Journal of Machine Learning Research 23 (2022) 1-50 Submitted 11/21; Revised 8/22; Published 9/22

Vector-Valued Least-Squares Regression under Output Regularity Assumptions

Luc Brogat-Motte LUC.MOTTE@TELECOM-PARIS.FR
LTCI, Télécom Paris,
IP Paris, France

Alessandro Rudi ALESSANDRO.RUDI@INRIA.FR
INRIA, Paris, France,
École Normale Supérieure, Paris, France
PSL Research, France

Céline Brouard CELINE.BROUARD@INRAE.FR
INRAE, Toulouse, France
MIAT, Toulouse, France
Université de Toulouse, France

Juho Rousu JUHO.ROUSU@AALTO.FI
Department of Computer Science,
Aalto University, Espoo, Finland

Florence d'Alché-Buc FLORENCE.DALCHE@TELECOM-PARIS.FR
LTCI, Télécom Paris,
IP Paris, France

Editor: Corinna Cortes

Abstract

We propose and analyse a reduced-rank method for solving least-squares regression problems with infinite dimensional output. We derive learning bounds for our method, and study under which setting statistical performance is improved in comparison to full-rank method. Our analysis extends the interest of reduced-rank regression beyond the standard low-rank setting to more general output regularity assumptions. We illustrate our theoretical insights on synthetic least-squares problems. Then, we propose a surrogate structured prediction method derived from this reduced-rank method. We assess its benefits on three different problems: image reconstruction, multi-label classification, and metabolite identification.

Keywords: reduced-rank regression, structured prediction, statistical learning theory, kernel methods

1. Introduction

Learning vector-valued functions plays a key role in a large variety of fields such as economics (Lütkepohl, 2013), physics, computational biology, where multiple variables have to be predicted simultaneously. As opposed to solving multiple single regression problems, the interest of vector-valued regression lies on the ability to take into account the dependence structure among the output variables by appropriate regularization (see for instance

Micchelli and Pontil, 2005; Baldassarre et al., 2012; Álvarez et al., 2012; Lim et al., 2015) or by imposing a low-rank assumption (Anderson, 1951; Izenman, 1975; Velu and Reinsel, 2013). Regarding the infinite dimensional output case, besides functional output regression (Kadri et al., 2016), the motivation for vector-valued regression mainly comes from the application of surrogate approaches in Structured Output Prediction (Weston et al., 2003; Geurts et al., 2006; Kadri et al., 2013; Brouard et al., 2016b; Ciliberto et al., 2020). In order to learn a model to predict an output with some discrete structure, surrogate approaches embed the structured output variable into a Hilbert space and thus boil down to vector-valued regression with a potentially infinite dimensional output space. At prediction time, decoding allows to return a prediction in the original structured output space. Image completion (Weston et al., 2003), label ranking (Korba et al., 2018) and graph prediction (Brouard et al., 2016a) are all examples of structured prediction tasks that can be handled by surrogate approaches.

One way to implement infinite dimensional output regression consists in learning in vector-valued Reproducing Kernel Hilbert Spaces (vv-RKHS) (Micchelli and Pontil, 2005). In particular, regularized least-squares estimators in vv-RKHS enjoy strong theoretical guarantees (see Caponnetto and De Vito, 2007). However complex tasks such as structure prediction very often involve a limited amount of training data compared to the complexity of the input and output data. To overcome this issue, the structure of the target output can be leveraged. This is typically the goal of reduced-rank approaches (Mukherjee and Zhu, 2011; Luise et al., 2019).

In this paper, our aim is to improve upon the regularized least-squares estimators by imposing a rank constraint on the least-squares estimator. Our contributions are three-fold.

As a first contribution, we introduce a novel reduced-rank estimator for vector-valued least-squares regression in the general case of infinite dimensional outputs. Denoting $\mathcal{Y}$ a Hilbert space and $\mathcal{X}$ a Polish space, we consider the following relationship between the input variable and the output variable:

$$y = h^*(x) + \epsilon, \tag{1}$$

where the pair of random vectors (x, y) takes its values in $\mathcal{X} \times \mathcal{Y}$, $\epsilon \in \mathcal{Y}$ is a random noise independent of x with expectation $\mathbb{E}[\epsilon] = 0$ and $h^* : \mathcal{X} \to \mathcal{Y}$ is a measurable function. Assuming we have already an estimator $\hat{h} : \mathcal{X} \to \mathcal{Y}$ of h^* built from a training i.i.d. sample $(x_i, y_i)_{i=1}^n$, we propose to learn a linear operator $\hat{P}$ of rank p, for $p \in \mathbb{N}^*$ allowing to project $\hat{h}(x)$ onto $Z \subset \mathcal{Y}$ with $\dim(Z) \leq p$ giving rise to the following new estimator:

$$x \mapsto \hat{P}\hat{h}(x).$$

This novel estimator generalizes the reduced-rank kernel ridge regression estimator proposed by Mukherjee and Zhu (2011) to the infinite dimensional case.

The second contribution of this paper is to study the proposed least-squares estimator under output regularity assumptions and provide excess-risk bounds. We assume that h^* belongs to a vector-valued reproducing kernel Hilbert Space, namely $h^* = H\phi(.)$ with $H \in \mathcal{Y} \otimes \mathcal{H}_x, \|H\|_{\text{HS}} < +\infty$, and $\phi : \mathcal{X} \to \mathcal{H}_x$ is a canonical map associated to a scalar-valued kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The difficulty of the learning problem in Eq. (1) can be characterized by standard complexity measures. For instance, the capacity condition measures the regularity of the features in terms of eigenvalue decay rate of the covariance operator $C = \mathbb{E}[\phi(x) \otimes \phi(x)]$, and the source condition measures the regularity of H in terms of alignment of H^*H with C (Caponnetto and De Vito, 2007; Ciliberto et al., 2020; Varre et al., 2021). The more regular the problem is, the better are the statistical guarantees. In this work, we consider regularity assumptions on the outputs of the learning problem. We measure the eigenvalue decay rates of the covariance operator $\mathbb{E}[h^*(x) \otimes h^*(x)]$, and $\mathbb{E}[\epsilon \otimes \epsilon]$, and also the alignment of HH^* with HCH^*.

The third contribution of this paper is a novel structured prediction method, which leverages our reduced-rank estimator in the surrogate regression problem. The proposed approach makes use of both an input and an output kernel. In this case, the resulting surrogate regression problem's output space is thus a reproducing kernel Hilbert space. The least-squares analysis allows to prove the the statistical and computational interest of the structured prediction method. In particular, consistency and learning rates for our structured prediction method are given. Moreover, we show by an extensive empirical study on different real world structured prediction tasks that the proposed approach improves upon full rank and state-of-the art structured prediction approaches.

Outline. The paper is organized as follows. In Section 2, we provide a novel reduced-rank method for solving vector-valued least-squares problems. In Section 3, we give learning bounds for the proposed least-squares estimator. Then, we study under which setting this method improves the statistical and computational performance. In particular, our analysis includes and extends the interest of reduced-rank regression beyond the standard setting of reduced-rank regression where the optimum is assumed to be low-rank, and the noise homogeneous in $\mathcal{Y}$. In Section 4, we show how the proposed estimator can be advantageously used in structured prediction with surrogate methods. We give an excess-risk bound for the resulting structured predictor, inherited from our least-squares theoretical analysis. In Section 5, we illustrate our theoretical analysis on synthetic least-squares problems. We empirically show the benefit of the method in structured prediction on three different problems: image reconstruction, multi-label classification, and metabolite identification.

2. Problem Setting and Proposed Estimator

In this section, we introduce the learning setting of vector-valued least-squares regression. Then, we give background on kernel ridge regression. Finally, we present the reduced-rank least-squares estimator proposed in this work.

Vector-valued least-squares regression. We consider the problem of estimating a function $h : \mathcal{X} \to \mathcal{Y}$ with values in a separable Hilbert space $\mathcal{Y}$ with norm $\|.\|_{\mathcal{Y}}$, given a finite set $\{(x_i, y_i)_{i=1}^n\}$ independently drawn from an unknown distribution ρ on $\mathcal{X} \times \mathcal{Y}$, minimizing the expected risk

$$R(h) = \mathbb{E}_\rho[\|h(x) - y\|_{\mathcal{Y}}^2]. \tag{2}$$

The solution is given by $h^*(x) := \mathbb{E}_{\rho(y|x)}[y]$. We define the noise ϵ as the random variable defined by the following equation

$$y = h^*(x) + \epsilon. \tag{3}$$

In practice, solving (2) requires the choice of an hypothesis space $\mathcal{H}$. In this work, we consider reproducing kernel Hilbert space (RKHS).

Reproducing kernel Hilbert spaces. Given a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, one can build a Hilbert space $\mathcal{H}_x$ of scalar-valued functions $\mathcal{H}_x$, called the associated RKHS of k, defined by the completion $\mathcal{H}_x = \overline{\operatorname{span}\{k(x,.) \,|\, x \in \mathcal{X}\}}$ according to the norm induced by the scalar product $\langle k(x,.), k(x',.)\rangle_{\mathcal{H}_x} := k(x,x')$. There is a one-to-one relation between a kernel k and its associated RKHS (Aronszajn, 1950). A crucial tool is the representer theorem which allows to solve in practice regularized empirical risk minimization problems over RKHS (Wahba, 1990; Schölkopf et al., 2001).

Vector-valued reproducing kernel Hilbert spaces. The theory of vector-valued RKHSs (vv-RKHSs) extends the theory of real-valued RKHS by enabling to build Hilbert spaces of vector-valued functions (Senkene and Tempel'man, 1973; Micchelli and Pontil, 2005; Carmeli et al., 2010). We note A^* the adjoint of any operator A. An operator-valued kernel is an application $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ with values in the set of bounded linear operator on $\mathcal{Y}$, satisfying the two following properties: $K(x,x') = K(x',x)^*$ and $\sum_{i,j=1}^n \langle K(x_i, x_j')y_i,\, y_j\rangle_{\mathcal{Y}} \geq 0$ for any $n \in \mathbb{N}^*$, $(x_1,y_1),\ldots,(x_n,y_n) \in \mathcal{X} \times \mathcal{Y}$. Then, akin to scalar-valued kernel, one can build a Hilbert space $\mathcal{H}$ of vector-valued function from $\mathcal{X}$ to $\mathcal{Y}$, called the associated RKHS of K, defined by the completion $\mathcal{H} = \overline{\operatorname{span}\{K(x,.)y \,|\, (x,y) \in \mathcal{X} \times \mathcal{Y}\}}$ according to the norm induced by the scalar product $\langle K(x,.)y, K(x',.)y'\rangle_{\mathcal{H}} := \langle K(x,x')y, y'\rangle_{\mathcal{Y}}$. There is a one-to-one relation between a kernel K and its associated vv-RKHS. Learning with operator-valued kernels is also possible thanks to representer theorems (Micchelli and Pontil, 2005).

Kernel ridge regression. The kernel ridge regression method (KRR) considers the estimator minimizing the following empirical objective

$$\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \|h(x_i) - y_i\|_{\mathcal{Y}}^2 + \lambda \|h\|_{\mathcal{H}}^2 \tag{4}$$

where $\mathcal{H}$ is the RKHS associated to an operator-valued kernel K. In this work, we consider kernel of the form $K(x,x') = k(x,x')I_{\mathcal{Y}}$, where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite scalar-valued kernel on $\mathcal{X}$. In this case, the solution of the problem above can be computed in closed-form as follows:

$$\hat{h}(x) = \sum_{i=1}^n \alpha_i(x) y_i, \qquad \text{with } \alpha(x) = (K + n\lambda)^{-1} k_x \tag{5}$$

where $K = (k(x_i,x_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$, and $k_x = (k(x,x_i))_{i=1}^n \in \mathbb{R}^n$.

Related works in reduced-rank regression. Reduced-rank (or low-rank) estimators are estimators whose predictions $\hat{y} \in \mathcal{Y}$ lie in a linear subspace $Z \subset \mathcal{Y}$, estimated from the data. Reduced-rank regression methods have been proposed for both linear models (Izenman, 1975) and non parametric models (Mukherjee and Zhu, 2011; Foygel et al., 2012; Rabusseau and Kadri, 2016; Luise et al., 2019). Two ways of building reduced-rank estimators have been proposed so far. A first way consists in imposing small rank constraints on the estimated linear operator (Izenman, 1975; Mukherjee and Zhu, 2011; Rabusseau and

Kadri, 2016): on other words, the obtained estimators can be written as full-rank estimators that has been projected with estimated projection operators for a chosen rank p. Among those works devoted to finite dimensional vector-valued regression, the contribution of Rabusseau and Kadri (2016) differs in many ways. They consider a tensor output (the constraint is thus a multilinear rank constraint) and also provide learning bounds. Another way to address reduced-rank regression is to use nuclear norm (or trace norm) penalization as a convex relaxation to rank penalization as developed in (Romera-Paredes et al., 2013; Foygel et al., 2012; Luise et al., 2019). It is worth mentioning that only Luise et al. (2019) tackle an infinite dimensional vector valued-regression problem and provide a statistical study. More precisely, in terms of statistical guarantees, Rabusseau and Kadri (2016) and Luise et al. (2019) show improved constants in learning bounds when using reduced-rank regression, in comparison with full-rank, in their respective settings.

Proposed least-squares estimator. We introduce a non-parametric estimator belonging to the family of reduced-rank estimators. Let $\lambda_1, \lambda_2 > 0$ and $p \in \mathbb{N}^*$. Let $\mathcal{P}_p$ be the set of the orthogonal projections from $\mathcal{Y}$ to $\mathcal{Y}$ of rank p. We note $\hat{h}_\lambda$ a KRR estimator defined using with the training sample $(x_i, y_i)_{i=1}^n$ and a regularization parameter $\lambda > 0$.
Ideally, we would propose the reduced-rank estimator $x \mapsto P\hat{h}_{\lambda_2}(x)$ where P is the operator defined as follows:

$$P := \underset{\mathsf{P} \in \mathcal{P}_p}{\arg\min}\ \mathbb{E}[\|\mathsf{P}h^*(x) - h^*(x)\|_{\mathcal{Y}}^2]. \tag{6}$$

Nevertheless, P is unknown, so we replace it by the following empirical estimator

$$\hat{P}_{\lambda_1} := \underset{\mathsf{P} \in \mathcal{P}_p}{\arg\min}\ \frac{1}{n}\sum_{i=1}^n \|\mathsf{P}\hat{h}_{\lambda_1}(x_i) - \hat{h}_{\lambda_1}(x_i)\|_{\mathcal{Y}}^2, \tag{7}$$

based on a KKR estimator $\hat{h}_{\lambda_1}$ of h^*, with possibly $\lambda_1 \neq \lambda_2$. Eventually, this approximation gives rise to the following proposition for our reduced-rank estimator with hyperparameters $(p, \lambda_1, \lambda_2)$:

$$x \mapsto \hat{P}_{\lambda_1}\hat{h}_{\lambda_2}(x). \tag{8}$$

Remark 2.1 *Note that P is the projection onto the span of the p eigenvectors of the covariance operator $\mathbb{E}[h^*(x) \otimes h^*(x)]$ corresponding to the p greatest eigenvalues. Similarly, $\hat{P}_{\lambda_1}$ is the projection onto the span of the p eigenvectors of the empirical covariance operator $\frac{1}{n}\sum_{i=1}^n \hat{h}_{\lambda_1}(x_i) \otimes \hat{h}_{\lambda_1}(x_i)$ corresponding to the p greatest eigenvalues.*

The proposed estimator allows to cope with any separable Hilbert output space $\mathcal{Y}$ (potentially infinite dimensional), which is of practical interest (See Section 4). Furthermore, efficient and theoretically grounded approximation methods for KRR and kernel principal component analysis (Rudi et al., 2015; Rudi and Rosasco, 2017; Sterge et al., 2020) can be straightforwardly leveraged to alleviate the computation of this estimator. For sake of simplicity, in the remainder of the paper, except when it is necessary, we omit the dependency in λ_1 and λ_2 and use notations $\hat{h}$ and $\hat{P}$.

$\mathcal{X}$	input space
$\mathcal{Y}$	regression output space
$\mathcal{Z}$	structured output space
ρ	probability distribution on $\mathcal{X} \times \mathcal{Y}$
$\Vert\cdot\Vert_{\mathcal{Y}}$	norm of the Hilbert space $\mathcal{Y}$
n/n_{te}	number of training data/test data
h^*	least-squares optimum $x \to \mathbb{E}_{\rho(y\vert x)}[y]$
Δ	structured loss $\Delta : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}^+$
f^*	structured prediction optimum $x \to \arg\min_{\hat{z}\in\mathcal{Z}} \mathbb{E}_{\rho(z\vert x)}[\Delta(z, \hat{z})]$
k	positive definite kernel on $\mathcal{X}$
$\mathcal{H}_x$	RKHS associated to k
$\mathcal{H}$	vv-RKHS associated to $K(x, x') = k(x, x')I_{\mathcal{Y}}$
$\mathcal{P}_p$	space of orthogonal projections from $\mathcal{Y}$ to $\mathcal{Y}$ with rank p
P	$\arg\min_{\mathsf{P}\in\mathcal{P}_p} \mathbb{E}[\Vert\mathsf{P}h^*(x) - h^*(x)\Vert_{\mathcal{Y}}^2]$
A^*	adjoint of A
$A \preceq B$	$\forall u, \langle u, Au\rangle \leq \langle u, Bu\rangle$
$\mu_p(A)$	p-th eigenvalue of A sorted in decreasing order
$\Vert\cdot\Vert_{\text{HS}}$	Hilbert-Schmidt norm
$\Vert\cdot\Vert_{\infty}$	operator norm
$a \otimes b$	defined such as $\forall x, a \otimes bx = \langle b,\, x\rangle a$
$S_p(A)$	$\sum_{k=1}^{p} \mu_k(A)$

Table 1: Notations

Remark 2.2 *The proposed estimator can be seen as a generalization of the reduced-rank estimator defined in (Mukherjee and Zhu, 2011) for finite dimensional vector-valued to the infinite dimensional output case and when λ_1 and λ_2 are not necessarily equal. In this work, we additionally provide learning bounds by leveraging the linear structure of the noise ϵ and those of the outputs $h^*(x)$.*

Notations are gathered in Table 1.

