Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2005-031
AIM-2005-015
CBCL-249

May 16, 2005

Risk Bounds for Regularized Least-squares Algorithm with Operator-valued Kernels

Ernesto De Vito, Andrea Caponnetto

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Risk bounds for regularized least-squares algorithm with operator-valued kernels

Ernesto De Vito [a] Andrea Caponnetto [b]

[a] *Dipartimento di Matematica, Università di Modena e Reggio Emilia, Via Campi 213/B, 41100 Modena, Italy and I.N.F.N., Sezione di Genova, Via Dodecaneso 33, 16146 Genova, Italy*

[b] *C.B.C.L.,McGovern Institute, Massachusetts Institute of Technology, Bldg.E25-201, 45 Carleton St., Cambridge, MA 02142*
and Dipartimento di Informatica e Scienza dell'Informazione, Università degli Studi di Genova, Via Dodecaneso 35,16146 Genova, Italy

Abstract

We show that recent results in [3] on risk bounds for regularized least-squares on reproducing kernel Hilbert spaces can be straightforwardly extended to the vector-valued regression setting. We first briefly introduce central concepts on operator-valued kernels, then we show how risk bounds can be expressed in terms of a generalization of *effective dimension*.

This report describes research done at the Center for Biological & Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Dept. of Brain & Cognitive Sciences, and which is affiliated with the Computer Sciences & Artificial Intelligence Laboratory (CSAIL), as well as in the Dipartimento di Informatica e Scienze dell'Informazione (DISI) at University of Genoa, Italy.

This research was sponsored by grants from: Office of Naval Research (DARPA) Contract No. MDA972-04-1-0037, Office of Naval Research (DARPA) Contract No. N00014-02-1-0915, National Science Foundation (ITR/SYS) Contract No. IIS-0112991, National Science Foundation (ITR) Contract No. IIS-0209289, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218693, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218506, and National Institutes of Health (Conte) Contract No. 1 P20 MH66239-01A1.

Additional support was provided by: Central Research Institute of Electric Power Industry (CRIEPI), Daimler-Chrysler AG, Compaq/Digital Equipment Corporation, Eastman Kodak Company, Honda R&D Co., Ltd., Industrial Technology Research Institute (ITRI), Komatsu Ltd., Eugene McDermott Foundation, Merrill-Lynch, NEC Fund, Oxygen, Siemens Corporate Research, Inc., Sony, Sumitomo Metal Industries, and Toyota Motor Corporation.

This research has been partially funded by the FIRB Project ASTAA and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

1 Introduction

This work presents an extension to multi-task learning of our recent results [3] on risk estimates for regularized least-squares (RLS) with reproducing kernel Hilbert spaces (RKHS). Recently various papers, [18], [1],[10],[8], have addressed the problem of multi-task learning using kernel techniques. For instance [18] employs two kernels one on the inputs and one on the outputs, in order to represent similarity measures on the corresponding spaces. The underlying similarity measures are supposed to capture some inherent regularity of the phenomenon under investigation and should be chosen according to the available prior knowledge. On the contrary in [1] the prior knowledge is coded by a single kernel on the space of input-output couples, and a generalization of standard support vector machines is proposed. It was in [10], [8] that for the first time in the learning theory literature it was pointed out that particular scalar kernels defined on input-output couples can be profitably mapped onto operator-valued kernels [2] defined on the input space.

It is well known [2] that the machinery of RKHS can be elegantly extended to cope with vector-valued functions using operator-valued kernels, so one would expect that kernels methods for single-task learning can be adapted to multi-task learning in this extended RKHS framework. In fact we show that the risk bounds obtained in [3] can be straightforwardly rephrased in this more general setting.

The paper is organized as follows. In sections 2 we recall very briefly the main concepts of statistical learning theory with vector-valued outputs and define the RLS algorithm in this framework. In section 3 we fix the notations extending the familiar formalism of reproducing kernel Hilbert spaces to the operator-valued case. We also introduce the assumptions on the hypothesis space and on the probability measure from which the samples are drawn. Furthermore we prove some preliminary results on the structure of RLS estimators and concentration of measure for vector valued random variables. Finally in section 4 we prove the probabilistic upper bound for the excess risk of RLS estimators using a generalized effective dimension.

2 Learning from examples

We first briefly introduce some basic concepts of statistical learning theory in the regression setting for vector-valued outputs (for details see [16], [9], [13], [4], [10] and references therein).
In the framework of learning from examples there are two sets of variables: the input space X and the output space Y which we will assume to be a separable Hilbert space. The relation between the input $x \in X$ and the output $y \in Y$ is described by a probability distribution $\rho(x, y) = \nu(x)\rho(y|x)$ on $X \times Y$, where ν is the marginal distribution on X and $\rho(\cdot|x)$ is the conditional distribution of y given $x \in X$. The distribution ρ is known only through a sample $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_\ell, y_\ell))$, called *training set*, drawn independently and identically distributed (i.i.d.) according

to ρ. Given the sample $\mathbf{z}$, the aim of learning theory is to find a function $f_{\mathbf{z}} : X \to Y$ such that $f_{\mathbf{z}}(x)$ is a good estimate of the output y when a new input x is given. The function $f_{\mathbf{z}}$ is called *estimator* and the map providing $f_{\mathbf{z}}$, for any training set $\mathbf{z}$, is called *learning algorithm*.

Given a function $f : X \to Y$, the ability of f to describe the distribution ρ is measured by its *expected risk* defined as

$$I[f] = \int_{X \times Y} \|f(x) - y\|_Y^2 \, d\rho(x, y),$$

and the regression function

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

is the minimizer of the expected risk over the space of all the measurable Y-valued functions on X. In this sense f_ρ can be seen as the ideal estimator of the distribution probability ρ. However, the regression function cannot be reconstructed exactly since only a finite, possibly small, set of examples $\mathbf{z}$ is given.

