On spectral windows in supervised learning from data

Giorgio Gnecco, Marcello Sanguineti

Department of Communications, Computer, and System Sciences (DIST), University of Genoa, Via Opera Pia 13, 16145 Genova, Italy

A R T I C L E   I N F O

Article history:
Received 10 June 2009
Received in revised form 20 August 2010
Accepted 24 August 2010
Available online 20 September 2010
Communicated by P.M.B. Vitányi

Keywords:
Analysis of algorithms
Learning from data
Regularization
Suboptimal solutions
Empirical error functionals
Probabilistic estimates

A B S T R A C T

For Tikhonov regularization in supervised learning from data, the effect on the regularized solution of a joint perturbation of the regression function and the data is investigated. Spectral windows in the finite-sample and population cases are compared via probabilistic estimates of the differences between regularized solutions.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

For a nonempty set $X \subseteq \mathbb{R}^d$ and a probability measure $\rho$ on $X \times \mathbb{R}$, Statistical Learning Theory [27] - a branch of Computational Learning Theory [19] - models the supervised learning problem as the minimization of the expected error functional

$$E(f) = \int_{X \times \mathbb{R}} (f(x) - y)^2 d\rho,$$

where $f : X \to \mathbb{R}$ belongs to a suitable space $\mathcal{H}$ of functions, called hypothesis space. We assume that there exists $N > 0$ such that $y \in [-N, N]$ and that $\rho$ is nondegenerate (i.e., nonempty open subsets of $X \times [-N, N]$ have strictly positive measure) and has a nondegenerate marginal probability measure on $X$, denoted by $\nu$.

Usually, $\rho$ is unknown and one has at disposal, for a positive integer $m$, a data sample

$$z = (x, y) = \{(x_i, y_i) \in X \times \mathbb{R}, \ i = 1, \ldots, m\},$$

where $(x_i, y_i)$, $i = 1, \ldots, m$, are random variables independent and identically distributed (i.i.d.) according to $\rho$. The information provided by the data sample can be exploited to minimize a suitable approximation of the expected error functional, instead of the expected error itself.

Typically, the problem of supervised learning from data is ill-posed [6] and regularization [26] can be used to cope with this drawback. A widespread regularization approach consists in minimizing over $\mathcal{H}$ the regularized empirical error functional

$$E_k(f) + \gamma \Psi(f),$$

where $E_k : \mathcal{H} \to \mathbb{R}$, defined as

$$E_k(f) = \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2,$$

is called empirical error functional, $\Psi : \mathcal{H} \to \mathbb{R}$ is a functional called stabilizer, and $\gamma > 0$ is a regularization parameter. The parameter $\gamma$ controls the trade-off between the following two requirements: i) fitting to the data sample (via the value $E_k(f)$ of the empirical error associated with $f$) and ii) penalizing solutions $f$ that provide a large value of the stabilizer $\Psi(f)$. 

© 2010 Elsevier B.V. All rights reserved. doi:10.1016/j.ipl.2010.08.011
As hypothesis spaces, in this paper we consider Hilbert spaces of a special type, called reproducing kernel Hilbert spaces (RKHSs). They were formally defined by Aronszajn [2] and are commonly used in Statistical Learning Theory. RKHSs were introduced into applications closely related to learning by Parzen [22] and Wahba [30], and into learning theory by Cortes and Vapnik [7] and Girosi [11]. Often, norms on such spaces play the role of measures of various types of oscillations of input/output mappings. So, when a RKHS is used as a hypothesis space \( \mathcal{H} \), the choice \( \mathbf{Ψ}(\cdot) = \|\cdot\|^2_{\mathcal{H}} \) in the regularized empirical error functional allows one to enforce desired smoothness properties of the solution (see [17,18] and the references therein). This approach is an application to supervised learning from data of regularization techniques developed during the ’60s of the previous century for ill-posed inverse problems and known as Tikhonov regularization [4, pp. 68–78].