3. Theoretical analysis

In this section, we present a statistical analysis of the proposed estimator. We start, in Section 3.1, by giving the assumptions on the learning problem that we considered. Then, in Section 3.2, we provide learning bounds. Finally, in Section 3.3, we study under which setting reduced-rank regression is statistically and computationally beneficial.

3.1 Assumptions

Here, we introduce and discuss the main assumptions that we need in order to prove our results.

Assumption 1 (attainable case) *We assume that the solution h^* belongs to the RKHS associated to the kernel $K(x, x') = k(x, x')I_{\mathcal{Y}}$, i.e. there exists a linear operator H from $\mathcal{H}_x$*

to $\mathcal{Y}$ with $\|H\|_{\mathrm{HS}} < +\infty$ such that:

$$h^*(x) = H\phi(x). \tag{9}$$

This assumption states that the solution h^* indeed belongs to the chosen hypothesis space $\mathcal{H}$. It is a standard assumption in the learning theory (Ciliberto et al., 2020).

Assumption 2 (regularity of target's outputs) *The operator $M = \mathbb{E}[h^*(x) \otimes h^*(x)]$ satisfies the following property. There exists $\alpha \in [0,1]$ such that:*

$$c_1 := \mathrm{Tr}(M^\alpha) < +\infty. \tag{10}$$

Assumption 2 is always verified for $\alpha = 1$ (as $\mathrm{Tr}(M) \leq \|H\|_{\mathrm{HS}}^2\kappa^2$), and the smaller the α the faster is the eigenvalue decay of M. It quantifies the regularity of the target's outputs $h^*(x) \in \mathcal{Y}$. As a limiting case, when M is finite rank $\alpha = 0$. The capacity condition is a standard assumption for least-squares problems, which can be written $\mathrm{Tr}(C^r) < +\infty$ with $r \in [0,1]$, and that characterises instead the regularity of the features $\phi(x) \in \mathcal{H}_x$. Remark that it implies the Assumption 2 to hold with at least $\alpha \leq r$, but $\alpha \ll r$ is possible.

Assumption 3 (output source condition) *The operators H and $C = \mathbb{E}[\phi(x) \otimes \phi(x)]$ satisfy the following property. There exists $\beta \in [0,1]$, $c_2 > 0$ such that:*

$$HH^* \preceq c_2 M^{1-\beta}. \tag{11}$$

Assumption 3 is always verified for $\beta = 1$ (as $\|H\|_\infty < +\infty$), and the smaller the β the stricter the assumption is. It quantifies the alignment of the left-singular vectors of H with the main components of M. The source condition is a standard assumption for least-squares problems, which can be written $H^*H \preceq aC^{1-r}$ with $r \in [0,1], a > 0$, and that quantifies instead the alignment of the right-singular vectors of H with the main components of C (See, e.g. Ciliberto et al., 2020; Caponnetto and De Vito, 2007). The Assumption 3 allows to show a fast convergence rate of $\hat{P}$. In general, Assumption 3 can be maximum ($\beta = 0$) while the source condition is arbitrarily weak ($r = 1$).

Assumption 4 (diffuse noise and concentrated signal) *The operators M and $E = \mathbb{E}[\epsilon \otimes \epsilon]$ satisfy the following property. There exists $\gamma \in [0,1]$, $c_3 > 0$ such that*

$$c_3 M^{1-\gamma} \preceq E. \tag{12}$$

Assumption 4 quantifies the alignment of the main components of E and M, and the greater the γ the more the noise is diffuse in comparison to the signal. As a limiting case, when $\gamma \to 1$, then $\sigma^2 I_{\mathcal{Y}} \preceq E$ with a certain $\sigma^2 > 0$, which is only possible in finite dimension (e.g. $E = \sigma^2 I_{\mathcal{Y}}$, homogeneous noise commonly assumed in low-rank regression).

Example 1 (finite-rank example) *The standard low-rank regression setting (See Figure 1 left) corresponds to $\mathcal{Y} = \mathbb{R}^d$, $C = \sigma_c^2 I_{\mathcal{H}_x}$ with $\sigma_c^2 > 0$, $H = \sum_{i=1}^r v_i \otimes u_i$ with $r \in \mathbb{N}^*$, $E = \sigma_\epsilon^2 I_{\mathcal{Y}}$ with $\sigma_\epsilon^2 > 0$, $(u_i)_i$, $(v_i)_i$ being orthonormal bases (ONB) of respectively $\mathcal{H}_x$ and $\mathcal{Y}$. In this case, the assumptions are verified with $\alpha = 0, \beta = 0, \gamma = 1$.*

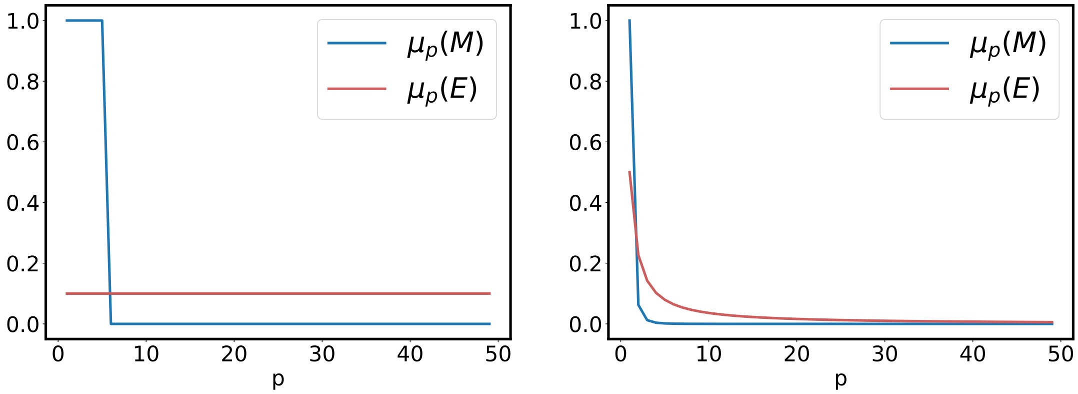

Figure 1: Illustration of finite-rank setting with $r = 5$, $\sigma_c^2 = 1, \sigma_\epsilon^2 = 0.1$ (Left) and polynomial setting with $r_c = 3/2, r_h = 5/4, r_e = 8/7$ (Right). We plot $p \to \mu_p(M) = \langle v_p, M v_p \rangle_{\mathcal{Y}}$ and $p \to \mu_p(E) = \langle v_p, E v_p \rangle_{\mathcal{Y}}$.

Example 2 (polynomial example) *In this paper, we study reduced-rank regression beyond low-rank setting. For instance, we can consider polynomial forms (See Figure 1 right) for* $C = \sum_{i=1}^{+\infty} i^{-r_c} u_i \otimes u_i$, $H = \sum_{i=1}^{+\infty} i^{-r_h} v_i \otimes u_i$, $E = 0.5 \times \sum_{i=1}^{+\infty} i^{-r_e} v_i \otimes v_i$, *with* $(u_i)_i$ *and* $(v_i)_i$ *being (ONB) of* $\mathcal{H}_x$ *and* $\mathcal{Y}$*, respectively. In this case, the assumptions are verified with* $\alpha = \frac{2}{2r_h + r_c}$, $c_1 = \operatorname{Tr}(M^\alpha) < 2$, $\beta = \frac{r_c}{2r_h + r_c}$, $\gamma = 1 - \frac{r_e}{2r_h + r_c}$.

3.2 Main Result

Now, we present the main result of this work which is Theorem 1. Under Assumptions 1, 2, 3, 4, it provides a bound on the proposed estimator's excess-risk for a chosen $p = \operatorname{rank}(\hat{P})$.

Theorem 1 (Learning bounds) *Let* $\hat{P}\hat{h}$ *be the proposed estimator in Eq.* (8) *with* $\operatorname{rank}(\hat{P}) = p$*, built from* n *independent couples* $(x_i, y_i)_{i=1}^n$ *drawn from* ρ*. Let* $\delta \in [0, 1]$*. Under the Assumptions 1, 2, 3, 4, there exists constants* $c_4, c_5, c_8 > 0$, $n_0 \in \mathbb{N}^*$ *defined in the proof, and independent of* p, n, δ*, such that, if* $\mu_{p+1}(M) \geq c_8 \log^8(\frac{8}{\delta}) n^{-\frac{1}{\beta+1}}$ *and* $n \geq n_0$*, then with probability at least* $1 - 3\delta$*,*

$$\mathbb{E}_x[\|\hat{P}\hat{h}(x) - h^*(x)\|_{\mathcal{Y}}^2]^{1/2} \leq \left(c_4 \sqrt{p} n^{-1/4} + c_5 S_p(E)^{1/4}\right) n^{-1/4} \log(n/\delta) + \sqrt{3c_1} \mu_{p+1}(M)^{1/2(1-\alpha)} \tag{13}$$

with $S_p(E) = \sum_{i=1}^p \mu_i(E)$.

The bound is the sum of two terms: the first one increases with p, the second one decreases with p. When $p = o(\sqrt{n})$, the first term is dominated by a term proportional to $S_p(E)^{1/4} \log(n/\delta) n^{-1/4}$, which should be compared to the dominating term of the kernel ridge estimator's bound $\operatorname{Tr}(E)^{1/4} n^{-1/4}$ (cf. Lemma 13): instead of the total amount of noise $\operatorname{Tr}(E)$, the reduced-rank estimator only incurs the quantity within the p main components of E, plus a logarithmic term in n. The second term of the sum decays w.r.t p at the speed of the eigenvalue decay rates of $\mathbb{E}_x[h^*(x) \otimes h^*(x)]$, modulo an exponent $1-\alpha$. Finally, the condition $\mu_{p+1}(M) \geq c_8 n^{-\frac{1}{\beta+1}}$ stems from the estimation error of P, and can translate into the existence of a plateau threshold p^* from which the second term cannot decrease anymore (See Rudi et al. (2013)). Hence, the stronger is Assumption 3, the faster is the estimation of $\hat{P}$ and the divergence rate of the plateau threshold. We give here a sketch of the proof for the Theorem 1. The complete proof is detailed in Appendix A.

Sketch of the proof. The proof consists in decomposing the excess-risk of the estimator $\hat{P}\hat{h}$ as follows.

$$\mathbb{E}_x[\|\hat{P}\hat{h}(x) - h^*(x)\|_{\mathcal{Y}}^2]^{1/2} \leq \underbrace{\mathbb{E}_x[\|\hat{P}\hat{h}(x) - \hat{P}h^*(x)\|_{\mathcal{Y}}^2]^{1/2}}_{\text{regression error on a subspace}} + \underbrace{\mathbb{E}_x[\|\hat{P}h^*(x) - h^*(x)\|_{\mathcal{Y}}^2]^{1/2}}_{\text{reconstruction error}}. \quad (14)$$

Then each right-hand term is bounded using a dedicated lemma given in the Appendix A. Lemma 7 bounds the regression error on the subspace defined by $\hat{P}$ (akin to a variance). Lemma 11 bounds the reconstruction error (akin to a bias). We exploit techniques and schemes similar to those used in (Rudi et al., 2013; Rudi and Rosasco, 2017; Ciliberto et al., 2016, 2020; Luise et al., 2019) in order to prove these lemmas. Namely, L^2-norms of functions in $\mathcal{H}$ are expressed as Hilbert-Schmidt norms of Hilbert-Schmidt operators in $\mathcal{Y} \otimes \mathcal{H}_x$. Relevant norms decompositions lead to study the deviation of the sample operators from the true operators $\mathbb{E}[y \otimes \phi(x)]$ and $\mathbb{E}[\phi(x) \otimes \phi(x)]$. For this purpose, Bernstein's inequalities for the operator norm, or the Hilbert-Schmidt norm, of random operators between separable Hilbert spaces are applied (Tropp, 2012). The previously introduced assumptions of Section 3.1 play an important role in the proof of Lemma 11, allowing to obtain faster learning rate for $\hat{P}$.

Remark 3.1 (Independence assumption on $\phi(x)$ and ϵ) *In this work, we assume that $\phi(x)$ is independent of ϵ. This allows to keep a clear exposition of the proofs, by performing lighter mathematical derivations. Nevertheless, such assumptions is not exploited by the proposed method, and similar results hold without this assumption as we discuss in Appendix A.7.*

3.3 Polynomial Eigenvalue Decay Rates

In this subsection, we discuss under which setting reduced-rank ridge regression can be statistically and computationally advantageous in comparison to standard full-rank ridge regression. For this purpose, we apply Theorem 1 considering polynomial eigenvalue decay rates for M and E.

Assumption 5 (polynomial eigenvalue decay rates) *M and E have polynomial eigenvalue decay rates with parameter $s > 1$ and $e > 1$, if there exist constants $a, A, b, B > 0$ such that:*

$$ap^{-s} \leq \mu_p(M) \leq Ap^{-s}, \quad (15)$$

$$bp^{-e} \leq \mu_p(E) \leq Bp^{-e}. \quad (16)$$

Parameters s and e characterize the shapes of the signal's and noise's distributions in $\mathcal{Y}$, and provide information complementary to the total amounts of variance $\mathrm{Tr}(M)$ and $\mathrm{Tr}(E)$. Moreover, notice that Assumption 5 does not require an exact polynomial decay of the eigenvalues $\mu_k \propto k^{-r}$. In particular, one can define a measure of distortion of $\mu_k(M)$ and $\mu_k(E)$ from exact polynomial decays as the values $\frac{A}{a}$ and $\frac{B}{b}$, respectively. The greater are these ratios the greater are the distortions.

Remark 3.2 (Assumptions relationship) *Assumption 5 implies that Assumption 2 holds with* $c_1 = \mathrm{Tr}(M^{\frac{2}{s}})$, *and Assumption 4 holds with* $\gamma = 1 - \frac{e}{s}$ *and* $c_3 = A^{e/s}b^{-1}$.

Under the Assumptions 1, 3, and 5 we derive the following corollary from Theorem 1 in the special case of polynomial eigenvalue decay rates.

Corollary 2 (Learning bounds (polynomial decay rates)) *Let* $\delta \in\ ]0,1]$, $n \geq n_0$. *Under Assumptions 1, 3, and 5, assuming* $\frac{B}{b} \leq \theta$ *with* $\theta \geq 1$, *then by taking only*

$$p = c_9(\log^8(\frac{8}{\delta}))^{-\frac{1}{s}} n^{\frac{1}{(\beta+1)s}}, \tag{17}$$

we have with probability at least $1 - 3\delta$:

$$\mathbb{E}_x[\|\hat{P}\hat{h}(x) - h^*(x)\|_{\mathcal{Y}}^2]^{1/2} \ \leq\ c_{10}(s,e)\ \log^{5/4}(\frac{n}{\delta})\, n^{-1/4} \ +\ c_{11}(e)\, n^{-\frac{1}{2}\frac{1-2/s}{1+\beta}}\ \log^8(\frac{8}{\delta}), \tag{18}$$

where $c_{10}(s,e) = \tilde{c}_{10}\left(\frac{e(e-1)}{s}\right)^{1/4}\left(1 + \log\left(\frac{e}{e-1}\right)\right)$, $c_{11}(e) = \tilde{c}_{11}\left(1 + \log\left(\frac{e}{e-1}\right)\right)$. $\tilde{c}_{10}$, $\tilde{c}_{11}$, n_0, *are constants independent of* n, δ, s, e, *and* c_9 *is a constant independent of* n, δ, *defined in the proofs.*

As a first remark, note that the chosen components number p of order $\mathcal{O}(n^{\frac{1}{(\beta+1)s}})$ is significantly smaller than n when s is big (concentrated signal). For instance, $s = 2$ yields at most to $p = O(\sqrt{n})$. Then, notice that the bound is the sum of two terms. The first term is decaying in $O(n^{-1/4})$ modulo a logarithm term in n, and its multiplicative constant can be arbitrarily small when e is small (spread noise), as $c_{10}(s,e) \xrightarrow[e \to 1^+]{} 0$. The decreasing rate of the second term varies within the open interval $]0, 1/2[$. The greater is s and the smaller is β, the better is the rate.

Comparison with full-rank estimator's bound. The bound provided in Eq. (18) sheds light on the role of M and E's shapes, flat ($s, e \to 1^+$) or concentrated ($s, e \to +\infty$), in the performance of the reduced-rank estimator. At the opposite, remark that the full-rank ridge estimator's bound is dominated by a term of the form $c(\kappa + \|H\|_{\mathrm{HS}})\,\mathrm{Tr}(E)n^{-1/4}\log(\frac{4}{\delta})$ with $c > 0$ a constant independent of n, δ, s, e (See Lemma 13). So, the ridge estimator is not impacted by the shapes of M and E, but is only affected by the total amounts of signal $\|H\|_{\mathrm{HS}}$, and noise $\mathrm{Tr}(E)$.

Favorable settings for reduced-rank. Which situations are favorable to the proposed reduced-rank method? To simplify the discussion, let us not consider the terms $(1 + \log(e/(e-1)))$ appearing in c_{10}, c_{11}. If s is big enough and β small enough then the right term of (18) is $o(n^{-1/4})$ (e.g. $s = 6$, $\beta = 0$ gives $\mathcal{O}(n^{-1/3})$). So, for n big enough, it remains to compare the left term of the bound with the dominating term of the ridge bound. When e becomes close to 1^+ the left term can be arbitrarily smaller than the ridge bound, because $c_{10}(s, e) \to 0$, while $c \operatorname{Tr}(E)$ is unchanged. Let be $q \in \mathbb{N}^*$. For the following family of settings:

$$\beta < 1 - \frac{4}{s}, \qquad\qquad e \in]1, e^*(n, q)] \tag{19}$$

with $e^*(n, q) = \sup\{e \,/\, c_{10}(s, e) < \frac{c \operatorname{Tr}(E)^{1/4}}{q \log^{5/4}(n)}\}$, the reduced-rank bound is q times smaller than the full-rank one, when n is big enough.
This gain is obtained because the projection yields to an important noise reduction and a small increase in bias. This can be think as a direct generalization of the low-rank regression setting.