To overcome this problem, in the framework of the regularized least squares algorithm [17], [12], [4], [20], a Hilbert space $\mathcal{H}$ of real functions on X is fixed and the estimator $f_{\mathbf{z}}{}^{\lambda}$ is defined as the solution of the regularized least squares problem,

$$\min_{f \in \mathcal{H}} \{ \frac{1}{\ell} \sum_{i=1}^{\ell} \|f(x_i) - y_i\|_Y^2 + \lambda \|f\|_{\mathcal{H}} \}, \tag{1}$$

where λ is a positive parameter to be chosen in order to ensure that the discrepancy.

$$I[f_{\mathbf{z}}{}^{\lambda}] - \inf_{f \in \mathcal{H}} I[f]$$

is small with hight probability. Since ρ is unknown, the above difference is studied by means of a probabilistic bound $B(\lambda, \ell, \eta)$, which is a function depending on the regularization parameter λ, the number ℓ of examples and the confidence level $1 - \eta$, such that

$$\mathbf{P} \left[I[f_{\mathbf{z}}{}^{\lambda}] - \inf_{f \in \mathcal{H}} I[f] \leq B(\lambda, \ell, \eta) \right] \geq 1 - \eta.$$

In particular, the learning algorithm is *consistent* if it is possible to choose the regularization parameter, as a function of the available data $\lambda = \lambda(\ell, \mathbf{z})$, in such a way that

$$\lim_{\ell \to +\infty} \mathbf{P} \left[I[f_{\mathbf{z}}{}^{\lambda(\ell, \mathbf{z})}] - \inf_{f \in \mathcal{H}} I[f] \geq \epsilon \right] = 0, \tag{2}$$

for every $\epsilon > 0$. The above convergence in probability is usually called *(weak) consistency* of the algorithm (see [5] for a discussion on the different kind of consistencies).

3 Notations and preliminary results

In this section we state the notations, we set the main assumptions and we prove some preliminary results.

We assume that the input space X is a Polish space and the output space Y is a real separable Hilbert space. We let Z be the product space $X \times Y$, which is a Polish space too. The assumptions on X and Y will avoid measurability problems.
The space of bounded linear operators on Y with the uniform norm $\|\cdot\|_{\mathcal{L}(\mathcal{H})}$ will be denoted by $\mathcal{L}(Y)$, and $\mathcal{L}_2(Y)$ will be the separable Hilbert space of Hilbert-Schmidt operators on Y with scalar product

$$\langle A, B\rangle_{\mathcal{L}_2(Y)} = \mathrm{Tr}(B^* A)$$

and norm

$$\|A\|_{\mathcal{L}_2(Y)} = \sqrt{\mathrm{Tr}(A^* A)} \geq \|A\|_{\mathcal{L}(\mathcal{H})},$$

where Tr denotes the trace and * the adjoint (similar notation we use by replacing Y with other Hilbert spaces).

We first discuss the assumptions on the space $\mathcal{H}$.

Hypothesis 1 *The space $\mathcal{H}$ is a* separable *Hilbert space with reproducing kernel*

$$K : X \times X \to \mathcal{L}_2(Y) \subset \mathcal{L}(Y)$$

such that

$$X \times X \ni (x,t) \mapsto \langle K(x,t)v, w\rangle_Y \text{ is measurable } \forall v, w \in Y \tag{3}$$

$$\|K(x,x)\|_{\mathcal{L}_2(Y)} \leq \kappa \qquad \forall x \in X \tag{4}$$

for some $\kappa > 0$.

We recall that $\mathcal{H}$ is a real Hilbert space of functions $f : X \to Y$ satisfying the following *reproducing* property

$$f(x) = K_x^* f \qquad f \in \mathcal{H},\ x \in X, \tag{5}$$

where $K_x : Y \to \mathcal{H}$ is the bounded operator

$$K_x v = K(\cdot, x)v \qquad v \in Y. \tag{6}$$

and (5) gives

$$K_t^* K_x = K(t,x) \in \mathcal{L}_2(Y) \qquad \forall x, t \in X. \tag{7}$$

Moreover, given $x \in X$ the operator

$$T_x = K_x K_x^* \in \mathcal{L}_2(\mathcal{H}), \tag{8}$$

is a positive Hilbert-Schmidt operator and (8) ensures

$$\|T_x\|_{\mathcal{L}(\mathcal{H})} \leq \|T_x\|_{\mathcal{L}_2(\mathcal{H})} = \|K(x,x)\|_{\mathcal{L}_2(Y)} \leq \kappa. \tag{9}$$

If $Y = \mathbb{R}$, the space $\mathcal{L}_2(Y)$ reduces to $\mathbb{R}$, $K_x \in \mathcal{H}$, and $K_x^* f = \langle f, K_x \rangle_{\mathcal{H}}$, so that $\mathcal{H}$ is the reproducing kernel Hilbert space with kernel K [2]. The theory can be extended to vector valued functions [14]. In particular the space $\mathcal{H}$ is uniquely defined by its kernel in the sense that, given a kernel $K : X \times X \to \mathcal{L}(Y)$ such that

$$\langle K(x,t)v, w\rangle_Y = \langle K(t,x)w, v\rangle_Y$$
$$\sum_{i=1}^{n} \langle K(x_i,x_j)v_j, v_i\rangle_Y \geq 0$$

there is a unique Hilbert space of functions $f : X \to Y$ satisfying (5).

The assumption that the kernel K takes values in the Hilbert space $\mathcal{L}_2(Y) \subset \mathcal{L}(Y)$ simplifies the theory and is enough for our purposes.
The condition that $\mathcal{H}$ is separable, which is essential in the following, is not ensured by the assumptions on the kernel K. However, if (3) is replaced by the stronger condition

$$X \times X \ni (x,t) \mapsto \langle K(x,t)v, w\rangle_Y \text{ is continuous } \forall v, w \in Y,$$

the fact that X and Y are separable imply that $\mathcal{H}$ is separable, too.
As shown by Proposition 2 below, (3) and (9) are the minimal requirement to ensure that any $f \in \mathcal{H}$ has a finite expected risk for all probability measure satisfying (10).

Let now ρ be a probability measure on Z. By ρ_X we will denote the marginal distribution of ρ on X and by $\rho(\cdot|x)$ the conditional distribution on Y given $x \in X$, both existing since Z is a Polish space (see Teo 10.2.2 [6]). Let $L^2(Z,\rho,Y)$ be the Hilbert space of functions $\phi : Z \to Y$ that are square-integrable with respect to ρ, and denote by $\|\cdot\|_\rho$ and $\langle\cdot,\cdot\rangle_\rho$ the corresponding norm and scalar product. Similar notation we use for $L^2(X,\rho_X,Y)$.
Since ρ is a a probability measure, $L^2(X,\rho_X,Y)$ can be regarded as a closed subspace of $L^2(Z,\rho,Y)$ and the corresponding orthogonal projection Q is

$$(Q\phi)(x) = \int_Y \phi(x,y)\ d\rho(y|x) \qquad \phi \in L^2(Z,\rho,Y).$$

Finally, the expected risk with respect to ρ of a measurable function $f : X \to Y$ is

$$I[f] = \int_Z \|f(x) - y\|_Y^{\,2}\ d\rho(x,y).$$

We are now ready to state the hypothesis on ρ.