In this paper, we investigate the supervised learning technique based on Tikhonov regularization in terms of spectral windows [4, Section V], introduced in the study of inverse problems to evaluate the robustness of regularization algorithms against noise in the data. For their use in learning from data, see [21,14,13,16] and the references therein. Our analysis takes the hint from some results derived in [29]. We investigate the effects, on the regularized solution, of a joint perturbation of the regression function and the data. Then we compare the spectral windows in the finite-sample and population cases. In both contexts, the results are expressed in terms of probabilistic upper bounds on the RKHS norms of the differences between regularized solutions.

The paper is organized as follows. Section 2, after summarizing definitions and notations, deals with Tikhonov regularization in supervised learning from data and with spectral windows. Section 3 reviews some available probabilistic estimates for supervised learning in the presence of noisy data. Sections 4 and 5 contain our main original contributions. They regard (i) the extensions of the estimates in Section 3 to the case in which perturbations are present both in the regression function and in the data and (ii) the comparison of spectral windows in the finite-sample and population cases. Section 6 discusses some connections with other works.

2. Spectral windows and regularization in supervised learning

We consider as hypothesis spaces Reproducing Kernel Hilbert Spaces (RKHSs). A RKHS is a Hilbert space \((\mathcal{H}_K(X), \|\cdot\|_{\mathcal{H}_K(X)})\) formed by real-valued functions defined on a nonempty set \(X\), such that for every \(u \in X\) the evaluation functional \(\mathcal{F}_u\), defined for every \(f \in \mathcal{H}_K(X)\) as \(\mathcal{F}_u(f) = f(u)\), is bounded [2]. By the Riesz Representation Theorem [10, p. 200], for every \(u \in X\) there exists a unique element \(K_u \in \mathcal{H}_K(X)\), called the representor of \(u\), such that for every \(f \in \mathcal{H}_K(X)\) the reproducing property

\[
\mathcal{F}_u(f) = \langle f, K_u \rangle_{\mathcal{H}_K(X)} \tag{1}
\]

holds.

RKHSs can be characterized in terms of kernels. A positive-semidefinite kernel is a symmetric function \(K : X \times X \to \mathbb{R}\) such that for every positive integer \(m\), every \((w_1, \ldots, w_m) \in \mathbb{R}^m\), and every \((u_1, \ldots, u_m) \in X^m\)

\[
\sum_{i,j=1}^m w_i w_j K(u_i, u_j) \geq 0. \tag{2}
\]

If the inequality (2) holds strictly when not all the \(w_i\) are zero, then we have a positive-definite kernel.\(^1\) For a kernel \(K : X \times X \to \mathbb{R}\), a positive integer \(m\), and a vector \(x = (x_1, \ldots, x_m) \in X^m\), the \(m \times m\) Gram matrix of the kernel \(K\) with respect to \(x\) is defined as

\[
K(x_i, x_j).
\]

Then, a kernel is positive semidefinite if and only if for every positive integer \(m\) its \(m \times m\) Gram matrices with respect to every \(x \in X^m\) are positive semidefinite; it is positive definite if and only if \(K(x)\) is positive definite whenever the vector \(x\) has no repeated entries. Every kernel \(K : X \times X \to \mathbb{R}\) generates an RKHS \(\mathcal{H}_K(X)\) defined as the completion of the linear span of the set \(\{K_u : u \in X\}\) with the inner product

\[
\langle K_u, K_v \rangle_{\mathcal{H}_K(X)} = K(u, v)
\]

and the induced norm \(\|\cdot\|_{\mathcal{H}_K(X)}\) (see, e.g., [2] and [3, p. 81]).

In order to exploit some results from [29], we assume that \(X \subset \mathbb{R}^d\) is compact and \(K\) is a continuous kernel. We denote by \(\mathbb{E}^m\) the Euclidean \(m\)-space with inner product

\[
\langle v, w \rangle_{\mathbb{E}^m} = \frac{1}{m} \sum_{i=1}^m v_i w_i \tag{3}
\]

and by \(C(\mathcal{H}_K(X))\) the Banach space of bounded linear operators from \(\mathcal{H}_K(X)\) to \(\mathcal{H}_K(X)\).