In the following corollary, we duly show that, despite the $(1 + \log(e/(e-1)))$ terms, one can find settings $(n, s, e) \in \mathbb{N}^* \times \mathbb{R}^+ \times \mathbb{R}^+$ such that the learning bound (18) is arbitrarily smaller than the kernel ridge estimator's one under the same assumptions on the learning problem.

Corollary 3 (Statistical gain of reduced-rank regression) *Let $\delta \in]0, 1]$ and $\epsilon > 0$. If $\beta < 1$, then there exists a setting $s, e > 1$, $n \in \mathbb{N}^*$, such that, under the assumptions of Corollary 2, with probability at least $1 - 3\delta$,*

$$\mathbb{E}_x[\|\hat{P}\hat{h}(x) - h^*(x)\|^2_{\mathcal{Y}}]^{\frac{1}{2}} \leq \epsilon \times \operatorname{Tr}(E)^{1/4} \times n^{-1/4}. \tag{20}$$

Proof We exhibit such a setting (n, s, e). We choose (s, β) such that $\beta < 1 - \frac{4}{s}$. One can check that in this case $c_{11} n^{-\frac{1}{2}\frac{1-2/s}{1+\beta}} \log(n/\delta) = o(n^{-1/4})$, and also $c_{10} \left(\frac{e\theta}{\zeta(e)s} \times \log^5(\frac{n}{\delta})\right)^{1/4} n^{-1/4} = o(n^{-1/4})$ (when $e \to 1^+, n \to +\infty$, with $e \geq 1 + \frac{1}{n^a}$ for any $a > 0$). So, taking n big enough we obtain the desired inequality. ■

Corollary 3 shows that a significant statistical gain is possible using reduced-rank regression, even if the support of $h^*(x)$ covers the entire output space $\mathcal{Y}$, i.e. beyond the standard low-rank setting. Besides the statistical gain, reducing the rank of the predictions' space is of interest for reducing the computational complexity at prediction time.

As it will be presented in the application to structured prediction (See Section 4), decoding predictions in surrogate approaches or simply computing mean squared errors require to calculate inner products between the predictions provided by the regression estimator and elements of the output space. In the following lemma, we analyze the complexity in time of such computations. Note that the same complexity holds for computing distances between predictions and elements of the output space. We consider the setting where the dimension of $\mathcal{Y}$ is bigger than n (e.g. infinite).

Corollary 4 (Computational gain of reduced-rank regression) *Let $\hat{h} : \mathcal{X} \to \mathcal{Y}$ be a kernel ridge estimator trained on n points. Let $\hat{P} : \mathcal{Y} \to \mathcal{Y}$ be a projection operator of rank p. Given N output points $(y_i)_{i=1}^N$, computing the inner products $\left(\langle \hat{P}\hat{h}(x), y_i\rangle_{\mathcal{Y}}\right)_{i=1}^N$ has a time and space complexity of order $\mathcal{O}(p(N+n))$ while computing the inner products $\left(\langle \hat{h}(x), y_i\rangle_{\mathcal{Y}}\right)_{i=1}^N$ has a time complexity $\mathcal{O}(nN)$.*

Proof In order to compute $\left(\langle \hat{P}\hat{h}(x), y_i\rangle_{\mathcal{Y}}\right)_{i=1}^N$ one needs to compute

$$\underbrace{\alpha(x)^T}_{(1,n)} \underbrace{(UY_{tr})^T}_{(n,p)} \underbrace{UY}_{(p,N)} \tag{21}$$

with $\alpha(x) = (K + n\lambda I)^{-1} k_x$, $k_x = (k(x, x_1), \ldots, k(x, x_n))$, $U = \sum_{i=1}^p e_i \otimes u_i$, where $(u_i)_{i=1}^p$ is an orthogonal basis of the range of $\hat{P}$, $(e_i)_{i=1}^p$ an orthogonal basis of $\mathbb{R}^p$, and Y_{tr} is the operator with the n training output points as columns, Y the operator with the N output points as columns. This costs $p(N+n)$ in time and space complexity. In order to compute the $\left(\langle \hat{h}(x), y_i\rangle_{\mathcal{Y}}\right)_{i=1}^N$ one needs to compute

$$\underbrace{\alpha(x)^T}_{(1,n)} \underbrace{K^y}_{(n,N)} \tag{22}$$

with K^y the gram matrix between the n training points and N output points for the kernel $k_y(y, y') = \langle y, y'\rangle_{\mathcal{Y}}$. This costs nN in time and space complexity. ■

Corollary 4 shows that a significant computational gain is possible when $N \gg p$ and $n \gg p$, as in this case $p(N+n) \ll nN$. Combining this result with Corollary 3 we conclude that, under the output regularity assumptions made, the proposed method offers both statistical and computational gains by projecting the ridge estimator onto an estimated linear subspace.

Remark 3.3 (Consequences for finite dimensional $\mathcal{Y}$.) *The obtained results are not limited to the infinite dimensional setting and are still valuable when $\mathcal{Y} = \mathbb{R}^d$. One can notice that in the finite dimensional case Assumptions 2, 3, and 4 are always verified choosing the best exponents $\alpha = \beta = 0, \gamma = 1$ (if $M, E \succ 0$), but it is at the price of very large constants c_1, c_2 and very small c_3, which make the bounds very large. In fact, it amounts to using the rough inequalities $\mathrm{Tr}(A) \leq d \times \|A\|_\infty$ and $A \preceq \frac{\mu_1(B)}{\mu_d(B)} B$ for any bounded operators A, B, thereby loosing information on the shape of M and E. At the opposite, choosing α, β, γ such that the constants c_1, c_2, c_3 remain close to 1 allows to obtain finer bounds, taking into account the signal/noise configuration, closed to the observed behaviors.*

Take-home message. The proposed reduced-rank regression estimator enjoys a statistical gain under more general assumptions than standard low-rank assumptions. As parameter λ, the rank p acts as a regularization parameter whose impact should disappear when the size of the training sample increases, i.e. $p \xrightarrow[n \to +\infty]{} +\infty$. The settings where the proposed method performs better than the kernel ridge estimator require faster eigenvalue decay rates for $\mathbb{E}[h^*(x) \otimes h^*(x)]$ than for $\mathbb{E}[\epsilon \otimes \epsilon]$ (concentrated signal/diffuse noise). But this is not sufficient: Assumption 3 with a sufficiently small β ($\beta < 1 - \frac{4}{s}$) is also necessary to ensure a fast enough estimation of P. Last but not least, reducing the predicted outputs' dimension can also yield to substantial computational gains.

4. Application to Structured Prediction

In this section, we develop an application of the reduced-rank estimator to structured prediction. The novel method fits into the generic framework of surrogate approaches for structured prediction and exploits an infinite dimensional embedding by the mean of a kernel. We describe the algorithm and give learning bounds for the proposed structured prediction estimator.

4.1 Surrogate Reduced-Rank Estimator for Structured Prediction

Structured prediction consists in solving a supervised learning task where the output variable is a structured object. Denoting $\mathcal{Z}$ the structured output space, a structured loss $\Delta : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ measures the discrepancy between a true output and a predicted output. The goal of structured prediction is to minimize the following expected risk:

$$R(f) = \mathbb{E}_\rho[\Delta(f(x), z)], \tag{23}$$

over a class of functions $f : \mathcal{X} \to \mathcal{Z}$, using a finite set $(x_i, z_i)_{i=1}^n$ independently drawn from an unknown distribution ρ on $\mathcal{X} \times \mathcal{Z}$. In other words, if we note $f^* : \mathcal{X} \to \mathcal{Z}$ the minimizer of $R(f)$, the aim of learning is therefore to get an estimator $\hat{f}$ of f^* based on the finite sample $(x_i, z_i)_{i=1}^n$ with the best possible statistical properties.

A surrogate approach: Output Kernel Regression We consider here the case when Δ is defined as a metric induced by a positive definite kernel k_z acting over the structured output space $\mathcal{Z}$:

$$\Delta(z, z') = \|\psi(z) - \psi(z')\|_{\mathcal{H}_z}^2. \tag{24}$$

This boils down to embedding objects of $\mathcal{Z}$ into the Reproducing Kernel Hilbert Space associated to k_z using the canonical feature map $\psi : \mathcal{Z} \to \mathcal{H}_z$ associated to k_z, and then consider the square loss over $\mathcal{H}_z$. Relying on the abundant literature about kernels on structured objects (Gärtner, 2003), this class of losses covers a wide variety of structured prediction problems.

However, learning directly f through ψ still raises an issue and a simple way to overcome it consists in seeking instead a **surrogate** model $h : \mathcal{X} \to \mathcal{H}_z$ able to predict the embedded objects in the infinite dimensional space $\mathcal{H}_z$ and leverage the kernel trick in the output space. This approach is referred as Output Kernel Regression (OKR) (Weston et al., 2003; Geurts et al., 2006; Brouard et al., 2016b). The original structured prediction problem is then replaced by the following surrogate vector-valued regression problem stated in terms of the surrogate true risk:

$$\min_{h : \mathcal{X} \to \mathcal{H}_z} \mathbb{E}_\rho[\|h(x) - \psi(z)\|_{\mathcal{H}_z}^2]. \tag{25}$$

Assume h^* is the function $x \to \mathbb{E}_y[\psi(z)|x]$ (solution of Eq. (25)). Then at prediction time, one can retrieve a prediction in the original space $\mathcal{Z}$ through an appropriate decoding function $d : \mathcal{H}_z \to \mathcal{Z}$:

$$z^* = f^{**}(x) := d \circ h^*(x) := \underset{z \in \mathcal{Z}}{\arg\min} \ \|h^*(x) - \psi(z)\|^2_{\mathcal{H}_z}. \tag{26}$$

The overall approach is illustrated on Fig. 2.

$$\begin{tikzcd}
 & \mathcal{Y} = \mathcal{H}_z \arrow[d, bend left, "d"] \\
\mathcal{X} \arrow[ur, "h \in \mathcal{H}"] \arrow[r, "f \ = d \circ h"'] & \mathcal{Z} \arrow[u, "\psi"']
\end{tikzcd}$$

Figure 2: Schematic illustration of OKR.

Ciliberto et al. (2016) have proved that f^{**} solves exactly the original structured prediction problem, i.e. $f^{**} = f^*$. Fo that purpose, they have shown that $\Delta(z, z') = \|\psi(z) - \psi(z')\|^2_{\mathcal{H}_z}$ belongs to the wide family of Structure Encoding Loss Functions (SELF), as it can be written $z, z' \to \langle \gamma(z), \theta(z') \rangle_\nu$ with $\nu = \mathcal{H}_z \oplus \mathbb{R} \oplus \mathbb{R}$, $\gamma(z) = (\sqrt{2}\psi(z), \|\psi(z)\|^2_{\mathcal{H}_z}, 1)$, and $\gamma(z) = (-\sqrt{2}\psi(z'), 1, \|\psi(z')\|^2_{\mathcal{H}_z})$.

Moreover, when providing an estimator $\hat{h}$ of h^* using the training sample $(x_i, z_i)_{i=1}^n$, we benefit from the so called comparison inequality from Ciliberto et al. (2016)

$$R(\hat{f}) - R(f^*) \leq c \times \mathbb{E}_x[\|\hat{h}(x) - h^*(x)\|^2_{\mathcal{H}_z}]^{1/2}, \tag{27}$$

where $\hat{f} = d \circ \hat{h}$ and the constants c and Q are defined as: $c = 2\sqrt{2Q^2 + Q^4 + 1}$, and $Q = \sup_z \|\psi(z)\|_{\mathcal{H}_z}$.

Reduced-rank regression in structured prediction. The OKR problem depicted in Eq. (25) can be solved in various hypothesis spaces and trees-based approaches (Geurts et al., 2006) as well as kernel methods (Weston et al., 2003; Geurts et al., 2006; Brouard et al., 2011; Kadri et al., 2013; Laforgue et al., 2020) have been developed so far to tackle it. We focus here on Input Output Kernel Regression (IOKR), a method that exploits operator-valued kernels (Brouard et al., 2016b) and assumes that h belongs to a vv-RKHS. In particular, IOKR-ridge solves the kernel ridge regression problem in Eq. (4) with the following choice s: the output space is $\mathcal{Y} := \mathcal{H}_z$, the chosen operator-valued kernel writes as $K(x, x') = k(x, x')I_{H_z}$,and the hypothesis space$\mathcal{H}$ is the vv-RKHS associated to K. Instantiating Eq. 5, the solution to IOKR-ridge writes as:

$$\hat{h}(x) = \sum_{i=1}^n \alpha_i(x)\psi(z_i), \tag{28}$$

where α_i's are defined according Eq. 5.

In this section, we propose to solve the surrogate problem in Eq. (25) using our reduced-rank estimator based on the IOKR-ridge estimator. This gives rise to the definition of a novel structured output prediction $\hat{f}$:

$$\hat{f}(x) := \operatorname*{arg\,min}_{z \in \mathcal{Z}} \|\hat{P}\hat{h}(x) - \psi(z)\|^2_{\mathcal{H}_z}. \tag{29}$$

Because of the comparison inequality Eq. (27), the resulting structured predictor directly benefits from the learning bound on the least-squares problem.

Theorem 5 (Excess-risk bound for the structured predictor) *Let $\delta \in]0, 1]$, $n \geq n_0$. Under Assumptions 1, 3, and 5, assuming $\frac{B}{b} \leq \theta$ with $\theta \geq 1$, then by taking only*

$$p = c_9(\log^8(\frac{8}{\delta}))^{-\frac{1}{s}} n^{\frac{1}{(\beta+1)s}} \tag{30}$$

then with probability at least $1 - 3\delta$

$$R(\hat{f}) - R(f^*) \;\leq\; c \times \Big(c_{10}(s,e)\, \log^{5/4}(\frac{n}{\delta})\, n^{-1/4} \;+\; c_{11}(e)\, n^{-\frac{1}{2}\frac{1-2/s}{1+\beta}}\, \log^8(\frac{8}{\delta})\Big) \tag{31}$$

where $c_{10}(s,e) = \tilde{c}_{10}\left(\frac{e(e-1)}{s}\right)^{1/4}\left(1 + \log\left(\frac{e}{e-1}\right)\right)$, $c_{11}(e) = \tilde{c}_{11}\left(1 + \log\left(\frac{e}{e-1}\right)\right)$. $\tilde{c}_{10}$, $\tilde{c}_{11}$, n_0, are constants independent of n, δ, s, e and c_9 is a constant independent of n, δ, defined in the proofs.

The bound provided in Theorem 5 is similar to the one of Corollary 2 modulo the multiplicative constant c, and thus the interpretation is the same. In particular, when s is sufficiently big and e, β sufficiently small, we can obtain a significant statistical gain in comparison to the not projected estimator, as shown in Corollary 3.

4.2 Algorithms and Complexity Analysis

To define the final reduced-rank IOKR-ridge estimator $\hat{f}$, one has to apply Algorithm 4.2 to compute all the parameters of $\hat{P}\hat{h}$ necessary to the decoding phase described in Algorithm 4.2.

Complexity in time At decoding/prediction time, one needs to compute n_{te} times the prediction $\hat{f}(x_i)$, for the testing data points $(x_i)_{i=1}^{n_{te}}$. Each prediction requires to calculate the distances in Eq. (26). This is made possible by using the kernel trick, avoiding to compute the infinite dimensional vectors $\hat{h}(x)$ and $\psi(z)$. These computations cost $\mathcal{O}(n_{te}n|\mathcal{Z}|)$ in time, where n and $|\mathcal{Z}| \in \mathbb{N}^*$ are the size of the training data set and the number of output candidates, respectively. Note that $|\mathcal{Z}|$ is typically very big in structured prediction. For instance, in multilabel classification with d labels $|\mathcal{Z}| = \{0,1\}^d = 2^d$. In practice, one often chooses a subset of $\mathcal{Z}$ as a candidate set. Hence, the decoding phase badly scales with n, and in general is computationally expensive. Because of the projection onto a finite dimensional space, the proposed method can significantly alleviate these computations. When using $\hat{P}\hat{h}$ with $\hat{P}$ of rank p, the decoding time complexity reduces to $\mathcal{O}(n_{te}p|\mathcal{Z}|)$ as shown in Corollary 4. Furthermore, the training phase consists in a matrix inversion for computing $\hat{h}$ plus a singular value decomposition for computing $\hat{P}$. Hence, the time complexity of the training algorithm without approximation is $\mathcal{O}(2n^3)$. It can still be reduced using efficient and theoretically grounded approximation methods for KRR and kernel principal component analysis developed in (Rudi et al., 2015; Rudi and Rosasco, 2017; Sterge et al., 2020).