Hypothesis 2 *The probability measure ρ on Z satisfies*

$$\int_Z \|y\|_Y{}^2 \, d\rho(x,y) < +\infty \tag{10}$$

$$\mathrm{Tr}\left(\int_X T_x \, d\rho_X(x)\right) < +\infty, \tag{11}$$

and there are $f_{\mathcal{H}} \in \mathcal{H}$ *and* $M > 0$ *such that*

$$I[f_{\mathcal{H}}] = \inf_{f \in \mathcal{H}} I[f] \tag{12}$$

$$|y - f_{\mathcal{H}}(x)|^2 \leq M \quad \text{a.}s. \tag{13}$$

If (10) is not satisfied, then $I[f] = +\infty$ for all $f \in \mathcal{H}$ and learning problem does not make sense. If it holds $I[f]$ is finite for all $f \in L^2(X, \rho_X, Y)$.
In general $f_{\mathcal{H}}$ is not unique as an element of $\mathcal{H}$, but two different solutions are equal almost everywhere, see (17) below.
If the regression function

$$f_\rho = \int_Y y \, d\rho(y|x) = Qy$$

belongs to $\mathcal{H}$, clearly $f_{\mathcal{H}} = f_\rho$. However, in general the existence of $f_{\mathcal{H}}$ is a weaker condition than $f_\rho \in \mathcal{H}$, for example, if $\mathcal{H}$ is finite dimensional $f_{\mathcal{H}}$ always exists.
Proposition 1 will prove that the integral in (11) always converges to a positive Hilbert-Schmidt operator T, see (15) below, so (11) states that T is in fact trace class. Condition (11), (12) and (13) are needed to prove the upper bound (28).

We now study some mathematical properties of the expected risk and of the regularized least square algorithm.
Let $A : \mathcal{H} \to L^2(Z, \rho, Y)$ be the linear operator

$$(Af)(x,y) = K_x^* f \quad \forall (x,y) \in Z.$$

Equation (5) implies that the action of A on an element f is simply

$$(Af)(x,y) = f(x),$$

that is, A is the canonical inclusion of $\mathcal{H}$ into $L^2(Z, \rho, Y)$, where the variable y is *dumb* and functions are identified ρ-almost everywhere. So A could be not injective and $\mathcal{H}$ could be not closed in $L^2(Z, \rho, Y)$, since $\|f\|_{\mathcal{H}}$ is different from $\|f\|_\rho$.
The main properties of the operator A are summarized in the following proposition.

Proposition 1 *If* $\mathcal{H}$ *satisfies Hypothesis 1 and* ρ *is a probability measure,* A *is a bounded operator from* $\mathcal{H}$ *into* $L^2(Z, \rho, Y)$*, the adjoint* $A^* : L^2(Z, \rho, Y) \to \mathcal{H}$ *is*

$$A^*\phi = \int_Z K_x \; \phi(x,y) \, d\rho(x,y) = \int_X K_x \; (Q\phi)(x) \, d\rho_X(x), \tag{14}$$

where the integral converges in $\mathcal{H}$*, and* A^*A *is the Hilbert-Schmidt operator on* $\mathcal{H}$

$$T = \int_X T_x \, d\rho_X(x), \tag{15}$$

where the integral converges in $\mathcal{L}_2(\mathcal{H})$, and

$$\|T\|_{\mathcal{L}(\mathcal{H})} \leq \|T\|_{\mathcal{L}_2(\mathcal{H})} \leq \kappa. \tag{16}$$

PROOF. The proof is standard for $Y = \mathbb{R}$ and it can easily extended to the vector case.
First we prove that any function $f \in \mathcal{H}$ is measurable and bounded. Since Y is separable, it is enough to prove that the function

$$x \mapsto \langle f(x), v\rangle_Y = \langle f, K_x v\rangle_{\mathcal{H}}$$

is measurable for all $v \in Y$. If $f = K_t w$ for some $t \in X$ and $w \in Y$, the claim follows by (7) and (3). Since (5) ensures that the set

$$\{K_t w \,|\, t \in X,\ w \in Y\}$$

is total in $\mathcal{H}$, the measurability for arbitrary f follows by density. Finally, (9) gives

$$\|f(x)\|_Y{}^2 = \langle T_x f, f\rangle_{\mathcal{H}} \leq \|f\|_{\mathcal{H}}{}^2 \|T_x\|_{\mathcal{L}(\mathcal{H})} \leq \kappa \|f\|_{\mathcal{H}}{}^2.$$

Since ρ is a probability measure, then $f \in L^2(Z, \rho, Y)$ and A is a linear operator from $\mathcal{H}$ to $L^2(Z, \rho, Y)$ with $\|Af\|_\rho \leq \sqrt{\kappa}\|f\|_{\mathcal{H}}$, so that A is bounded.
We now prove (14). Indeed, given $\phi \in L^2(Z, \rho, Y)$, (3) ensures that the function

$$(x, y) \mapsto \langle K_x \phi(x, y), f\rangle_{\mathcal{H}} = \langle \phi(x, y), f(x)\rangle_{\mathcal{H}}$$

is measurable for all $f \in \mathcal{H}$. Since $\mathcal{H}$ is separable, the map

$$Z \ni (x, y) \mapsto K_x \phi(x, y) \in \mathcal{H}$$

is measurable as map taking values in $\mathcal{H}$ and (4) gives

$$\|K_x \phi(x, y)\|_{\mathcal{H}} = \sqrt{\langle K_x^* K_x \phi(x, y), \phi(x, y)\rangle_Y} \leq \sqrt{\kappa}\|\phi(x, y)\|_Y$$

for all $(x, y) \in Z$. Since ρ is finite, ϕ is in $L^1(Z, \rho, Y)$ and, hence, $(x, y) \mapsto K_x \phi(x, y)$ is integrable, as a vector valued map. Finally, for all $f \in \mathcal{H}$,

$$\int_Z \langle K_x \phi(x, y), f\rangle_{\mathcal{H}}\, d\rho(x, y) = \langle \phi, Af\rangle_\rho = \langle A^* \phi, f\rangle_{\mathcal{H}}\,,$$

so the first part of (14) holds and the second one is a consequence of the definition of Q.
Reasoning as above, it follows that the map

$$X \ni x \mapsto T_x \in \mathcal{L}_2(\mathcal{H})$$

is integrable as function taking values in $\mathcal{L}_2(\mathcal{H})$. In particular, $T \in \mathcal{L}_2(\mathcal{H})$ and (15) is a consequence of (14). The bound (16) follows from (9).

The role of the operator A in the context of learning theory is clear observing that for all $f \in \mathcal{H}$

$$I[f] = \|Af - y\|_\rho{}^2 \quad ,$$

where y denotes both the variable and the function $(x, y) \mapsto y$, which belongs to $L^2(Z, \rho, Y)$ by (10). So the following result holds.