For every \(\gamma > 0\), we consider the following Tikhonov-regularized supervised learning problem\(^2\) [8].

Problem \(T_{\gamma}\): find

\[
\min_{f \in \mathcal{H}_K(X)} \left\{ \mathcal{E}_K(f) + \gamma \|f\|^2_{\mathcal{H}_K(X)} \right\}. \tag{4}
\]

The following result (see, e.g., [8, Proposition 8, p. 42]) is called the Representor Theorem for Tikhonov regularization in supervised learning from data.

Theorem 1. Let \(X \subset \mathbb{R}^d\) be a nonempty compact set, \(K : X \times X \to \mathbb{R}\) a positive semi-definite continuous and symmetric kernel, \(m\) a positive integer, \(z = (x, y)\) a data sample of size \(m\), and \(\gamma > 0\) a regularization parameter. Then, the solution to Problem \(T_{\gamma}\) is given by

\[
f_K^\gamma(z) = \frac{1}{m} \sum_{i=1}^m c_{\gamma, i} K_u(z_i), \tag{5}
\]

\(^1\) Some authors call “positive-definite” the kernels for which (2) is satisfied and “strictly positive-definite” those for which (2) holds strictly when not all the \(w_i\) are zero; see, e.g., [23, Definition 2.1].

\(^2\) Whenever we write “argmin”, we implicitly suppose that the minimum exists. When it does not, we mean that we are interested, for a given \(\epsilon > 0\), in an \(\epsilon\)-near minimum point. For example, when the minimum in (4) is not achieved, it means that we search for \(f_{\gamma} \in \mathcal{H}_K(X)\) such that \(\mathcal{E}_K(f_{\gamma}) + \|f_{\gamma}\|^2_{\mathcal{H}_K(X)} < \inf_{f \in \mathcal{H}_K(X)} \mathcal{E}_K(f) + \gamma \|f\|^2_{\mathcal{H}_K(X)} + \epsilon.\)
where $c_γ = (c_{γ,1}, \ldots, c_{γ,m})^T$ is the unique solution to the well-posed linear system

$$\left( \frac{1}{m} K[x] + γ I_m \right) c_γ = y.$$  \hspace{1cm} (6)

Insights into Theorem 1 can be obtained by interpreting (5) and (6) in terms of spectral windows (see [14] and the references therein). Suppose that the data vector $y$ is perturbed by the additive quantity $Δy$ given by an eigenvector $y_λ$ of the matrix $\frac{1}{m} K[x]$ associated with the positive eigenvalue $λ$. Then, setting

$$W'_γ(λ) = \frac{λ}{λ + γ}.$$  \hspace{1cm} (7)

of the solution and

$$Δc_γ = c_{γ,λ} = \frac{1}{(λ + γ)} y_λ = \frac{1}{λ + γ} y_λ,$$  \hspace{1cm} (8)

of the coefficient vector. Note that we have separated the contributions to the regularized solution due to $1/λ$, which corresponds to the absence of regularization, and $W'_γ(λ)$, which plays the role of the spectral window. Fig. 1 compares the qualitative behaviors of the spectral windows $W'_γ(λ)$ for different values of $γ > 0$.

If the data are corrupted by noise, then the spectral window measures how much in computing the regularized solution each spectral component of the noise is filtered out, with respect to the absence of regularization (which corresponds to the product between $1/λ$ and a spectral window identically equal to 1). By [8, Corollary 4, p. 35], the reproducing property (1), (3), and (7), we get

$$\|f_{z,λ}\|_H(x) ≤ \frac{1}{λ} s_K W_γ(λ) \|y_λ\|_\mathbb{H},$$  \hspace{1cm} (9)

where $s_K = \sup_{x \in \mathbb{K}} \|K_{x}(x)\|_H(x) = \sup_{x \in \mathbb{K}} \sqrt{K(x,x)}$.

