA
Ding-Xuan Zhou MAZHOU@CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong, China

Editor: Francis Bach

Abstract

We consider a learning algorithm generated by a regularization scheme with a concave regularizer for the purpose of achieving sparsity and good learning rates in a least squares regression setting. The regularization is induced for linear combinations of empirical features, constructed in the literatures of kernel principal component analysis and kernel projection machines, based on kernels and samples. In addition to the separability of the involved optimization problem caused by the empirical features, we carry out sparsity and error analysis, giving bounds in the norm of the reproducing kernel Hilbert space, based on a priori conditions which do not require assumptions on sparsity in terms of any basis or system. In particular, we show that as the concave exponent q of the concave regularizer increases to 1, the learning ability of the algorithm improves. Some numerical simulations for both artificial and real MHC-peptide binding data involving the ℓ^q regularizer and the SCAD penalty are presented to demonstrate the sparsity and error analysis.

Keywords: Sparsity, concave regularizer, reproducing kernel Hilbert space, regularization with empirical features, ℓ^q-penalty, SCAD penalty.

1. Introduction

Kernel methods provide efficient learning algorithms for analyzing nonlinear features, processing complex data, and studying data structures or relations. One may use a (unknown) probability measure ρ_X to model the distribution and structures of data on a compact metric space X (input space) and a Mercer kernel $K : X \times X \to \mathbb{R}$ to quantify by its value $K(x, u)$ similarities between two data points x and u. Then some ideas of kernel methods may be understood (Cristianini and Shawe-Taylor, 2000) in terms of eigenfunctions $\{\phi_i\}$ of the integral operator L_K defined by $L_K(f) = \int_X K(\cdot, x)f(x)d\rho_X(x)$ on the reproducing kernel Hilbert space (RKHS) $(\mathcal{H}_K, \|\cdot\|_K)$ of functions on X induced by the

kernel K. These eigenfunctions can be used to represent a feature map and provide insightful, generally nonlinear, features regarding a particular learning problem. As the data distribution ρ_X is unknown, one needs to learn or approximate the features from a data set $\mathbf{x} = \{x_i\}_{i=1}^m \subset X$ and then carries out learning tasks based on the learned data dependent approximate features.

Here we are interested in a class of data dependent features $\{\phi_i^{\mathbf{x}}\}_{i=1}^\infty$ on X, called empirical features, constructed from the data set $\mathbf{x}$ and the kernel K. They have been used in kernel principal component analysis (Schölkopf et al., 1998), kernel ridge regression (Cristianini and Shawe-Taylor, 2000; Hastie et al., 2001), kernel projection machines (Blanchard et al., 2004), and spectral algorithms (Lo Gerfo et al., 2008; Caponnetto and Yao, 2010). They are defined by means of an *empirical integral operator* $L_K^{\mathbf{x}}$ on $\mathcal{H}_K$ expressed as

$$L_K^{\mathbf{x}} f = \frac{1}{m} \sum_{i=1}^m f(x_i) K_{x_i}, \qquad f \in \mathcal{H}_K, \tag{1}$$

where $K_x := K(\cdot, x)$ is a function in $\mathcal{H}_K$ for $x \in X$. It can be seen from the reproducing property $f(x_i) = \langle f, K_{x_i} \rangle_K$ that the operator $L_K^{\mathbf{x}}$ is symmetric, positive and of rank at most m. Denote $\{(\lambda_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})\}_i$ the normalized eigenpairs of $L_K^{\mathbf{x}}$ with (possibly multiple) eigenvalues $\lambda_1^{\mathbf{x}} \ge \lambda_2^{\mathbf{x}} \ge \cdots \ge \lambda_m^{\mathbf{x}} \ge 0 = \lambda_{m+1}^{\mathbf{x}} = \cdots$, then the eigenfunctions $\{\phi_i^{\mathbf{x}}\}_i$ form an orthonormal basis of $\mathcal{H}_K$ and they are called *empirical features*.

In this paper we consider some empirical feature-based regularization schemes in a regression setting and study sparsity of these learning algorithms when the regularizer is a concave function. Here the output space is $Y = \mathbb{R}$. With a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in (X \times Y)^m$, the learning algorithm producing the output function

$$f^{\mathbf{z}} = \sum_{i=1}^\infty c_i^{\mathbf{z}} \phi_i^{\mathbf{x}} \tag{2}$$

is given in terms of its coefficient sequence $c^{\mathbf{z}} = (c_i^{\mathbf{z}})_{i=1}^\infty$ by the regularization scheme

$$c^{\mathbf{z}} = \arg\min_{c \in \ell^2} \left\{ \frac{1}{m} \sum_{i=1}^m \left(\sum_{j=1}^\infty c_j \phi_j^{\mathbf{x}}(x_i) - y_i \right)^2 + \gamma \sum_{j=1}^\infty \Omega(|c_j|) \right\}, \tag{3}$$

where $\gamma > 0$ is a regularization parameter and $\Omega : [0, \infty) \to [0, \infty)$ is a nonzero *concave* function satisfying $\Omega(0) = 0$. We shall show under some regularity assumptions that the above learning algorithm has *strong sparsity* in the sense that with confidence, the number of nonzero coefficients in the expression (2) is of order $O(m^{\theta_{sp}})$ with $0 < \theta_{sp} < 1$, much smaller than the sample size m.

The scheme (3) with special forms of regularizers can be found in the literature of kernel methods. When the regularization on the sequence $c = (c_j)_j$ is replaced by the restriction $c_j = 0$ for $j > N$, the scheme is the kernel principal component regression (Schölkopf et al., 1998) or spectral cut-off algorithm (Lo Gerfo et al., 2008; Caponnetto and Yao, 2010) where detailed error analysis can be found. The case $\Omega(|c|) = |c|^2$ corresponds to the kernel ridge regression (Cristianini and Shawe-Taylor, 2000; Hastie et al., 2001) with error analysis well conducted in a large literature (Caponnetto and De Vito, 2007; Bauer et al., 2007; Smale

and Zhou, 2007). The kernel projection machines can be expressed (Blanchard et al., 2004) by taking Ω to be the indicator function of the set $(0,\infty)$ and $\sum \Omega(|c_j|)$ to be the number of nonzero terms in the sequence c, hence correspond to the classical variable subset selection method. These algorithms were applied and analyzed for classification and regression in (Zwald, 2005; Zwald and Blanchard, 2006; Blanchard and Zwald, 2008).

A main choice of the regularizer in scheme (3) is $\Omega(|c|) = |c|^q$ with $0 < q < 2$. It can be viewed as a kernel version of the classical bridge regression (Frank and Friedman, 1993) which has advantages in some applications. To describe more details, we express the empirical features explicitly in terms of eigenpairs of the kernel (Gramian) matrix $\mathbb{K} := (K(x_i, x_j))_{i,j=1}^m$ (see e.g. Schölkopf et al. (1998); Guo and Zhou (2012)): if $\widehat{\lambda}_i^{\mathbf{x}} > 0$ is the i-th largest eigenvalue of $\mathbb{K}$ with a corresponding normalized eigenvector $\widehat{\mu}_i \in \mathbb{R}^m$, then $\lambda_i^{\mathbf{x}} = \widehat{\lambda}_i^{\mathbf{x}}/m$ and $\phi_i^{\mathbf{x}} = \sum_{j=1}^m (\widehat{\mu}_i)_j K_{x_j}/\sqrt{\widehat{\lambda}_i^{\mathbf{x}}}$. In particular, when $X \subset \mathbb{R}^n$ and K is the linear kernel $K(x,y) = x \cdot y$, we know that $\phi_i^{\mathbf{x}}$ is exactly the i-th principal component of the data matrix $A_{\mathbf{x}} = [x_1, \ldots, x_m]^T \in \mathbb{R}^{m\times n}$ and $\mathbb{K}$ is the kernel matrix $\mathbb{K} = A_{\mathbf{x}} A_{\mathbf{x}}^T$. So the scheme (3) may be viewed as regularized kernel principal component analysis (RKPCA). Moreover, a large statistical literature with the linear kernel on $\mathbb{R}^n$ reveals advantages of various methods (Frank and Friedman, 1993): principal component regression and ridge regression perform well in reducing variances when many variables together collectively effect the response with no small variable subset standing out. In particular, ridge regression (with $q = 2$ in $\Omega(|c|) = |c|^q$) has the best performance when a prior distribution of the regression vector in a Bayesian framework is Gaussian or rotationally invariant setting no preference for any particular directions. A Gaussian process interpretation can be used to understand some advantages of the kernel ridge regression. On the other hand, the variable subset selection method (with $q = 0$) has an optimal performance when the prior distribution puts the entire probability mass on the variable axes, only a few variables have influences on the response, but no information as to which ones is available. Bridge regression may have advantages when the prior distribution is concentrated along some favored directions. It also provides ways for automatic variable selection, for optimizing the power index $q \in (0,2)$ and expanding the model selection criterion by estimating jointly the optimal values of q and γ. As an extension to deal with nonlinear features in RKHSs, it is expected that the kernel bridge regression included in (3) has the same flexibility and some advantages, which will be simulated for real MHC-peptide binding data in subsection 5.2 and discussed in our sparsity and error analysis.

A crucial property of empirical features is their orthogonality with respect to the discrete measure $\frac{1}{m}\sum_{i=1}^m \delta_{x_i}$ stated as $\frac{1}{m}\sum_{i=1}^m \phi_j^{\mathbf{x}}(x_i)\phi_l^{\mathbf{x}}(x_i) = \delta_{j,l}\lambda_j^{\mathbf{x}}$. This is a classical fact and simplifies the empirical error term in (3) as $\frac{1}{m}\sum_{i=1}^m \left(\sum_{j=1}^\infty c_j \phi_j^{\mathbf{x}}(x_i) - y_i\right)^2 = \sum_{i=1}^m \lambda_i^{\mathbf{x}} c_i^2 - 2\sum_{i=1}^m \lambda_i^{\mathbf{x}} S_i^{\mathbf{z}} c_i + \frac{1}{m}\sum_{i=1}^m y_i^2$, where $S_i^{\mathbf{z}}$ is a number defined in terms of the sample $\mathbf{z}$ as

$$S_i^{\mathbf{z}} = \begin{cases} \frac{1}{m\lambda_i^{\mathbf{x}}} \sum_{j=1}^m y_j \phi_i^{\mathbf{x}}(x_j), & \text{if } \lambda_i^{\mathbf{x}} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

This simplification easily implies that the optimization problem (3) can be solved separately for each coefficient c_i, and $c_i^{\mathbf{z}} = 0$ for $i \geq m+1$. So we may replace the summations in (2) and (3) by those up to m (we keep them for the convenience of the proofs).

Theorem 1 *Let $\Omega : [0,\infty) \to [0,\infty)$, $\gamma > 0$ and $\mathbf{z} \in (X \times Y)^m$. Then a sequence $c^{\mathbf{z}} = (c_i^{\mathbf{z}})_{i=1}^{\infty}$ is a solution to (3) if and only if for each i, $c_i^{\mathbf{z}}$ is a minimizer of the univariate function defined by*

$$h_i(c) = h_{\lambda_i^{\mathbf{x}}, S_i^{\mathbf{z}}, \gamma, \Omega}(c) = \lambda_i^{\mathbf{x}}(c - S_i^{\mathbf{z}})^2 + \gamma\Omega(|c|), \qquad c \in \mathbb{R}. \tag{5}$$

2. Main Results on Sparsity and Error Analysis

The main purpose of this paper is to show that both strong sparsity and fast learning rate can be achieved by the learning algorithm (2) when the regularizing function Ω in (3) is concave. We describe the main ideas in this section and will provide detailed general analysis in Section 4 while some numerical simulations for both artificial and real data will be presented in Section 5.

2.1 Concave regularizing functions

The concavity of regularizing functions plays a central role in achieving sparsity in this paper. It has the following nice property.

Theorem 2 *If $\Omega : [0,\infty) \to [0,\infty)$ is a nonzero continuous concave function satisfying $\Omega(0) = 0$, then $\Omega(1) > 0$, and that*

$$\Omega(c) \geq \Omega(1)c, \qquad \forall c \in (0,1] \tag{6}$$

and

$$\Omega(c) \leq \Omega(1)c, \qquad \forall c \in [1,\infty). \tag{7}$$

Theorem 2 is part of Proposition 10 in Section 3 which will give more properties for concave regularizing functions.

Note that (6) is a lower bound for Ω on $(0,1]$. Our error bounds will be presented by means of the asymptotic behavior of the concave regularizing function Ω near the origin, which is characterized by a concave exponent $q \in [0,1]$.

Definition 3 *We say that a concave regularizing function Ω has a concave exponent $q \in [0,1]$ if there is a positive constant C_Ω^* such that*

$$\Omega(c) \leq C_\Omega^* c^q, \qquad \forall c \in (0,1]. \tag{8}$$

Theorem 2 tells us that the concave exponent q in (8) is at most 1. We also know from Proposition 10 in Section 3 that (8) is always true with $q = 0$ and $C_\Omega^* = \Omega(1)$. Sharper error bounds with better q are possible. The following are two such families of concave regularizing functions: ℓ^q-regularizer ($0 < q \leq 1$) which is well studied for bridge regression (Frank and Friedman, 1993; Fu and Knight, 2000; Liu et al., 2007; Xu et al., 2012), and SCAD penalties (Fan and Li, 2001).

Example 1 *Let $0 < q \leq 1$ and $\Omega : [0,\infty) \to [0,\infty)$ be the ℓ^q-regularizer given by $\Omega(c) = c^q$. Then (8) is satisfied with $C_\Omega^* = 1$.*

Example 2 *Let $b > 2$ and $\Omega : [0,\infty) \to [0,\infty)$ be a SCAD penalty given as a concave continuous function by $\Omega(0) = 0$ and*

$$\Omega'(c) = \begin{cases} 1, & \text{for } 0 < c < 1, \\ \frac{c-b}{1-b}, & \text{for } 1 < c < b, \\ 0, & \text{for } c > b. \end{cases}$$

Then $\Omega(c) = c$ for $c \in [0,1]$, $\Omega(c) = \frac{1+b}{2} - \frac{(c-b)^2}{2(b-1)}$ for $c \in (1,b]$ and $\Omega(c) = \frac{1+b}{2}$ for $c \in (b,\infty)$. Hence (8) is satisfied with $q = 1$ and $C_\Omega^* = 1$. Moreover, we have $\Omega(c) \le \frac{1+b}{2}$ for every $c \in [1,\infty)$.

In our results for sparsity and error analysis, we shall use a general power index $q \in [0,1]$ instead of the universal choice of $q = 0$.

2.2 Sparsity and learning rates

Throughout the paper, we assume that the sample set $\mathbf{z}$ is drawn independently according to a Borel probability measure ρ on $X \times Y$ and that for some constant $M > 0$, $|y| \le M$ almost surely. The regression function in our regression setting is defined as a function f_ρ on X given by

$$f_\rho(x) = \int_Y y d\rho(y|x), \qquad x \in X,$$

where $\rho(\cdot|x)$ is the conditional measure induced by ρ at $x \in X$. The regularity assumption we shall take for the regression function is

$$f_\rho = L_K^r(g_\rho) \quad \text{for some } r > 0 \text{ and } g_\rho \in \mathcal{H}_K. \tag{9}$$

Here L_K is a compact, self-adjoint and positive operator on $\mathcal{H}_K$ having eigenpairs $\{(\lambda_i, \phi_i)\}_i$ with the eigenvalues $\{\lambda_i\}$ forming a nonincreasing sequence tending to 0 and and eigenfunctions $\{\phi_i\}$ an orthonormal basis of $\mathcal{H}_K$. Its r-th power L_K^r is given by $L_K^r(\sum_i c_i\phi_i) = \sum_i c_i\lambda_i^r\phi_i$ and assumption (9) means $f_\rho = \sum_i d_i\lambda_i^r\phi_i$ for some sequence $\{d_i\} \in \ell^2$ representing $g_\rho = \sum_i d_i\phi_i$. The exponent r in (9) measures the decay of the coefficients $\{d_i\lambda_i^r\}$ of f_ρ with respect to the orthonormal basis $\{\phi_i\}$ of $\mathcal{H}_K$, and thereby the regularity of the regression function f_ρ.

Let us illustrate our general analysis for strong sparsity and learning rates by two special cases, derived from Corollary 16 (with $\alpha_1 = \alpha_2 = \alpha$) and Corollary 17 (with $\beta_1 = \beta_2 = \beta$) in Section 4, for which the eigenvalues of the integral operator L_K decay polynomially or exponentially.