Algorithm 1 Reduced-rank IOKR-ridge - Training phase

```
    **Input:** $K_x, K_z \in \mathbb{R}^{n\times n}$, $\lambda \geq 0$, $p \in \mathbb{N}^*$
    **KRR estimation:** $W = (K_x + n\lambda I)^{-1} \in \mathbb{R}^{n\times n}$
    **Subspace estimation:**
    $K_h = WK_xK_zK_xW \in \mathbb{R}^{n\times n}$
    $\beta = \begin{bmatrix} | & & | \\ \frac{u_1}{\sqrt{\mu_1}} & \dots & \frac{u_p}{\sqrt{\mu_p}} \\ | & & | \end{bmatrix} \in \mathbb{R}^{n\times p} \leftarrow SVD(K_h) = \sum_{l=1}^{n} \mu_l u_l u_l^T$
    **Training outputs projection:**
    $K_{zh} = K_zWK_x \in \mathbb{R}^{n\times n}$
    $UY = K_{zh}\beta \in \mathbb{R}^{n\times p}$
    **Return:** $W$ (KRR coefficients), $\beta$ (projection coefficients), $UY$ (projected training outputs)
```

Algorithm 2 Reduced-rank IOKR-ridge - Decoding phase

```
    **Input:** $k_x^{te} \in \mathbb{R}^n$, $Z_{candidates} \in \mathbb{R}^{n_c\times d}$, $UY \in \mathbb{R}^{n\times p}$, $W \in \mathbb{R}^{n\times n}$
    **Output candidates projection:**
    $K_{zh} = WK_xK_z^{tr/c} \in \mathbb{R}^{n\times n_c}$
    $UY_c = K_{zh}\beta \in \mathbb{R}^{n_c\times p}$
    **Distances computation:**
    $\alpha = Wk_x^{te} \in \mathbb{R}^n$
    $Uh_{te} = UY^T\alpha \in \mathbb{R}^p$
    $S := ”\langle \hat{P}\hat{h}(x_{te}),\, \psi(Z_{candidates})\rangle_{\mathcal{H}_z}” = (Uh_{te})^T UY_c \in \mathbb{R}^{n_c}$
    $N := ”\|\psi(Z_{candidates})\|^2_{\mathcal{H}_z}” = (K_z(z,z))_{z\in Z_{candidates}} \in \mathbb{R}^{n_c}$
    $D = N - 2S$
    **1-NN prediction :**
    $\hat{i} = \arg\min_{i\in[1,n_c]} D_i$
    $\hat{z} = Z_{candidates}[\hat{i}] \in \mathcal{Y}$
    **Return:** $\hat{z}$ (prediction)
```

Algorithm	IOKR	Reduced-rank IOKR
Training	$\mathcal{O}(n^3)$	$\mathcal{O}(2n^3)$
Decoding	$\mathcal{O}(n_{te}n\lvert\mathcal{Z}\rvert)$	$\mathcal{O}(n_{te}p\lvert\mathcal{Z}\rvert)$

Table 2: Time complexity of IOKR versus reduced-rank IOKR.

5. Numerical Experiments

We now carry out experiments with the methods proposed in this work. In Section 5.1, we illustrate our theoretical insights on synthetic least-squares problems. In Section 5.2, we test the proposed structured prediction method on three different problems: image reconstruction, multi-label classification, and metabolite identification.

5.1 Reduced-rank regression: statistical gain and importance of Assumption 3

We illustrate, on synthetic least-squares problems, the theoretical insights, given in Subsection 3.3. For $d = 300$, $\mathcal{X} = \mathcal{H}_x = \mathcal{Y} = \mathbb{R}^d$, we choose $\mu_p(C) = \frac{1}{\sqrt{p}}$, $\mu_p(E) = \frac{0.2}{p^{1/10}}$. We draw randomly the eigenvector associated to each eigenvalue. We draw $H_0 \in \mathbb{R}^{d\times d}$ with independently drawn coefficients from the standard normal distribution. We consider two different optimums $H = H_0$ ($\beta = 1$) and $H = (H_0CH_0)H_0$ ($\beta = 1/3$). Then, we generate $n \in [10^2, \ldots, 5\times 10^3], n_{val} = 1000, n_{test} = 1000$ couples (x, y) such that $x \sim \mathcal{N}(0, C)$, $\epsilon \sim \mathcal{N}(0, E)$, and $y = Hx + \epsilon$. We select the hyper-parameters of the three estimators $\hat{h}$, $P\hat{h}$, and $\hat{P}\hat{h}$ in logarithmic grids, with the best validation MSE. On the Figure 3 we plot the test MSE obtain by the three estimators for various p and n, and for the two different optimums $H = H_0$ (left) and $H = (H_0CH_0)H_0$ (right). There exists for both H (left/right) a minimum MSE w.r.t p for $P\hat{h}$ below the MSE of $\hat{h}$ when n is big enough: P offers a valuable regularization of $\hat{h}$. Moreover, we observe that the selected p increases when n increases with a decreasing gain, following the provided bounds' behavior. Furthermore, we observe that because of the estimation error of $\hat{P}$, there is no gain for $\hat{P}\hat{h}$ when $H = H_0$, while when $H = (H_0CH_0)H_0$ there is a gain for n big enough. This illustrates the faster convergence rate of $\hat{P}$ when β is small.

5.2 Experiments on Structured Prediction

In this section, we assess the performance of the reduced-rank IOKR estimator calculated using Algorithms 4.2 and 4.2 proposed in Section 4 on three real-world structured prediction tasks: image reconstruction, multi-label classification, and metabolite identification. Our experiments show how reduced-rank regression can be advantageously used for surrogate methods in structured prediction in order to improve both statistical and computational aspects. In these experiments, we choose $\lambda_1 = \lambda_2$ in order to reduce the quantity of hyperparameters.

State of the art approaches For each task, we compared our reduced-rank method to relevant existing SOTA approaches. SPEN Belanger and McCallum (2016), a neural network learned by minimizing the structured hinge loss, is an Energy-Based Model (EBM), considered as a strong benchmark in the literature. Contrary to surrogate approaches,

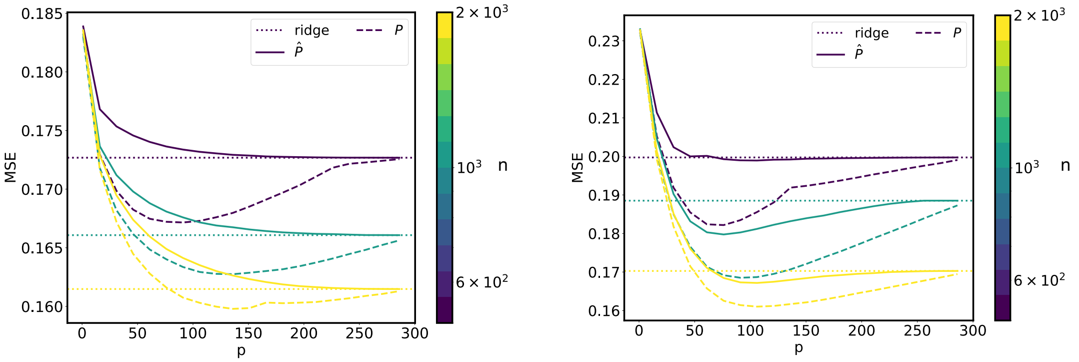

Figure 3: Test MSE w.r.t p (x axis) and the quantity of training data n (color bar), obtained with the optimal projection P and its estimation $\hat{P}$, for various output source condition. (Left) Output source condition $\beta = 1$, $H = H_0$. (Right) Output source condition $\beta = 1/3$, $H = (H_0 C H_0) H_0$.

EBM involves the computation of the decoding phase during the training phase. Kernel Dependency Estimation (KDE) (Weston et al., 2003) shares with IOKR the use of kernels in the input and output space with the following differences: in KDE, Kernal PCA is used to decompose the output feature vectors into p orthogonal directions. Kernel ridge regression is then used for learning independently the mapping between the input feature vectors and each direction. By applying KPCA on the outputs KDE aims at estimating the linear subspace of the output embedding $\psi(y)$ while the proposed reduced-rank estimator aims at estimating the linear subspace of the $h^*(x)$. Additionally, for the multi-label classification problem, we choose the exact setting of previous benchmark experiments (See for instance, (Gygli et al., 2017; Lin et al., 2014)) and thus benefited from the collected results and comparison with other methods.

5.2.1 Image Reconstruction

Problem and data set. The goal of the image reconstruction problem provided by Weston et al. (2003) is to predict the bottom half of a USPS handwritten postal digit (16 x 16 pixels), given its top half. The data set contains 7291 training labeled images and 2007 test images.

Experimental setting. As in Weston et al. (2003) we used as target loss an RBF loss $\|\psi(y) - \psi(y')\|^2_{\mathcal{H}_y}$ induced by a Gaussian kernel k and visually chose the kernel's width $\sigma^2_{output} = 10$, looking at reconstructed images of the method using the ridge estimator (i.e. without reduced-rank estimation). We used a Gaussian input kernel of width σ^2_{input}. For the pre-image step, we used the same candidate set for all methods constituted with all the 7291 training bottom half digits. We considered $\lambda := \lambda_1 = \lambda_2$ for the proposed method. The hyper-parameters for all tested methods (including $\sigma^2_{input}, \lambda, p$, and SPEN layers' sizes) have been selected using logarithmic grids via 5 repeated random sub-sampling validation (80%/20%).

Reduced-rank estimator for surrogate problem. We start by evaluating the performance of the reduced-rank estimator in solving the Hilbert space valued least-squares problem described in Eq. (25). We plot on Figure 4 the test mean squared error of our estimator, and of the ridge estimator, w.r.t the quantity of training data n from $n = 500$ to $n = 7000$. We observe that the reduced-rank estimator ($p < +\infty$) always performs better than the kernel ridge estimator ($p = +\infty$). Nevertheless, we see that this gain is smaller for small n or big n. This is a typical behavior observed in our experiments, which can be interpreted as a difficulty in estimating $\hat{P}$ when n is small, and the diminishing usefulness of regularization when n increase. Indeed p can be thought of as a regularization parameter exploiting a different regularity assumption than λ, but whose action, similarly to λ, should decrease when n increases, such that $p \to +\infty$ when $n \to +\infty$.

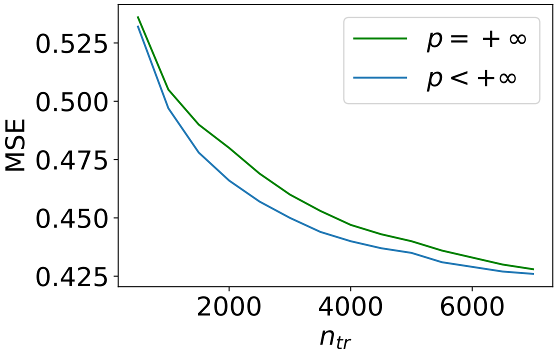

Figure 4: Test MSE of the proposed reduced-rank estimator ($p < +\infty$), and of the ridge estimator ($p = +\infty$) w.r.t n on the USPS problem.

Comparison with SOTA methods. Then, in a second experiment, we compare the structured predictor (see Eq. (29)) using reduced-rank estimation, to state-of-the-art methods: SPEN (Belanger and McCallum, 2016), IOKR (Brouard et al., 2016b), and Kernel Dependency Estimation (KDE) (Weston et al., 2003). We fix $n = 1000$ where the reduced-rank estimation seems helpful, according to Figure 4. For SPEN we employed the standard architecture and training method described in the corresponding article (cf. supplements for more details). We evaluated the results in term of RBF loss (e.g. Gaussian kernel loss), as in Weston et al. (2003). The obtained results are given in Table 3. Firstly, we see that SPEN obtains worse results than KDE, IOKR, and reduced-rank IOKR. Furthermore, note that the number of hyperparameters for SPEN (architecture and optimization) is usually larger than reduced-rank IOKR. Finally, notice that IOKR correspond to the proposed method with $p = +\infty$. Hence, this shows the benefit of exploiting output regularity thanks to reduced-rank estimation in structured prediction.

Method	RBF loss	p
SPEN	0.801 ± 0.011	128
KDE	0.764 ± 0.011	64
IOKR	0.751 ± 0.011	∞
Reduced-rank IOKR	**0.734** ± 0.011	64

Table 3: Test mean losses and standard errors for the proposed method, IOKR, KDE, and SPEN on the USPS digits reconstruction problem where $n = 1000$, and $n_{test} = 2007$.

5.2.2 Multi-label Classification

Problem and data set. Bibtex and Bookmarks (Katakis et al., 2008) are tag recommendation problems, in which the objective is to propose a relevant set of tags (e.g. url, description, journal volume) to users when they add a new Bookmark (webpage) or Bibtex entry to the social bookmarking system Bibsonomy. Corel5k is an image data set and the goal of this application is to annotate these images with keywords. Information on these data sets is given in Table 4.

data set	n	n_{te}	$n_{features}$	n_{labels}	$\bar{l}$
Bibtex	4880	2515	1836	159	2.40
Bookmarks	60000	27856	2150	208	2.03
Corel5k	4500	499	37152	260	3.52

Table 4: Multi-label data sets description. $\bar{l}$ denotes the averaged number of labels per point.

Experimental setting. For all multi-label experiments we used a Gaussian input and output kernels with widths σ^2_{input} and $\sigma^2_{output} = \bar{l}$, where $\bar{l}$ is the averaged number of labels per point. As candidate sets we used all the training output data. We measured the quality of predictions using example-based F1 score. We selected the hyper-parameters λ and p in logarithmic grids.

Comparison with SOTA methods. We compare our method with several multi-label and structured prediction approaches including IOKR (Brouard et al., 2016b), logistic regression (LR) trained independently for each label (Lin et al., 2014), a two-layer neural network with cross entropy loss (NN) by (Belanger and McCallum, 2016), the multi-label approach PRLR (Posterior-Regularized Low-Rank) (Lin et al., 2014), the energy-based model SPEN (Structured Prediction Energy Networks) (Belanger and McCallum, 2016) as well as DVN (Deep Value Networks) (Gygli et al., 2017). The results in Table 5 show that surrogate methods (first two lines) can compete with state-of-the-art dedicated multilabel methods on the standard data sets Bibtex and Bookmarks. With Bookmarks ($n/n_{te} = 60000/27856$) we used a Nyström approximation with 15000 anchors when computing $\hat{h}$

to reduce the training complexity, and we learned $\hat{P}$ only with a subset of 12000 training data. $\hat{h}$ decoding took about 56 minutes, and $\hat{P}\hat{h}$ decoding less than 4 minutes. With a drastically smaller amount of time, $\hat{P}\hat{h}$ (first line) achieves the same order of magnitude of F1 as $\hat{h}$ (line two) at a lower cost (see Table 6) and still has better performance than all other competitors.

Method	Bibtex	Bookmarks
Reduced-rank IOKR	43.8	39.1
IOKR	44.0	**39.3**
LR	37.2	30.7
NN	38.9	33.8
SPEN	42.2	34.4
PRLR	44.2	34.9
DVN	**44.7**	37.1

Table 5: Tag prediction from text data. F_1 score of reduced-rank IOKR compared to state-of-the-art methods. LR (Lin et al., 2014), NN (Belanger and McCallum, 2016), SPEN (Belanger and McCallum, 2016), PRLR (Lin et al., 2014), DVN (Gygli et al., 2017). Results are taken from the corresponding articles.

	IOKR	Reduced-rank IOKR
Bibtex	2s/13s	15s/4s
Bookmarks	465s/3371s	617s/214s
USPS	0.1s/9s	0.4s/1s

Table 6: Fitting/Decoding computation time of IOKR compared to our method (in seconds)

Small training data regime. We evaluate the reduced-rank structured predictor in a setting where only a small number of training examples is known. For this setting, we consider only the 2000 first couples (x_i, y_i) of each multi-label data set as training set. Hyper-parameters have been selected using 5 repeated random sub-sampling validation (80%/20%) and the same λ was used for IOKR. The results of this comparison are given in Table 7. We observe that the proposed reduced-rank structured predictor obtains higher F1 scores than the one using kernel ridge regression in this setup. This highlights the interest of our method in a setting where the data set is small in comparison to the difficulty of the task.

About the selected rank p. We selected the rank p with integer logarithmic scales, ensuring that the selected dimensions were always smaller than the maximal one of the grids. From Table 7 to Table 5, the selected dimension p for Bibtex/Bookmarks are 80/30, then 130/200. In Table 7 recall that we used a reduced number of training couples. Interpreting

	Bibtex	Bookmarks	Corel5k
n	2000	2000	2000
n_{te}	2515	2500	499
IOKR	35.9	22.9	13.7
Reduced-rank IOKR	**39.7**	**25.9**	**16.1**

Table 7: Test F_1 score of reduced-rank IOKR and IOKR on different multi-label problems in a small training data regime.

p as a regularisation parameter, we see that when n increases then the p increases, i.e. the rank regularisation decreases.

5.2.3 Metabolite Identification

Problem and data set. An important problem in metabolomics is to identify the small molecules, called metabolites, that are present in a biological sample. Mass spectrometry is a widespread method to extract distinctive features from a biological sample in the form of a tandem mass (MS/MS) spectrum. The goal of this problem is to predict the molecular structure of a metabolite given its tandem mass spectrum. The molecular structures of the metabolites are represented by fingerprints, that are binary vectors of length $d = 7593$. Each value of the fingerprint indicates the presence or absence of a certain molecular property. Labeled data are expensive to obtain, and despite the problem complexity only $n = 6974$ labeled data are available. State-of-the-art results for this problem have been obtained with the IOKR method by Brouard et al. (2016a). The median size of the candidate sets is 292, and the biggest candidate set is of size 36918. Hence, the metabolite identification data set is characterized by high-dimensional complex outputs, a small training set, and a very large number of candidates.

Experimental setting. We adopt a similar numerical experimental protocol (5-CV Outer/4-CV Inner loops) than in Brouard et al. (2016a), probability product input kernel for mass spectra, and Gaussian-Tanimoto output kernel on the molecular fingerprints (with parameter $\sigma^2 = 1$). We selected the hyper-parameters λ, p in logarithmic grids using nested cross-validation with 5 outer folds and 4 inner folds.

Improved prediction with reduced-rank estimation . We compare the proposed reduced-rank structured predictor with SPEN, and with the state-of-the art method on this problem IOKR (which corresponds to our method with $p = +\infty$). The result are given in Table 8. We observe that reduced-rank IOKR improved upon plain IOKR, in this context of supervised learning with complex outputs and a small training data set.