3. Probabilistic evaluation of the noise levels

When for every positive integer $m$ the sampling points $x_1, \ldots, x_m$ are fixed a priori, the functions $K_{x_i}(\cdot)$ and the Gram matrix $K[x]$ are fixed, too, so one has a sample-independent definition of the spectral components of the noise on the data vector $y$. However, in the more usual case of a Gram matrix $K[x]$ associated with i.i.d. samples, which we shall consider in the following, all these quantities have a sample-dependent definition, so it is natural to wonder whether, for a sufficiently large sample size $m$, the effects of a perturbation in the data $y$ can be well-described in terms of population (i.e., sample-independent) quantities. An answer to this question is provided by Propositions 3 and 4 in Sections 4 and 5, respectively. In this section, we review some results from [29,28], which we shall exploit in their proofs.

Following [29], we define the following operators. Recall that $ν$ denotes the marginal probability measure on $X$.

- $A : \mathcal{H}_K(X) \rightarrow L^2(X, ν)$ such that, for every $f \in \mathcal{H}_K(X)$ and $x \in X$,

  $$(Af)(x) = \langle f, K_x(x) \rangle_{\mathcal{H}_K(X)} = f(x),$$

  is the canonical embedding of $\mathcal{H}_K(X)$ into $L^2(X, ν)$.

- $A_x : \mathcal{H}_K(X) \rightarrow \mathbb{H}$ such that, for every $f \in \mathcal{H}_K(X)$ and $i \in \{1, \ldots, m\}$,

  $$(A_x f)_{i} = \langle f, K_{x_i}(x) \rangle_{\mathcal{H}_K(X)} = f(x_i).$$

- $A^{*} : L^2(X, ν) \rightarrow \mathcal{H}_K(X)$ such that, for every $ϕ \in L^2(X, ν)$ and $x \in X$,

  $$(A^* ϕ)(x) = \int_{X} ϕ(y) K_y(x) \, dv(y).$$

- $A^*_x : \mathbb{H} \rightarrow \mathcal{H}_K(X)$ such that, for every $y \in \mathbb{H}$ and $x \in X$,

  $$(A^*_x y)(x) = \frac{1}{m} \sum_{i=1}^{m} y_{i} K_{x_i}(x).$$

- $A^* A : \mathcal{H}_K(X) \rightarrow \mathcal{H}_K(X)$ such that, for every $f \in \mathcal{H}_K(X)$ and $x \in X$,

  $$(A^* A f)(x) = \int_{X} \langle f, K_y(x) \rangle_{\mathcal{H}_K(X)} K_y(x) \, dv(y).$$

- $A^*_x A_x : \mathcal{H}_K(X) \rightarrow \mathcal{H}_K(X)$ such that, for every $f \in \mathcal{H}_K(X)$ and $x \in X$,

  $$(A^*_x A_x f)(x) = \frac{1}{m} \sum_{i=1}^{m} \langle f, K_{x_i}(x) \rangle_{\mathcal{H}_K(X)} K_{x_i}(x).$$

- $K_{\mathcal{H}_K} = A A^* : L^2(X, ν) \rightarrow L^2(X, ν)$ such that, for every $ϕ \in L^2(X, ν)$ and $x \in X$,
\((AA^*\phi)(x) = \int \phi(y)K(x, y)\,dv(y)\).

- \(A_xA_x^* : \mathbb{E}^m \to \mathbb{E}^m\) such that, for every \(y \in \mathbb{E}^m\) and \(i \in \{1, \ldots, m\}\),
  \[
  (A_xA_x^*y)_i = \left\langle \frac{1}{m} \sum_{j=1}^m y_jK_{x_j}, K_{x_i} \right\rangle_{\mathcal{H}_K(X)} = \frac{1}{m} \langle K(x)\rangle_i.
  \]

For every sample \(z = (x, y)\) of size \(m\) and every \(f \in \mathcal{H}_K(X)\), the empirical error functional can be written in terms of the operator \(A_x\) as

\[
E_{z}(f) = \|A_x f - y\|_{\mathbb{E}^m}^2.
\]

Since

\[
(A_x^*A_x + \gamma I_m)^{-1}A_x^* = A_x^*(A_xA_x^* + \gamma I_m)^{-1}
\]

(see, e.g., [9, Section 2.3]), it is easy to see that (5) and (6) are equivalent to

\[
f^{(x)}_z = (A_x^*A_x + \gamma I)^{-1}A_x^*y. \tag{10}
\]

The function defined for every \(x \in X\) as \(g(x) = E[y|x]\) is called regression function and we suppose that \(g \in \mathcal{L}^2(X, \nu)\).