Theorem 4 *Assume (9) with $r > \frac{1}{2}$, and that Ω has a concave exponent $q \in [0,1]$ with (8) valid. Suppose that for some positive constants D_1, D_2 and α, the eigenvalues $\{\lambda_i\}$ of L_K decay polynomially as*

$$D_1 i^{-\alpha} \le \lambda_i \le D_2 i^{-\alpha}, \qquad \forall i \in \mathbb{N} \tag{10}$$

with $2\alpha \max\{r,1\} > 1$. Let $0 < \delta < 1$. If we choose

$$\gamma = C_1(D_2/\lambda_1)^{r+1}\left(\log\frac{4m}{\delta}\right)^{1+2r} m^{-\frac{1+r}{1+2r}}, \tag{11}$$

then with confidence $1-\delta$ *we have*

$$c_i^{\mathbf{z}} = 0, \qquad \forall\ m^{\theta_{sp}} + 1 \le i \le m \quad \textit{with } \theta_{sp} = \frac{1}{\alpha(1+2r)} < 1 \tag{12}$$

and

$$\|f^{\mathbf{z}} - f_\rho\|_K \le C_2 \left(\log \frac{4m}{\delta}\right)^{1+2r} m^{-\theta_{rate}}, \qquad \theta_{rate} = \frac{\alpha \min\{4r, 4r(2-q)\} - 2(2-q)}{4(2r+1)(2-q)\alpha},$$

where C_1 *and* C_2 *are constants independent of* m *or* δ *(to be specified in the proof).*

The eigenvalue decay condition (10) is typical for Sobolev smooth kernels on domains in Euclidean spaces, with the power index α depending on the smoothness of the kernel (Reade, 1984).

The regularity assumptions (9) and (10) impose restrictions on the concave exponent $q \in [0,1]$. To see this, we express $g_\rho = \sum_i d_i \phi_i$ with $(d_i)_i \in \ell^2$ and $f_\rho = L_K^r(g_\rho) = \sum_i \lambda_i^r d_i \phi_i$. A natural requirement for f_ρ corresponding to the ℓ^q-regularizer is $(\lambda_i^r d_i)_i \in \ell^q$. Imposing this uniformly with respect to the coefficient sequence $(d_i)_i$ is the same as the boundedness from ℓ^2 to ℓ^q of the diagonal operator D_{λ^r} associated with the fixed non-indreasing sequence $(\lambda_i^r)_i$. This problem together with asymptotic behaviors of the entropy numbers of D_{λ^r} has been widely studied in the literature of function spaces and approximation theory (Edmunds and Triebel, 1996; Kühn, 2008) and the boundedness can be characterized by the condition

$$(\lambda_i^r)_i \in \ell^s \text{ with } \frac{1}{s} = \frac{1}{q} - \frac{1}{2}. \tag{13}$$

Under the eigenvalue decay assumption (10), the characterization condition (13) is equivalent to $\sum_{i=1}^\infty i^{-\alpha r s} = \sum_{i=1}^\infty i^{-\frac{2\alpha r q}{2-q}} < \infty$, which can be stated as

$$q > \frac{2}{2\alpha r + 1}. \tag{14}$$

Thus the concave exponent q is tailored to the regularity assumption and the eigenvalue decay, and a larger regularity index r leads to a wider range of the concave exponent q.

Combining the regularity assumption (9) and the eigenvalue decay condition (10) has been an approach for error analysis of learning algorithms. In particular, the minimax rates of convergence in the $L^2_{\rho_X}$ metric was derived in (Caponnetto and De Vito, 2007) under these conditions with the restrictions $\alpha > 1$ and $0 < r \le \frac{1}{2}$. Moreover, the well-known regularized least squares regression (RLS) scheme

$$f^{\mathbf{z}} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \gamma \|f\|_K^2 \right\} \tag{15}$$

achieves these rates in probability as $\|f^{\mathbf{z}} - f_\rho\|^2_{L^2_{\rho_X}} = O(m^{-\frac{\alpha(2r+1)}{\alpha(2r+1)+1}})$. Error estimates in the $\mathcal{H}_K$ metric provide error analysis for the distribution mismatch problem (where the distribution for predictions might be different from the sampling distribution ρ_X) and

for sampling processes with nonidentical distributions (Smale and Zhou, 2009; Zhou, 2003). Such estimates for the RLS algorithm (15) were conducted in (Smale and Zhou, 2007; Bauer et al., 2007) where the learning rates are $\|f^{\mathbf{z}}-f_\rho\|_K = O(m^{-\frac{r}{2r+2}})$ under the same restriction $0 < r \leq \frac{1}{2}$ and the maximum exponent is $\frac{1}{6}$ when $r = \frac{1}{2}$. The maximum exponent for the RLS algorithm (15) cannot be improved further for $r > \frac{1}{2}$ and this is called a saturation effect in the theory of inverse problems (Bauer et al., 2007).

As pointed out in (Bauer et al., 2007; Lo Gerfo et al., 2008), spectral cut-off algorithms do not suffer from the saturation phenomenon. Theorem 4 confirms this advantage for the algorithm (2) in the range $r > \frac{1}{2}$ (the range $0 < r \leq \frac{1}{2}$ is covered by Corollary 16 in Section 4). To be specific, let $\frac{1}{2} < q \leq 1$ and $r \geq \frac{1}{4q-2}$. Then the power index θ_{rate} for the learning rate in Theorem 4 is

$$\theta_{rate} = \frac{2r\alpha - 2 + q}{2(2r+1)(2-q)\alpha} \tag{16}$$

which becomes larger as the regularity index r increases, and can be arbitrarily close to $\frac{1}{2}$ when r is large enough (f_ρ is smooth enough) and $q = 1$. This applied to the case when Ω is the SCAD penalty given in Example 2. Even in the range $0 < q \leq \frac{1}{2}$, for a sufficiently large r, the power index θ_{rate} in Theorem 4 can be arbitrarily close to $\frac{1}{2(2-q)}$.

The estimate (12) for sparsity in Theorem 4 tells us that with confidence, the output function $f^{\mathbf{z}} = \sum c_i^{\mathbf{z}} \phi_i^{\mathbf{z}}$ has at most $m^{\theta_{sp}}$ nonzero coefficients with a sparsity exponent $\theta_{sp} < 1$, a small proportion of the m coefficients in the expression (2). Moreover, θ_{sp} decreases, leading to better sparsity, as r increases. Note that by our analysis, the restriction (14) is the only influence of the concave exponent q for the sparsity.

Theorem 5 *Assume (9) with $r > \frac{1}{2}$, and that Ω has a concave exponent $q \in [0, 1]$ with (8) valid. Suppose that for some positive constants D_1, D_2 and β, the eigenvalues $\{\lambda_i\}$ of L_K decay exponentially as*

$$D_1\beta^{-i} \leq \lambda_i \leq D_2\beta^{-i}, \qquad \forall i \in \mathbb{N}. \tag{17}$$

Let $0 < \delta < 1$. If we choose γ as (11), then with confidence $1 - \delta$ we have

$$c_i^{\mathbf{z}} = 0, \qquad \forall\ \frac{\log(m+1)}{(1+2r)\log\beta} + 1 \leq i \leq m \tag{18}$$

and

$$\|f^{\mathbf{z}} - f_\rho\|_K \leq C_2 \left(\log\frac{4m}{\delta}\right)^{2r+1} m^{-\theta_{rate}}, \qquad \theta_{rate} = \min\left\{\frac{r}{(2-q)(1+2r)}, \frac{(2-q)r}{(2-q)(1+2r)}\right\},$$

where C_2 is a constant independent of m or δ (to be specified in the proof).

Remark 6 *The eigenvalue decay condition (17) is typical for analytic kernels on domains in Euclidean spaces (Reade, 1984). When the regularity index r is large enough, the power index θ_{rate} for the learning rate is $\frac{1}{2(2-q)} - \epsilon$ with an arbitrarily small $\epsilon > 0$. So the learning rate depends on the concave exponent q, better as q increases. On the other hand, (18) tells us that with confidence, the output function $f^{\mathbf{z}} = \sum c_i^{\mathbf{z}} \phi_i^{\mathbf{z}}$ has at most $\frac{\log(m+1)}{(1+2r)\log\beta}$ nonzero coefficients, a logarithmic proportion of the m coefficients in the expression (2).*

2.3 Minimax lower bound

The learning rate stated in Theorem 4 is close to be optimal when r is large. One might use some existing methods for dealing with the $L^2_{\rho_X}$ error in the literature (Yang and Barron, 1999; Bauer et al., 2007; Caponnetto and De Vito, 2007; DeVore et al., 2004; Suzuki et al., 2012; Raskutti et al., 2012; Steinwart et al., 2009) to give lower bounds. Here we focus on the error in the $\mathcal{H}_K$-metric and present a minimax lower bound. Denote $\kappa = \max_{x\in X}\sqrt{K(x,x)}$.

Definition 7 *Let $\mathcal{P}(\alpha, r, M, R, D_1, D_2)$ be the set of all Borel probability measures ρ on $X \times Y$ such that the regularity assumption (9) is satisfied with $\|g_\rho\|_K \le R$, (10) holds true, and the conditional measure $\rho(\cdot|x)$ is supported on $[-M, M]$ for almost all $x \in X$.*

Theorem 8 *Let α, r, R, D_1, D_2 be positive constants and $M \ge 4\kappa^{r+\frac{1}{2}}R$. Let $f^{\mathbf{z}} \in \mathcal{H}_K$ be the output of an arbitrary learning algorithm based on the sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$. Then for every $0 < \delta < 1$, there exists a positive constant $C_{\delta,\alpha,r,M,R,D_1,D_2}$ such that*

$$\lim_{m\to\infty}\inf_{f^{\mathbf{z}}}\sup_{\rho\in\mathcal{P}(\alpha,r,M,R,D_1,D_2)}\mathbb{P}_{\mathbf{z}\sim\rho^m}\left\{\|f^{\mathbf{z}} - f_\rho\|_K \ge C_{\delta,\alpha,r,M,R,D_1,D_2}m^{-\frac{\alpha r}{\alpha(1+2r)+1}}\right\} \ge 1-\delta.$$

The proof of Theorem 8 follows from a more general result to be given in Appendix B. The power index $\frac{\alpha r}{\alpha(2r+1)+1}$ for the minimax lower bound stated in Theorem 8 corresponds to the upper bound index (16) in Theorem 4 for a smoother regularity class with $r' = r + \frac{2-q}{2\alpha}$. This shows the gap between our upper bound and the minimax lower bound. It would be interesting to derive minimax rates of convergence in the $\mathcal{H}_K$-metric which can be achieved by the learning algorithm (2) with $\Omega(c) = c^q$ for $0 < q \le 1$.

2.4 Connections to ridge regression and some other learning algorithms

The classical RLS algorithm (15) can be stated as the scheme (2) by taking the regularizer $\Omega(c) = c^2$ corresponding to the ridge regression. This follows from a representer theorem for (15), the identities $\mathrm{span}\{K_{x_j}\}_{j=1}^m = \mathrm{span}\{\phi_j^{\mathbf{x}}\}_{j=1}^m$ and $\left\|\sum_{j=1}^\infty c_j\phi_j^{\mathbf{x}}\right\|_K^2 = \sum_{j=1}^\infty |c_j|^2$.

The regularizer $\Omega(c) = c^q$ with $0 < q \le 2$ correspond to the bridge regression. When $1 < q \le 2$, this regularizer is convex instead of being concave. It has the special property that $\Omega'_+(0) = 0$ where $\Omega'_+(c)$ denotes the right-side derivative of Ω at $c \in [0, \infty)$. This leads to the observation that sparsity is hardly achieved for the learning algorithm (2) associated with such a convex regularizer.

Theorem 9 *Let $\Omega : [0, \infty) \to [0, \infty)$, $\gamma > 0$ and $\Omega(0) = 0$. If $\Omega'_+(0) = 0$, then for each i, $c_i^{\mathbf{z}}$ vanishes if and only if either $\lambda_i^{\mathbf{x}} = 0$ or $S_i^{\mathbf{z}} = 0$.*

An elastic net learning algorithm (Zou and Hastie, 2005) can be introduced by taking the regularizer in (3) as

$$\Omega^{en}(c) = c + \zeta c^2, \tag{19}$$

where $\zeta > 0$ is an elastic net parameter controlling the proportion of the ℓ^2-norm square in the regularizer Ω^{en}. Though the regularizer Ω^{en} is strictly convex, it does not satisfy the assumption $\Omega'_+(0) = 0$ in Theorem 9. When ζ is small, this regularizer is actually close to

the ℓ^1-penalty. Hence we would expect that the corresponding learning algorithm with a strictly convex regularizer has strong sparsity. This is beyond the discussion in this paper.

Let us mention that the learning scheme (2) is closely related to spectral algorithms (Lo Gerfo et al., 2008; Caponnetto and Yao, 2010) which can be stated in terms of the empirical features $\{\phi_j^{\mathbf{x}}\}_{j=1}^m$ and a filter function $g_\gamma : [0,1] \to \mathbb{R}$ as

$$f^{\mathbf{z}} = \sum_{j=1}^m \sqrt{\lambda_j^{\mathbf{x}}/m} \left(\sum_{i=1}^m (\widehat{\mu}_j)_i y_i \right) g_\gamma(\lambda_j^{\mathbf{x}}) \phi_j^{\mathbf{x}},$$

where $\{(m\lambda_j^{\mathbf{x}}, \widehat{\mu}_j)\}$ are the normalized eigenpairs of the kernel matrix $\mathbb{K}$.

Our analysis relies heavily on the special form of the least squares loss, as seen from Theorem 1. It would be interesting to establish similar analysis for schemes associated with other loss functions such as those in the minimum error entropy principle, at least when the scaling parameter is large (Hu et al., 2015).

3. Properties of Concave Regularizing Functions

In this section we give some properties of concave regularizing functions, and then estimate the solution $c^{\mathbf{z}}$ to (3) by means of the explicit expression stated in Theorem 1.

Proposition 10 *Let $\Omega : [0,\infty) \to [0,\infty)$ be a nonzero continuous concave function satisfying $\Omega(0) = 0$. Then it has the following properties.*

(a) The function Ω is nondecreasing on $[0,\infty)$, and $\Omega(c) > 0$ for $c \in (0,\infty)$. The right-hand derivative Ω'_+ is well defined, nonincreasing, finite, and nonnegative on $(0,\infty)$. At the origin, $\Omega'_+(0) \in (0,\infty]$.

(b) We have $\Omega(c) \geq \Omega(1)c$ for $c \in [0,1]$, and $\Omega(c) \leq \Omega(1)c$ for $c \in [1,\infty)$.

(c) There holds $\Omega(a+b) \leq \Omega(a) + \Omega(b)$ for any $a, b > 0$.

(d) The positive function $\frac{\Omega(c)}{c}$ defined on $(0,\infty)$ is nonincreasing and satisfies $\lim_{c\to 0^+} \frac{\Omega(c)}{c} = \Omega'_+(0)$.

(e) The positive function $\frac{\Omega(c)}{c^2}$ defined on $(0,\infty)$ is continuous and strictly decreasing from $\lim_{c\to 0^+} \frac{\Omega(c)}{c^2} = +\infty$ to $\lim_{c\to\infty} \frac{\Omega(c)}{c^2} = 0$.

Proposition 10 will be proved in Appendix C.

For our analysis, we need the following two auxiliary functions.

Definition 11 *Define an auxiliary function $\Omega^* : (0,\infty) \to (0,\infty)$ of a positive function Ω as*

$$\Omega^*(\lambda) = \inf_{c\in(0,\infty)} \left\{ \frac{\Omega(c)}{c} + \lambda c \right\}, \qquad \lambda \in (0,\infty).$$

Define another auxiliary function $\widetilde{\Omega} : (0,\infty) \to (0,\infty)$ as

$$\widetilde{\Omega}(\lambda) = \arg\sup \left\{ c \in (0,\infty) : \frac{\Omega(c)}{c^2} \geq \lambda \right\}, \qquad \lambda \in (0,\infty).$$

Remark 12 *The value $-\Omega^*(\lambda)$ is exactly equal to the value at the point $-\lambda$ of the conjugate function of $\frac{\Omega(c)}{c}$ defined in the literature of optimization.*

We can now estimate the solution $c^{\mathbf{z}}$ to (3) in terms of $S_i^{\mathbf{z}}, \lambda_i^{\mathbf{x}}$ and γ, by means of the explicit expression stated in Theorem 1.

Theorem 13 *Let $\gamma > 0$ and $\Omega : [0,\infty) \to [0,\infty)$ be a nonzero continuous concave function satisfying $\Omega(0) = 0$.*

(a) Both functions Ω^* and $\widetilde{\Omega}$ are well-defined and positive on $(0,\infty)$. The function Ω^* is nondecreasing while $\widetilde{\Omega}$ is non-increasing.