6. Conclusion

In this paper, we proposed a novel reduced-rank regression estimator in the case of regularized least squares regression with infinite dimensional outputs and gave excess-risk bounds

Method	MSE	Tanimoto-Gaussian loss	Top-k accuracies $k=1 \mid k=5 \mid k=10$
SPEN	–	0.537 ± 0.008	25.9% | 54.1% | 64.3%
IOKR	0.781 ± 0.002	0.463 ± 0.009	29.6% | 61.1% | 71.0%
Reduced-rank IOKR	$\mathbf{0.766 \pm 0.003}$	$\mathbf{0.459 \pm 0.010}$	**30.0%** | **61.5%** | **71.4%**

Table 8: Test mean losses and standard errors for the metabolite identification problem. SPEN MSE in $\mathcal{H}_z$ is not defined as predictions are directly done in $\mathcal{Z}$.

under general output regularity assumptions. In particular, we characterized a family of situations where reduced-rank regression is statistically and computationally beneficial. We used the proposed reduced-rank regression for structured prediction, and derived theoretical guarantees on the resulting estimator. Experiments on structured prediction problems confirm the advantages in practice of the approach.

Acknowledgments

The first and last authors are funded by the French National Research Agency (ANR) through ANR-18-CE23-0014 APi (Apprivoiser la Pré-image) and the Télécom Paris Chair DSAIDIS. This work was supported by the French government under the management of the Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). We also acknowledge support from the European Research Council (grants REAL 947908). This research was partially funded by Academy of Finland (grants 334790, 339421, and 345802).

Appendix A. Proofs of the Learning Bounds

In this section we prove Theorem 1 and Corollary 2. The proofs are organized as follows:

- Appendix A.1 introduces some necessary notations and definitions.
- Appendix A.2 provides the proof for bounding $\mathbb{E}[\|\hat{P}(\hat{h}(x) - h^*(x))\|^2_{\mathcal{Y}}]$ (Lemma 7).
- Appendix A.3 provides the proof for bounding $\mathbb{E}[\|\hat{P}h^*(x) - h^*(x)\|^2_{\mathcal{Y}}]$ (Lemma 11).
- Appendix A.4 provides the proof for bounding $\mathbb{E}[\|\hat{P}\hat{h}(x) - h^*(x)\|^2_{\mathcal{Y}}]$ (Theorem 1) using Lemmas 7 and 11.
- Appendix A.5 provides the proof for the Corollary 2 using Theorem 1.
- Appendix A.6 gives some technical results used in the proofs.
- Appendix A.7 discusses the assumption that $\phi(x)$ and ϵ are independent.

A.1 Notations and Definitions

In the following we consider $\mathcal{X}$ to be a Polish space, and $\mathcal{Y}$ a separable Hilbert space. We define here the ideal operators that we will use in the following

- The feature map $\phi : \mathcal{X} \to \mathcal{H}_x$, $\forall x \in \mathcal{X}$, $\phi(x) = k(x, .)$, with $\|\phi(x)\|_{\mathcal{H}_x} \leq \kappa$ with $\kappa > 0$.
- The target $h^*(.) \in \mathcal{H} = \mathbb{E}_{y|.}(y)$, and $Q > 0$ such that $\forall y \in \mathcal{Y}$, $\|y\|_{\mathcal{Y}} \leq Q$.
- $S : f \in \mathcal{H}_x \to \langle f, \phi(.)\rangle_{\mathcal{H}_x} \in L^2(\mathcal{X}, \rho_{\mathcal{X}})$
- $Z : y \in \mathcal{Y} \to \langle y, h^*(.)\rangle_{\mathcal{Y}} \in L^2(\mathcal{X}, \rho_{\mathcal{X}})$

and their empirical counterparts

- The KRR estimator $\hat{h}(.) \in \mathcal{H}$ trained with n couples $(x_i, y_i)_{i=1}^n$
- $S_n : f \in \mathcal{H}_x \to \frac{1}{\sqrt{n}}(\langle f, \phi(x_i)\rangle_{\mathcal{H}_x})_{1\leq i\leq n} \in \mathbb{R}^n$
- $Z_n : y \in \mathcal{Y} \to \frac{1}{\sqrt{n}}(\langle y, y_i\rangle_{\mathcal{Y}}))_{1\leq i\leq n} \in \mathbb{R}^n$

From there, we can define the following covariance operators

- $C = \mathbb{E}_x[\phi(x) \otimes \phi(x)] = S^*S$
- $V = \mathbb{E}_y[y \otimes y]$
- $M = \mathbb{E}_x[h^*(x) \otimes h^*(x)]$
- $Z^*S = \mathbb{E}_{x,y}[y \otimes \phi(x)]$

and their empirical counterparts

- $C_n = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) \otimes \phi(x_i)$

- $V_n = \frac{1}{n}\sum\limits_{i=1}^{n} y_i \otimes y_i$
- $M_n = \frac{1}{n}\sum\limits_{i=1}^{n} \hat{h}(x_i) \otimes \hat{h}(x_i)$
- $Z_n^* S_n = \frac{1}{n}\sum\limits_{i=1}^{n} y_i \otimes \phi(x_i)$

From Lemmas 16 and 17 in Ciliberto et al. (2016) we recall that we have

- $h^*(.) = H\phi(.)$ with $H = Z^*SC^\dagger \in \mathcal{Y} \otimes \mathcal{H}_x$
- $\hat{h}(.) = H_n\phi(.)$ with $H_n = Z_n^* S_n(C_n + \lambda I)^{-1} \in \mathcal{Y} \otimes \mathcal{H}_x$
- $M = HCH^*$
- $M_n = H_n C_n H_n^*$

A.2 KRR Error on a Subspace

In this subsection we prove a bound on the kernel ridge regression error on the subspace defined by $\hat{P}$:

$$\mathbb{E}_x[\|\hat{P}\hat{h}(x) - \hat{P}h^*(x)\|_{\mathcal{Y}}^2]^{1/2} = \|\hat{P}(H_n - H)S^*\|_{\mathrm{HS}}. \tag{32}$$

Equation (32) is obtained by definition of the operators H_n, H, S (see e.g. Ciliberto et al. (2016)).

In order to bound (32), one can not directly apply standard learning bounds for kernel ridge estimator on the learning problem $(x, \hat{P}y)$ with $(x, y) \sim \rho$, as $\hat{P}$ depends on the training data. That is why we will decompose (32) as

$$\|\hat{P}(H_n - H)S^*\|_{\mathrm{HS}} \leq \|\hat{P}(A + tI)^{1/2}\|_{\mathrm{HS}} \times \|(A + tI)^{-1/2}(H_n - H)S^*\|_\infty \tag{33}$$

with a well chosen operator $A : \mathcal{Y} \to \mathcal{Y}$.

As a first step, we give a bound on the KRR estimator excess-risk with respect to the operator norm.

Lemma 6 (Bound $\|(H_n - H)S^*\|_\infty$) *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded kernel with $\forall x \in \mathcal{X}, k(x, x) \leq \kappa^2$. Let ρ be a distribution on $\mathcal{X} \times \mathcal{Y}$ such that its marginal w.r.t y is supported on the ball $\|y\|_{\mathcal{Y}} \leq Q$. Let $\hat{h} = H_n\phi(.)$ be the KRR estimator trained with n independent couples drawn from ρ, and regularization parameter $\lambda_2 > \frac{9\kappa^2}{n}\log(\frac{n}{\delta})$. Let $\delta \in [0, 1]$. Then, under Assumption 1, $H_nS^* - HS^* = A_1 + A_2$, with*

$$A_1 := Z_n^* S_n (C_n + \lambda_2 I)^{-1} S^* - HC_n(C_n + \lambda_2 I)^{-1} S^* \tag{34}$$

$$A_2 := HC_n(C_n + \lambda_2 I)^{-1} S^* - HS^* \tag{35}$$

and with probability at least $1 - 2\delta$

$$\|A_1\|_\infty \leq \sqrt{\frac{24\eta(Q^2 + \|E\|_\infty \lambda_2^{-1}\kappa^2)}{n}} + \frac{8\kappa Q\eta}{3\sqrt{\lambda_2}n}; \qquad \|A_2\|_\infty \leq \sqrt{2}\sqrt{\lambda_2}\|H\|_\infty \tag{36}$$

with $\eta = \log(\frac{4(\frac{2\,\mathrm{Tr}(C)}{\lambda_2} + \frac{\mathrm{Tr}(E)}{\|E\|_\infty})}{\delta})$, $E = \mathbb{E}[\epsilon \otimes \epsilon]$, $\epsilon = y - h^*(x)$, $R = \|H\|_{\mathrm{HS}}$.

Proof

Decomposition. The decomposition $H_nS^* - HS^* = A_1 + A_2$ is obtained noticing that we have $H_n = Z_n^*S_n(C_n + \lambda_2 I)^{-1}$ (See section A.1).

1. Bound $\|A_1\|_\infty$. We have

$$\|A_1\|_\infty \leq \|(Z_n^*S_n - HC_n)(C + \lambda_2 I)^{-1/2}\|_\infty \times \|(C + \lambda_2 I)^{1/2}(C_n + \lambda_2 I)^{-1}S^*\|_\infty \tag{37}$$

1.1 Bound $\|(Z_n^*S_n - HC_n)(C + \lambda_2 I)^{-1/2}\|_\infty$. We define

$$\xi_i = \epsilon_i \otimes \phi(x_i)(C + \lambda_2 I)^{-1/2} \tag{38}$$

with $\epsilon_i = y_i - h^*(x_i)$. In this way,

$$\|(Z_n^*S_n - HC_n)(C + \lambda_2 I)^{-1/2}\|_\infty = \|\frac{1}{n}\sum_{i=1}^{n}\xi_i\|_\infty \tag{39}$$

We aim at applying the Bernstein inequality given in Theorem 14 to the random linear operator ξ. So, we define

$$T = 2\kappa Q\lambda_2^{-1/2} \geq \|\xi_i\|_\infty, \tag{40}$$

$$\sigma^2 = \max(\|\mathbb{E}[\xi\xi^*]\|_\infty, \|\mathbb{E}[\xi^*\xi]\|_\infty), \tag{41}$$

$$d = \operatorname{Tr}(\mathbb{E}[\xi^*\xi] + \mathbb{E}[\xi\xi^*])/\sigma^2. \tag{42}$$

Note that $\|\epsilon\| \leq \|y\|_\mathcal{Y} + \|h^*(x)\|_\mathcal{Y} \leq 2Q$, and $\|\phi(x)\| \leq \kappa$. Then, we have

$$\|\mathbb{E}[\xi\xi^*]\|_\infty = \|\mathbb{E}[\epsilon_i \otimes \epsilon_i \times \langle\phi(x_i), (C + \lambda_2 I)^{-1}\phi(x_i)\rangle_{\mathcal{H}_x}]\|_\infty \tag{43}$$

$$\leq \|\mathbb{E}[\epsilon \otimes \epsilon]\|_\infty \times \frac{\kappa^2}{\lambda_2}. \tag{44}$$

and

$$\|\mathbb{E}[\xi^*\xi]\|_\infty = \|(C + \lambda_2 I)^{-1/2}C(C + \lambda_2 I)^{-1/2}\|_\infty \times \mathbb{E}[\|\epsilon\|_\mathcal{Y}^2] \tag{45}$$

$$\leq 4Q^2. \tag{46}$$

Moreover, if $\lambda_2 < \|C\|_\infty$,

$$d \leq \frac{\operatorname{Tr}(\mathbb{E}[\xi^*\xi])}{\|\mathbb{E}[\xi^*\xi]\|_\infty} + \frac{\operatorname{Tr}(\mathbb{E}[\xi\xi^*])}{\|\mathbb{E}[\xi\xi^*]\|_\infty} \leq \frac{2\operatorname{Tr}(C)}{\lambda_2} + \frac{\operatorname{Tr}(E)}{\|E\|_\infty}. \tag{47}$$

Thus, by applying the Bernstein inequality given in Theorem 14, we have

$$\|(Z_n^*S_n - HC_n)(C + \lambda_2 I)^{-1/2}\|_\infty \leq \sqrt{\frac{2\eta(4Q^2 + \|E\|_\infty\kappa^2\lambda_2^{-1})}{n}} + \frac{4\kappa Q\lambda_2^{-1/2}\eta}{3n} \tag{48}$$

where $\eta = \log(\frac{4(\frac{2\operatorname{Tr}(C)}{\lambda_2} + \frac{\operatorname{Tr}(E)}{\|E\|_\infty})}{\delta})$, $E = \mathbb{E}[\epsilon \otimes \epsilon]$.

1.2 Bound $\|(C+\lambda_2 I)^{1/2}(C_n+\lambda_2 I)^{-1}S^*\|_\infty$. We apply Lemma B.6 in Ciliberto et al. (2020), with $\lambda_2 \geq \frac{9\kappa^2}{n}\log(\frac{n}{\delta})$, and get with probability at least $1-\delta$,

$$\|(C+\lambda_2 I)^{1/2}(C_n+\lambda_2 I)^{-1}S^*\|_\infty \leq \|(C_n+\lambda_2 I)^{-1/2}(C+\lambda_2)^{1/2}\|_\infty^2 \leq 2. \tag{49}$$

Finally, we have

$$\|A_1\|_\infty \leq \sqrt{\frac{24\eta(Q^2+\|E\|_\infty\kappa^2\lambda_2^{-1})}{n}} + \frac{8\kappa Q\lambda_2^{-1/2}\eta}{3n}. \tag{50}$$

Bound $\|A_2\|_\infty$. We have

$$\begin{aligned}\|A_2\|_\infty &= \|H(C_n(C_n+\lambda_2 I)^{-1}-I)S^*\|_\infty & (51)\\ &= \|H(-\lambda_2(C_n+\lambda_2 I)^{-1})S^*\|_\infty & (52)\\ &\leq \lambda_2\|H\|_\infty\|(C_n+\lambda_2 I)^{-1}S^*\|_\infty & (53)\end{aligned}$$

and

$$\begin{aligned}\|(C_n+\lambda_2 I)^{-1}S^*\|_\infty &\leq \lambda_2^{-1/2}\|(C_n+\lambda_2 I)^{-1/2}S^*\|_\infty & (54)\\ &= \lambda_2^{-1/2}\|(C_n+\lambda_2 I)^{-1/2}C^{1/2}\|_\infty & (55)\\ &\leq \lambda_2^{-1/2}\|(C_n+\lambda_2 I)^{-1/2}(C+\lambda_2)^{1/2}\|_\infty & (56)\\ &\leq \sqrt{2}\lambda_2^{-1/2} & (57)\end{aligned}$$

because $\|(C_n+\lambda_2 I)^{-1/2}(C+\lambda_2)^{1/2}\|_\infty^2 \leq 2$ from Equation (49).

Finally, we have

$$\|A_2\|_\infty = \sqrt{2}\sqrt{\lambda_2}\|H\|_\infty. \tag{58}$$

Conclusion. The bound on $\|(H_n-H)S^*\|_\infty$ is obtained by summing the two bounds on $\|A1\|_\infty$ and $\|A2\|_\infty$. ■

We are now ready to prove a bound on the excess-risk of the ridge estimator on the random subspace defined by $\hat{P}$, namely $\|\hat{P}(H_n-H)S^*\|_{\mathrm{HS}}$.

Lemma 7 (KRR excess-risk on a subspace) *Let $k : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be a bounded kernel with $\forall x\in\mathcal{X}, k(x,x)\leq\kappa^2$. Let ρ be a distribution on $\mathcal{X}\times\mathcal{Y}$ such that its marginal w.r.t y is supported on the ball $\|y\|_{\mathcal{Y}}\leq Q$. Let $\hat{h}$ be the KRR estimator trained with n independent couples drawn from ρ. Let $\delta\in[0,1]$. Define $S_p(E)=\sum_{i=1}^p \mu_i(E)$. Then, under the*

Assumptions 1, 3, 4, taking for n big enough $\lambda_2 = \max(S_p(E)^{1/2}n^{-1/2}, n^{-1}, \frac{9\kappa^2}{n}\log(\frac{n}{\delta}))$, *then with probability at least* $1-2\delta$

$$\mathbb{E}_x[\|\hat{P}\hat{h}(x) - \hat{P}h^*(x)\|_{\mathcal{Y}}^2]^{1/2} \leq \Big(c_4\sqrt{p}n^{-1/4} + c_5 S_p(E)^{1/4}\Big)n^{-1/4}\log(n/\delta)$$

with $c_4 = (7Q + 4\kappa Q + 2\|H\|_{\mathrm{HS}}(1+3\kappa))(1+c_6)$, $c_5 = 10\sqrt{(1+c_6)}\kappa\|E\|_\infty^{1/2} + 2\|H\|_{\mathrm{HS}}$, $c_6 = \log(8(\frac{\mathrm{Tr}(C)}{\|E\|_\infty^{1/2}} + \frac{\mathrm{Tr}(E)}{\|E\|_\infty}))$.

Proof

Decomposition. We decompose $\|\hat{P}(H_n - H)S^*\|_{\mathrm{HS}}$ as follows

$$\|\hat{P}(H_n - H)S^*\|_{\mathrm{HS}} \leq \|\hat{P}A_1\|_{\mathrm{HS}} + \|\hat{P}A_2\|_{\mathrm{HS}} \tag{59}$$

with A_1, A_2 defined above in Lemma 6. Then, let be $t_1, t_2 > 0$,

$$\begin{aligned}\|\hat{P}A_1\|_{\mathrm{HS}} &\leq \|\hat{P}(E+t_1I)^{1/2}\|_{\mathrm{HS}} \times \|(E+t_1I)^{-1/2}A_1\|_\infty &(60)\\ &= \mathrm{Tr}(\hat{P}(E+t_1I))^{1/2} \times \|(E+t_1I)^{-1/2}A_1\|_\infty &(61)\\ &\leq \sqrt{S_p(E) + pt_1} \times \|(E+t_1I)^{-1/2}A_1\|_\infty. &(62)\end{aligned}$$

and similarly

$$\|\hat{P}A_2\|_{\mathrm{HS}} \leq \sqrt{S_p(HH^*) + pt_2} \times \|(HH^* + t_2I)^{-1/2}A_2\|_\infty. \tag{63}$$

Sketch of the following proof. We are going to bound $\|(E+t_1I)^{-1/2}A_1\|_\infty$ and $\|(HH^*+t_2I)^{-1/2}A_2\|_\infty$, using the Lemma 6 two times. This is done noticing that $\|(E+t_1I)^{-1/2}A_1\|_\infty$ is exactly the error "part A_1" of the KRR estimator trained with data $(x_i, (E+t_1I)^{-1/2}y)_{i=1}^n$, trying to solve the least-squares problem : $(E+t_1I)^{-1/2}y = (E+t_1I)^{-1/2}H\phi(x) + (E+t_1I)^{-1/2}\epsilon$. The same trick is used for $\|(HH^*+t_2I)^{-1/2}A_2\|_\infty$. In the two cases, we compute then the resulting modified constants in the bound because of these left linear operators multiplications.