Note that the expression

\[
f^{(y)} = (A^*A + \gamma I)^{-1}A^*g
\]

is equivalent to

\[
f^{(y)}(\cdot) = \int \chi_{\gamma}(y)K(\cdot, y)\,dv(y), \tag{11}
\]

where \(\chi_{\gamma} \in \mathcal{L}^2(X, \nu)\) solves

\[
(\mathcal{K} + \gamma I)\chi_{\gamma} = g. \tag{12}
\]

For the reader’s convenience, we report [29, Theorem 3] and [28, Proposition 3.2] as the next Propositions 1 and 2, respectively.

**Proposition 1.** Let \(P\) be the a-priori probability with respect to the random draw of the i.i.d. sample \(z = (x, y)\), \(\psi(t) = \frac{1}{2}(t + \sqrt{t^2 + 4\gamma})\), \(\tau \in (0, 1)\), \(\delta_1(m, \tau) = \frac{N_x}{2}\psi\left(\frac{8}{m} \ln \frac{4}{\tau}\right)\), and \(\delta_2(m, \tau) = \frac{1}{2}\psi\left(\frac{8}{m} \ln \frac{4}{\tau}\right)\). Then

\[
P\left\{ \|A^* g - A_x^*y\|_{\mathcal{H}_K(X)} \leq \delta_1(m, \tau) \right\}
\]

and

\[
\|A^*A - A_x^*A_x\|_{\mathcal{L}(\mathcal{H}_K(X))} \leq \delta_2(m, \tau) \geq 1 - \tau.
\]

**Proposition 2.** Let \(M, \eta_1, \eta_2, \gamma > 0\), \(\|y\|_{\mathbb{E}^m} \leq M\), \(\|A^* g - A_x^*y\|_{\mathcal{H}_K(X)} \leq \eta_1\), and \(\|A^*A - A_x^*A_x\|_{\mathcal{L}(\mathcal{H}_K(X))} \leq \eta_2\). Then

\[
\|f^{(y)} - f^{(y)}_z\|_{\mathcal{H}_K(X)} \leq \frac{M}{2\gamma^{3/2}} \eta_2 + \frac{1}{\gamma} \eta_1.
\]

Propositions 1 and 2 imply that the upper bound

\[
\|f^{(y)} - f^{(y)}_z\|_{\mathcal{H}_K(X)} \leq \frac{N}{2\gamma^{3/2}} \delta_2(m, \tau) + \frac{1}{\gamma} \delta_1(m, \tau) \tag{13}
\]

holds with probability at least \(1 - \tau\) with respect to the random draw of the i.i.d. sample \(z\). So, with increasing probability with the sample size \(m\), the sample-dependent regularized solution \(f^{(y)}_z\) obtained by (5) and (6) is close, in the norm of \(\mathcal{H}_K(X)\), to the sample-independent one \(f^{(y)}\).

**4. Joint perturbation of the regression function and the data**

Let \(\Delta g \in \mathcal{H}_K(X)\) be a perturbation of the regression function. According to the interpretation of (5) in terms of spectral windows, discussed in Section 2, we are interested in comparing the quantities

\[
\Delta f^{(y)}_z = (A_x^*A_x + \gamma I)^{-1}A_x^*\Delta y \tag{14}
\]

and

\[
\Delta f^{(y)} = (A^*A + \gamma I)^{-1}A^*\Delta g \tag{15}
\]

when the perturbations \(\Delta y\) and \(\Delta g\) are related by

\[
\Delta y = A_x\Delta g
\]

(i.e., \((\Delta y)_i = \Delta g(x)_i, i = 1, \ldots, m\)).