(b) Let $i \in \mathbb{N}$. If

$$|S_i^{\mathbf{z}}| < \frac{\Omega^*(\frac{\lambda_i^{\mathbf{x}}}{\gamma})}{2\frac{\lambda_i^{\mathbf{x}}}{\gamma}}, \tag{20}$$

then $c_i^{\mathbf{z}} = 0$. If $|S_i^{\mathbf{z}}| > \frac{\Omega^*(\frac{\lambda_i^{\mathbf{x}}}{\gamma})}{2\frac{\lambda_i^{\mathbf{x}}}{\gamma}}$, then $c_i^{\mathbf{z}}$ has the same sign as $S_i^{\mathbf{z}}$ and satisfies $|S_i^{\mathbf{z}}| - \widetilde{\Omega}\left(\frac{\lambda_i^{\mathbf{x}}}{\gamma}\right) \le |c_i^{\mathbf{z}}| \le |S_i^{\mathbf{z}}|$.

(c) Let $d^{\mathbf{x}} \le m$ be the rank of the Gramian matrix $\mathbb{K}$. Then $\lambda_i^{\mathbf{x}} = 0$ if and only if $i > d^{\mathbf{x}}$. Hence $c_i^{\mathbf{z}} = 0$ for $i > d^{\mathbf{x}}$.

Proof (a) The first statement follows easily from the definitions of the auxiliary functions and Proposition 10.

(b) Since $\gamma > 0$, when $\lambda_i^{\mathbf{x}} = 0$ or $S_i^{\mathbf{z}} = 0$, our statement follows from Theorem 1. So we consider the case that $\lambda_i^{\mathbf{x}} > 0$ and $S_i^{\mathbf{z}} \neq 0$. By symmetry we only need to prove our statement for the case $S_i^{\mathbf{z}} > 0$.

With $\lambda_i^{\mathbf{x}} > 0$ and $S_i^{\mathbf{z}} > 0$, we find that the left-side derivative of the function h_i is $(h_i)'_-(c) = 2\lambda_i^{\mathbf{x}}(c - S_i^{\mathbf{z}}) - \gamma\Omega'_+(|c|) < 0$ for $c \in (-\infty, 0]$, hence all its possible minimizers are achieved on $[0,\infty)$. Let us consider the difference function $h_i(c) - h_i(0)$ for $c > 0$ and factorize it as

$$h_i(c) - h_i(0) = c g_i(c), \quad \text{where} \quad g_i(c) := \gamma\frac{\Omega(c)}{c} + \lambda_i^{\mathbf{x}} c - 2\lambda_i^{\mathbf{x}} S_i^{\mathbf{z}}. \tag{21}$$

If $|S_i^{\mathbf{z}}| < \frac{\Omega^*(\frac{\lambda_i^{\mathbf{x}}}{\gamma})}{2\frac{\lambda_i^{\mathbf{x}}}{\gamma}}$, then $\inf_{c>0} g_i(c) > 0$ which implies $h_i(c) - h_i(0) = c g_i(c) > 0$ for every $c > 0$. Hence in this case h_i has the only minimizer at $0 = c_i^{\mathbf{z}}$.

If $|S_i^{\mathbf{z}}| > \frac{\Omega^*(\frac{\lambda_i^{\mathbf{x}}}{\gamma})}{2\frac{\lambda_i^{\mathbf{x}}}{\gamma}}$, then $\gamma \inf_{c>0}\left\{\frac{\Omega(c)}{c} + \frac{\lambda_i^{\mathbf{x}}}{\gamma} c\right\} < 2\lambda_i^{\mathbf{x}} S_i^{\mathbf{z}}$ meaning that $\inf_{c>0} g_i(c) < 0$. It follows that a minimizer c_* of the function g_i on $(0,\infty)$ satisfies $g_i(c_*) < 0$. Hence $h_i(c_*) - h_i(0) = c_* g_i(c_*) < 0$. So 0 is not a minimizer of h_i.

Since Ω is nondecreasing on $[0,\infty)$, we know that h_i is strictly increasing on $(S_i^{\mathbf{z}}, \infty)$. Hence the minimizer $c_i^{\mathbf{z}}$ of h_i satisfies $0 < c_i^{\mathbf{z}} \le S_i^{\mathbf{z}}$. We also know from $h_i(c_i^{\mathbf{z}}) \le h_i(S_i^{\mathbf{z}})$ that

$$h_i(c_i^{\mathbf{z}}) = \lambda_i^{\mathbf{x}}\,(c_i^{\mathbf{z}} - S_i^{\mathbf{z}})^2 + \gamma\Omega(c_i^{\mathbf{z}}) \le h_i(S_i^{\mathbf{z}}) = \lambda_i^{\mathbf{x}}\,(S_i^{\mathbf{z}} - S_i^{\mathbf{z}})^2 + \gamma\Omega(S_i^{\mathbf{z}}) = \gamma\Omega(S_i^{\mathbf{z}}).$$

Express $S_i^{\mathbf{z}}$ as $S_i^{\mathbf{z}} - c_i^{\mathbf{z}} + c_i^{\mathbf{z}}$. Proposition 10 (c) yields $\Omega(S_i^{\mathbf{z}}) = \Omega(S_i^{\mathbf{z}} - c_i^{\mathbf{z}} + c_i^{\mathbf{z}}) \leq \Omega(c_i^{\mathbf{z}}) + \Omega(S_i^{\mathbf{z}} - c_i^{\mathbf{z}})$. It follows that

$$\lambda_i^{\mathbf{x}} \left(c_i^{\mathbf{z}} - S_i^{\mathbf{z}}\right)^2 \leq \gamma \Omega(S_i^{\mathbf{z}} - c_i^{\mathbf{z}}).$$

Therefore,

$$\frac{\Omega(S_i^{\mathbf{z}} - c_i^{\mathbf{z}})}{\left(S_i^{\mathbf{z}} - c_i^{\mathbf{z}}\right)^2} \geq \frac{\lambda_i^{\mathbf{x}}}{\gamma}.$$

By the definition of the function $\widetilde{\Omega}$, this implies that $S_i^{\mathbf{z}} - c_i^{\mathbf{z}} \leq \widetilde{\Omega}\left(\frac{\lambda_i^{\mathbf{x}}}{\gamma}\right)$. This proves the range of $c_i^{\mathbf{z}}$ and verifies out second statement.

(c) It is well-known (e. g. Guo and Zhou (2012)) that the first $d^{\mathbf{x}}$ eigenvalues of the matrix $\mathbb{K}$ are given by $\{m\lambda_i^{\mathbf{x}}\}_{i=1}^{d^{\mathbf{x}}}$ while $\lambda_i^{\mathbf{x}} = 0$ for $i \geq d^{\mathbf{x}} + 1$. So $\lambda_i^{\mathbf{x}} = 0$ if and only if $i > d^{\mathbf{x}}$. In this case, condition (20) is satisfied and by the conclusion in part (b), $c_i^{\mathbf{z}} = 0$. The proof of Theorem 13 is thus complete. ■

4. General Analysis for Sparsity and Error Bounds

In this section we present a general result on sparsity and error bounds for the learning algorithm (2) generated by the regularization scheme (3) based on empirical features and concave regularizing functions. To this end, we need the following bounds for the auxiliary functions Ω^* and $\widetilde{\Omega}$.

Lemma 14 *If $\Omega : [0,\infty) \to [0,\infty)$ is a nonzero continuous concave function satisfying $\Omega(0) = 0$, then there exists a positive constant $C_{\Omega,1}$ such that*

$$\Omega^*(\lambda) \geq C_{\Omega,1} \min\{\sqrt{\lambda}, 1\}, \qquad \forall \lambda > 0. \tag{22}$$

If moreover, Ω has a concave exponent $q \in [0,1]$ with (8) valid, then there exists a positive constant $C_{\Omega,2}$ such that

$$\widetilde{\Omega}(\lambda) \leq C_{\Omega,2} \max\left\{\left(\frac{1}{\lambda}\right)^{1/(2-q)}, \frac{1}{\lambda}\right\} \qquad \forall \lambda > 0. \tag{23}$$

Proof For $c \in (0,1]$, we apply Proposition 10 (d) and find

$$\frac{\Omega(c)}{c} + \lambda c \geq \Omega(1) + \lambda c \geq \Omega(1) \geq \Omega(1) \min\{\sqrt{\lambda}, 1\}.$$

For $c \in (1,\infty)$, we have $\Omega(c) \geq \Omega(1)$. Then $\frac{\Omega(c)}{c} + \lambda c \geq \frac{\Omega(1)}{c} + \lambda c \geq 2\sqrt{\Omega(1)\lambda} \geq 2\sqrt{\Omega(1)} \min\{\sqrt{\lambda}, 1\}$. Thus (22) holds with $C_{\Omega,1} = \max\{\Omega(1), 2\sqrt{\Omega(1)}\}$.

To prove (23), we let $\lambda \in (0,\infty)$. Denote $\widetilde{\Omega}(\lambda)$ as c^*. We know from the definition of $\widetilde{\Omega}(\lambda)$ that $\frac{\Omega(c^*)}{(c^*)^2} \geq \lambda$.

When $c^* \leq 1$, we use condition (8) and find $\Omega(c^*) \leq C_\Omega^* (c^*)^q$. But $\Omega(c^*) \geq \lambda (c^*)^2$. So $C_\Omega^* (c^*)^q \geq \lambda (c^*)^2$ and $c^* \leq (C_\Omega^*)^{1/(2-q)} \left(\frac{1}{\lambda}\right)^{1/(2-q)}$.

When $c^* > 1$, we apply (7) in Theorem 2 to c^* and obtain $\lambda \le \frac{\Omega(c^*)}{(c^*)^2} \le \frac{\Omega(1)c^*}{(c^*)^2} \le \frac{\Omega(1)}{c^*}$ and thereby $c^* \le \frac{\Omega(1)}{\lambda}$. Combing the above two cases, we know that (23) is valid with $C_{\Omega,2} = \max\{(C_\Omega^*)^{1/(2-q)}, \Omega(1)\}$. This proves the lemma. ■

Theorem 15 *Assume (9) with $r > 0$, and that Ω has a concave exponent $q \in [0,1]$ with (8) valid. If $0 < \delta \le 1$ and for some $1 \le p \le m$, the regularization parameter γ satisfies*

$$\gamma \ge \begin{cases} C_1 \left(\log \frac{4m}{\delta}\right)^{1+2r} \left(\max\left\{\frac{\lambda_p}{\lambda_1}, \frac{1}{\sqrt{m}}\right\}\right)^{r+1}, & \text{if } 0 < r \le \frac{1}{2}, \\ C_1 \left(\log \frac{4m}{\delta}\right)^{1+2r} \max\left\{\left(\frac{\lambda_p}{\lambda_1}\right)^{r+\frac{1}{2}}, \frac{1}{\sqrt{m}}\right\} \left(\max\left\{\frac{\lambda_p}{\lambda_1}, \frac{1}{\sqrt{m}}\right\}\right)^{\frac{1}{2}}, & \text{if } r > \frac{1}{2}, \end{cases} \tag{24}$$

then with confidence $1-\delta$ we have

$$c_i^{\mathbf{z}} = 0, \quad \forall i = p+1, \ldots, m$$

and

$$\begin{aligned} \|f^{\mathbf{z}} - f_\rho\|_K \;\le\; & C_{\Omega,2}\sqrt{p}\left\{\left(\frac{2\gamma}{\lambda_p}\right)^{\frac{1}{2-q}} + \frac{2\gamma}{\lambda_p}\right\} + \|g_\rho\|_K \lambda_p^r + C_3 \frac{\sqrt{p}\log\frac{4m}{\delta}}{\sqrt{m}} \lambda_p^{\min\{-1/2, r-1\}} \\ & + C_4 \lambda_p^{\min\{r-1,0\}} \left(\sum_{i=p+1}^{\infty} \lambda_i^{2\max\{r,1\}}\right)^{1/2}, \end{aligned} \tag{25}$$

where $C_1 \ge 1$, C_3 and C_4 are constants independent of γ, p, δ, or m.

The detailed proof of Theorem 15 will be given in Appendix A where the constants C_1, C_3 and C_4 will be specified explicitly. Here we outline the ideas of the proof by referring to three lemmas, Lemmas 18, 19 and 20 to be given in Appendix A, for estimating three quantities $|\lambda_i^{\mathbf{x}} - \lambda_i|$, $\sqrt{\lambda_i^{\mathbf{x}}}|S_i^{\mathbf{z}} - \langle f_\rho, \phi_i^{\mathbf{x}}\rangle_K|$ and $\sqrt{\lambda_i^{\mathbf{x}}}|\langle f_\rho, \phi_i^{\mathbf{x}}\rangle_K|$.

Step 1. To achieve the desired sparsity, we apply (22) in Lemma 14 and know that for verifying condition (20) in Theorem 13, it is sufficient to show that for $i \ge p+1$,

$$|S_i^{\mathbf{z}}| < \frac{C_{\Omega,1}}{2} \frac{\min\{\sqrt{\lambda_i^{\mathbf{x}}/\gamma}, 1\}}{\lambda_i^{\mathbf{x}}/\gamma}$$

or equivalently,

$$\sqrt{\lambda_i^{\mathbf{x}}}|S_i^{\mathbf{z}}| < \frac{C_{\Omega,1}}{2} \min\left\{\sqrt{\gamma}, \gamma/\sqrt{\lambda_i^{\mathbf{x}}}\right\}. \tag{26}$$

Step 2. Our desired bound (26) is verified by estimating

$$\lambda_i^{\mathbf{x}} \le |\lambda_i^{\mathbf{x}} - \lambda_i| + \lambda_i$$

by Lemma 18 and the decay of $\{\lambda_i\}$, and estimating

$$\sqrt{\lambda_i^{\mathbf{x}}}|S_i^{\mathbf{z}}| \le \sqrt{\lambda_i^{\mathbf{x}}}|S_i^{\mathbf{z}} - \langle f_\rho, \phi_i^{\mathbf{x}}\rangle_K| + \sqrt{\lambda_i^{\mathbf{x}}}|\langle f_\rho, \phi_i^{\mathbf{x}}\rangle_K|$$

by Lemma 19 and Lemma 20.

Step 3. To prove the error bound (25), we expand the error function $f^{\mathbf{z}} - f_\rho$ with respect to the orthonormal basis $\{\phi_i^{\mathbf{x}}\}$ of $\mathcal{H}_K$ and express the norm as

$$\|f^{\mathbf{z}} - f_\rho\|_K^2 = \sum_{i\in\mathbb{N}} \left(\langle f^{\mathbf{z}} - f_\rho, \phi_i^{\mathbf{x}}\rangle_K\right)^2 = \sum_{i\in\mathbb{N}} \left(c_i^{\mathbf{z}} - \langle f_\rho, \phi_i^{\mathbf{x}}\rangle_K\right)^2.$$

Split

$$c_i^{\mathbf{z}} - \langle f_\rho, \phi_i^{\mathbf{x}}\rangle_K = \{c_i^{\mathbf{z}} - S_i^{\mathbf{z}}\} + \left\{S_i^{\mathbf{z}} - \langle f_\rho, \phi_i^{\mathbf{x}}\rangle_K\right\}.$$

While the term $|c_i^{\mathbf{z}} - S_i^{\mathbf{z}}|$ can be bounded by $\widetilde{\Omega}\left(\frac{\lambda_i^{\mathbf{x}}}{\gamma}\right)$ according to Theorem 13, the other term will be expressed as

$$|S_i^{\mathbf{z}} - \langle f_\rho, \phi_i^{\mathbf{x}}\rangle_K| = \frac{\sqrt{\lambda_i^{\mathbf{x}}}|S_i^{\mathbf{z}} - \langle f_\rho, \phi_i^{\mathbf{x}}\rangle_K|}{\sqrt{\lambda_i^{\mathbf{x}}}}.$$

Step 4. We can control the denominator of the above expression by introducing a set with large $\lambda_i^{\mathbf{x}}$ as $\mathcal{S} := \{i \in \{1,\dots,p\}, \lambda_i^{\mathbf{x}} > \lambda_p/2\}$, and then bound the expression by Lemma 19. This together with our previous estimate Guo and Zhou (2012) for the terms involving $i \in \mathbb{N}\setminus\mathcal{S}$ finally yields the desired error bound.

Let us demonstrate how to apply our general analysis in Theorem 15 by two special cases where the eigenvalues of the integral operator L_K decay polynomially and exponentially.