Proposition 2 *If Hypotheses 1 and (10) hold, $f_{\mathcal{H}} \in \mathcal{H}$ is a minimizer of the expected risk $I[\cdot]$ if and only if it satisfies*

$$Tf_{\mathcal{H}} = A^*y. \tag{17}$$

and

$$I[f] - I[f_{\mathcal{H}}] = \|A(f - f_{\mathcal{H}})\|_\rho^{\,2} = \left\|\sqrt{T}(f - f_{\mathcal{H}})\right\|_{\mathcal{H}}^{\,2} \quad \forall f \in \mathcal{H}\ . \tag{18}$$

Moreover, for $\lambda > 0$, a unique minimizer f^λ of the regularized expected risk

$$I[f] + \lambda \|f\|_{\mathcal{H}}^{\,2}$$

exists and is given by

$$f^\lambda = (T + \lambda)^{-1} A^* y = (T + \lambda)^{-1} T f_{\mathcal{H}}. \tag{19}$$

PROOF. The result is well known in the framework of linear inverse problems [7] and we report it for completeness. Since the expect risk is convex, $f_{\mathcal{H}}$ is a minimizer if and only if the derivative of $I[f]$ is zero, that is,

$$\langle Af, Af_{\mathcal{H}} - y\rangle_\rho = 0 \quad \forall f \in \mathcal{H} \tag{20}$$

and (17) follows.
Given $f \in \mathcal{H}$

$$\begin{aligned} I[f] - I[f_{\mathcal{H}}] &= \|Af - y\|_\rho^{\,2} - \|Af_{\mathcal{H}} - y\|_\rho^{\,2} \\ &= \|A(f - f_{\mathcal{H}})\|_\rho^{\,2} + 2\langle A(f - f_{\mathcal{H}}), Af_{\mathcal{H}} - y\rangle_\rho \\ &= \|A(f - f_{\mathcal{H}})\|_\rho^{\,2} \end{aligned}$$

since the second term is zero due to (20). Let $A = U\sqrt{T}$ be the polar decomposition. Since U is a partial isometry from the closure of the range of $\sqrt{T}$ onto the closure of the range of A

$$\|A(f - f_{\mathcal{H}})\|_\rho = \left\|\sqrt{T}(f - f_{\mathcal{H}})\right\|_{\mathcal{H}}.$$

Finally, (19) follows taking the derivative be equal to zero.

Clearly, A, T, $f_{\mathcal{H}}$ and f^λ depend on ρ and, if it is needed, we write explicitly this dependence.
In particular, given $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_\ell, y_\ell)) \in Z^\ell$, consider the empirical measure

$$\rho_{\mathbf{z}} = \frac{1}{\ell}\sum_{i=1}^{\ell} \delta_{(x_i, y_i)} \qquad (\rho_{\mathbf{z}})_X = \frac{1}{\ell}\sum_{i=1}^{\ell} \delta_{x_i},$$

where $\delta_{(x,y)}$ is the Dirac measure at point $(x, y) \in Z$. Since $\rho_{\mathbf{z}}$ is finitely supported, any element $\mathbf{w} \in L^2(Z, \rho_{\mathbf{z}})$ is uniquely defined by ℓ vectors

$$w_i = \mathbf{w}(x_i, y_i) \in Y \qquad i = 1, \ldots, \ell$$

with the condition that $w_i = w_j$ whenever $(x_i, y_i) = (x_j, y_j)$ and the scalar product becomes

$$\left\langle \mathbf{w}, \mathbf{w}' \right\rangle_{L^2(Z,\rho_{\mathbf{z}})} = \frac{1}{\ell} \sum_{i=1}^{\ell} \left\langle w_i, w'_i \cdot \right\rangle_Y$$

In the following we let $A_{\mathbf{z}} = A_{\rho_{\mathbf{z}}}$, $T_{\mathbf{x}} = T_{\rho_{\mathbf{z}X}}$ and ${f_{\mathbf{z}}}^{\lambda} = f^{\lambda}_{\rho_{\mathbf{z}}}$.
Since $\rho_{\mathbf{z}}$ has a finite support, (10) reduces to the condition $\mathbf{y} \in L^2(Z, \rho_{\mathbf{z}})$, which is clearly satisfied, so Propositions 1 and 2 become

$$(A_{\mathbf{z}} f)_i = K^*_{x_i} f = f(x_i) \quad \forall i = 1, \ldots, \ell \tag{21}$$

$${A_{\mathbf{z}}}^* \mathbf{w} = \frac{1}{\ell} \sum_{i=1}^{\ell} K_{x_i} w_i \quad \mathbf{w} \in L^2(Z, \rho_{\mathbf{z}}) \tag{22}$$

$$T_{\mathbf{x}} := {A_{\mathbf{z}}}^* A_{\mathbf{z}} = \frac{1}{\ell} \sum_{i=1}^{\ell} T_{x_i} \tag{23}$$

$$f^{\lambda}_{\mathbf{z}} = (T_{\mathbf{x}} + \lambda)^{-1} {A_{\mathbf{z}}}^* \mathbf{y}, \tag{24}$$

where ${f_{\mathbf{z}}}^{\lambda}$ is the unique minimizer of the regularized empirical error

$$\frac{1}{\ell} \sum_{i=1}^{\ell} {\|f(x_i) - y_i\|_Y}^2 + \lambda {\|f\|_{\mathcal{H}}}^2.$$

The following technical lemma will be used in the proof of Theorem 5.

Lemma 3 *If ρ satisfies (11), i.e. T is trace class, then T_x is trace class for ρ_X-almost all $x \in X$.*

PROOF. Let $(e_k)_{k \geq 1}$ be a basis of $\mathcal{H}$. For all $k \geq 1$ the functions

$$x \mapsto \langle T_x e_k, e_k \rangle_{\mathcal{H}}$$

are positive and meausurable by (3) and

$$\int_X \sum_{k=1}^{n} \langle T_x e_k, e_k \rangle_{\mathcal{H}} \, d\rho_X(x) = \sum_{k=1}^{n} \left\langle \left(\int_X T_x \, d\rho_X(x) \right) e_k, e_k \right\rangle_{\mathcal{H}} \leq \mathrm{Tr}(T) < +\infty.$$

Clearly $\sum_{k=1}^{n} \langle T_x e_k, e_k \rangle_{\mathcal{H}}$ convergences to $\mathrm{Tr}\, T_x$, which is finite for almost all $x \in X$ by monotone convergence theorem.

Finally, we need the following probabilistic inequality due to [11].