The effect of a joint perturbation of the regression function \(g\) and the data \(y\) was not considered in [29,28]. It may arise in learning problems as a consequence, e.g., of changing the measurement devices used in acquiring the data (in practice, this can be seen as a perturbation of the probability measure \(\rho\)). For instance, each device may introduce a different measurement error (here modeled as a bias \(\Delta g(x)\) for each device and each \(x \in X\)).

The next Proposition 3 extends Proposition 1 to the case of a joint perturbation of \(g\) and \(y\). In both Propositions 1 and 3, the samples \((x_i, y_i)\) are i.i.d. generated according to the probability measure \(\rho\), but in the second case each \(y_i, i = 1, \ldots, m\), is perturbed by the corresponding \(\Delta g(x_i)\) (so, for every \(x \in X\), the regression function \(g(x)\) is translated by the quantity \(\Delta g(x)\)).

**Proposition 3.** Let \(\varepsilon > 0\), \(\tau \in (0, 1)\) and \(\delta_3(m, \tau, \varepsilon) = \delta_1(m, \tau) + \varepsilon\delta_2(m, \tau)\), where \(\delta_1(m, \tau)\) and \(\delta_2(m, \tau)\) are defined as in Proposition 1. Then

\[
P \left\{ \sup_{\Delta g \in \mathcal{B}_\varepsilon(\mathcal{H}_K(X))} \|A^* (g + \Delta g) - A_x^* (y + \Delta y)\|_{\mathcal{H}_K(X)} \right. 
\]

\[
\leq \delta_3(m, \tau, \varepsilon), \|A^*A - A_x^*A_x\|_{\mathcal{L}(\mathcal{H}_K(X))} \leq \delta_2(m, \tau) \geq 1 - \tau,
\]

where \(\Delta y = A_x\Delta g\).

**Proof.** (i) By the triangle inequality, we get

\[
\|A^* (g + \Delta g) - A_x^* (y + \Delta y)\|_{\mathcal{H}_K(X)}
\]

\[
\leq \|A^* g - A_x^* y\|_{\mathcal{H}_K(X)} + \|A^* \Delta g - A_x^* \Delta y\|_{\mathcal{H}_K(X)}. \tag{16}
\]
By Proposition 1, the first term in the right-hand side of (16) is bounded from above by $\delta_1(m, \tau)$, with probability at least $1 - \tau$ with respect to the random draw of the i.i.d. sample $z$. As to the second term, exploiting the fact that $\mathcal{A}$ is the canonical embedding of $H_K(X)$ into $L^2(X, \nu)$, we get $\Delta g = \mathcal{A} \Delta g$. This, combined with $\Delta y = \mathcal{A}_x \Delta g$, gives

$$
\left\|\mathcal{A}^* \Delta g - \mathcal{A}_x^* \Delta y\right\|_{H_K(X)} = \left\|\mathcal{A}^* \Delta g - \mathcal{A}^*_x \mathcal{A} \Delta g\right\|_{H_K(X)} \\
\leq \left\|\mathcal{A}^* A - \mathcal{A}^*_x \mathcal{A}\right\|_{L^2(H_K(X))} \left\|\Delta g\right\|_{H_K(X)}.
$$

(17)

We conclude by combining (16), the bound on $\left\|\mathcal{A}^* A - \mathcal{A}^*_x \mathcal{A}\right\|_{L^2(H_K(X))}$ from Proposition 1, and the definition of $\delta_2(m, \tau, \varepsilon)$. \boxdot

Note that Proposition 3 implies

$$
P\left(\left\|\mathcal{A}^* (g + \Delta g) - \mathcal{A}^*_x (y + \Delta y)\right\|_{H_K(X)} \leq \delta_3(m, \tau, \varepsilon)\right)
$$

and

$$
\left\|\mathcal{A}^* A - \mathcal{A}^*_x \mathcal{A}\right\|_{L^2(H_K(X))} \leq \delta_2(m, \tau) \geq 1 - \tau,
$$

where $\Delta y = \mathcal{A}_x \Delta g$ and $\Delta g$ is a random element of $B_\mathbb{K}(\left\|\cdot\right\|_{H_K(X)})$. So, the analysis holds even if $\Delta g$ is a priori unknown but satisfies $\Delta g \left\|_{H_K(X)} \leq \varepsilon$.