Corollary 16 *Assume (9) with $r > 0$, and that Ω has a concave exponent $q \in [0,1]$ with (8) valid. Suppose that for some positive constants D_1, D_2, $\alpha_1 \geq \alpha_2$, the eigenvalues $\{\lambda_i\}$ of L_K decay polynomially as*

$$D_1 i^{-\alpha_1} \leq \lambda_i \leq D_2 i^{-\alpha_2}, \qquad \forall i \in \mathbb{N} \tag{27}$$

with $2\alpha_2 \max\{r,1\} > 1$. Let $0 < \delta < 1$. If we choose

$$\gamma = C_1 (D_2/\lambda_1)^{r+1} \left(\log\frac{4m}{\delta}\right)^{1+2r} m^{-\min\{\frac{1+r}{2}, \frac{1+r}{1+2r}\}}, \tag{28}$$

then with confidence $1-\delta$ we have

$$c_i^{\mathbf{z}} = 0 \qquad \forall\ m^{\frac{1}{\alpha_2 \max\{2,1+2r\}}} + 1 \leq i \leq m \tag{29}$$

and

$$\|f^{\mathbf{z}} - f_\rho\|_K \leq C_2 \left(\log\frac{4m}{\delta}\right)^{1+2r} m^{-\theta_{rate}},$$

where $\theta_{rate} = \min\{\theta_1, \theta_2\}$ with

$$\begin{aligned}
\theta_1 &= \begin{cases} \frac{2(r+1)\alpha_2 - 2\alpha_1 - 2 + q}{4(2-q)\alpha_2}, & \text{if } 0 < r \leq 1/2, \\ \frac{2(2r+2)\alpha_2 - 4\alpha_1 - 2(2-q)}{4(2r+1)(2-q)\alpha_2}, & \text{if } r > 1/2, \end{cases} \\
\theta_2 &= \frac{2\alpha_2 r - 1 - 2(\alpha_1 - \alpha_2)\max\left\{1-r, \frac{1}{2}\right\}}{2\alpha_2 \max\{2, 1+2r\}},
\end{aligned}$$

and C_1 and C_2 are constants independent of m or δ (given explicitly in the proof).

Proof Denote $\mu = \max\{2, 1+2r\}$. Take $p = \lceil m^{\frac{1}{\alpha_2\mu}} \rceil$, the smallest integer greater than or equal to $m^{\frac{1}{\alpha_2\mu}}$. Then we have

$$m^{\frac{1}{\alpha_2\mu}} \leq p \leq 2m^{\frac{1}{\alpha_2\mu}}$$

and by (27),

$$\lambda_p \leq D_2 p^{-\alpha_2} \leq D_2 m^{-\frac{\alpha_2}{\alpha_2\mu}} = D_2 m^{-\frac{1}{\mu}}.$$

It follows that $\frac{\lambda_p}{\lambda_1} \leq \frac{D_2}{\lambda_1} m^{-\frac{1}{\mu}} \leq \frac{D_2}{\lambda_1}\frac{1}{\sqrt{m}}$ for $0 < r \leq \frac{1}{2}$ and for $r > \frac{1}{2}$, there holds $(\lambda_p/\lambda_1)^{r+\frac{1}{2}} \leq (D_2/\lambda_1)^{r+\frac{1}{2}}\frac{1}{\sqrt{m}}$. Note that (27) implies $\lambda_1 \leq D_2$. Then (24) is satisfied if we choose γ by (28). Hence the conclusion of Theorem 15 holds true. In particular, the statement (29) about the sparsity follows from the choice of p. What is left is to bound the right-hand side of (25) by estimating the four terms separately.

The first term of (25) can be estimated by bounding $\frac{2\gamma}{\lambda_p}$ from the choice of γ and the lower bound of λ_p as

$$\frac{2\gamma}{\lambda_p} \leq \frac{2C_1}{D_1}(D_2/\lambda_1)^{r+1}\left(\log\frac{4m}{\delta}\right)^{1+2r} m^{-\min\{\frac{1+r}{2},\frac{1+r}{1+2r}\}}\left(2m^{\frac{1}{\alpha_2\mu}}\right)^{\alpha_1}.$$

Observe that

$$\min\left\{\frac{1+r}{2},\frac{1+r}{1+2r}\right\} - \frac{\alpha_1}{\alpha_2\mu} = \begin{cases} \frac{1+r}{2} - \frac{\alpha_1}{2\alpha_2} = \frac{(r+1)\alpha_2-\alpha_1}{2\alpha_2}, & \text{if } 0 < r \leq 1/2, \\ \frac{1+r}{1+2r} - \frac{\alpha_1}{(2r+1)\alpha_2} = \frac{2(2r+2)\alpha_2-4\alpha_1}{4(2r+1)\alpha_2}, & \text{if } r > 1/2. \end{cases}$$

Therefore,

$$C_{\Omega,2}\sqrt{p}\left\{\left(\frac{2\gamma}{\lambda_p}\right)^{\frac{1}{2-q}} + \frac{2\gamma}{\lambda_p}\right\} \leq C_{\Omega,2}\sqrt{2}2^{\alpha_1+2}\frac{C_1}{D_1}(D_2/\lambda_1)^{r+1}\left(\log\frac{4m}{\delta}\right)^{1+2r} m^{-\theta_1},$$

where

$$\theta_1 = \begin{cases} \frac{(r+1)\alpha_2-\alpha_1}{2(2-q)\alpha_2} - \frac{1}{4\alpha_2} = \frac{2(r+1)\alpha_2-2\alpha_1-2+q}{4(2-q)\alpha_2}, & \text{if } 0 < r \leq 1/2, \\ \frac{2(2r+2)\alpha_2-4\alpha_1-2(2-q)}{4(2r+1)(2-q)\alpha_2}, & \text{if } r > 1/2. \end{cases}$$

The second term of (25) can be estimated by the choice of γ and the upper bound of λ_p as

$$\|g_\rho\|_K \lambda_p^r \leq \|g_\rho\|_K D_2^r p^{-\alpha_2 r} \leq \|g_\rho\|_K D_2^r m^{-\frac{2\alpha_2 r}{2\alpha_2\mu}} \leq \|g_\rho\|_K D_2^r m^{-\theta_2},$$

where θ_2 is the power index defined in the statement of the theorem.

The third term of (25) can be estimated by the choice of p and the lower bound of λ_p as

$$\begin{aligned} & C_3\frac{\sqrt{p}\log\frac{4m}{\delta}}{\sqrt{m}}\lambda_p^{\min\{-1/2,r-1\}} \\ & \leq C_3 D_1^{\min\{-1/2,r-1\}} p^{\frac{1}{2}-\alpha_1\min\{-1/2,r-1\}}\left(\log\frac{4m}{\delta}\right)m^{-1/2} \\ & \leq C_3\sqrt{2}\left(2^{-\alpha_1}D_1\right)^{\min\{-1/2,r-1\}}\left(\log\frac{4m}{\delta}\right)m^{-\frac{1}{2}+\frac{1}{\alpha_2\mu}\left(\frac{1}{2}-\alpha_1\min\left\{-\frac{1}{2},r-1\right\}\right)}. \end{aligned}$$

From the identity $\mu - 2r = \max\{2, 2r+1\} - 2r = 2\max\{1-r, 1/2\}$, we see that the power index of m equals

$$\begin{aligned}
&-\frac{1}{2} + \frac{1}{\alpha_2 \mu}\left(\frac{1}{2} + \alpha_1 \max\left\{\frac{1}{2}, 1-r\right\}\right) \\
&= -\frac{1}{2\alpha_2\mu}\left(2\alpha_2 r - 1 + \alpha_2(\mu - 2r) - 2\alpha_1 \max\{1/2, 1-r\}\right) \\
&= -\frac{1}{2\alpha_2\mu}\left(2r\alpha_2 - 1 - 2(\alpha_1 - \alpha_2)\max\{1/2, 1-r\}\right)
\end{aligned}$$

which is exactly $-\theta_2$.

Turn to the last term of (25). By the restriction $2\alpha_2 \max\{r,1\} > 1$, we can bound the series as

$$\begin{aligned}
\sum_{i=p+1}^{\infty} \lambda_i^{2\max\{r,1\}} &\leq \sum_{i=p+1}^{\infty} D_2^{2\max\{r,1\}} i^{-2\alpha_2 \max\{r,1\}} \\
&\leq D_2^{2\max\{r,1\}} \int_p^{\infty} x^{-2\alpha_2 \max\{r,1\}}\, \mathrm{d}x = \frac{D_2^{2\max\{r,1\}} p^{1-2\alpha_2\max\{r,1\}}}{2\alpha_2 \max\{r,1\} - 1}.
\end{aligned}$$

This combining with the choice of p and the lower bound of λ_p yields

$$\begin{aligned}
&C_4 \lambda_p^{\min\{r-1,0\}} \left(\sum_{i=p+1}^{\infty} \lambda_i^{2\max\{r,1\}}\right)^{1/2} \\
&\leq \frac{C_4 D_1^{\min\{r-1,0\}} D_2^{\max\{r,1\}}}{\sqrt{2\alpha_2 \max\{r,1\} - 1}} m^{\frac{1}{2\alpha_2\mu}(1 - 2\alpha_2\max\{r,1\} - 2\alpha_1 \min\{r-1,0\})}.
\end{aligned}$$

But

$$\begin{aligned}
&\frac{1}{2\alpha_2\mu}\left(1 - 2\alpha_2 \max\{r,1\} - 2\alpha_1 \min\{r-1, 0\}\right) \\
&= -\frac{1}{2\alpha_2\mu}\left(2r\alpha_2 - 1 - 2(\alpha_1 - \alpha_2)\max\{0, 1-r\}\right) \leq -\theta_2.
\end{aligned}$$

So we can combine this bound with the above estimates for the first three terms of (25) and verify the learning rate stated in the theorem by taking

$$\begin{aligned}
C_2 = \quad & C_{\Omega,2}\sqrt{2}2^{\alpha_1+2}\frac{C_1}{D_1}(D_2/\lambda_1)^{r+1} + \|g_\rho\|_K D_2^r \\
& + C_3\sqrt{2}\left(2^{-\alpha_1} D_1\right)^{\min\{-1/2, r-1\}} + \frac{C_4 D_1^{\min\{r-1,0\}} D_2^{\max\{r,1\}}}{\sqrt{2\alpha_2\max\{r,1\} - 1}}.
\end{aligned}$$

The proof of Corollary 16 is complete. ■

Corollary 17 *Assume (9) with $r > 0$, and that Ω has a concave exponent $q \in [0,1]$ with (8) valid. Suppose that for some positive constants D_1, D_2, $\beta_1 \geq \beta_2$, the eigenvalues $\{\lambda_i\}$ of L_K decay exponentially as*

$$D_1 \beta_1^{-i} \leq \lambda_i \leq D_2 \beta_2^{-i}, \qquad \forall i \in \mathbb{N}. \tag{30}$$

Let $0 < \delta < 1$. If we choose γ as (28), then with confidence $1-\delta$ we have

$$c_i^{\mathbf{z}} = 0, \qquad \forall\ \frac{\log(m+1)}{\max\{2, 1+2r\}\log\beta_2} + 1 \leq i \leq m \tag{31}$$

and

$$\|f^{\mathbf{z}} - f_\rho\|_K \leq C_2 \left(\log\frac{4m}{\delta}\right)^{2r+1} m^{-\theta_{rate}},$$

where

$$\theta_{rate} = \frac{1}{(2-q)\max\{2,1+2r\}} \min\left\{1 + r - \frac{\log\beta_1}{\log\beta_2}, (2-q)r - \frac{(2-q)\log(\beta_1/\beta_2)}{\log\beta_2} \max\left\{1-r, \frac{1}{2}\right\}\right\}$$

and C_2 is a constant independent of m or δ (specified in the proof).

Proof Take $\mu = \max\{2, 1+2r\}$ and $p = \lceil \frac{\log(m+1)}{\mu\log\beta_2} \rceil$. Then

$$\frac{\log(m+1)}{\mu\log\beta_2} \leq p < 1 + \frac{\log(m+1)}{\mu\log\beta_2},$$

which implies $m^{1/\mu} \leq \beta_2^p \leq \beta_1^p \leq \beta_1 (2m)^{\frac{\log\beta_1}{\mu\log\beta_2}}$. Hence $\frac{\lambda_p}{\lambda_1} \leq \frac{D_2}{\lambda_1}\beta_2^{-p} \leq \frac{D_2}{\lambda_1}\frac{1}{\sqrt{m}}$ for $0 < r \leq \frac{1}{2}$ and for $r > \frac{1}{2}$, there holds $(\lambda_p/\lambda_1)^{r+\frac{1}{2}} \leq (D_2/\lambda_1)^{r+\frac{1}{2}}\frac{1}{\sqrt{m}}$. Then (24) is satisfied and the conclusion of Theorem 15 holds true. The statement (31) about the sparsity follows from the choice of p, and the next step is to estimate the four summing terms of the error bound (25).

For the first term, we notice

$$\frac{2\gamma}{\lambda_p} \leq \frac{2C_1}{D_1}(D_2/\lambda_1)^{r+1}\left(\log\frac{4m}{\delta}\right)^{1+2r} m^{-\min\{\frac{1+r}{2}, \frac{1+r}{1+2r}\}} \beta_1 (2m)^{\frac{\log\beta_1}{\mu\log\beta_2}}.$$

So the first term can be bounded as

$$C_{\Omega,2}\sqrt{p}\left\{\left(\frac{2\gamma}{\lambda_p}\right)^{\frac{1}{2-q}} + \frac{2\gamma}{\lambda_p}\right\} \leq C_{\Omega,2}C_5\sqrt{\log(2m)}\left(\log\frac{4m}{\delta}\right)^{1+2r} m^{-\frac{1}{2-q}\min\{\frac{1+r}{2}, \frac{1+r}{1+2r}\} + \frac{\log\beta_1}{(2-q)\mu\log\beta_2}},$$

where C_5 is the constant given by

$$C_5 = 2\sqrt{\frac{1}{\log 2} + \frac{1}{\mu\log\beta_2}} \max\left\{\frac{2C_1}{D_1}(D_2/\lambda_1)^{r+1}\beta_1 2^{\frac{\log\beta_1}{\mu\log\beta_2}}, 1\right\}.$$

The second term of (25) is easy to handle:

$$\|g_\rho\|_K \lambda_p^r \le \|g_\rho\|_K D_2^r \beta_2^{-pr} \le \|g_\rho\|_K D_2^r m^{-r/\mu}.$$

The third term of (25) can be estimated by the choice of p and the lower bound of λ_p as

$$C_3 \frac{\sqrt{p}\log\frac{4m}{\delta}}{\sqrt{m}} \lambda_p^{\min\{-1/2,r-1\}} \le C_6 \sqrt{\log(2m)} \left(\log\frac{4m}{\delta}\right) m^{-\frac{1}{2}+\frac{\log\beta_1}{\mu\log\beta_2}\max\{1/2,1-r\}},$$

where

$$C_6 = C_3 \sqrt{\frac{1}{\log 2} + \frac{1}{\mu\log\beta_2}} D_1^{\min\{-1/2,r-1\}} \left(\beta_1 2^{\frac{\log\beta_1}{\mu\log\beta_2}}\right)^{\max\{1/2,1-r\}}.$$

Observe that $\max\{1/2, 1-r\} + r = \mu/2$. The power index of m equals

$$\begin{aligned} &-\frac{1}{2} + \frac{\log\beta_1}{\mu\log\beta_2}\max\{1/2,1-r\} = -\frac{r}{\mu} + \frac{r\log\beta_2 - \frac{\mu}{2}\log\beta_2 + \frac{\mu}{2}\log\beta_1 - r\log\beta_1}{\mu\log\beta_2} \\ &= -\frac{r}{\mu} + \frac{\log\frac{\beta_1}{\beta_2}}{\mu\log\beta_2}\max\{1/2,1-r\}. \end{aligned}$$

Finally, we bound the series in the last term of (25) and find

$$\begin{aligned} &C_4 \lambda_p^{\min\{r-1,0\}} \left(\sum_{i=p+1}^{\infty} \lambda_i^{2\max\{r,1\}}\right)^{1/2} \\ &\le C_4 (\beta_1/D_1)^{\max\{1-r,0\}} (2m)^{\frac{\log\beta_1}{\mu\log\beta_2}\max\{1-r,0\}} D_2^{\max\{r,1\}} \left(\sum_{i=p+1}^{\infty} \beta_2^{-2i\max\{r,1\}}\right)^{1/2} \\ &\le C_4 \left(\frac{\beta_1}{D_1}\right)^{\max\{1-r,0\}} (2m)^{\frac{\log\beta_1}{\mu\log\beta_2}\max\{1-r,0\}} D_2^{\max\{r,1\}} \frac{\beta_2^{-p\max\{r,1\}}}{\sqrt{\beta_2^{2\max\{r,1\}}-1}} \\ &\le C_4 \left(\frac{\beta_1}{D_1}\right)^{\max\{1-r,0\}} \frac{D_2^{\max\{r,1\}}}{\sqrt{\beta_2^{2\max\{r,1\}}-1}} (2m)^{\frac{\max\{1-r,0\}(\log\beta_1)}{\mu\log\beta_2}} m^{-\frac{\max\{r,1\}}{\mu}}. \end{aligned}$$

Note that the power index of m is

$$\frac{\max\{1-r,0\}\log\beta_1}{\mu\log\beta_2} - \frac{\max\{r,1\}}{\mu} = \frac{\max\{1-r,0\}\log(\beta_1/\beta_2) - r\log\beta_2}{\mu\log\beta_2}.$$

Then the desired learning rate is verified by observing $\min\{\frac{1+r}{2}, \frac{1+r}{1+2r}\}\mu = 1+r$ and taking

$$C_2 = C_{\Omega,2}C_5 + \|g_\rho\|_K D_2^r + C_6 + C_4 \left(\frac{\beta_1}{D_1} 2^{\frac{\log\beta_1}{\mu\log\beta_2}}\right)^{\max\{1-r,0\}} \frac{D_2^{\max\{r,1\}}}{\sqrt{\beta_2^{2\max\{r,1\}}-1}}.$$

The proof of Corollary 17 is complete. ■

5. Simulations

In this section we give some simulations for both artificial and real data. We demonstrate that with either the ℓ^q-regularizer or the SCAD penalty, RKPCA is comparable with the regularized least squares in learning error, and achieves satisfactory sparsity.