Proposition 4 *Let $(\Omega, \mathcal{F}, P)$ be a probability space and ξ be a random variable on Ω taking value in a real separable Hilbert space $\mathcal{K}$. Assume that there are two positive constants H and σ such that*

$$\|\xi(\omega)\|_{\mathcal{K}} \leq \frac{H}{2} \quad \text{a.}s \tag{25}$$

$$\mathbb{E}[\|\xi\|_{\mathcal{K}}^2] \leq \sigma^2. \tag{26}$$

Let $\ell \in \mathbb{N}$ and $0 < \eta < 1$, then

$$\mathbf{P}^{\ell}\left[(\omega_1, \ldots, \omega_\ell) \in \Omega^\ell \mid \left\| \frac{1}{\ell}\sum_{i=1}^{\ell} \xi(\omega_i) - \mathbb{E}[\xi] \right\|_{\mathcal{K}} \leq 2\left(\frac{H}{\ell} + \frac{\sigma}{\sqrt{\ell}}\right)\log\frac{2}{\eta}\right] \geq 1 - \eta. \tag{27}$$

PROOF. It is just a restatement of Th. 3.3.4 of [19] (see also [15]). Consider the probability space $(\Omega^\ell, \mathcal{F}^\ell, P^\ell)$ and the set of independent random variables with zero mean $\xi_i(\omega_1, \ldots, \omega_\ell) = \xi(\omega_i) - \mathbb{E}[\xi]$ defined on Ω^ℓ. The fact that ξ_i are i.i.d and conditions (25), (26) ensure that

$$\|\xi_i\|_{\mathcal{K}} \leq H \quad \text{a.}s$$
$$\mathbb{E}[\|\xi_i\|_{\mathcal{K}}^2] \leq \sigma^2,$$

so that, for all $m \geq 2$ it holds

$$\sum_{i=1}^{\ell} \mathbb{E}[\|\xi_i\|_{\mathcal{K}}^m] \leq \frac{1}{2} m! B^2 H^{m-2},$$

with $B^2 = \ell\sigma^2$. So Th. 3.3.4 of [19] can be applied and it ensures

$$\mathbf{P}\left[\frac{1}{\ell}\left\|\sum_{i=1}^{\ell}(\xi(z_i) - \mathbb{E}[\xi])\right\| \geq \frac{xB}{\ell}\right] \leq 2\exp\left(-\frac{x^2}{2(1 + xHB^{-1})}\right).$$

for all $x \geq 0$. Letting $\delta = \frac{xB}{\ell}$, we get the equation

$$\frac{1}{2}(\frac{\ell\delta}{B})^2 \frac{1}{1 + \ell\delta HB^{-2}} = \frac{\ell\delta^2\sigma^{-2}}{2(1 + \delta H\sigma^{-2})} = \log\frac{2}{\eta},$$

since $B^2 = \ell\sigma^2$. Defining $t = \delta H\sigma^{-2}$

$$\frac{\ell\sigma^2}{2H^2}\frac{t^2}{1+t} = \log\frac{2}{\eta}.$$

The inverse of the function $\frac{t^2}{1+t}$ is the function $g(t) = \frac{1}{2}(t + \sqrt{t^2 + 4t})$ so

$$\left\| \frac{1}{\ell}\sum_{i=1}^{\ell} \xi(z_i) - \mathbb{E}[\xi] \right\|_{\mathcal{K}} \leq \frac{\sigma^2}{H}\, g\left(\frac{2H^2}{\ell\sigma^2}\log\frac{2}{\eta}\right)$$

with probability greater than $1 - \eta$. The thesis follows observing that $g(t) \leq t + \sqrt{t}$ and $2\log\frac{2}{\eta} \geq \sqrt{2\log\frac{2}{\eta}} \geq 1$.

4 Risk bound

The aim of this section is to give a probabilistic upper bound on the expect risk of the solution given by the regularized least square algorithm. The bound depends on the number of examples ℓ, the regularization parameter and some a priori information on the probability distribution ρ.

In the following, we assume that the space $\mathcal{H}$ and the probability distribution ρ satisfy Hypotheses 1 and 2, we fix a parameter $\lambda > 0$ and we define

(1) the *residual*

$$\mathcal{A}(\lambda) = \left\| f^\lambda - f_\mathcal{H} \right\|_\rho^2 = \left\| \sqrt{T}(f^\lambda - f_\mathcal{H}) \right\|_\mathcal{H}^2,$$

where T is given by (15), f^λ by (19) and $f_\mathcal{H}$ by (12);

(2) the *reconstruction error*

$$\mathcal{B}(\lambda) = \left\| f^\lambda - f_\mathcal{H} \right\|_\mathcal{H}^2;$$

(3) the *effective dimension*

$$\mathcal{N}(\lambda) = \mathrm{Tr}[(T+\lambda)^{-1}T],$$

where the trace is finite due to (11).

In the framework of learning $\mathcal{A}(\lambda)$ is called *approximation error*, whereas in the framework of approximation theory $\sqrt{\mathcal{B}(\lambda)}$ is the approximation error. To avoid confusion we follow the notation of inverse problems.

We are now ready to state our main result on the upper bound.

Theorem 5 *Let $\mathbf{z} \in Z^\ell$ be a training set drawn i.i.d according to ρ and $f_\mathbf{z}{}^\lambda \in \mathcal{H}$ the corresponding estimator given by (24). With probability greater than $1-\eta$, $0 < \eta < 1$,*

$$I[f_\mathbf{z}{}^\lambda] - I[f_\mathcal{H}] \leq C_\eta \left(\mathcal{A}(\lambda) + \frac{\kappa^2 \mathcal{B}(\lambda)}{\ell^2 \lambda} + \frac{\kappa \mathcal{A}(\lambda)}{\ell \lambda} + \frac{\kappa M}{\ell^2 \lambda} + \frac{M \mathcal{N}(\lambda)}{\ell} \right) \tag{28}$$

provided that

$$\ell \geq \frac{C_\eta \kappa}{2\lambda} \max(\mathcal{N}(\lambda), \sqrt{2/C_\eta}) \tag{29}$$

where $C_\eta = 128 \log^2(8/\eta)$.