1. Bound $\|(E+t_1I)^{-1/2}A_1\|_\infty$. We apply Lemma 6 on the KRR estimator trained with $(x_i, (E+t)^{-1/2}y_i)$.

We have

$$\begin{aligned}&\|(E+t_1I)^{-1/2}E(E+t_1I)^{-1/2}\|_\infty \leq 1 &(64)\\ &\|(E+t_1I)^{-1/2}y\| \leq t_1^{-1/2}Q &(65)\\ &\|(E+t_1I)^{-1/2}H\|_{\mathrm{HS}} \leq t_1^{-1/2}\|H\|_{\mathrm{HS}}. &(66)\end{aligned}$$

Furthermore, if $\|E\|_\infty \geq t_1$, we have

$$\begin{aligned}\frac{\mathrm{Tr}(E(E+t_1)^{-1})}{\|E(E+t_1)^{-1}\|_\infty} &= \mathrm{Tr}(E(E+t_1)^{-1})\frac{\|E\|_\infty + t}{\|E\|_\infty} &(67)\\ &\leq 2\,\mathrm{Tr}(E(E+t_1)^{-1}) &(68)\\ &\leq 2\,\mathrm{Tr}(E)t_1^{-1}. &(69)\end{aligned}$$

Thus we get with probability at least $1-2\delta$

$$\|(E+t_1)^{-1/2}A_1\|_\infty \leq \sqrt{\frac{24\eta(Q^2t_1^{-1}+\lambda_2^{-1}\kappa^2\|E\|_\infty)}{n}} + \frac{8\kappa Q t_1^{-1/2}\eta}{3\sqrt{\lambda_2}n}. \tag{70}$$

with $\eta = \log(\frac{8(\frac{\mathrm{Tr}(C)}{\lambda_2}+\frac{\mathrm{Tr}(E)}{t_1})}{\delta})$, $E=\mathbb{E}[\epsilon\otimes\epsilon]$, $\epsilon = y-h^*(x)$.

2. Bound $\|(HH^*+t_2I)^{-1/2}A_2\|_\infty$. We apply Lemma 6 on the KRR estimator trained with $(x_i,(HH^*+t_2)^{-1/2}y_i)$. We have

$$\|(HH^*+t_2I)^{-1/2}H\|_\infty = \|(HH^*+t_2I)^{-1}HH^*\|_\infty^{1/2} \leq 1. \tag{71}$$

So,

$$\|(HH^*+t_2I)^{-1/2}A_2\|_\infty \leq \sqrt{2}\sqrt{\lambda_2} \tag{72}$$

Conclusion. We conclude by summing the bound. We have

$$\begin{aligned}\|\hat{P}(H_n-H)S^*\|_{\mathrm{HS}} \leq{}& \sqrt{S_p(E)+pt_1} \times \left(\sqrt{\frac{24\eta(Q^2t_1^{-1}+\lambda_2^{-1}\kappa^2\|E\|_\infty)}{n}} + \frac{8\kappa Q t_1^{-1/2}\eta}{3\sqrt{\lambda_2}n}\right)\\ &+ \sqrt{S_p(HH^*)+pt_2} \times \left(\sqrt{\lambda_2}\sqrt{2}\right).\end{aligned}$$

Taking $t_1 = p^{-1}S_p(E) \leq \|E\|_\infty$, and $t_2 = p^{-1}S_p(HH^*)$, we get

$$\begin{aligned}\|\hat{P}(H_n-H)S^*\|_{\mathrm{HS}} \leq{}& \sqrt{\frac{48\eta(Q^2p+2S_p(E)\lambda_2^{-1}\kappa^2\|E\|_\infty)}{n}} + \frac{4\kappa Q\sqrt{p}\eta}{\sqrt{\lambda_2}n}\\ &+ 2\sqrt{S_p(HH^*)}\sqrt{\lambda_2}.\end{aligned}$$

Now, taking $\lambda_2 = \max(S_p(E)^{1/2}n^{-1/2}, n^{-1}, \frac{9\kappa^2}{n}\log(\frac{n}{\delta}))$, we get

$$\begin{aligned}\|\hat{P}(H_n-H)S^*\|_{\mathrm{HS}} \leq{}& 7\sqrt{\frac{\eta Q^2 p}{n}} + 7\sqrt{\frac{2\eta S_p(E)\lambda_2^{-1}\kappa^2\|E\|_\infty}{n}} + \frac{4\kappa Q\sqrt{p}\eta}{\sqrt{\lambda_2}n}\\ &+ 2\|H\|_{\mathrm{HS}}\sqrt{\lambda_2}\\ \leq{}& 7\sqrt{\frac{\eta Q^2 p}{n}} + 7\sqrt{\frac{2\eta S_p(E)^{1/2}\kappa^2\|E\|_\infty}{n^{1/2}}} + \frac{4\kappa Q\sqrt{p}\eta}{n^{1/2}}\\ &+ 2\|H\|_{\mathrm{HS}}\left(S_p(E)^{1/4}n^{-1/4} + n^{-1/2} + 3\kappa n^{-1/2}\log^{1/2}(\frac{n}{\delta})\right)\\ \leq{}& \Big[\left(7\sqrt{\eta}Q + 4\kappa Q\eta + 2\|H\|_{\mathrm{HS}}(1+3\kappa\log(\frac{n}{\delta}))\right)\sqrt{p}n^{-1/4}\\ &+ \left(10\sqrt{\eta}\kappa\|E\|_\infty^{1/2} + 2\|H\|_{\mathrm{HS}}\right)S_p(E)^{1/4}\Big]n^{-1/4}\\ \leq{}& \left(c_4\sqrt{p}n^{-1/4} + c_5 S_p(E)^{1/4}\right)n^{-1/4}\log(n/\delta)\end{aligned}$$

with $c_4 = (7Q + 4\kappa Q + 2\|H\|_{\mathrm{HS}}(1+3\kappa))(1+c_6)$, $c_5 = 10\sqrt{(1+c_6)}\kappa\|E\|_\infty^{1/2} + 2\|H\|_{\mathrm{HS}}$, $c_6 = \log(8(\frac{\mathrm{Tr}(C)}{\|E\|_\infty^{1/2}} + \frac{\mathrm{Tr}(E)}{\|E\|_\infty}))$, as $\eta \leq c_6 + \log(n/\delta) \leq (c_6+1)\log(n/\delta)$ if $p \leq n$.

■

A.3 Supervised Subspace Learning

In this subsection we prove a bound on the supervised reconstruction error:

$$\mathbb{E}_x[\|\hat{P}h^*(x) - h^*(x)\|_{\mathcal{Y}}^2]^{1/2} = \|(\hat{P} - I)M^{1/2}\|_{\mathrm{HS}}. \tag{73}$$

We use the proof scheme of Rudi et al. (2013) for subspace learning, retaking also the Lemma 8 restated just below. The novelty to deal with is that the random variable, whose reconstruction error is minimized here, is $h^*(x)$. The unknown $h^*(x_i)$ are estimated via our supervised subspace learning method (7) thanks to the couples $(x_i, y_i)_{i=1}^n$. This leads to additional derivations in our proofs.

We start by restating the Lemma 3.6 from Rudi et al. (2013) in a convenient form for our purposes.

Lemma 8 (Convergence of covariance operators) *Let $\mathcal{X}, \mathcal{Y}$ be two Hilbert spaces, $H \in \mathcal{Y} \otimes \mathcal{X}$, $A = \mathbb{E}_x[Hx \otimes Hx]$, $(x_i)_{i=1}^n$ i.i.d from a distribution ρ on $\mathcal{X}$ supported on the unit ball, $A_n = \frac{1}{n}\sum_{i=1}^n Hx_i \otimes Hx_i$, $B \in \mathcal{Y} \otimes \mathcal{Y}$ any positive semidefinite operator, $\frac{9}{n}\log(\frac{n}{\delta}) \leq t \leq \|A\|_\infty$, then with probability at least $1-\delta$ it is*

$$\sqrt{\frac{2}{3}} \leq \|(A + B + tI)^{\frac{1}{2}}(A_n + B + tI)^{-\frac{1}{2}}\|_\infty \leq \sqrt{2}$$

Proof By defining $B_n = (A + B + tI)^{-\frac{1}{2}}(A - A_n)(A + B + tI)^{-\frac{1}{2}}$, we have

$$\|(A + B + tI)^{\frac{1}{2}}(A_n + B + tI)^{-\frac{1}{2}}\|_\infty = \|(I - B_n)^{-1}\|_\infty^{1/2} \tag{74}$$

and B is positive semidefinite so

$$\|B_n\|_\infty = \|(A + B + tI)^{-\frac{1}{2}}(A - A_n)(A + B + tI)^{-\frac{1}{2}}\|_\infty \tag{75}$$

$$\leq \|(A + tI)^{-\frac{1}{2}}(A - A_n)(A + tI)^{-\frac{1}{2}}\|_\infty \tag{76}$$

Now, by applying Lemma 3.6 from Rudi et al. (2013), we get with probability at least $1-\delta$, if $\frac{9}{n}\log(\frac{n}{\delta}) \leq t \leq \|A\|_\infty$

$$\|B_n\|_\infty \leq \frac{1}{2}. \tag{77}$$

We conclude by observing that

$$\frac{1}{\sqrt{1 + \|B_n\|_\infty}} \leq \|(I - B_n)^{-1}\|_\infty^{1/2} \leq \frac{1}{\sqrt{1 - \|B_n\|_\infty}}. \tag{78}$$

■

The two following lemmas handle the estimation of $M = HCH^* = \mathbb{E}[h^*(x) \otimes h^*(x)]$ in our supervised subspace learning method. In particular, here is exploited the Assumption 3, whose the divergence rate of the plateau threshold p_{max}, from which the error remains constant (See Rudi et al. (2013)), depends on.

Lemma 9 *Let be* $\xi > 0, \delta \in [0,1]$. *Under Assumptions 1, 3, taking*

$$t \geq n^{-\frac{1}{\beta+1}} (\xi/2)^{-\frac{4}{\beta+1}} \left(4\kappa(Q+\kappa R)\left(1 + 2\kappa\|M\|_{\infty}^{\frac{1}{4}(1-\beta)}\right)\log^2\frac{8}{\delta} + c_2^{1/2}\right)^{\frac{4}{\beta+1}} \tag{79}$$

and

$$\lambda_1 = t^{-\frac{1-\beta}{2}} n^{-1/2}$$

is enough to achieve with probability at least $1-\delta$

$$\|(HCH^* + tI)^{-\frac{1}{2}}(H_n - H)S^*\|_\infty \leq \xi. \tag{80}$$

Proof We note for convenience $A = (HCH^* + tI)^{-\frac{1}{2}}$. We proceed as in the proof of Lemma 18 and Theorem 5 in (Ciliberto et al., 2016) (showing a learning bound for the kernel ridge estimator). However, we monitor the action of A, and we use Assumption 3, in order to obtain the best bound w.r.t t and n, decreasing fast when n and t increase. We have

$$\begin{aligned} \|A(H_n - H)S^*\|_\infty &= \|AZ_n^* S_n (C_n + \lambda_1)^{-1} S^* - AZ^*\|_\infty & (81)\\ &\leq \text{(I)} + \text{(II)} + \text{(III)} & (82) \end{aligned}$$

with

$$\begin{aligned} \text{(I)} &= \|AZ_n^* S_n (C_n + \lambda_1)^{-1} - AZ^* S (C_n + \lambda_1)^{-1} S^*\|_\infty \\ &\leq \sqrt{\frac{1}{t}} \times \|Z_n^* S_n (C_n + \lambda_1)^{-1} - Z^* S (C_n + \lambda_1)^{-1} S^*\|_{\text{HS}} \\ \text{(II)} &= \|AZ^* S (C_n + \lambda_1)^{-1} S^* - AZ^* S (C + \lambda_1)^{-1} S^*\|_\infty \\ &\leq \sqrt{\frac{1}{t}} \times \|Z^* S (C_n + \lambda_1)^{-1} S^* - Z^* S (C + \lambda_1)^{-1} S^*\|_{\text{HS}} \\ \text{(III)} &= \|AZ^* S (C + \lambda_1)^{-1} S^* - AZ^*\|_\infty \end{aligned}$$

Bound (III). From Assumption 1 we have $Z^* = HS^*$, and

$$\begin{aligned} \text{(III)} &= \|AZ^*(S(C + \lambda_1)^{-1} S^* - I)\|_\infty & (83)\\ &= \|AHS^*(S(C + \lambda_1)^{-1} S^* - I)\|_\infty & (84)\\ &= \|AH(S^* - \lambda_1 (C + \lambda_1)^{-1} - S^*)\|_\infty & (85)\\ &= \lambda_1 \|AH(C + \lambda_1)^{-1} S^*\|_\infty & (86)\\ &\leq \|AH\|_\infty \times \lambda_1 \|(C + \lambda_1)^{-1} S^*\|_\infty & (87)\\ &\leq \|AH\|_\infty \times \sqrt{\lambda_1}. & (88) \end{aligned}$$

Using Assumption 3 we have

$$\begin{aligned}
\|AH\|_\infty &= \|(M+tI)^{-\frac{1}{2}}H\|_\infty && (89)\\
&\leq \|(HCH^*+tI)^{-\frac{1}{2}}c_2^{1/2}M^{(1-\beta)/2}\|_\infty && (90)\\
&\leq c_2^{1/2} \times t^{-\frac{\beta}{2}}. && (91)
\end{aligned}$$

Bound (I) and (II). We bound (I) and (II), as in Ciliberto et al. (2016) (Lemma 18).

Conclusion. This leads to the following bound with probability at least $1-\delta$:

$$\|A(H_n-H)S^*\|_\infty \leq 4\kappa\frac{Q+\kappa R}{\sqrt{\lambda_1 n t}}\left(1+\sqrt{\frac{4\kappa^2}{\lambda_1\sqrt{n}}}\right)\log^2\frac{8}{\delta}+c_2^{1/2}\sqrt{\lambda_1}t^{-\frac{\beta}{2}}. \tag{92}$$

Now, choosing $\lambda_1 = \frac{t^{-\frac{1}{2}(1-\beta)}}{\sqrt{n}}$, if $t \leq \|M\|_\infty$, we obtain

$$\begin{aligned}
\|A(H_n-H)S^*\|_\infty &\leq \left(4\kappa(Q+\kappa R)\left(1+2\kappa t^{\frac{1}{4}(1-\beta)}\right)\log^2\frac{8}{\delta}+c_2^{1/2}\right)n^{-1/4}t^{-\frac{1}{4}(\beta+1)} && (93)\\
&\leq \left(4\kappa(Q+\kappa R)\left(1+2\kappa\|M\|_\infty^{\frac{1}{4}(1-\beta)}\right)\log^2\frac{8}{\delta}+c_2^{1/2}\right)n^{-1/4}t^{-\frac{1}{4}(\beta+1)} && (94)
\end{aligned}$$

Hence, taking $t \geq n^{-\frac{1}{\beta+1}}(\xi/2)^{-\frac{4}{\beta+1}}\left(4\kappa(Q+\kappa R)\left(1+2\kappa\|M\|_\infty^{\frac{1}{4}(1-\beta)}\right)\log^2\frac{8}{\delta}+c_2^{1/2}\right)^{\frac{4}{\beta+1}}$ is enough to achieve

$$\|A(H_n-H)S^*\|_\infty \leq \xi. \tag{95}$$

■

We combine Lemmas 8 and 9 to finally prove a concentration bound for $H_nC_nH_n^*$ deviating from HCH^*.

Lemma 10 (Convergence of the supervised covariance M_n) *Let be $\delta \in [0,1]$. Under Assumptions 1, 3, and defining*

$$B_n = (HCH^*+tI)^{-\frac{1}{2}}(H_nC_nH_n^*-HCH^*)(HCH^*+tI)^{-\frac{1}{2}}$$

if $t \geq c_8\log^8(\frac{8}{\delta})n^{-\frac{1}{\beta+1}}$, $n \geq n_0$ (constant independent of δ), then with probability $1-2\delta$

$$\|B_n\|_\infty \leq \frac{1}{2}$$

with $c_8 = (\xi/2)^{-\frac{4}{\beta+1}}\left(4\kappa(Q+\kappa R)\left(1+2\kappa\|M\|_\infty^{\frac{1}{4}(1-\beta)}\right)+c_2^{1/2}\right)^{\frac{4}{\beta+1}}$ and $\xi = \frac{1}{14}$, $n_0 \in \mathbb{N}^*$ constant defined in the proof.