5.1 Simulation on artificial data

We start with a simulation on artificial data. For simplicity we take $X = [0, 1]$. Let ρ be a Borel probability measure on $X \times Y$ to be specified later and $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ be a sample of size m divisible by 5. We divide $\mathbf{z}$ evenly into five disjoint subsets $\mathbf{z} = \cup_{j=1}^5 \mathbf{z}_j$, and do 5-fold cross-validation to select the parameter γ^* from a geometric sequence $\{10^{-10}, \cdots, 10^{-2}\}$ of length 60, to minimize the root-mean-square error (RMSE). Here, with a fixed γ the RMSE score is defined by

$$\mathcal{E}_{\mathsf{RMSE},\mathbf{z}}(\gamma) = \left(\sum_{j=1}^{5} \sum_{(x,y)\in \mathbf{z}\setminus\mathbf{z}_j} \left(f_\gamma^{\mathbf{z}_j}(x) - y\right)^2 \right)^{1/2}. \tag{32}$$

Then RKPCA is trained with γ^* on $\mathbf{z}$ and outputs $f^{\mathbf{z}}$. Sparsity is evaluated by the percentage of the non-zero coefficients in (2). The prediction performance is evaluated with the oracle RMSE defined by

$$\mathcal{E}_{\mathsf{RMSE},f_\rho}(f^{\mathbf{z}}) = \left(\int_0^1 \left(f^{\mathbf{z}}(x) - f_\rho(x)\right)^2 \mathrm{d}x \right)^{1/2}, \tag{33}$$

where the integral is computed with 1000 equispaced points.

First, we simulate with the Gaussian kernel

$$K_{\mathsf{G}} = \exp\left(-\frac{(x-y)^2}{0.6^2}\right).$$

We use the regression model $f_\rho(x) = \mathrm{e}^{-(x-1/3)^2/0.7^2}$. Let ρ_X be the uniform distribution on $[0, 1]$, and $\rho(\cdot|x)$ be the uniform distribution on $[f_\rho(x) - 0.1, f_\rho(x) + 0.1]$. The simulation is summarized in Table 1. We find that the behavior of the SCAD penalty is comparable on this data set with the penalty $\Omega(|c|) = |c|$, and despite of very strong sparsity, RKPCA achieves comparable precision with that of RLS.

Next, we simulate with the Sobolev kernel

$$K_{\mathsf{S}}(x, y) = \mathrm{e}^{-|x-y|}.$$

The regression function is set as $f_\rho(x) = |2x - 1|^\tau$. The marginal and conditional probabilities ρ_X and $\rho(\cdot|x)$ are defined as above. We use $\tau = 1$, 1.5, 2.5, and 4.5. Note that in addition to $\mathcal{E}_{\mathsf{RMSE},f_\rho}$, the RKHS norm is now also easy to compute as another measurement. In fact, one has for $f, g \in \mathcal{H}_{K_{\mathsf{S}}}$,

$$2\langle f, g\rangle_{K_{\mathsf{S}}} = f(0)g(0) + f(1)g(1) + \int_0^1 f(t)g(t)\,\mathrm{d}t + \int_0^1 f'(t)g'(t)\,\mathrm{d}t.$$

The simulation is summarized in Table 2, from which we have the following observations:

sample size	RKPCA				RLS
	$q = 1$	$q = 2/3$	$q = 1/3$	SCAD	
100	3.5(2.0)%	3.2(1.7)%	3.2(1.9)%	3.3(1.5)%	100(0)%
	0.013(0.006)	0.012(0.005)	0.012(0.007)	0.011(0.005)	0.012(0.005)
300	1.4(1.2)%	1.2(1.4)%	1.1(0.7)%	1.1(0.5)%	100(0)%
	0.007(0.004)	0.007(0.006)	0.007(0.003)	0.006(0.002)	0.007(0.003)
1000	0.4(0.4)%	0.4(0.3)%	0.3(0.2)%	0.4(0.2)%	100(0)%
	0.004(0.002)	0.004(0.002)	0.004(0.002)	0.004(0.001)	0.004(0.002)

Table 1: A simulation with Gaussian kernel. Here SCAD and $q = 1$, 2/3, and 1/3 stand for RKPCA with penalty SCAD (as defined in Example 2, and we set $b = 2.5$) and $\Omega(|c|) = |c|^q$ respectively. The scores of RLS are also listed for comparison. In each cell, the top percentage gives the proportion of the non-zero coefficients, and the bottom score is $\mathcal{E}_{\mathsf{RMSE},f_\rho}$ as defined in (33). Each simulation is repeated 100 times. We present the mean scores in the table, and give the sample standard deviation in parentheses.

(a) The sparsity and learning error of RKPCA with the SCAD penalty is again comparable on this data set to that with the penalty $\Omega(|c|) = |c|$. This shows that the expression of the SCAD penalty near the origin (the same as that for $\Omega(|c|) = |c|$) and the concave exponent q play a crucial role in its performance.

(b) Compared with RLS, RKPCA achieves very strong sparsity while its approximation ability with $\Omega(|c|) = |c|$ in terms of the RKHS metric is consistently better. This might be caused by the orthogonality of the empirical features in the RKHS. The learning ability in terms of the root-mean-square error defined by (33) is comparable.

5.2 Simulation on MHC-peptide binding data

We apply RKPCA to the quantitative Immune Epitope Database (IEDB) benchmark data of human leukocyte antigen (HLA)–peptide binding affinities, introduced in (Nielsen et al., 2008). Nielsen and Lund (2009) developed an artificial neural network-based algorithm called NN-align, which gave on this data set the state-of-the-art prediction in 2009. Later, Shen et al. (2012) designed a string kernel denoted in their paper by $\hat{K}^3$, and applied it with the regularized least squares (RLS), which produced better prediction than NN-align on the same data set. We use this $\hat{K}^3$ in RKPCA, and show that RKPCA achieves some sparsity in addition to the precision comparable with that in (Shen et al., 2012).

Here are more details of our simulation. The quantitative IEDB benchmark data set in (Nielsen et al., 2008) as mentioned above, consists of 14 groups, each containing the affinities of a set of peptides to a specific HLA allele. We use the 14 groups separately. Now fix an allele a and denote $X = \mathscr{P}_a$ the set of peptides given in the data set. For $p \in \mathscr{P}_a$, the affinity $y_p \in [0, 1] \subset Y = \mathbb{R}$ is a real number (see Nielsen and Lund (2009); Shen et al. (2012)). We divide $\mathscr{P}_a$ into 5 disjoint subsets $\mathscr{P}_a = \cup_{j=1}^5 \mathscr{P}_a^j$, following exactly the division in (Nielsen and Lund, 2009) and (Shen et al., 2012), for a 5-fold cross-validation. In the jth cross-validation round ($j = 1, \cdots, 5$), we take $\mathscr{P}_a^j$ as testing data and $\mathscr{P}_a \backslash \mathscr{P}_a^j$ as training data. Within the training data, another 5-fold cross-validation is employed to select the

parameter γ_j^* in (3), from a geometric sequence $\{10^{-8}, \cdots, 10^{-2}\}$ of length 60 to minimize the RMSE score defined in (32). Then RKPCA is trained on $\mathscr{P}_a \backslash \mathscr{P}_a^j$ with γ_j^* to predict the affinities on $\mathscr{P}_a^j$. After all the five rounds, each peptide $p \in \mathscr{P}_a$ has a predicted affinity $\tilde{y}_p$ obtained during the jth round where $\mathscr{P}_a^j \ni p$. Note that $\tilde{y}_p$ may not always fall in $[0, 1]$, and might be projected back onto $[0, 1]$ to increase precision. However we do not adopt the projection, for being consistent and comparable with (Shen et al., 2012) where they did not either. Since there is no oracle information, we use

$$\mathcal{E}_{\mathsf{RMSE},a} = \left(\frac{1}{\#\mathscr{P}_a} \sum_{p \in \mathscr{P}_a} (\tilde{y}_p - y_p)^2 \right)^{1/2} \tag{34}$$

as the RMSE score. A lower RMSE score indicates a better performance.

The area under the receiver operating characteristic (ROC) curve (AUC), defined as

$$\mathcal{E}_{\mathsf{AUC},a} = \frac{\#\left\{(p, p') : p \in \mathscr{P}_{a,B}, p' \in \mathscr{P}_{a,N}, \tilde{y}_p > \tilde{y}_{p'}\right\}}{(\#\mathscr{P}_{a,B})\,(\#\mathscr{P}_{a,N})} \in [0, 1], \tag{35}$$

is another performance index. Here $\mathscr{P}_{a,B} = \{p \in \mathscr{P}_a : y_p > 0.426\}$ and $\mathscr{P}_{a,N} = \mathscr{P}_a \backslash \mathscr{P}_{a,B}$ are the sets of binding peptides and non-binding ones respectively, with the threshold 0.426 used in (Nielsen and Lund, 2009). A higher AUC score indicates a better performance. The above scores (34) and (35) are used in (Shen et al., 2012). See also (Nielsen and Lund, 2009) for details.

We test the RKPCA with $\Omega(c) = |c|^q$, where q is set to be 1, 2/3, and 1/3 in three separated tests, and with the SCAD penalty. For defining $\hat{K}^3$, the Hadamard power index is fixed to be 0.11387 for simplicity, as suggested in (Shen et al., 2012).

The simulation is summarized in Table 3, from which we have the following observations:

(a) In terms of AUC on this real data set, RLS (Shen et al., 2012) has better performance than NN-align (Nielsen and Lund, 2009). The improvement is 0.55% on average, with better AUC scores for 9 out of 14 test groups while the score difference is always at the second significant figure. RKPCA with $\Omega(c) = |c|$ has even slightly better performance, giving an improvement of 0.11% on average, and better AUC scores for 8 out of 14 test groups with the score difference always at the third significant figure only. Improvements in (Shen et al., 2012) and in our simulation seem to be small, but we regard the results to be valuable because this data set has been well investigated in the immunological literature and any improvement is difficult. In particular, the dissimilarity metric BLOSUM62-2 among the 20 basic amino-acids, based on which the string kernel $\hat{K}^3$ is constructed in (Shen et al., 2012), was obtained in a very tight form after long-term effort and a vast biological literature (see, e.g., Henikoff and Henikoff (1992)).

(b) Sparsity and error bounds in terms of both AUC and root-mean-square error for the simulation with the SCAD penalty is almost the same on this real data set as that with $\Omega(|c|) = |c|$, verifying again the role of the concave exponent $q = 1$.

sample size	τ	RKPCA				RLS
		$q = 1$	$q = 2/3$	$q = 1/3$	SCAD	
100	1.0	14.2(7.5)%	9.7(4.4)%	8.1(3.7)%	16.0(8.0)%	100(0)%
		0.026(0.007)	0.026(0.006)	0.029(0.006)	0.025(0.005)	0.025(0.005)
		0.685(0.318)	0.761(0.401)	1.033(0.772)	0.738(0.378)	0.801(0.185)
	1.5	16.9(8.7)%	11.2(6.9)%	9.2(6.4)%	17.7(8.4)%	100(0)%
		0.026(0.007)	0.027(0.005)	0.030(0.006)	0.026(0.006)	0.027(0.008)
		0.780(0.384)	0.885(0.653)	1.006(0.966)	0.805(0.378)	0.908(0.228)
	2.5	22.5(10.3)%	14.2(6.7)%	13.1(11.7)%	20.4(8.8)%	100(0)%
		0.028(0.007)	0.031(0.008)	0.033(0.007)	0.029(0.007)	0.029(0.006)
		1.086(0.545)	1.217(0.678)	1.601(1.653)	1.124(1.235)	1.195(0.380)
	4.5	26.6(10.6)%	17.8(8.1)%	17.8(11.6)%	26.2(9.8)%	100(0)%
		0.033(0.007)	0.036(0.010)	0.039(0.010)	0.036(0.011)	0.035(0.010)
		1.483(0.515)	1.758(0.814)	2.488(1.979)	1.623(0.882)	1.685(0.385)
300	1.0	5.4(1.6)%	3.8(1.3)%	3.0(1.6)%	6.1(2.8)%	100(0)%
		0.015(0.003)	0.016(0.003)	0.018(0.003)	0.016(0.002)	0.016(0.002)
		0.503(0.073)	0.604(0.264)	0.865(1.084)	0.568(0.207)	0.652(0.104)
	1.5	6.4(2.3)%	4.1(1.4)%	3.2(1.2)%	6.0(2.1)%	100(0)%
		0.016(0.003)	0.017(0.003)	0.019(0.004)	0.016(0.003)	0.016(0.003)
		0.589(0.164)	0.666(0.242)	0.824(0.687)	0.578(0.148)	0.708(0.117)
	2.5	7.7(2.5)%	5.2(1.9)%	4.0(1.2)%	7.3(2.2)%	100(0)%
		0.018(0.004)	0.019(0.003)	0.021(0.003)	0.018(0.003)	0.018(0.002)
		0.802(0.166)	0.946(0.463)	1.044(0.683)	0.759(0.139)	0.966(0.152)
	4.5	10.2(2.7)%	6.9(2.0)%	5.2(1.2)%	9.8(3.1)%	100(0)%
		0.020(0.004)	0.022(0.003)	0.024(0.003)	0.020(0.004)	0.021(0.003)
		1.142(0.218)	1.372(0.475)	1.495(0.768)	1.164(0.537)	1.382(0.223)
1000	1.0	2.0(0.6)%	1.4(0.5)%	1.0(0.3)%	2.0(0.5)%	100(0)%
		0.009(0.001)	0.010(0.001)	0.011(0.002)	0.009(0.001)	0.010(0.001)
		0.434(0.085)	0.484(0.180)	0.533(0.368)	0.421(0.039)	0.570(0.114)
	1.5	2.3(0.7)%	1.5(0.4)%	1.2(0.3)%	2.4(0.6)%	100(0)%
		0.010(0.001)	0.010(0.002)	0.011(0.002)	0.010(0.001)	0.010(0.001)
		0.467(0.068)	0.516(0.122)	0.583(0.325)	0.477(0.066)	0.612(0.103)
	2.5	2.8(0.6)%	1.8(0.4)%	1.4(0.3)%	3.0(0.7)%	100(0)%
		0.011(0.001)	0.012(0.001)	0.013(0.002)	0.011(0.001)	0.011(0.001)
		0.642(0.090)	0.711(0.207)	0.846(0.533)	0.647(0.085)	0.781(0.085)
	4.5	3.8(0.9)%	2.4(0.4)%	1.9(0.4)%	3.7(0.8)%	100(0)%
		0.012(0.002)	0.013(0.001)	0.015(0.002)	0.012(0.002)	0.013(0.001)
		0.950(0.155)	0.998(0.184)	1.254(0.912)	0.931(0.108)	1.163(0.119)

Table 2: A simulation with Sobolev kernel. Here SCAD and $q = 1$, $2/3$, and $1/3$ stand for RKPCA with penalty SCAD (as defined in Example 2, and we set $b = 2.5$) and $\Omega(|c|) = |c|^q$ respectively. The scores of RLS are also listed for comparison. In each cell, the top percentage gives the proportion of the non-zero coefficients, the middle score is $\mathcal{E}_{\mathsf{RMSE},f_\rho}$ as defined in (33), and the bottom score gives the RKHS distance of $f^{\mathbf{z}}$ to f_ρ. Each simulation is repeated 100 times. We present the mean scores in the table, and give the sample standard deviation in parentheses.