5. Spectral windows for the finite-sample and population cases

In this section, we compare the effects of the spectral windows in the finite-sample and population cases. In order to simplify the comparison, we assume that the kernel $K$ is positive definite, so with probability 1 the matrix $\frac{1}{m} K[x]$ is nonsingular (as the marginal probability measure $\nu$ on $X$ is nondegenerate).

The next Proposition 4(i) provides a probabilistic upper bound on $\left\|\Delta f^Y - \Delta f^Z\right\|_{H_K(X)}$ when $\Delta y$ is an eigenvector of $\frac{1}{m} K[x]$ and $\Delta g$ is related to $\Delta y$ in a suitable way. Proposition 4(ii) gives a similar probabilistic upper bound when $\Delta g$ is an eigenfunction of $\mathcal{K}$ and $\Delta y$ is suitably related to $\Delta g$.

**Proposition 4.** Let $K$ be a positive-definite kernel. Then the following hold.

(i) Suppose that $\Delta y = y_\lambda$, where $y_\lambda$ is an eigenvector of the matrix $\frac{1}{m} K[x]$ associated with the positive eigenvalue $\lambda$, and let $\Delta g = \mathcal{A}^*_x (A_x A^*)_x^{-1} y_\lambda$. Then

$$
\left\|\Delta f^Y - \Delta f^Z\right\|_{H_K(X)} \leq \frac{\left\|\Delta y\right\|_{\mathbb{E}^m}}{2\gamma^{3/2}} \delta_2(m, \tau) + \frac{1}{\sqrt{\lambda}} \left\|\Delta y\right\|_{\mathbb{E}^m} \delta_2(m, \tau)
$$

(18)

with probability at least $1 - \tau$ with respect to the random draw of the i.i.d. sample $z$.

(ii) Suppose that $\Delta g = g_\lambda$, where $g_\lambda$ is an eigenfunction of the symmetric operator $\mathcal{K} = A^* A$, associated with the positive eigenvalue $\lambda$, and let $\Delta y = \mathcal{A}_x g_\lambda$. If $\lambda > \delta_2(m, \tau)$, then

$$
\left\|\Delta f^Y - \Delta f^Z\right\|_{H_K(X)} \leq \frac{\left\|\Delta y\right\|_{\mathbb{E}^m}}{2\gamma^{3/2}} \delta_2(m, \tau) + \frac{1}{\sqrt{\lambda}} \left\|\Delta y\right\|_{\mathbb{E}^m} \delta_2(m, \tau)
$$

(19)

with probability at least $1 - \tau$ with respect to the random draw of the i.i.d. sample $z$.

**Proof.** (i) By the definition, $\Delta g = \mathcal{A}^*_x (A_x A^*)_x^{-1} y_\lambda \in H_K(X)$ and

$$
\left\|\Delta y\right\|_{\mathbb{E}^m} \leq \left\|(A_x A^*)_x^{-1} y_\lambda - (A_x A^*)_x^{-1} y_\lambda \right\|_{H_K(X)}
$$

$$
= \left\|(A_x A^*)_x^{-1} y_\lambda - (A_x A^*)_x^{-1} y_\lambda \right\|_{\mathbb{E}^m}
$$

$$
= \frac{1}{\sqrt{\lambda}} \left\|\Delta y\right\|_{\mathbb{E}^m}.
$$

(20)

Then the estimate (18) follows by (17), the bound on $\left\|\mathcal{A}^* A - \mathcal{A}^*_x \mathcal{A}\right\|_{L^2(H_K(X))}$ from Proposition 1, and Proposition 2 (with $y^*, g^*, M, \eta_1$ and $\eta_2$ replaced by $\Delta y, \Delta g, \Delta y^*, \Delta g^*$).