Proof We decompose in 7 terms the difference of products, then we will bound each associated term in $\|B_n\|_\infty$.

$$\begin{aligned} H_nC_nH_n^* - HC_nH^* &= (H_n - H)CH^* \quad (i) \\ &+ HC(H_n - H)^* \quad (ii) \\ &+ (H_n - H)C(H_n - H)^* \quad (iii) \\ &+ (H_n - H)(C_n - C)H^* \quad (iv) \\ &+ H(C_n - C)(H_n - H)^* \quad (v) \\ &+ (H_n - H)(C_n - C)(H_n - H)^* \quad (vi) \\ &+ H(C_n - C)H^* \quad (vii) \end{aligned}$$

Bound (i) **and** (ii)**.**

$$\begin{aligned} \|(HCH^* + tI)^{-\frac{1}{2}}HC(H_n - H)^*(HCH^* + tI)^{-\frac{1}{2}}\|_\infty \leq \|(HCH^* + tI)^{-\frac{1}{2}}HS^*\|_\infty \\ \times \|(HCH^* + tI)^{-\frac{1}{2}}(H_n - H)S^*\|_\infty \end{aligned}$$

But:

$$\begin{aligned} \|(HCH^* + tI)^{-\frac{1}{2}}HS^*\|_\infty &= \|(HCH^* + tI)^{-\frac{1}{2}}HS^*SH^*(HCH^* + tI)^{-\frac{1}{2}}\|_\infty^{1/2} \\ &= \|(HCH^* + tI)^{-\frac{1}{2}}HCH^*(HCH^* + tI)^{-\frac{1}{2}}\|_\infty^{1/2} \\ &\leq 1 \end{aligned}$$

And from Lemma 9, defining $c_8 = (\xi/2)^{-\frac{4}{\beta+1}}\left(4\kappa(Q + \kappa R)\left(1 + 2\kappa\|M\|_\infty^{\frac{1}{4}(1-\beta)}\right) + c_2^{1/2}\right)^{\frac{4}{\beta+1}}$, $\xi = 14$,if $t \geq c_8\log^8(\frac{8}{\delta})n^{-\frac{1}{\beta+1}}$ we get with probability at least $1 - \delta$

$$\|(HCH^* + tI)^{-\frac{1}{2}}(H_n - H)S^*\|_\infty \leq \frac{1}{14}$$

Bound (iii)**.** As for (i) and (ii), from Lemma 9 we have

$$\begin{aligned} \|(HCH^* + tI)^{-\frac{1}{2}}(H_n - H)C(H_n - H)^*(HCH^* + tI)^{-\frac{1}{2}}\|_\infty &\leq \|(HCH^* + tI)^{-\frac{1}{2}}(H_n - H)S^*\|_\infty^2 \\ &\leq \frac{1}{14^2} \leq \frac{1}{14}. \end{aligned}$$

Bound (iv) **and** (v)**.** We decompose

$$\begin{aligned} &\|(HCH^* + tI)^{-\frac{1}{2}}(H_n - H)(C_n - C)H^*(HCH^* + tI)^{-\frac{1}{2}}\|_\infty \leq \\ &\|(HCH^* + tI)^{-\frac{1}{2}}(H_n - H)C_t^{1/2}\|_\infty \times \|C_t^{-1/2}(C_n - C)C_t^{-1/2}\|_\infty \times \|C_t^{1/2}H^*(HCH^* + tI)^{-\frac{1}{2}}\|_\infty. \end{aligned}$$

We bound

$$\begin{aligned} \|(HCH^* + tI)^{-\frac{1}{2}}(H_n - H)C_t^{1/2}\|_\infty &= \|(HCH^* + tI)^{-\frac{1}{2}}(H_n - H)C_t(H_n - H)^*(HCH^* + tI)^{-\frac{1}{2}}\|_\infty^{1/2} \\ &\leq \|(HCH^* + tI)^{-\frac{1}{2}}(H_n - H)S^*\|_\infty + t^{1/2}\|(HCH^* + tI)^{-\frac{1}{2}}(H_n - H)\|_\infty \\ &\leq \|(HCH^* + tI)^{-\frac{1}{2}}(H_n - H)S^*\|_\infty + \|H_n - H\|_\infty \end{aligned}$$

and similarly,

$$\begin{aligned}\|C_t^{1/2}H^*(HCH^*+tI)^{-\frac{1}{2}}\|_\infty &\leq \|(HCH^*+tI)^{-\frac{1}{2}}HS^*\|_\infty + t^{1/2}\|(HCH^*+tI)^{-\frac{1}{2}}cH\|_\infty \\ &\leq \|(HCH^*+tI)^{-\frac{1}{2}}HS^*\|_\infty + \|H\|_\infty \\ &\leq 1+\|H\|_\infty\end{aligned}$$

finally we obtain

$$\begin{aligned}&\|(HCH^*+tI)^{-\frac{1}{2}}(H_n-H)(C_n-C)H^*(HCH^*+tI)^{-\frac{1}{2}}\|_\infty \leq \\ &\left(\|(HCH^*+tI)^{-\frac{1}{2}}(H_n-H)S^*\|_\infty + \|H_n-H\|_\infty\right) \\ &\times \|C_t^{-1/2}(C_n-C)C_t^{-1/2}\|_\infty \\ &\times (1+\|H\|_\infty)\,.\end{aligned}$$

From Lemma 9, $\|(HCH^*+tI)^{-\frac{1}{2}}(H_n-H)(C_n-C)H^*(HCH^*+tI)^{-\frac{1}{2}}\|_\infty \leq 1/14$ if $t \geq c_8 \log^8(\frac{8}{\delta})n^{-\frac{1}{\beta+1}}$. From Lemma 15, $\|H_n-H\|_\infty \leq 2\log^8(\frac{8}{\delta})R$ if $n \geq n_1$ with n_1 a constant independent of δ. So, defining $u=(1/14+2R)\times(1+R)$. Now, using Lemma 3.6 from Rudi et al. (2013), we can have $\|C_t^{-1/2}(C_n-C)C_t^{-1/2}\|_\infty \leq 1/14 \times u^{-1}\log^{-8}(\frac{8}{\delta})$ if $t \geq a_1\frac{\log n/\delta}{n}$ with $a_1>0$ a constant independent of δ. We conclude that

$$\|(HCH^*+tI)^{-\frac{1}{2}}(H_n-H)(C_n-C)H^*(HCH^*+tI)^{-\frac{1}{2}}\|_\infty \leq \frac{1}{14}.$$

Bound (vi)**.** Similarly as for (v), we have

$$\begin{aligned}&\|(HCH^*+tI)^{-\frac{1}{2}}(H_n-H)(C_n-C)(H_n-H)^*(HCH^*+tI)^{-\frac{1}{2}}\|_\infty \leq \\ &\left(\|(HCH^*+tI)^{-\frac{1}{2}}(H_n-H)S^*\|_\infty + \|H_n-H\|_\infty\right)^2 \\ &\times \|C_t^{-1/2}(C_n-C)C_t^{-1/2}\|_\infty\end{aligned}$$

and, if $t \geq a_2\frac{\log n/\delta}{n}$, with a_2 a constant independent of δ, we also have

$$\|(HCH^*+tI)^{-\frac{1}{2}}(H_n-H)(C_n-C)(H_n-H)^*(HCH^*+tI)^{-\frac{1}{2}}\|_\infty \leq 1/14.$$

Bound (vii)**.** As previously, from Lemma 3.6 from Rudi et al. (2013), there exists a constant $a_3>0$ such that with probability at least $1-\delta$ if $t \geq a_3\frac{\log n/\delta}{n}$, with $a_3>0$, we have

$$\|(HCH^*+tI)^{-\frac{1}{2}}H(C_n-C)H^*(HCH^*+tI)^{-\frac{1}{2}}\|_\infty \leq 1/14.$$

Conclusion. But there exists n_0 independent of δ such that $\forall n \geq n_0 \geq n_1$, $c_8\log^8(\frac{8}{\delta})n^{-\frac{1}{\beta+1}} \geq \max(a_1,a_2,a_3)\frac{\log n/\delta}{n}$. So, we conclude that, if $t \geq c_8\log^8(\frac{8}{\delta})n^{-\frac{1}{\beta+1}}$, and $n \geq n_0$,

$$\|B_n\|_\infty \leq \frac{1}{2}.$$

■

We are now ready to prove the main result of this section. We prove a bound on the reconstruction error of $\hat{P}$ when reconstructing the $h^*(x)$, namely $\mathbb{E}_x[\|\hat{P}h^*(x)-h^*(x)\|^2_{\mathcal{Y}}]^{1/2}$.

Lemma 11 (Supervised subspace learning) *Let $(x_i, y_i)_{i=1}^n$ be drawn independently from a probability measure ρ and $(y_i)_{i=1}^m$ be drawn independently from the marginal ρ w.r.t y with support in the ball $\|y\|_{\mathcal{Y}} \leq Q$. Let $\hat{P}$ be the estimated projection in the proposed method. Then, under Assumptions 1, 2 and 3, there exist constants $c_8 > 0, n_0 \in \mathbb{N}^*$, such that, if $\mu_{p+1}(M) \geq c_8 \log^8(\frac{8}{\delta}) n^{-\frac{1}{\beta+1}}$, $n \geq n_0$, $\lambda_1 = \mu_{p+1}(M)^{-\frac{1-\beta}{2}} n^{-\frac{1}{2}}$, then with probability at least $1 - 3\delta$*

$$\mathbb{E}_x[\|\hat{P}h^*(x) - h^*(x)\|_{\mathcal{Y}}^2]^{\frac{1}{2}} \leq \sqrt{3c_1}\mu_{p+1}(M)^{1/2(1-\alpha)}.$$

Proof We have (See Proposition C.4. in Rudi et al. (2013)):

$$\mathbb{E}_x[\|\hat{P}h^*(x) - h^*(x)\|]^{1/2} = \|(\hat{P} - I)M_c^{\frac{1}{2}}\|_{\mathrm{HS}}^2 \tag{96}$$

Then, as in the proofs of Rudi et al. (2013), we split (96) into three parts, and bound each term,

$$\|(\hat{P} - I)M^{\frac{1}{2}}\|_{\mathrm{HS}} \leq \underbrace{\|(M + tI)^{\frac{1}{2}}(M_n + tI)^{-\frac{1}{2}}\|_{\infty}}_{\mathcal{A}} \times \underbrace{(\mu_{p+1}(M_n) + t)^{\frac{1}{2}}}_{\mathcal{B}} \times \underbrace{\|(M + tI)^{-\frac{1}{2}}M^{\frac{1}{2}}\|_{\mathrm{HS}}}_{\mathcal{C}}$$

Bound $\mathcal{A} = \|(M + tI)^{\frac{1}{2}}(M_n + tI)^{-\frac{1}{2}}\|_{\infty}$. We have:

$$\begin{aligned}\|(M + tI)^{\frac{1}{2}}(M_n + tI)^{-\frac{1}{2}}\|_{\infty} &= \|(M + tI)^{\frac{1}{2}}(M_n + tI)^{-1}(M + tI)^{\frac{1}{2}}\|_{\infty}^{1/2} \\ &= \|(I - B_n)^{-1}\|_{\infty}^{1/2}\end{aligned}$$

with $B_n = (M + tI)^{-1/2}(M - M_n)(M + tI)^{-1/2}$. So, if $\|B_n\|_{\infty} < 1$,

$$\frac{1}{\sqrt{1 + \|B_n\|_{\infty}}} \leq \|(M + tI)^{\frac{1}{2}}(M_n + tI)^{-\frac{1}{2}}\|_{\infty} \leq \frac{1}{\sqrt{1 - \|B_n\|_{\infty}}}$$

Then applying Lemma 10, if $t \geq c_8 \log^8(\frac{8}{\delta}) n^{-\frac{1}{\beta+1}}$, with probability $1 - 3\delta$ it is

$$\sqrt{\frac{2}{3}} \leq \|(M + tI)^{\frac{1}{2}}(M_n + tI)^{-\frac{1}{2}}\|_{\infty} \leq \sqrt{2}$$

Bound $\mathcal{B} = (\mu_{p+1}(M_n) + t)^{\frac{1}{2}}$. $\sqrt{\frac{2}{3}} \leq \|(M + tI)^{\frac{1}{2}}(M_n + tI)^{-\frac{1}{2}}\|_{\infty}$ is equivalent to $M_n + t \preceq \frac{3}{2}M_n + t$ (by Lemma B.2 point 4 in (Rudi et al., 2013)). Then, $\forall k \in \mathbb{N}^*, \mu_k(M_n + t) \leq \frac{3}{2}\mu_k(M + t)$, so we have

$$\sqrt{\mu_{p+1}(M_n) + t} \leq \sqrt{\frac{3}{2}}\sqrt{\mu_{p+1}(M) + t}. \tag{97}$$

Bound $\mathcal{C} = \|(M + tI)^{-\frac{1}{2}}M^{\frac{1}{2}}\|_{\mathrm{HS}}$. We have

$$\begin{aligned}\mathcal{C}^2 &= \mathrm{Tr}(M(M + t)^{-1}) & (98)\\ &= \mathrm{Tr}(M^{\alpha}M^{1-\alpha}(M + t)^{-1}) & (99)\\ &\leq \mathrm{Tr}(M^{\alpha})\|M^{1-\alpha}(M + t)^{-1}\|_{\infty} & (100)\\ &\leq c_1 \times t^{-\alpha} \quad \text{(from Assumption 2 and Young's inequality for products).} & (101)\end{aligned}$$

Finally, we get the following upper bound.

$$\mathbb{E}_x[\|\hat{P}h^*(x) - h^*(x)\|]^{1/2} \leq \sqrt{3}\sqrt{\mu_{p+1}(M) + t} \times c_1^{1/2} \times t^{-\alpha/2} \tag{102}$$

Taking $t = \mu_{p+1}(M)$, which is possible if $\mu_{p+1}(M) \geq c_8 \log^8(\frac{8}{\delta})n^{-\frac{1}{\beta+1}}$, we get

$$\mathbb{E}_x[\|\hat{P}h^*(x) - h^*(x)\|]^{1/2} \leq \sqrt{3c_1}\mu_{p+1}(M)^{1/2(1-\alpha)}. \tag{103}$$

We get the wanted upper bound.

■

A.4 Theorem

In this subsection we give the main result of this paper which is a learning bound for the proposed method. That is we bound:

$$\mathbb{E}_x[\|\hat{P}\hat{h}(x) - h^*(x)\|^2_{\mathcal{Y}}]. \tag{104}$$

The proof consists in decomposing this excess-risk in two terms, as in equation (27), then bounding each term applying the two lemmas previously proved.

Theorem 1 (Learning bounds) *Let $\hat{P}\hat{h}$ be the proposed estimator in Eq.* (8) *with* $\operatorname{rank}(\hat{P}) = p$*, built from n independent couples $(x_i, y_i)_{i=1}^n$ drawn from ρ. Let $\delta \in [0, 1]$. Under the Assumptions 1, 2, 3, 4, there exists constants $c_4, c_5, c_8 > 0$, $n_0 \in \mathbb{N}^*$ defined in the proof, and independent of p, n, δ, such that, if $\mu_{p+1}(M) \geq c_8 \log^8(\frac{8}{\delta})n^{-\frac{1}{\beta+1}}$ and $n \geq n_0$, then with probability at least $1 - 3\delta$,*

$$\mathbb{E}_x[\|\hat{P}\hat{h}(x) - h^*(x)\|^2_{\mathcal{Y}}]^{1/2} \leq \left(c_4\sqrt{p}n^{-1/4} + c_5 S_p(E)^{1/4}\right)n^{-1/4}\log(n/\delta) + \sqrt{3c_1}\mu_{p+1}(M)^{1/2(1-\alpha)} \tag{13}$$

with $S_p(E) = \sum_{i=1}^p \mu_i(E)$.

Proof We decompose the excess-risk as follows

$$\mathbb{E}_x[\|\hat{P}\hat{h}(x) - h^*(x)\|^2_{\mathcal{Y}}]^{1/2} \leq \underbrace{\mathbb{E}_x[\|\hat{P}\hat{h}(x) - \hat{P}h^*(x)\|^2_{\mathcal{Y}}]^{1/2}}_{\text{regr. error on a subspace}} + \underbrace{\mathbb{E}_x[\|\hat{P}h^*(x) - h^*(x)\|^2_{\mathcal{Y}}]^{1/2}}_{\text{reconstruction error}}. \tag{105}$$

We apply the Lemmas 7 and 11, and we get, if $\mu_{p+1}(M) \geq c_8 \log^8(\frac{8}{\delta})n^{-\frac{1}{\beta+1}}$, with probability at least $1 - 3\delta$:

$$\mathbb{E}_x[\|P\hat{h}(x) - h^*(x)\|^2_{\mathcal{Y}}]^{1/2} \leq \left(c_4\sqrt{p}n^{-1/4} + c_5 S_p(E)^{1/4}\right)n^{-1/4}\log(n/\delta) + \sqrt{3c_1}\mu_{p+1}(M)^{1/2(1-\alpha)}. \tag{106}$$

with $c_4 = (7Q + 4\kappa Q + 2\|H\|_{\mathrm{HS}}(1 + 3\kappa))(1 + c_6)$, $c_5 = 10\sqrt{(1 + c_6)}\kappa\|E\|_\infty^{1/2} + 2\|H\|_{\mathrm{HS}}$, $c_6 = \log(8(\frac{\operatorname{Tr}(C)}{\|E\|_\infty^{1/2}} + \frac{\operatorname{Tr}(E)}{\|E\|_\infty}))$.

■

A.5 Corollary

In this subsection we derive from the Theorem 1 a corollary in the case where M and E have polynomial eigenvalue decay rates. This allows to explicit the optimal quantity of components p, and also obtaining a condition on the decay rates $s, e > 1$ in order to obtain a statistical gain.