Allele a	$\#\mathscr{P}_a$	NN-align	RLS	RKPCA			
				$q=1$	$q=2/3$	$q=1/3$	SCAD
DRB1*0101	5166	–	–	74.65%	59.30%	60.81%	74.66%
		–	0.18660	0.18690	0.18746	0.18830	0.18694
		0.836	0.85707	0.85651	0.85512	0.85306	0.85637
DRB1*0301	1020	–	–	88.04%	71.84%	56.47%	86.00%
		–	0.18497	0.18476	0.18495	0.18551	0.18483
		0.816	0.82813	0.82995	0.82950	0.82714	0.83008
DRB1*0401	1024	–	–	72.39%	60.16%	61.40%	73.36%
		–	0.24055	0.24089	0.24202	0.24277	0.24152
		0.771	0.78431	0.78023	0.77697	0.77505	0.77839
DRB1*0404	663	–	–	70.55%	57.84%	57.88%	71.12%
		–	0.20702	0.20797	0.20918	0.20878	0.20796
		0.818	0.81425	0.81695	0.81134	0.80801	0.81701
DRB1*0405	630	–	–	81.47%	69.56%	63.06%	78.85%
		–	0.20069	0.20037	0.20017	0.20076	0.20048
		0.781	0.79296	0.79837	0.79929	0.79791	0.79799
DRB1*0701	853	–	–	98.65%	91.76%	86.96%	98.65%
		–	0.21944	0.21826	0.21840	0.21849	0.21826
		0.841	0.83440	0.83883	0.83918	0.83916	0.83883
DRB1*0802	420	–	–	96.85%	93.75%	87.98%	96.90%
		–	0.19666	0.19555	0.19557	0.19572	0.19557
		0.832	0.83538	0.83968	0.83938	0.83749	0.83968
DRB1*0901	530	–	–	73.11%	53.35%	50.94%	74.15%
		–	0.25398	0.25563	0.25653	0.25784	0.25593
		0.616	0.66591	0.66293	0.66273	0.66163	0.66177
DRB1*1101	950	–	–	94.61%	83.82%	80.21%	94.61%
		–	0.20776	0.20799	0.20802	0.20780	0.20799
		0.823	0.83703	0.83679	0.83680	0.83706	0.83678
DRB1*1302	498	–	–	84.99%	72.64%	62.25%	81.28%
		–	0.22569	0.22518	0.22540	0.22578	0.22496
		0.831	0.80410	0.80479	0.80439	0.80303	0.80533
DRB1*1501	934	–	–	75.80%	64.94%	74.79%	77.89%
		–	0.23268	0.23318	0.23401	0.23419	0.23313
		0.758	0.76436	0.76258	0.76086	0.76058	0.76219
DRB3*0101	549	–	–	92.94%	89.57%	87.52%	92.49%
		–	0.15945	0.15932	0.15916	0.15911	0.15934
		0.844	0.80228	0.80504	0.80546	0.80622	0.80509
DRB4*0101	446	–	–	96.75%	81.28%	76.18%	96.75%
		–	0.20809	0.20765	0.20838	0.20834	0.20765
		0.811	0.81057	0.81096	0.80791	0.80713	0.81098
DRB5*0101	924	–	–	100.00%	99.95%	98.76%	100.00%
		–	0.23038	0.23045	0.23045	0.23046	0.23045
		0.797	0.80568	0.80549	0.80550	0.80557	0.80549
Average		–	–	85.77%	74.98%	71.80%	85.48%
		–	0.21100	0.21101	0.21141	0.21170	0.21107
		0.7982	0.80260	0.80351	0.80246	0.80136	0.80328

Table 3: Comparison of sparsity and error. Each cell consists of the average of proportions of the non-zero coefficients in the five rounds of test (the top percentage), RMSE defined by (34) (the middle number), and AUC defined by (35) (the bottom number). We cite the scores of NN-align from (Nielsen and Lund, 2009) and that of RLS from (Shen et al., 2012).

Acknowledgments

We would like to sincerely thank the referees for their constructive suggestions and comments. The work described in this paper was supported by the Research Grants Council of Hong Kong [Project No. CityU 105011] and by the National Natural Science Foundation of China under Grants 11461161006 and 11471292. The corresponding author is Ding-Xuan Zhou.

Appendix A. Proof of Theorem 15

In this appendix, we prove our general result on sparsity and error bounds stated in Theorem 15.

The following three lemmas are needed for proving Theorem 15. The first one is cited from (Zwald and Blanchard, 2006). See also (Koltchinskii and Giné, 2000; Guo and Zhou, 2012).

Lemma 18 *(a) We have*

$$\sum_{i=1}^{\infty}(\lambda_i - \lambda_i^{\mathbf{x}})^2 \le \|L_K - L_K^{\mathbf{x}}\|_{HS}^2. \tag{36}$$

(b) For any $0 < \delta < 1$, with confidence $1-\delta$ we have

$$\|L_K - L_K^{\mathbf{x}}\|_{HS} \le \frac{4\kappa^2 \log\frac{2}{\delta}}{\sqrt{m}}. \tag{37}$$

The second lemma needed for proving Theorem 15 improves our previous estimate $\|\{\lambda_i^{\mathbf{x}}|S_i^{\mathbf{z}} - \langle f_\rho, \phi_i^{\mathbf{x}}\rangle_K|\}\|_{\ell^2} \le \frac{8M\kappa\log\frac{2}{\delta}}{\sqrt{m}}$ given in (Guo and Zhou, 2012) for the case of ℓ^1-penalty. The significant improvement we make here is to reduce the power of $\lambda_i^{\mathbf{x}}$ from 1 to $\frac{1}{2}$. Hence a different method for the proof is needed.

Lemma 19 *Let $f_\rho \in \mathcal{H}_K$. For $0 < \delta < 1$, with confident $1-\delta$ we have*

$$\sqrt{\lambda_i^{\mathbf{x}}}\left|S_i^{\mathbf{z}} - \langle f_\rho, \phi_i^{\mathbf{x}}\rangle_K\right| \le \frac{2\sqrt{2}M}{\sqrt{m}}\sqrt{\log\frac{2m}{\delta}}, \qquad \forall\ i \in \mathbb{N}. \tag{38}$$

Proof When $\lambda_i^{\mathbf{x}} = 0$, (38) is obvious. When $i \ge m+1$, $\lambda_i^{\mathbf{x}} = 0$ since the rank of $L_K^{\mathbf{x}}$ is not greater than m. For any fixed $\lambda_i^{\mathbf{x}} > 0$, denote

$$\eta_j = \frac{y_j - f_\rho(x_j)}{\sqrt{m}}, \qquad a_j = \frac{\phi_i^{\mathbf{x}}(x_j)}{\sqrt{m\lambda_i^{\mathbf{x}}}}.$$

Then by the definition of $S_i^{\mathbf{z}}$, $\sqrt{\lambda_i^{\mathbf{x}}}\left(S_i^{\mathbf{z}} - \langle f_\rho, \phi_i^{\mathbf{x}}\rangle_K\right) = \sum_{j=1}^m \eta_j a_j$. Also, $\sum_{j=1}^m a_j^2 = 1$. Since $|y| \le M$ almost surely, we have $|a_j\eta_j| \le 2M|a_j|/\sqrt{m}$ almost surely. By Hoeffding's inequality, we have for any $\varepsilon > 0$,

$$\mathbb{P}\left\{\left|\sum_{j=1}^m a_j\eta_j\right| \ge \varepsilon\right\} \le 2\exp\left(-\frac{m\varepsilon^2}{8M^2}\right).$$

Taking the union of the above at most m events, we know that

$$\mathbb{P}\left\{\max_{i=1,\ldots,m}\left|\sqrt{\lambda_i^{\mathbf{x}}}\left(S_i^{\mathbf{z}}-\langle f_\rho,\phi_i^{\mathbf{x}}\rangle_K\right)\right|\geq\varepsilon\right\}\leq 2m\exp\left(-\frac{m\varepsilon^2}{8M^2}\right).$$

One completes the proof by taking $\varepsilon>0$ to be the positive solution to the equation $2m\exp(-\frac{m\varepsilon^2}{8M^2})=\delta$. ■

Lemma 20 *Let $I\subset\mathbb{N}$ be a finite index set. If $f_\rho=L_K^r g_\rho$ for some $g_\rho\in\mathcal{H}_K$, then when $0<r<1/2$,*

$$\left(\sum_{i\in I}(\sqrt{\lambda_i^{\mathbf{x}}}\langle f_\rho,\phi_i^{\mathbf{x}}\rangle_K)^2\right)^{1/2}\leq 2^r\|g_\rho\|_K(\#I)^{(1-2r)/4}\|L_K-L_K^{\mathbf{x}}\|_{\mathrm{HS}}^{(1+2r)/2}$$
$$+2^r\|g_\rho\|_K\left(\sum_{i\in I}(\lambda_i^{\mathbf{x}})^{1+2r}\right)^{1/2},\tag{39}$$

and when $r\geq 1/2$,

$$\left(\sum_{i\in I}(\sqrt{\lambda_i^{\mathbf{x}}}\langle f_\rho,\phi_i^{\mathbf{x}}\rangle_K)^2\right)^{1/2}\leq\sqrt{2}\lambda_1^{r-\frac{1}{2}}\|g_\rho\|_K\|L_K-L_K^{\mathbf{x}}\|_{\mathrm{HS}}$$
$$+2^r\|g_\rho\|_K\left(\sum_{i\in I}(\lambda_i^{\mathbf{x}})^{1+2r}\right)^{1/2}.\tag{40}$$

Proof Let $g_\rho=\sum_{j=1}^\infty d_j\phi_j$ with $\{d_j\}\in\ell^2$. Then $\|\{d_j\}\|_{\ell^2}=\|g_\rho\|_K$ and $f_\rho=\sum_{j=1}^\infty\lambda_j^r d_j\phi_j$. For $i\in I$, since whenever $\lambda_i^{\mathbf{x}}=0$, $\sqrt{\lambda_i^{\mathbf{x}}}\langle f_\rho,\phi_i^{\mathbf{x}}\rangle_K=0$, without loss of generality we assume $\lambda_i^{\mathbf{x}}>0$. Then we expand $\sqrt{\lambda_i^{\mathbf{x}}}\langle f_\rho,\phi_i^{\mathbf{x}}\rangle_K$ as

$$\sqrt{\lambda_i^{\mathbf{x}}}\langle f_\rho,\phi_i^{\mathbf{x}}\rangle_K=\left(\sum_{j:\lambda_j\geq 2\lambda_i^{\mathbf{x}}}+\sum_{j:\lambda_j<2\lambda_i^{\mathbf{x}}}\right)\sqrt{\lambda_i^{\mathbf{x}}}\lambda_j^r d_j\langle\phi_j,\phi_i^{\mathbf{x}}\rangle_K.\tag{41}$$

The second sum in (41) is easy to handle:

$$\left|\sum_{j:\lambda_j<2\lambda_i^{\mathbf{x}}}\sqrt{\lambda_i^{\mathbf{x}}}\lambda_j^r d_j\langle\phi_j,\phi_i^{\mathbf{x}}\rangle_K\right|\leq 2^r(\lambda_i^{\mathbf{x}})^{(1+2r)/2}\|\{d_j\}\|_{\ell^2}\left(\sum_{j=1}^\infty\langle\phi_j,\phi_i^{\mathbf{x}}\rangle_K^2\right)^{1/2}$$
$$=2^r\|g_\rho\|_K(\lambda_i^{\mathbf{x}})^{(1+2r)/2}.\tag{42}$$

When $r\geq 1/2$ and $\lambda_j\geq 2\lambda_i^{\mathbf{x}}$, since $\sqrt{\lambda_i^{\mathbf{x}}\lambda_j}\leq\frac{\lambda_j}{\sqrt{2}}\leq\sqrt{2}(\lambda_j-\lambda_i^{\mathbf{x}})$, the first sum in (41) can be bounded as

$$\left|\sum_{j:\lambda_j\geq 2\lambda_i^{\mathbf{x}}}\sqrt{\lambda_i^{\mathbf{x}}}\lambda_j^r d_j\langle\phi_j,\phi_i^{\mathbf{x}}\rangle_K\right|\leq\sum_{j:\lambda_j\geq 2\lambda_i^{\mathbf{x}}}\lambda_j^{r-\frac{1}{2}}\sqrt{2}\left|(\lambda_j-\lambda_i^{\mathbf{x}})\langle\phi_j,\phi_i^{\mathbf{x}}\rangle_K d_j\right|$$
$$\leq\sqrt{2}\lambda_1^{r-\frac{1}{2}}\|g_\rho\|_K\left(\sum_{j=1}^\infty(\lambda_j-\lambda_i^{\mathbf{x}})^2\langle\phi_j,\phi_i^{\mathbf{x}}\rangle_K^2\right)^{1/2}.\tag{43}$$

When $0 < r < 1/2$ and $\lambda_j \geq 2\lambda_i^{\mathbf{x}}$, we observe that

$$\frac{\sqrt{\lambda_i^{\mathbf{x}}}\lambda_j^r}{|\lambda_j - \lambda_i^{\mathbf{x}}|^{r+\frac{1}{2}}} \leq \frac{\lambda_j^{r+\frac{1}{2}}/\sqrt{2}}{(\lambda_j/2)^{r+\frac{1}{2}}} = 2^r.$$

So in this case the first sum in (41) can also be bounded as

$$\begin{aligned}
&\left| \sum_{j:\lambda_j \geq 2\lambda_i^{\mathbf{x}}} \sqrt{\lambda_i^{\mathbf{x}}}\lambda_j^r d_j \left\langle \phi_j, \phi_i^{\mathbf{x}} \right\rangle_K \right| \leq \sum_{j:\lambda_j \geq 2\lambda_i^{\mathbf{x}}} 2^r |\lambda_j - \lambda_i^{\mathbf{x}}|^{r+\frac{1}{2}} \left| d_j \left\langle \phi_j, \phi_i^{\mathbf{x}} \right\rangle_K \right| \\
&\leq 2^r \left(\sum_j d_j^2 \right)^{1/2} \left(\sum_j \left| (\lambda_j - \lambda_i^{\mathbf{x}}) \left\langle \phi_j, \phi_i^{\mathbf{x}} \right\rangle_K \right|^{(r+\frac{1}{2})\frac{2}{r+\frac{1}{2}}} \right)^{\frac{r+\frac{1}{2}}{2}} \left(\sum_j \left| \left\langle \phi_j, \phi_i^{\mathbf{x}} \right\rangle_K \right|^{(\frac{1}{2}-r)\frac{2}{\frac{1}{2}-r}} \right)^{\frac{\frac{1}{2}-r}{2}} \\
&= 2^r \|g_\rho\|_K \left(\sum_{j=1}^{\infty} (\lambda_j - \lambda_i^{\mathbf{x}})^2 \left\langle \phi_j, \phi_i^{\mathbf{x}} \right\rangle_K^2 \right)^{(1+2r)/4}. \qquad (44)
\end{aligned}$$

By (41), (42), and the triangle inequality,

$$\begin{aligned}
\left(\sum_{i\in I} |\sqrt{\lambda_i^{\mathbf{x}}} \left\langle f_\rho, \phi_i^{\mathbf{x}} \right\rangle_K|^2 \right)^{1/2} &\leq \left(\sum_{i\in I} \left(\sum_{j:\lambda_j \geq 2\lambda_i^{\mathbf{x}}} \sqrt{\lambda_i^{\mathbf{x}}}\lambda_j^r d_j \left\langle \phi_j, \phi_i^{\mathbf{x}} \right\rangle_K \right)^2 \right)^{1/2} \\
&\quad + 2^r \|g_\rho\|_K \left(\sum_{i\in I} (\lambda_i^{\mathbf{x}})^{1+2r} \right)^{1/2}. \qquad (45)
\end{aligned}$$

We denote the first term of the right-hand side of (45) as Υ. The definition of the Hilbert-Schmidt norm tells us that

$$\|L_K - L_K^{\mathbf{x}}\|_{HS}^2 = \sum \|(L_K - L_K^{\mathbf{x}})\phi_i^{\mathbf{x}}\|_K^2 = \sum_{i,j=1}^{\infty} (\lambda_j - \lambda_i^{\mathbf{x}})^2 \left\langle \phi_i^{\mathbf{x}}, \phi_j \right\rangle_K^2. \qquad (46)$$

So when $r \geq 1/2$, (43) and (46) give

$$\begin{aligned}
\Upsilon &\leq \sqrt{2}\lambda_1^{r-\frac{1}{2}} \|g_\rho\|_K \left(\sum_{i\in I} \sum_{j=1}^{\infty} (\lambda_j - \lambda_i^{\mathbf{x}})^2 \left\langle \phi_j, \phi_i^{\mathbf{x}} \right\rangle_K^2 \right)^{1/2} \\
&\leq \sqrt{2}\lambda_1^{r-\frac{1}{2}} \|g_\rho\|_K \|L_K - L_K^{\mathbf{x}}\|_{\mathrm{HS}},
\end{aligned}$$

which proves (40). When $0 < r < 1/2$, by (44), (46) and Hölder's inequality, we have

$$\Upsilon \leq 2^r \|g_\rho\|_K \left(\sum_{i\in I} \left(\sum_{j=1}^{\infty} (\lambda_j - \lambda_i^{\mathbf{x}})^2 \left\langle \phi_j, \phi_i^{\mathbf{x}} \right\rangle_K^2 \right)^{(1+2r)/2} \right)^{1/2}$$

$$\begin{aligned} &\leq 2^r\|g_\rho\|_K \left(\sum_{i\in I}\sum_{j=1}^{\infty}(\lambda_j-\lambda_i^{\mathbf{x}})^2\left\langle\phi_j,\phi_i^{\mathbf{x}}\right\rangle_K^2\right)^{(1+2r)/4}\left(\sum_{i\in I}1^{\frac{2}{1-2r}}\right)^{(1-2r)/4}\\ &\leq 2^r\|g_\rho\|_K\|L_K-L_K^{\mathbf{x}}\|_{\mathrm{HS}}^{(1+2r)/2}(\#I)^{(1-2r)/4}, \end{aligned}$$

which verifies (39). The proof of the lemma is complete. ■

We are now in a position to prove Theorem 15.