PROOF. We split the proof in several steps. Let λ, η and ℓ as in the statement of the theorem.
Step 1: Given a training set $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in Z^\ell$, (18) gives

$$I[f_\mathbf{z}{}^\lambda] - I[f_\mathcal{H}] = \left\| \sqrt{T}(f_\mathbf{z}{}^\lambda - f_\mathcal{H}) \right\|_\mathcal{H}^2.$$

As usual,

$$f_{\mathbf{z}}{}^{\lambda} - f_{\mathcal{H}} = (f_{\mathbf{z}}{}^{\lambda} - f^{\lambda}) + (f^{\lambda} - f_{\mathcal{H}})$$

and (19), (24) give

$$\begin{aligned} f_{\mathbf{z}}{}^{\lambda} - f^{\lambda} &= \left((T_{\mathbf{x}} + \lambda)^{-1} A_{\mathbf{z}}{}^{*}\mathbf{y}\right) - \left((T + \lambda)^{-1} A^{*} y\right) \\ &= (T_{\mathbf{x}} + \lambda)^{-1} \left\{(A_{\mathbf{z}}{}^{*}\mathbf{y} - A^{*} y) + (T - T_{\mathbf{x}})(T + \lambda)^{-1} A^{*} y\right\} \\ (\text{ Eq. (17) }) &= (T_{\mathbf{x}} + \lambda)^{-1} \left\{(A_{\mathbf{z}}{}^{*}\mathbf{y} - T_{\mathbf{x}} f_{\mathcal{H}} + T_{\mathbf{x}} f_{\mathcal{H}} - T f_{\mathcal{H}}) + (T - T_{\mathbf{x}}) f^{\lambda}\right\} \\ &= (T_{\mathbf{x}} + \lambda)^{-1} \left(A_{\mathbf{z}}{}^{*}\mathbf{y} - T_{\mathbf{x}} f_{\mathcal{H}}\right) + (T_{\mathbf{x}} + \lambda)^{-1} (T - T_{\mathbf{x}})(f^{\lambda} - f_{\mathcal{H}}). \end{aligned}$$

The inequality $\|f_1 + f_2 + f_3\|_{\mathcal{H}}{}^2 \leq 3(\|f_1\|_{\mathcal{H}}{}^2 + \|f_2\|_{\mathcal{H}}{}^2 + \|f_3\|_{\mathcal{H}}{}^2)$ implies

$$I[f_{\mathbf{z}}{}^{\lambda}] - I[f_{\mathcal{H}}] \leq 3\left(\mathcal{A}(\lambda) + \mathcal{S}_1(\lambda, \mathbf{z}) + \mathcal{S}_2(\lambda, \mathbf{z})\right) \tag{30}$$

where

$$\mathcal{S}_1(\lambda, \mathbf{z}) = \left\|\sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1} \left(A_{\mathbf{z}}{}^{*}\mathbf{y} - T_{\mathbf{x}} f_{\mathcal{H}}\right)\right\|_{\mathcal{H}}{}^2$$

$$\mathcal{S}_2(\lambda, \mathbf{z}) = \left\|\sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1} (T - T_{\mathbf{x}})(f^{\lambda} - f_{\mathcal{H}})\right\|_{\mathcal{H}}{}^2.$$

Step 2: probabilistic bound on $\mathcal{S}_2(\lambda, \mathbf{z})$. Clearly

$$\mathcal{S}_2(\lambda, \mathbf{z}) \leq \left\|\sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})}^2 \left\|(T - T_{\mathbf{x}})(f^{\lambda} - f_{\mathcal{H}})\right\|_{\mathcal{H}}{}^2. \tag{31}$$

Step 2.1: probabilistic bound on $\left\|\sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})}$. Assume that

$$\Theta(\lambda, \mathbf{z}) = \left\|(T + \lambda)^{-1}(T - T_{\mathbf{x}})\right\|_{\mathcal{L}(\mathcal{H})} \leq \frac{1}{2}, \tag{32}$$

then the Neumann series gives

$$\begin{aligned} \sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1} &= \sqrt{T}(T + \lambda)^{-1}(I - (T + \lambda)^{-1}(T - T_{\mathbf{x}}))^{-1} \\ &= \sqrt{T}(T + \lambda)^{-1} \sum_{n=0}^{+\infty} \left((T + \lambda)^{-1}(T - T_{\mathbf{x}})\right)^n \end{aligned}$$

so that

$$\begin{aligned} \left\|\sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})} &\leq \left\|\sqrt{T}(T + \lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})} \sum_{n=0}^{+\infty} \left\|(T + \lambda)^{-1}(T - T_{\mathbf{x}})\right\|_{\mathcal{L}(\mathcal{H})}^n \\ &\leq \frac{1}{2\sqrt{\lambda}} \frac{1}{1 - \Theta(\lambda, \mathbf{z})}, \end{aligned}$$

where, by spectral theorem, $\left\|\sqrt{T}(T + \lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})} \leq \frac{1}{2\sqrt{\lambda}}$. Inequality (32) now gives

$$\left\|\sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})} \leq \frac{1}{\sqrt{\lambda}}. \tag{33}$$

We claim that (29) implies (32) with probability greater than $1-\eta$. Indeed, let $\xi_1 : X \to \mathcal{L}_2(\mathcal{H})$ be the random variable

$$\xi_1(x) = (T+\lambda)^{-1} T_x.$$

Bound (9) and $\left\|(T+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})} \leq \frac{1}{\lambda}$ imply

$$\|\xi\|_{\mathcal{L}_2(\mathcal{H})} \leq \left\|(T+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})} \|T_x\|_{\mathcal{L}_2(\mathcal{H})} \leq \frac{\kappa}{\lambda} = \frac{H_1}{2}.$$

Lemma 3 ensures that T_x is trace class for almost all x and the inequality

$$\operatorname{Tr}(AB) \leq \|A\|_{\mathcal{L}(\mathcal{H})} \operatorname{Tr} B \tag{34}$$

(A positive bounded operator, B positive trace class operator) implies

$$\begin{aligned}
\mathbb{E}[\|\xi_1\|^2_{\mathcal{L}_2(\mathcal{H})}] &= \int_X \operatorname{Tr}\left(T_x\left(T_x^{\frac{1}{2}}(T+\lambda)^{-2}T_x^{\frac{1}{2}}\right)\right) d\rho_X(x) \\
&\leq \int_X \|T_x\|_{\mathcal{L}(\mathcal{H})} \operatorname{Tr}\left((T+\lambda)^{-2}T_x\right) d\rho_X(x) \\
(\ (9)\ ) &\leq \kappa \operatorname{Tr}\left((T+\lambda)^{-2}T\right) \\
&= \kappa \operatorname{Tr}\left((T+\lambda)^{-1}\left((T+\lambda)^{-\frac{1}{2}}T(T+\lambda)^{-\frac{1}{2}}\right)\right) \\
&\leq \kappa \left\|(T+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})} \operatorname{Tr}\left((T+\lambda)^{-1}T\right) \\
&\leq \frac{\kappa}{\lambda}\mathcal{N}(\lambda) = \sigma_1^2.
\end{aligned}$$