(ii) Since $A$ is the canonical embedding of $H_K(X)$ into $L^2(X, \nu)$, we get

$$
A^* A g_\lambda = A^* A g_\lambda = A^* A g_\lambda = \tilde{\lambda} g_\lambda.
$$

(21)

So, $g_\lambda$ is an eigenfunction of the symmetric operator $A^* A$, associated with the positive eigenvalue $\tilde{\lambda}$. Then

$$
\tilde{\lambda} \left\|g_\lambda\right\|_{H_K(X)}^2 = \left\|\mathcal{A}^*_x (A_x A^*)_x^{-1} y_\lambda\right\|_{H_K(X)}
$$

$$
= \left\|\mathcal{A}^*_x (A^* A - A_x^* A_x + A_x^* A_x) g_\lambda\right\|_{H_K(X)}
$$

$$
\leq \left\|A^* A - A_x^* A_x\right\|_{L^2(H_K(X))} \left\|g_\lambda\right\|_{H_K(X)}^2
$$

$$
+ \left\|\Delta y\right\|_{\mathbb{E}^m}^2.
$$

(22)

Hence, by (22) and the bound on $\left\|\mathcal{A}^* A - A_x^* A_x\right\|_{L^2(H_K(X))}$ from Proposition 1,

$$
\tilde{\lambda} \left\|g_\lambda\right\|_{H_K(X)}^2 \leq \delta_2(m, \tau) \left\|g_\lambda\right\|_{H_K(X)}^2 + \left\|\Delta y\right\|_{\mathbb{E}^m}^2
$$

if and only if

$$
\left\|\Delta y\right\|_{\mathbb{E}^m} \leq \frac{1}{\sqrt{\lambda}} \left\|\Delta y\right\|_{\mathbb{E}^m} \delta_2(m, \tau)
$$

with probability at least $1 - \tau$ with respect to the random draw of the i.i.d. sample $z$. Then we get the estimate (19) proceeding as at the end of the proof of (i), by applying Proposition 2 (with $y, g, M, \eta_1$ and $\eta_2$ replaced by $\Delta y, \Delta g, \Delta y^*, \Delta g^*$).

According to Proposition 4, for $\gamma, m, \lambda$ and $\tilde{\lambda}$ sufficiently large the effects of the spectral windows in the finite-sample and population cases are very similar (the right-hand sides of (18) and (19) are small with high probability).
6. Relationships with other works

In connection with Proposition 4, probabilistic upper bounds on the distance between the positive eigenvalues of the matrix $\frac{1}{m}K(x)$ and those of the operator $\bar{K}$ are given, e.g., in [25, formula (14) and Theorem 6]. Similar bounds may be obtained by combining [24, Corollary 5] and the infinite-dimensional extension [5, Theorem 1] of the Hoffman-Wielandt Inequality [20, Theorem 2.2]. Relations between the associated eigenvectors (eigenfunctions) may be derived by combining [1, Theorem 4] and the bound on $\|A^*A - A_k^*A_k\|_{L(\mathcal{H}_K(X))}$ provided by Proposition 1.

Extensions of Proposition 4 may be obtained for spectral windows associated with other kinds of regularizations, such as weight decay and Tikhonov regularization combined with weight decay [14]. To this end, one should modify the proof of Proposition 2 given in [28], by taking into account the specific expressions of the spectral windows associated with the two above-mentioned techniques.

Sparse suboptimal solutions to various regularized learning problems were investigated in [12,14,13,16,15,17,18]. The estimates derived therein may be combined with Proposition 2 (in the case of Tikhonov regularization) and its extensions mentioned above (for the other regularizations), in order to estimate the RKHS norms of the differences between the solutions to various regularized learning problems with an infinite number of samples and the respective sparse solutions to the finite-sample versions of the same problems.

References