Proof of Theorem 15. By Lemmas 18 and 19, we know that for any $0<\delta<\frac{1}{2}$ there exists a subset Z_δ of Z^m of measure at least $1-2\delta$ such that both (37) and (38) hold for each $\mathbf{z}\in Z_\delta$.

Let $\mathbf{z}\in Z_\delta$.

To prove $c_i^{\mathbf{z}}=0$ for $i=p+1,\ldots,m$, we show that condition (50) for γ, to be defined below which is the same as condition (24) in the statement of the theorem after scaling δ to $\delta/2$, implies (26) and thereby (20) in Theorem 13, according to (22) in Lemma 14. To this end, we estimate $\sqrt{\lambda_i^{\mathbf{x}}}|S_i^{\mathbf{z}}|$ and $\sqrt{\lambda_i^{\mathbf{x}}}|S_i^{\mathbf{z}}|\cdot\sqrt{\lambda_i^{\mathbf{x}}}$.

We first apply Lemma 20 to the set $I=\{i\}\subset\{p+1,\ldots,m\}$ and Lemma 18 and see that in either case of $0<r<1/2$ and $r\geq 1/2$, there holds

$$\begin{aligned} \sqrt{\lambda_i^{\mathbf{x}}}\left|\left\langle f_\rho,\phi_i^{\mathbf{x}}\right\rangle_K\right| &\leq \left(2^r+\sqrt{2}\lambda_1^{\max\{r-\frac{1}{2},0\}}\right)\|g_\rho\|_K\|L_K-L_K^{\mathbf{x}}\|_{\mathrm{HS}}^{\min\{\frac{1+2r}{2},1\}}+2^r\|g_\rho\|_K\left(\lambda_i^{\mathbf{x}}\right)^{\frac{1+2r}{2}}\\ &\leq C_1'\left(\lambda_i^{(1+2r)/2}+\left(\log\frac{2}{\delta}\right)^{(1+2r)/2}m^{-\min\{1/2,(1+2r)/4\}}\right), \end{aligned}$$

where

$$C_1'=\left(2^r+\sqrt{2}\lambda_1^{\max\{r-\frac{1}{2},0\}}\right)(2\kappa)^{\min\{1+2r,2\}}\|g_\rho\|_K+2^{2r+\frac{1}{2}}\|g_\rho\|_K(2\kappa+1)^{1+2r}.$$

This together with (38) in Lemma 19 gives

$$\begin{aligned} &\sqrt{\lambda_i^{\mathbf{x}}}|S_i^{\mathbf{z}}|\leq\sqrt{\lambda_i^{\mathbf{x}}}|S_i^{\mathbf{z}}-\left\langle f_\rho,\phi_i^{\mathbf{x}}\right\rangle_K|+\sqrt{\lambda_i^{\mathbf{x}}}\left|\left\langle f_\rho,\phi_i^{\mathbf{x}}\right\rangle_K\right|\\ &\leq C_1'\lambda_i^{r+\frac{1}{2}}+\left(2\sqrt{2}M+C_1'\right)\left(\log\frac{2m}{\delta}\right)^{(1+2r)/2}m^{-\min\{1/2,(1+2r)/4\}}\\ &\leq\left(2\sqrt{2}M+2C_1'\right)\left(\log\frac{2m}{\delta}\right)^{(1+2r)/2}\max\left\{\left(\max\left\{\lambda_p,\frac{1}{\sqrt{m}}\right\}\right)^{r+\frac{1}{2}},\frac{1}{\sqrt{m}}\right\}. \end{aligned}\tag{47}$$

It follows that the first inequality $\sqrt{\lambda_i^{\mathbf{x}}}|S_i^{\mathbf{z}}|<\frac{C_{\Omega,1}}{2}\sqrt{\gamma}$ of (26) is valid if γ satisfies

$$\gamma>\left(\frac{4\sqrt{2}M+4C_1'}{C_{\Omega,1}}\right)^2\left(\log\frac{2m}{\delta}\right)^{1+2r}\max\left\{\left(\max\left\{\lambda_p,\frac{1}{\sqrt{m}}\right\}\right)^{2r+1},\frac{1}{m}\right\}.\tag{48}$$

Then we estimate $\lambda_i^{\mathbf{x}}$ by Lemma 18 as

$$\sqrt{\lambda_i^{\mathbf{x}}}\leq\sqrt{\lambda_i}+\sqrt{|\lambda_i-\lambda_i^{\mathbf{x}}|}\leq(2\kappa+1)\sqrt{\log\frac{2}{\delta}}\left(\max\left\{\lambda_p,\frac{1}{\sqrt{m}}\right\}\right)^{\frac{1}{2}}.$$

Combining this with (47), we know that the second inequality $\sqrt{\lambda_i^{\mathbf{x}}}|S_i^{\mathbf{z}}| \cdot \sqrt{\lambda_i^{\mathbf{x}}} < \frac{C_{\Omega,1}}{2}\gamma$ of (26) is valid if γ satisfies

$$\gamma > \quad \frac{4\sqrt{2}M + 4C_1'}{C_{\Omega,1}} (2\kappa+1) \left(\log \frac{2m}{\delta}\right)^{1+r} \max\left\{ \left(\max\left\{\lambda_p, \frac{1}{\sqrt{m}}\right\}\right)^{r+1}, \right.$$
$$\left. \frac{1}{\sqrt{m}} \left(\max\left\{\lambda_p, \frac{1}{\sqrt{m}}\right\}\right)^{\frac{1}{2}} \right\}. \tag{49}$$

Now we can choose the constant C_1 from (48) and (49) by

$$C_1 = \max\left\{ \left(\frac{4\sqrt{2}M + 4C_1'}{C_{\Omega,1}}\right)^2 (1+\lambda_1)^{2r+1}, \frac{4\sqrt{2}M + 4C_1'}{C_{\Omega,1}} (2\kappa+1)(1+\lambda_1)^{2r+1}, 1 \right\}.$$

With this choice, we know that for γ satisfying

$$\gamma \geq \begin{cases} C_1 \left(\log \frac{2m}{\delta}\right)^{1+2r} \left(\max\left\{\frac{\lambda_p}{\lambda_1}, \frac{1}{\sqrt{m}}\right\}\right)^{r+1}, & \text{if } 0 < r \leq \frac{1}{2}, \\ C_1 \left(\log \frac{2m}{\delta}\right)^{1+2r} \max\left\{ \left(\frac{\lambda_p}{\lambda_1}\right)^{r+\frac{1}{2}}, \frac{1}{\sqrt{m}} \right\} \left(\max\left\{\frac{\lambda_p}{\lambda_1}, \frac{1}{\sqrt{m}}\right\}\right)^{\frac{1}{2}}, & \text{if } r > \frac{1}{2}, \end{cases} \tag{50}$$

both (48) and (49) are valid, which implies (26). Then by (22) in Lemma 14, we see that condition (20) in Theorem 13 is valid and hence $c_i^{\mathbf{z}} = 0$ for $i = p+1, \ldots, m$.

Now we turn to the desired error bound. Assume (50) for γ. Define an index set $\mathcal{S}$ by $\mathcal{S} = \{i \in \{1, \cdots, p\} : \lambda_i^{\mathbf{x}} > \lambda_p/2\}$.

When $i \in \{1, \cdots, p\}$ but $\lambda_i^{\mathbf{x}} \leq \lambda_p/2$, we check the process in proving (47) and see from the restriction $\lambda_i^{\mathbf{x}} \leq \lambda_p/2$ that condition (50) for γ ensures (26). Then by (22) in Lemma 14, we see that condition (20) is valid for i. Hence $c_i^{\mathbf{z}} = 0$ for $i \in \mathbb{N}\backslash\mathcal{S}$. So we can expand $\|f_\rho - f^{\mathbf{z}}\|_K$ with respect to the orthonormal basis $\{\phi_i^{\mathbf{x}}\}$ of $\mathcal{H}_K$ as

$$\|f_\rho - f^{\mathbf{z}}\|_K^2 = \sum_{i \in \mathbb{N}\backslash\mathcal{S}} \left(\langle f_\rho, \phi_i^{\mathbf{x}}\rangle_K\right)^2 + \sum_{i \in \mathcal{S}} \left(c_i^{\mathbf{z}} - \langle f_\rho, \phi_i^{\mathbf{x}}\rangle_K\right)^2. \tag{51}$$

For any $i \in \mathcal{S}$, we have $\lambda_i^{\mathbf{x}} > \lambda_p/2 > 0$ and

$$|c_i^{\mathbf{z}} - \langle f_\rho, \phi_i^{\mathbf{x}}\rangle_K| \leq |c_i^{\mathbf{z}} - S_i^{\mathbf{z}}| + \frac{\sqrt{\lambda_i^{\mathbf{x}}}|S_i^{\mathbf{z}} - \langle f_\rho, \phi_i^{\mathbf{x}}\rangle_K|}{\sqrt{\lambda_p/2}}.$$

Applying Theorem 13 (b), Lemma 14 and Lemma 19 gives

$$|c_i^{\mathbf{z}} - \langle f_\rho, \phi_i^{\mathbf{x}}\rangle_K| \leq C_{\Omega,2} \max\left\{ \left(\frac{2\gamma}{\lambda_p}\right)^{\frac{1}{2-q}}, \frac{2\gamma}{\lambda_p} \right\} + \frac{2\sqrt{2}M}{\sqrt{m\lambda_p/2}} \sqrt{\log \frac{2m}{\delta}}.$$

It follows that

$$\sqrt{\sum_{i \in \mathcal{S}} \left(c_i^{\mathbf{z}} - \langle f_\rho, \phi_i^{\mathbf{x}}\rangle_K\right)^2} \leq C_{\Omega,2}\sqrt{p} \left\{ \left(\frac{2\gamma}{\lambda_p}\right)^{\frac{1}{2-q}} + \frac{2\gamma}{\lambda_p} \right\} + \frac{4\sqrt{p}M}{\sqrt{m\lambda_p}} \sqrt{\log \frac{2m}{\delta}}.$$

To estimate the first sum in (51) we cite an estimate from Guo and Zhou (2012) for the quantity $\left(\sum_{i\in\mathbb{N}\setminus\mathcal{S}}(\langle f_\rho,\phi_i^{\mathbf{x}}\rangle_K)^2\right)^{1/2}$ which is independent of the regularizing function Ω and know that it can be bounded by

$$\|g_\rho\|_K\lambda_{p+1}^r+2^{\max\{r,1\}}\|g_\rho\|_K\lambda_p^{\min\{r-1,0\}}\left((\sum_{i=p+1}^{\infty}\lambda_i^{\max\{2r,2\}})^{\frac{1}{2}}+c_{r,\lambda_1}\|L_K-L_K^{\mathbf{x}}\|_{HS}\right),$$

where c_{r,λ_1} is the constant given by

$$c_{r,\lambda_1}=\begin{cases}2^{1+r}\lambda_1^{r-1}, & \text{if } r\geq 1,\\ 2, & \text{if } r<1.\end{cases}$$

Therefore

$$\begin{aligned}\|f^{\mathbf{z}}-f_\rho\|_K &\leq C_{\Omega,2}\sqrt{p}\left\{\left(\frac{2\gamma}{\lambda_p}\right)^{\frac{1}{2-q}}+\frac{2\gamma}{\lambda_p}\right\}+\|g_\rho\|_K\lambda_{p+1}^r+\frac{2\sqrt{p}M}{\sqrt{m\lambda_p}}\sqrt{\log\frac{2m}{\delta}}\\ &\quad+2^{\max\{r,1\}}\|g_\rho\|_K\lambda_p^{\min\{r-1,0\}}\left(\left(\sum_{i=p+1}^{\infty}\lambda_i^{\max\{2r,2\}}\right)^{1/2}+\frac{4c_{r,\lambda_1}\kappa^2\log\frac{2}{\delta}}{\sqrt{m}}\right)\\ &\leq C_{\Omega,2}\sqrt{p}\left\{\left(\frac{2\gamma}{\lambda_p}\right)^{\frac{1}{2-q}}+\frac{2\gamma}{\lambda_p}\right\}+\|g_\rho\|_K\lambda_{p+1}^r+C_3\frac{\sqrt{p}\log\frac{2m}{\delta}}{\sqrt{m}}\lambda_p^{\min\{-\frac{1}{2},r-1\}}\\ &\quad+C_4\lambda_p^{\min\{r-1,0\}}\left(\sum_{i=p+1}^{\infty}\lambda_i^{\max\{2r,2\}}\right)^{1/2},\end{aligned}$$

where

$$C_3=2M\lambda_1^{\max\{\frac{1}{2}-r,0\}}+2^{\max\{r,1\}+2}\|g_\rho\|_K c_{r,\lambda_1}\kappa^2\lambda_1^{\max\{r-\frac{1}{2},0\}}$$

and $C_4=2^{\max\{r,1\}}\|g_\rho\|_K$. After scaling δ to $\delta/2$, the proof of Theorem 15 is completed. ■

Appendix B. Minimax Lower Bounds

In this appendix, we derive a general minimax lower bound which includes Theorem 8 as a special case. First we define two sets of Borel probability measures.

Definition 21 *Let $\mathcal{P}(\alpha_1,\alpha_2,r,M,R,D_1,D_2)$ be the set of all Borel probability measures ρ on $X\times Y$ satisfying the following three conditions:*

1. *$|y|\leq M$ almost surely,*
2. *$f_\rho=L_K^r(g_\rho)$ for some $g_\rho\in\mathcal{H}_K$ with $\|g_\rho\|_K\leq R$,*
3. *$D_1 i^{-\alpha_1}\leq\lambda_i\leq D_2 i^{-\alpha_2}$ for each i.*

Let $\mathcal{P}(\beta_1, \beta_2, r, M, R, D_1, D_2)$ be the same as $\mathcal{P}(\alpha_1, \alpha_2, r, M, R, D_1, D_2)$ except that the last condition is replaced by $D_1\beta_1^{-i} \leq \lambda_i \leq D_2\beta_2^{-i}$ for each i.

For simplicity, we abbreviate these two sets as $\mathcal{P}(\alpha_1, \alpha_2, r)$ and $\mathcal{P}(\beta_1, \beta_2, r)$, respectively. Now we state the general minimax lower bound for the error in the $\mathcal{H}_K$-metric following the idea of (Caponnetto and De Vito, 2007). Our proof is mainly based on Lemma 2.9, Theorem 2.5 and the approach from (Tsybakov, 2009).