Observing that

$$\mathbb{E}[\xi_1] = T(T+\lambda)^{-1} \qquad \frac{1}{\ell}\sum_{i=1}^{\ell} \xi_1(x_i) = (T+\lambda)^{-1}T_{\mathbf{x}},$$

Proposition 4 applied to ξ_1 gives

$$\left\|(T+\lambda)^{-1}T_{\mathbf{x}} - T(T+\lambda)^{-1}\right\|_{\mathcal{L}_2(\mathcal{H})} \leq 2\log(6/\eta)\left(\frac{2\kappa}{\lambda\ell} + \sqrt{\frac{\kappa\mathcal{N}(\lambda)}{\lambda\ell}}\right)$$

with probability greater than $1-\eta/3$. Then for all $\ell \in \mathbb{N}$ satisfying (29)

$$\log(6/\eta)\left(\frac{2\kappa}{\lambda\ell} + \sqrt{\frac{\kappa\mathcal{N}(\lambda)}{\lambda\ell}}\right) \leq \frac{1}{8} + \frac{1}{8} \leq \frac{1}{4}$$

so that

$$\Theta(\lambda, \mathbf{z}) \leq \left\|(T+\lambda)^{-1}T_{\mathbf{x}} - T(T+\lambda)^{-1}\right\|_{\mathcal{L}_2(\mathcal{H})} \leq \frac{1}{2} \tag{35}$$

with probability greater than $1-\eta/3$.

Step 2.2: probabilistic bound on $\left\|(T-T_{\mathbf{x}})(f^\lambda - f_{\mathcal{H}})\right\|_{\mathcal{L}(\mathcal{H})}$. Let $\xi_2 : X \to \mathcal{H}$ be the random variable

$$\xi_2(x) = T_x(f^\lambda - f_{\mathcal{H}})$$

Bound (9) and the definition of $\mathcal{B}(\lambda)$ give

$$\|\xi_2(x)\|_{\mathcal{H}} \leq \|T_x\|_{\mathcal{L}(\mathcal{H})} \left\| f^\lambda - f_{\mathcal{H}} \right\|_{\mathcal{H}} \leq \kappa \sqrt{\mathcal{B}(\lambda)} = \frac{H_2}{2}.$$

Since T_x is a positive operator

$$\begin{aligned}
\mathbb{E}[\|\xi_2\|_{\mathcal{H}}^2] &= \int_X \left\langle T_x T_x^{\frac{1}{2}}(f^\lambda - f_{\mathcal{H}}), T_x^{\frac{1}{2}}(f^\lambda - f_{\mathcal{H}}) \right\rangle_{\mathcal{H}} d\rho_X(x) \\
&= \int_X \|T_x\|_{\mathcal{L}(\mathcal{H})} \left\langle T_x(f^\lambda - f_{\mathcal{H}}), f^\lambda - f_{\mathcal{H}} \right\rangle_{\mathcal{H}} d\rho_X(x) \\
&\leq \kappa \left\langle T(f^\lambda - f_{\mathcal{H}}), f^\lambda - f_{\mathcal{H}} \right\rangle_{\mathcal{H}} \\
&= \kappa \left\| \sqrt{T}(f^\lambda - f_{\mathcal{H}}) \right\|_{\mathcal{H}}^2 \\
&= \kappa \mathcal{A}(\lambda) = \sigma_2^2.
\end{aligned}$$

Observing that

$$\mathbb{E}[\xi_2] = T(f^\lambda - f_{\mathcal{H}}) \qquad \frac{1}{\ell} \sum_{i=1}^{\ell} \xi_2(x_i) = T_{\mathbf{x}}(f^\lambda - f_{\mathcal{H}}),$$

Proposition 4 applied to ξ_2 gives

$$\left\| (T - T_{\mathbf{x}})(f^\lambda - f_{\mathcal{H}}) \right\|_{\mathcal{H}} \leq 2 \log(6/\eta) \left(\frac{2\kappa \sqrt{\mathcal{B}(\lambda)}}{\ell} + \sqrt{\frac{\kappa \mathcal{A}(\lambda)}{\ell}} \right). \tag{36}$$

with probability greater than $1 - \eta/3$. Replacing (33), (36) in (31), for all $\ell \in \mathbb{N}$ satisfying (29) it holds

$$\mathcal{S}_2(\lambda, \mathbf{z}) \leq 8 \log^2(6/\eta) \left(\frac{4\kappa^2 \mathcal{B}(\lambda)}{\ell^2 \lambda} + \frac{\kappa \mathcal{A}(\lambda)}{\ell \lambda} \right) \tag{37}$$

with probability greater than $1 - 2\eta/3$.
Step 3: probabilistic bound on $\mathcal{S}_1(\lambda, \mathbf{z})$. Clearly

$$\mathcal{S}_1(\lambda, \mathbf{z}) \leq \left\| \sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}(T + \lambda)^{\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H})}^2 \left\| (T + \lambda)^{-\frac{1}{2}} \left(A_{\mathbf{z}}{}^* \mathbf{y} - T_{\mathbf{x}} f_{\mathcal{H}} \right) \right\|_{\mathcal{H}}^2. \tag{38}$$

Step 3.1: bound on $\left\| \sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}(T + \lambda)^{\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H})}$. Clearly,

$$\sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}(T + \lambda)^{\frac{1}{2}} = \sqrt{T}(T + \lambda)^{-\frac{1}{2}} \left\{ I - (T + \lambda)^{-\frac{1}{2}}(T - T_{\mathbf{x}})(T + \lambda)^{-\frac{1}{2}} \right\}^{-1}.$$

Spectral theorem ensures that $\left\| \sqrt{T}(T + \lambda)^{-\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H})} \leq 1$ so, reasoning as in Step 2.1,

$$\left\| \sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}(T + \lambda)^{\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H})} \leq 2 \tag{39}$$

provided that

$$\left\|(T+\lambda)^{-\frac{1}{2}}(T-T_{\mathbf{x}})(T+\lambda)^{-\frac{1}{2}}\right\|_{\mathcal{L}(\mathcal{H})} \leq \frac{1}{2}. \tag{40}$$

If $B = (T+\lambda)^{-\frac{1}{2}}(T-T_{\mathbf{x}})(T+\lambda)^{-\frac{1}{2}}$, then

$$\begin{aligned}
\|B\|^2_{\mathcal{L}_2(\mathcal{H})} &= \operatorname{Tr}\left((T+\lambda)^{-1}(T-T_{\mathbf{x}})(T+\lambda)^{-1}(T-T_{\mathbf{x}})\right) \\
&= \left\langle (T+\lambda)^{-1}(T-T_{\mathbf{x}}), \left((T+\lambda)^{-1}(T-T_{\mathbf{x}})\right)^* \right\rangle_{\mathcal{L}_2(\mathcal{H})} \\
&\leq \left\|(T+\lambda)^{-1}(T-T_{\mathbf{x}})\right\|_{\mathcal{L}_2(\mathcal{H})} \left\|\left((T+\lambda)^{-1}(T-T_{\mathbf{x}})\right)^*\right\|_{\mathcal{L}_2(\mathcal{H})} \\
&= \left\|(T+\lambda)^{-1}(T-T_{\mathbf{x}})\right\|^2_{\mathcal{L}_2(\mathcal{H})},
\end{aligned}$$

and, for all $\ell \in \mathbb{N}$ satisfying (29), (35) ensures that (40) holds with probability $1 - 2\eta/3$.