Theorem 22 *Assume $R > 0$ and $M \geq 4\kappa^{2r+1}R$. Let $f^{\mathbf{z}} \in \mathcal{H}_K$ be the output of an arbitrary learning algorithm based on the sample $\boldsymbol{z} = \{(x_i, y_i)\}_{i=1}^m$. Then for every $0 < \delta < \frac{1}{8}$, there exist positive constants τ_1, τ_2, independent of δ or m, such that*

$$\lim_{m\to\infty} \inf_{f^{\mathbf{z}}} \sup_{\rho\in\mathcal{P}(\alpha_1,\alpha_2,r)} \mathbb{P}_{\mathbf{z}\sim\rho^m} \left\{ \|f^{\mathbf{z}} - f_\rho\|_K \geq \tau_1 \delta^{\frac{\alpha_1 r}{\alpha_2(2r+1)+1}} m^{-\frac{\alpha_1 r}{\alpha_2(2r+1)+1}} \right\} \geq 1 - 2\delta \tag{52}$$

and

$$\lim_{m\to\infty} \inf_{f^{\mathbf{z}}} \sup_{\rho\in\mathcal{P}(\beta_1,\beta_2,r)} \mathbb{P}_{\mathbf{z}\sim\rho^m} \left\{ \|f^{\mathbf{z}} - f_\rho\|_K \geq \tau_2 \delta^{\frac{1}{2}} m^{-\frac{1}{2}} \sqrt{\log m} \right\} \geq 1 - 2\delta. \tag{53}$$

Proof First, we associate a probability measure $\rho_f \in \mathcal{P}(\alpha_1, \alpha_2, r)$ to a pair (μ, f) where μ is a Borel measure on Y such that the eigenvalues of the associated integral operator L_K satisfy $D_1 i^{-\alpha_1} \leq \lambda_i \leq D_2 i^{-\alpha_2}$, and $f = L_K^r g$ for some $g \in \mathcal{H}_K$ with $\|g\|_K \leq R$. Define a probability measure ρ_f by

$$d\rho_f(x, y) = \left[\frac{B + f(x)}{2B} d\delta_B(y) + \frac{B - f(x)}{2B} d\delta_{-B}(y) \right] d\mu(x),$$

where $B = 4\kappa^{2r+1}R$ and $d\delta_\xi$ denotes the Dirac delta with unit mass at ξ. By the reproducing property, $\|f\|_\infty \leq \kappa \|L_K^r g\|_K \leq \kappa^{2r+1}R = \frac{B}{4}$. It follows that ρ_f is a probability measure on $X \times Y$ with μ being the marginal distribution and f the regression function. Moreover, $M \geq 4\kappa^{2r+1}R$ ensures $|y| \leq M$ almost surely. Hence $\rho_f \in \mathcal{P}(\alpha_1, \alpha_2, r)$.

Then we construct a finite sequence $f_0, ..., f_N$ in the set $\{L_K^r g : g \in \mathcal{H}_K, \|g\|_K \leq R\}$ based on the Varshamov-Gilbert bound (Lemma 2.9 in (Tsybakov, 2009)) which asserts that for any integer $\gamma \geq 8$, there exists a set $\Theta = \{w_0, w_1, ..., w_N\} \subset \{0, 1\}^\gamma$ such that

1. $w_0 = (0, ..., 0)$.
2. For any $i \neq j$, $H(w_i, w_j) > \gamma/8$, where $H(\cdot, \cdot)$ is the Hamming distance.
3. $N \geq 2^{\gamma/8}$.

For $0 < \delta < \frac{1}{8}$, let γ be the smallest integer greater than or equal to $c_\delta m^{\frac{1}{\alpha_2(2r+1)+1}}$ with a constant $c_\delta > 0$ to be specified later. For $w_i = (w_i^{\gamma+1}, ..., w_i^{2\gamma}) \in \Theta$ with $i \in \{0, \ldots, N\}$, define $f_i = L_K^r g_i$ with

$$g_i = \sum_{k=\gamma+1}^{2\gamma} w_i^k R \gamma^{-\frac{1}{2}} \phi_k.$$

Note that $g_i \in \mathcal{H}_K$ and

$$\|g_i\|_K^2 = \left\| \sum_{k=\gamma+1}^{2\gamma} w_i^k R \gamma^{-\frac{1}{2}} \phi_k \right\|_K^2 = \sum_{k=\gamma+1}^{2\gamma} (w_i^k)^2 \gamma^{-1} R^2 \|\phi_k\|_K^2 \le R^2.$$

Hence $\{f_0, ..., f_N\} \subset \{L_K^r g : g \in \mathcal{H}_K, \|g\|_K \le R\}$, which implies $\{\rho_{f_0}, ..., \rho_{f_N}\} \subset \mathcal{P}(\alpha_1, \alpha_2, r)$.

Now we adopt Theorem 2.5 in (Tsybakov, 2009) to establish our desired lower bound.

Observe that for $0 \le i < j \le N$, the Kullback-Leibler distance $\mathcal{D}_{KL}(\rho_{f_i} \| \rho_{f_j})$ between ρ_{f_i} and ρ_{f_j} satisfies

$$\begin{aligned}
&\mathcal{D}_{KL}(\rho_{f_i} \| \rho_{f_j}) \\
&= \int_X \left\{ \frac{B + f_i(x)}{2B} \ln \left(1 + \frac{f_i(x) - f_j(x)}{B + f_j(x)} \right) + \frac{B - f_i(x)}{2B} \ln \left(1 - \frac{f_i(x) - f_j(x)}{B - f_j(x)} \right) \right\} d\mu(x) \\
&\le \frac{f_i(x) - f_j(x)}{2B} \left\{ \frac{B + f_i(x)}{B + f_j(x)} - \frac{B - f_i(x)}{B - f_j(x)} \right\} \\
&\le \frac{16}{15B^2} \lambda_\gamma^{2r+1} R^2 \gamma^{-1} \sum_{k=\gamma+1}^{2\gamma} (w_i^k - w_j^k)^2 \le \frac{D_2^{2r+1}}{15\kappa^{4r+2}} \gamma^{-\alpha_2(2r+1)},
\end{aligned}$$

which implies

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^N \mathcal{D}_{KL}((\rho_{f_i})^m \| (\rho_{f_0})^m) &\le \frac{D_2^{2r+1}}{15\kappa^{4r+2}} m \gamma^{-\alpha_2(2r+1)} \le \frac{D_2^{2r+1}}{15\kappa^{4r+2}} c_\delta^{-\alpha_2(2r+1)} m^{\frac{1}{\alpha_2(2r+1)+1}} \\
&\le \frac{D_2^{2r+1}}{15\kappa^{4r+2} c_\delta^{1+\alpha_2(2r+1)}} \gamma = \delta \log 2^{\gamma/8} \le \delta \log N
\end{aligned}$$

by taking

$$c_\delta = \left(\frac{8 D_2^{2r+1}}{15\kappa^{4r+2} \log 2} \right)^{1/(\alpha_2(2r+1)+1)} \delta^{-1/(\alpha_2(2r+1)+1)}.$$

On the other hand, for any $0 \le i < j \le N$.

$$\begin{aligned}
\|f_i - f_j\|_K^2 &= \|L_K^r (g_i - g_j)\|_K^2 \\
&= \sum_{k=\gamma+1}^{2\gamma} R^2 \gamma^{-1} \lambda_k^{2r} (w_i^k - w_j^k)^2 \\
&\ge R^2 \gamma^{-1} \lambda_{2\gamma}^{2r} \sum_{k=\gamma+1}^{2\gamma} (w_i^k - w_j^k)^2 \\
&= R^2 \gamma^{-1} \lambda_{2\gamma}^{2r} H(w_i, w_j) \\
&\ge \frac{R^2 \lambda_{2\gamma}^{2r}}{8} \\
&\ge 2^{-(2\alpha_1 r + 3)} R^2 D_1^{2r} \gamma^{-2\alpha_1 r} \\
&\ge 2\tau_1^2 \delta^{\frac{2\alpha_1 r}{\alpha_2(2r+1)+1}} m^{-\frac{2\alpha_1 r}{\alpha_2(2r+1)+1}}
\end{aligned}$$

for some constant $\tau_1 > 0$.

Therefore, as shown in (Tsybakov, 2009) we have

$$\inf_{f^{\mathbf{z}}} \sup_{\rho \in \mathcal{P}(\alpha_1,\alpha_2,r)} \mathbb{P}_{\mathbf{z}\sim\rho^m} \left\{ \|f^{\mathbf{z}} - f_\rho\|_K \geq \tau_1 \delta^{\frac{\alpha_1 r}{\alpha_2(2r+1)+1}} m^{-\frac{\alpha_1 r}{\alpha_2(2r+1)+1}} \right\}$$
$$\geq \frac{\sqrt{N}}{\sqrt{N}+1} \left(1 - 2\delta - \sqrt{\frac{2\delta}{\log N}} \right).$$

This completes the proof for the statement about $\mathcal{P}(\alpha_1, \alpha_2, r)$. The proof for the statement about $P(\beta_1, \beta_2, r)$ is similar. The proof of the theorem is complete. ■

Appendix C. Proof of Proposition 10

(a). Let $0 \leq \xi_1 < \xi_2 < \xi_3 < \xi_4$. For $i = 2$ or 3, one has

$$\Omega(\xi_i) \geq \frac{(\xi_{i+1} - \xi_i)\Omega(\xi_{i-1})}{\xi_{i+1} - \xi_{i-1}} + \frac{(\xi_i - \xi_{i-1})\Omega(\xi_{i+1})}{\xi_{i+1} - \xi_{i-1}}. \tag{54}$$

Let $i = 2$ in (54) to give

$$\begin{aligned} (\xi_2 - \xi_1)\Omega(\xi_3) &\leq (\xi_3 - \xi_1)\Omega(\xi_2) - (\xi_3 - \xi_2)\Omega(\xi_1) \\ &= \xi_3[\Omega(\xi_2) - \Omega(\xi_1)] - \xi_1\Omega(\xi_2) + \xi_2\Omega(\xi_1). \end{aligned} \tag{55}$$

If $\Omega(\xi_2) < \Omega(\xi_1)$, let $\xi_3 \to \infty$ to give $\Omega(\xi_3) \to -\infty$, which contradicts $\Omega([0,\infty)) \subset [0,\infty)$. So Ω is nondecreasing. Similarly we have $(\xi_3 - \xi_1)\Omega(\xi_2) \geq (\xi_3 - \xi_1 + \xi_1 - \xi_2)\Omega(\xi_1) + (\xi_2 - \xi_1)\Omega(\xi_3)$, so

$$\frac{\Omega(\xi_2) - \Omega(\xi_1)}{\xi_2 - \xi_1} \geq \frac{\Omega(\xi_3) - \Omega(\xi_1)}{\xi_3 - \xi_1} \geq 0. \tag{56}$$

If $\Omega(\xi_2) = 0$, since Ω is nondecreasing, $\Omega(\xi_1) = 0$ for all $0 \leq \xi_1 < \xi_2$, and (56) gives $\Omega(\xi_3) = 0$ for all $\xi_3 > \xi_2$, so we have $\Omega = 0$, a contradiction. Therefore $\Omega(c) > 0$ for $c > 0$.

From (56), the function $[\Omega(c) - \Omega(\xi_1)]/(c - \xi_1)$ of c is nonincreasing when $c > \xi_1$, so the right-hand derivative Ω'_+ is well-defined on $[0,\infty)$, taking values in $[0,\infty]$. We get from (55) that

$$\infty > \frac{\Omega(\xi_2) - \Omega(\xi_1)}{\xi_2 - \xi_1} \geq \frac{\Omega(\xi_3) - \Omega(\xi_2)}{\xi_3 - \xi_2}. \tag{57}$$

Let $\xi_3 \to \xi_2^+$ to give $\Omega'_+(\xi_2) < \infty$. Therefore $\Omega'_+(c)$ is finite for $c \in (0,\infty)$. Let $i = 3$. We have the analogue of (57),

$$\frac{\Omega(\xi_3) - \Omega(\xi_2)}{\xi_3 - \xi_2} \geq \frac{\Omega(\xi_4) - \Omega(\xi_3)}{\xi_4 - \xi_3}, \tag{58}$$

which, together with (57), gives that as $\xi_2 \to \xi_1^+$ and $\xi_4 \to \xi_3^+$, $\Omega'_+(\xi_1) \geq \Omega'_+(\xi_3)$. This proves that Ω'_+ is nonincreasing on $[0,\infty)$. If $\Omega'_+(0) = 0$, since $0 \leq \frac{\Omega(c) - \Omega(0)}{c - 0}$ is nonincreasing

for $c > 0$ as we proved before, we have $\Omega(c) = 0$ for all $c > 0$, a contradiction again. So $\Omega'_+(0) \in (0, \infty]$.

(b). Let $\xi_1 = 0$ and $\xi_3 = 1$, then (55) gives $\Omega(\xi_2) \geq \xi_2\Omega(1)$, so for all $c \in (0, 1]$, $\Omega(c) \geq c\Omega(1)$. In (55) let $\xi_1 = 0$ and $\xi_2 = 1$ to give $\Omega(\xi_3) \leq \xi_3\Omega(1)$, so for any $c \in [1, \infty)$, $\Omega(c) \leq \Omega(1)c$.

The properties stated in (c) and (d) follow immediately from the concavity of the function Ω.

(e). Write the function $\frac{\Omega(c)}{c^2}$ as $\frac{\Omega(c)}{c} \cdot \frac{1}{c}$. We see from (d) that this function is strictly decreasing on $(0, \infty)$. By (a), we obtain the limit $\lim_{c\to 0^+} \frac{\Omega(c)}{c^2} \geq \underline{\lim}_{c\to 0^+} \frac{\Omega(1)c}{c^2} = +\infty$. By (b), $\lim_{c\to\infty} \frac{\Omega(c)}{c^2} \leq \overline{\lim}_{c\to\infty} \frac{\Omega(1)c}{c^2} = 0$. The proof of Proposition 10 is complete. ■

References

F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23:52-72, 2007.

G. Blanchard, P. Massart, R. Vert, and L. Zwald. Kernel projection machine: a new tool for pattern recognition. *Advances in Neural Information Processing Systems*, 1649-1656, 2004.

G. Blanchard and L. Zwald. Finite dimensional projection for classification and statistical learning. *IEEE Transactions on Information Theory*, 54: 4169- 4182, 2008.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331-368, 2007.

A. Caponnetto and Y. Yao. Cross-validation based adaptation for regularization operators in learning theory. *Analysis and Applications*, 8:161-183, 2010.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 2005.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

R. DeVore, G. Kerkyacharian, D. Picard, and V. Temlyakov. Mathematical methods for supervised learning. *IMI Preprints*, 22:1-51, 2004.

D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers and Differential Operators*. Cambridge University Press, Cambridge, 1996.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348-1360, 2001.

I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35: 109–148, 1993.

W. Fu and K. Knight. Asymptotics for lasso-type estimators. *Annals of Statistics* 28:1356-1378, 2000.

X. Guo and D. X. Zhou. An empirical feature-based learning algorithm producing sparse approximations. *Applied and Computational Harmonic Analysis*, 32:389-400, 2012.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89:10915-9.

T. Hu, J. Fan, Q. Wu, and D. X. Zhou. Regularization schemes for minimum error entropy principle. *Analysis and Applications*, 13:437–455, 2015.

V. Koltchinskii and E. Giné. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6:113-167, 2000.

T. Kühn. Entropy numbers in sequence spaces with an application to weighted function spaces. *Journal of Approximation Theory*, 153:40–52, 2008.

Y. Liu, H. H. Zhang, C. Park, and J. Ahn. Support vector machines with adaptive Lq penalties. *Computational Statistics and Data Analysis*, 51:6380-6394, 2007.

L. Lo Gerfo, L. Rosasco, F. Odone, E. De Vito and A. Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20:1873-1897, 2008.

M. Nielsen and O. Lund. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics*, 10:296, 2009.

M. Nielsen, C. Lundegaard, T. Blicher, B. Peters, A. Sette, S. Justesen, S. Buus, and O. Lund. Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. *PLOS Computational Biology*, 4:e1000107, 2008.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13:389–427, 2012.

J.B. Reade. Eigenvalues of positive definite kernels II. *SIAM Journal on Mathematical Analysis*, 15:137-142, 1984.

J.B. Reade. Eigenvalues of analytic kernels. *SIAM Journal on Mathematical Analysis*, 15:133-136, 1984.

B. Schölkopf, A. Smola, and K. R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

W.J. Shen, H.S. Wong, Q.W. Xiao, X. Guo, and S. Smale. Introduction to the peptide binding problem of computational immunology: new results. *Foundations of Computational Mathematics*, 14:951–984, 2014.

S. Smale and D.X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153-172, 2007.

S. Smale and D. X. Zhou. Online learning with Markov sampling. *Analysis and Applications*, 7:87-113, 2009.

I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. *In Proceedings of the Annual Conference on Learning Theory*, 79-93, 2009.

T. Suzuki, R. Tomioka, and M. Sugiyama. Fast learning rate of multiple kernel learning: trade-off between sparsity and smoothness. *JMLR Workshop and Conference Proceedings*, 22: 1152-1183, 2012.

A. Tsybakov. *Introduction to Nonparametric Estimation*. New York, Springer, 2009.

Z. B. Xu, X. Y. Chang, F. M. Xu. L-1/2 regularization: A thresholding representation theory and a fast solver. *IEEE Transactions on Neural Networks and Learning Systems*, 23:1013-1027, 2012.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5): 1564–1599, 1999.

D. X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49:1743-1752, 2003.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67:301–320, 2005.

L. Zwald. Performances statistiques d'algorithmes d'apprentissage: Kernel Projection Machine et analyse en composantes principales à noyau. PhD thesis, Université Paris-Sud 11, 2005.

L. Zwald and G. Blanchard. On the convergence of eigenspaces in kernel principal component analysis. *In Advances in Neural Information Processing Systems 18 (Y. Weiss, B. Schölkopf, and J. Platt, eds.)*, 1649-1656. MIT Press, Cambridge, MA, 2006.