Step 3.2: bound on $\left\|(T+\lambda)^{-\frac{1}{2}}\left(A_{\mathbf{z}}{}^*\mathbf{y} - T_{\mathbf{x}} f_{\mathcal{H}}\right)\right\|_{\mathcal{H}}$. Let $\xi_3 : X \times Y \to \mathcal{H}$ be the random variable

$$\xi_3(x,y) = (T+\lambda)^{-\frac{1}{2}} K_x \left(y - f_{\mathcal{H}}(x)\right).$$

The definition of M and the polar decomposition of $K_x = \sqrt{T_x}U_x$, where U_x is a partial isometry, give

$$\|\xi_3(x,y)\|_{\mathcal{H}} \leq \left\|(T+\lambda)^{-\frac{1}{2}}\right\|_{\mathcal{H}} \|K_x\|_{\mathcal{L}(Y,\mathcal{H})} \sqrt{M} \leq \sqrt{\frac{\kappa M}{\lambda}} = \frac{H_3}{2}$$

almost surely. Let $P_{x,y} = \langle \cdot, y - f_{\mathcal{H}}(x)\rangle_Y (y - f_{\mathcal{H}}(x))$ with $\|P_{x,y}\|_{\mathcal{L}(Y)} = \|y - f_{\mathcal{H}}(x)\|_Y{}^2 \leq M$, then

$$\begin{aligned}
\mathbb{E}[\|\xi_3\|_{\mathcal{H}}{}^2] &= \int_{X\times Y} \operatorname{Tr}\left(K_x^*(T+\lambda)^{-1}K_x P_{x,y}\right) d\rho_X(x) \\
&\leq \int_X \|P_{x,y}\|_{\mathcal{L}(Y)} \operatorname{Tr}\left((T+\lambda)^{-1}T_x\right) d\rho_X(x) \\
&\leq M \operatorname{Tr}[(T+\lambda)^{-1}T] = M\mathcal{N}(\lambda) = \sigma_3^2,
\end{aligned}$$

where (34) is used replacing $\mathcal{H}$ with Y. Equation (17) gives

$$\mathbb{E}[\xi_3] = (T+\lambda)^{-\frac{1}{2}}\left(A^* y - T f_{\mathcal{H}}\right) = 0,$$

so Proposition 4 applied to ξ_3 ensures

$$\left\|(T+\lambda)^{-\frac{1}{2}}\left(A_{\mathbf{z}}{}^*\mathbf{y} - T_{\mathbf{x}} f_{\mathcal{H}}\right)\right\|_{\mathcal{H}} \leq 2\log(6/\eta)\left(\frac{2}{\ell}\sqrt{\frac{\kappa M}{\lambda}} + \sqrt{\frac{M\mathcal{N}(\lambda)}{\ell}}\right) \tag{41}$$

with probability greater than $1 - \eta/3$. Replacing (39), (41) in (38)

$$\mathcal{S}_1(\lambda, \mathbf{z}) \leq 32\log^2(6/\eta)\left(\frac{4\kappa M}{\ell^2\lambda} + \frac{M\mathcal{N}(\lambda)}{\ell}\right). \tag{42}$$

with probability greater than $1 - \eta$.
Replacing bounds (37), (42) in (30),

$$I[f_{\mathbf{z}}{}^{\lambda}] - I[f_{\mathcal{H}}] \leq 3\mathcal{A}(\lambda) + 8\log^2(6/\eta)\left(\frac{4\kappa^2\mathcal{B}(\lambda)}{\ell^2\lambda} + \frac{\kappa\mathcal{A}(\lambda)}{\ell\lambda} + \frac{16\kappa M}{\ell^2\lambda} + \frac{4M\mathcal{N}(\lambda)}{\ell}\right)$$

and (28) follows by bounding the numerical constants with 128.

Acknowledgments

We would like to thank T. Poggio and L. Rosasco for useful discussions and suggestions.

References

[1] Yasemin Altun, Ioannis Tsochantaridis, and Thomas Hofmann. Hidden markov support vector machines. In *20th International Conference on Machine Learning ICML-2004*, Washington DC, 2003.

[2] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[3] A. Caponnetto and E. De Vito. Fast rates for regularized least-squares algorithm. Technical report, Massachusetts Institute of Technology, Cambridge, MA, April 2005. CBCL/CSAIL Memo.

[4] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49 (electronic), 2002.

[5] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Number 31 in Applications of mathematics. Springer, New York, 1996.

[6] R. M. Dudley. *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.

[7] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[8] T. Evgeniou, C.A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. Technical report, Dept of Computer Science, UCL, January 2005. Research Note RN/04/20.

[9] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Adv. Comp. Math.*, 13:1–50, 2000.

[10] C.A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17:177–204, 2005.

[11] I. F. Pinelis and A. I. Sakhanenko. Remarks on inequalities for probabilities of large deviations. *Theory Probab. Appl.*, 30(1):143–148, 1985.

[12] T. Poggio and F. Girosi. A theory of networks for approximation and learning. In C. Lau, editor, *Foundation of Neural Networks*, pages 91–106. IEEE Press, Piscataway, N.J., 1992.

[13] B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[14] L. Schwartz. Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *J. Analyse Math.*, 13:115–256, 1964.

[15] I. Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4:1071–1105, 2003.

[16] V. N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.

[17] G. Wahba. *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[18] J. Weston, O. Chapelle, A. Elisseeff, B. Schoelkopf, and V. Vapnik. Kernel dependency estimation. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 873–880. MIT Press, Cambridge, MA, 2003.

[19] V. Yurinsky. *Sums and Gaussian vectors*, volume 1617 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1995.

[20] T. Zhang. Leave-one-out bounds for kenrel methods. *Neural Computation*, 13:1397–1437, 2003.