Corollary 12 (Learning bounds (polynomial decay rates)) *Let $\delta \in \ ]0,1]$, $n \geq n_0$. Under Assumptions 1, 3, and 5, assuming $\frac{B}{b} \leq \theta$ with $\theta \geq 1$, then by taking only*

$$p = c_9(\log^8(\frac{8}{\delta}))^{-\frac{1}{s}} n^{\frac{1}{(\beta+1)s}}, \tag{17}$$

we have with probability at least $1 - 3\delta$:

$$\mathbb{E}_x[\|\hat{P}\hat{h}(x) - h^*(x)\|_{\mathcal{Y}}^2]^{1/2} \ \leq \ c_{10}(s,e) \ \log^{5/4}(\frac{n}{\delta}) \, n^{-1/4} \ + \ c_{11}(e) \, n^{-\frac{1}{2}\frac{1-2/s}{1+\beta}} \ \log^8(\frac{8}{\delta}), \tag{18}$$

where $c_{10}(s,e) = \tilde{c}_{10}\left(\frac{e(e-1)}{s}\right)^{1/4}\left(1 + \log\left(\frac{e}{e-1}\right)\right)$, $c_{11}(e) = \tilde{c}_{11}\left(1 + \log\left(\frac{e}{e-1}\right)\right)$. $\tilde{c}_{10}$, $\tilde{c}_{11}$, n_0, are constants independent of n, δ, s, e, and c_9 is a constant independent of n, δ, defined in the proofs.

Proof The proof consists in applying the Theorem 1 in the specific case of polynomial eigenvalue decay rates. If $\mu_{p+1}(M) \geq c_8 \log^8(\frac{8}{\delta}) n^{-\frac{1}{\beta+1}}$, with probability at least $1 - 3\delta$:

$$\mathbb{E}_x[\|P\hat{h}(x) - h^*(x)\|_{\mathcal{Y}}^2]^{1/2} \leq \left(c_4\sqrt{p}n^{-1/4} + c_5 S_p(E)^{1/4}\right) n^{-1/4} \log(n/\delta) + \sqrt{3c_1}\mu_{p+1}(M)^{1/2(1-\alpha)}. \tag{107}$$

Bound $S_p(E)$. The polynomial eigenvalue decay assumption, give us that $\frac{a}{p^s} \leq \mu_p(M) \leq \frac{A}{p^s}$. So, Assumption 1 is verified with $\alpha = \frac{2}{s}$, and $c_1 = \operatorname{Tr}(M^\alpha) \leq \sum_i i^{-2} \times A^\alpha \leq 2A^\alpha$. Hence,

$$\sqrt{3c_1}\mu_{p+1}(M)^{1/2(1-\alpha)} \leq \frac{\sqrt{6A^\alpha}A^{1/2(1-\alpha)}}{p^{\frac{1}{\alpha}-1}} = \frac{\sqrt{6A}}{p^{\frac{s}{2}-1}}. \tag{108}$$

Moreover,

$$S_p(E) = \sum_{i=1}^{p} \mu_i(E) \leq B \sum_{i=1}^{p} i^{-e} \leq B(1 + \int_{x=1}^{p} x^{-e} dx) \leq \frac{B}{1 - e^{-1}} \times (1 - \frac{e^{-1}}{p^{e-1}}) \tag{109}$$

and using $(1 - 1/x) \leq \log(x) \leq x - 1$, we get

$$\begin{aligned} S_p(E) &\leq \frac{B}{1-e^{-1}} \times ((e-1)\log(p) + \log(e)) && (110)\\ &\leq \frac{B}{1-e^{-1}} \times ((e-1)\log(p) + (e-1)) && (111)\\ &= \frac{B}{1-e^{-1}} \times (e-1)(\log(p) + 1) && (112)\\ &\leq \frac{B}{1-e^{-1}} \times 2(e-1)\log(p) \quad (\text{if } p > 3) && (113)\\ &= 2Be\log(p). && (114)\end{aligned}$$

Now, taking $p = c_9(\log^8(\frac{8}{\delta}))^{-\frac{1}{s}} n^{\frac{1}{s(\beta+1)}}$, defining $c_9 = (\frac{c_8}{a})^{-\frac{1}{s}}$, ensures $\mu_p(M) \geq c_8 \log^8(\frac{8}{\delta}) n^{-\frac{1}{\beta+1}}$. Moreover, $B \leq \theta \times b \leq \theta \operatorname{Tr}(E)(\sum_{i=1}^{+\infty} i^{-e})^{-1} = \frac{\theta \operatorname{Tr}(E)}{\zeta(e)}$ by definition of the Riemann zeta function. So, using this defined p, we get,

$$S_p(E) \leq 2Be\left(\frac{1}{s}\log\left(\frac{a}{c_8}\right) + \frac{\log(n)}{s(\beta+1)}\right) \tag{115}$$

$$\leq \frac{2\theta \operatorname{Tr}(E) e \log(n)}{\zeta(e) s}\left(\log\left(\frac{a}{c_8}\right) + 1\right) \quad \text{(if } n > 3\text{)} \tag{116}$$

Bound $\sqrt{3c_1}\mu_{p+1}(M)^{1/2(1-\alpha)}$. Now, taking $p = c_9(\log^8(\frac{8}{\delta}))^{-\frac{1}{s}} n^{\frac{1}{s(\beta+1)}}$, defining $c_9 = (\frac{c_8}{a})^{-\frac{1}{s}}$, ensures $\mu_p(M) \geq c_8 \log^8(\frac{8}{\delta}) n^{-\frac{1}{\beta+1}}$. Using this defined p, we get

$$\sqrt{3c_1}\mu_{p+1}(M)^{1/2(1-\alpha)} \leq \sqrt{6A}(\frac{c_8}{a}\log^8(\frac{8}{\delta}))^{\frac{1}{2}(1-\frac{2}{s})} n^{-\frac{1}{2}\frac{1-\frac{2}{s}}{1+\beta}} \tag{117}$$

$$\leq \sqrt{6A}(\sqrt{\frac{c_8}{a}} + 1)\log^8(\frac{8}{\delta}) n^{-\frac{1}{2}\frac{1-\frac{2}{s}}{1+\beta}}. \tag{118}$$

Bound $\sqrt{p}n^{-1/2}$. Furthermore, one can check that $(\frac{1}{2} - \frac{1}{2s(\beta+1)}) > \frac{1}{2}\frac{1-2/s}{1+\beta}$, hence we have

$$\sqrt{p}n^{-1/2} \leq \left(\frac{a}{c_8}\right)^{1/2s} n^{-(\frac{1}{2} - \frac{1}{2s(\beta+1)})} \leq \left(\frac{a}{c_8} + 1\right) n^{-\frac{1}{2}\frac{1-2/s}{1+\beta}}. \tag{119}$$

Studying c_4, c_5, c_8, n_0 dependencies in s, e. In this work we study the behavior of the bound when the shape of E and M vary, i.e. when s and e vary. Therefore, it's important to make some derivations to studying c_4, c_5, c_8, n_0's dependencies in s and e. First, c_8, n_0 are independent of δ, s, e.

Then, observing that we have $\|E\|_\infty^{-1} = \mu_1(E)^{-1} \leq b^{-1} \leq \frac{\theta}{B} \leq \theta\frac{\zeta(e)}{\operatorname{Tr}(E)}$, leads to $c_6 \leq \log(8(\frac{\theta^{1/2}\operatorname{Tr}(C)}{\operatorname{Tr}(E)^{1/2}} + \theta)) + \log(\zeta(e))$. So, we have

$$c_4 = (7Q + 4\kappa Q + 2R(1+3\kappa))(1 + c_6) \tag{120}$$

$$\leq (\log(\zeta(e)) + 1)\left(1 + \log(8(\frac{\theta^{1/2}\operatorname{Tr}(C)}{\operatorname{Tr}(E)^{1/2}} + \theta))\right)(7Q + 4\kappa Q + 2R(1+3\kappa)) \tag{121}$$

and also

$$c_5 = 10\sqrt{(1 + c_6)}\kappa\|E\|_\infty^{1/2} + 2\|H\|_{\mathrm{HS}} \tag{122}$$

$$\leq (\log(\zeta(e) + 1))\left(1 + \log(8(\frac{\theta^{1/2}\operatorname{Tr}(C)}{\operatorname{Tr}(E)^{1/2}} + \theta))\right)\left(10\kappa\|E\|_{\mathrm{HS}}^{1/2} + 2\|H\|_{\mathrm{HS}}\right). \tag{123}$$

Conclusion. Thanks to the previous derivations we obtain the following bound

$$\mathbb{E}_x[\|P\hat{h}(x) - h^*(x)\|_{\mathcal{Y}}^2]^{1/2} \leq c_{10}(s,e)\log^{5/4}(\frac{n}{\delta})n^{-1/4} + c_{11}(e)n^{-\frac{1}{2}\frac{1-2/s}{1+\beta}}\log^8(\frac{8}{\delta})$$

with $c_{10}(s,e) = \tilde{c}_{10}(\log(\zeta(e))+1)\left(\frac{e}{\zeta(e)\times s}\right)^{1/4}$, $c_{11}(e) = \tilde{c}_{11}(\log(\zeta(e))+1)$. $\tilde{c}_{10}$ and $\tilde{c}_{11}$ are constants independent of n, δ, s, e, defined below

$$\tilde{c}_{10} = \left(1 + \log(8(\frac{\theta^{1/2}\operatorname{Tr}(C)}{\operatorname{Tr}(E)^{1/2}} + \theta))\right)\left(10\kappa\|E\|_{\mathrm{HS}}^{1/2} + 2\|H\|_{\mathrm{HS}}\right)\left(2\theta\operatorname{Tr}(E)(\log(\frac{a}{c_8})+1)\right)^{1/4} \tag{124}$$

$$\tilde{c}_{11} = \sqrt{6A}(\sqrt{\frac{c_8}{a}}+1) + \left(\frac{a}{c_8}+1\right)\left(1 + \log(8(\frac{\theta^{1/2}\operatorname{Tr}(C)}{\operatorname{Tr}(E)^{1/2}} + \theta))\right)(7Q + 4\kappa Q + 2R(1+3\kappa)). \tag{125}$$

The inequalities $\frac{1}{e-1} \le \zeta(e) \le \frac{e}{e-1}$ allow to conclude the proof.

■

A.6 Auxiliary Results

In this section, we give four auxiliary results:

- A bound on the KRR estimator which monitors the role of the total amount of noise $\operatorname{Tr}(E)$.
- A Bernstein inequality for bounded operator and the operator norm.
- A bound on $\|H_n - H\|_\infty$, used in the proof of Lemma 11.
- Some properties of Löwner's partial ordering

Lemma 13 (Full-rank KRR excess-risk) *Let $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be a bounded kernel with $\forall x \in \mathcal{X}, k(x,x) \le \kappa^2$. Let ρ be a distribution on $\mathcal{X}\times\mathcal{Y}$ such that its marginal w.r.t y is supported on the ball $\|y\|_{\mathcal{Y}} \le Q$. Let $\hat{h}$ be the KRR estimator trained with n independent couples drawn from ρ. Let $\delta \in [0,1]$. Then, under the assumption 1 and 3, taking*

$$\lambda_1 = \max\left(\frac{1}{n}, \frac{\|E^{1/2}\|_{\mathrm{HS}}}{\sqrt{n}}\right) \tag{126}$$

the following holds with probability at least $1-\delta$

$$\mathbb{E}_x[\|\hat{h}(x) - h^*(x)\|_{\mathcal{Y}}^2]^{\frac{1}{2}} \le C(p)n^{-\frac{1}{4}}\log\frac{4}{\delta} \tag{127}$$

with $C(p) = 10\left[\mathcal{O}(n^{-\frac{1}{4}}) + (\kappa+R)\|E^{1/2}\|_{\mathrm{HS}}^{\frac{1}{2}}\right]$, $R = \|H\|_{\mathrm{HS}}$.

Proof We follow the proofs of (Ciliberto et al., 2020) in order to derive a learning bound of the KRR estimator. We carefully monitor the role of the total amount of noise $\operatorname{Tr}(E)$.

We make appear the conditional variance by modifying the Proposition B.7 in (Ciliberto et al., 2020), with the following change from equation (B.55) to (B.58):

$$\begin{aligned}\mathbb{E}_x[\|C_\lambda^{-1/2}\phi(x)\|^2\sigma(x)^2] &\leq \frac{\kappa^2}{\lambda} \times \mathbb{E}_x[\sigma(x)^2] && (128)\\ &= \frac{\kappa^2}{\lambda} \times \mathbb{E}[\|\epsilon\|_{\mathcal{Y}}^2] && (129)\\ &= \frac{\kappa^2}{\lambda} \times \|E^{1/2}\|_{\text{HS}}^2 && (130)\end{aligned}$$

by defining the noise $\epsilon = \psi(y) - h^*(x)$, and $E = \mathbb{E}[\epsilon \otimes \epsilon]$.

Then, doing the same proof than Theorem B.8 from (Ciliberto et al., 2020), we get the following bound

$$\begin{aligned}\mathbb{E}_x[\|P\hat{h}(x) - Ph^*(x)\|_{\mathcal{Y}}^2]^{1/2} \leq \quad & \frac{8\kappa \log \frac{2}{\delta}}{\sqrt{\lambda} n} \times (Q + \kappa \|L_\lambda^{-1/2} Z\|_{\text{HS}})\\ & + \frac{1}{\sqrt{n}} \times \sqrt{64(d_{\text{eff}}(\lambda) \times \|E^{1/2}\|_{\text{HS}}^2 + \kappa^2\lambda\|L_\lambda^{-1}Z\|_{\text{HS}}^2) \log \frac{4}{\delta}}\\ & + 10 \times \lambda \|L_\lambda^{-1} Z\|_{\text{HS}}\end{aligned}$$

Now, using the assumption 1, we have

$$\begin{aligned}\|L_\lambda^{-1} Z\|_{\text{HS}} &= \|L_\lambda^{-1} S H^*\|_{\text{HS}} && (131)\\ &\leq \|L_\lambda^{-1} S\|_{\text{HS}} \times \|H\|_{\text{HS}} && (132)\\ &\leq \lambda^{-\frac{1}{2}} \times R && (133)\end{aligned}$$

and similarly $\|L_\lambda^{-1} Z\|_{\text{HS}} \leq R$. Moreover,

$$d_{\text{eff}}(\lambda) := \text{Tr}((C + \lambda I)^{-1} C) \leq \lambda^{-1}\kappa^2. \qquad (134)$$

So, we get

$$\begin{aligned}\mathbb{E}_x[\|\hat{h}(x) - h^*(x)\|_{\mathcal{Y}}^2]^{1/2} \leq \quad & \frac{\kappa(Q + \kappa R)}{\sqrt{\lambda} n} \times 10 \log \frac{4}{\delta}\\ & + \frac{1}{\sqrt{n}} \times \sqrt{(\lambda^{-1}\kappa^2\|E^{1/2}\|_{\text{HS}}^2 + \kappa^2 R)} \times 10 \log \frac{4}{\delta}\\ & + \lambda^{\frac{1}{2}} \times R \times 10 \log \frac{4}{\delta}\end{aligned}$$

Now, we define λ in order to minimize this bound, with

$$\lambda = \max\left(\frac{1}{n}, \frac{\|E^{1/2}\|_{\text{HS}}}{\sqrt{n}}\right)$$

so we obtain

$$\begin{aligned}\mathbb{E}_x[\|\hat{h}(x) - h^*(x)\|_{\mathcal{Y}}^2]^{1/2} \leq \quad & n^{-\frac{1}{4}} \times 10 \log \frac{4}{\delta} \times \Bigg[(\kappa(Q + \kappa R)) n^{-\frac{1}{4}} \\ & + \sqrt{\kappa^2 \|E^{1/2}\|_{\mathrm{HS}} + \kappa^2 R^2 n^{-\frac{1}{2}}} \\ & + \left(n^{-\frac{1}{4}} + \|E^{1/2}\|_{\mathrm{HS}}^{\frac{1}{2}} \right) \times R \Bigg].\end{aligned}$$

We conclude

$$\mathbb{E}_x[\|\hat{h}(x) - h^*(x)\|_{\mathcal{Y}}^2]^{1/2} \leq C(p) n^{-\frac{1}{4}} \log \frac{4}{\delta} \tag{135}$$

with

$$\begin{aligned}C(p) &= 10 \left[(\kappa(Q + \kappa R)) n^{-\frac{1}{4}} + \kappa \sqrt{\|E^{1/2}\|_{\mathrm{HS}} + R^2 n^{-\frac{1}{2}}} + \left(n^{-\frac{1}{4}} + \|E^{1/2}\|_{\mathrm{HS}}^{\frac{1}{2}} \right) R \right] \\ &= 10 \left[\mathcal{O}(n^{-\frac{1}{4}}) + (\kappa + R) \|E^{1/2}\|_{\mathrm{HS}}^{\frac{1}{2}} \right].\end{aligned}$$

■

Theorem 14 (Concentration inequality on the operator norm, Tropp (2012)(Theorem 7.3.2)) *Let ξ_i be independent copies of the random variable ξ with values in the space of bounded operators over a Hilbert space $\mathcal{H}$ such that $\mathbb{E}[\xi] = 0$. Let there be $R > 0$ such that $\|\xi\|_\infty \leq T$. Define $\sigma^2 = \max(\|\mathbb{E}[\xi\xi^*]\|_\infty, \|\mathbb{E}[\xi^*\xi]\|_\infty)$, and $d = \mathrm{Tr}(\mathbb{E}[\xi^*\xi] + \mathbb{E}[\xi\xi^*])/\sigma^2$. Then, if $\delta \in [0, 1]$, with probability at least $1 - \delta$*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \right\|_\infty \leq \sqrt{\frac{2\eta\sigma^2}{n}} + \frac{2T\eta}{3n} \tag{136}$$

where $\eta = \log(\frac{4d}{\delta})$.

Proof This theorem is a restatement of Theorem 7.3.2 of (Tropp, 2012) generalized to the separable Hilbert space case by means of the technique in Section 3.2 of (Minsker, 2017). ■

Lemma 15 (Bound $\|H_n - H\|_\infty$) *With probability at least $1 - 2\delta$ it is*

$$\|H_n - H\|_\infty \leq \frac{4 \log \frac{2}{\delta}}{\lambda_1 \sqrt{n}} (Q\kappa + \kappa^2 \|h_\psi^*\|_{\mathcal{H}}) + \|h_\psi^*\|_{\mathcal{H}}$$

Proof In order to bound $\|H_n - H\|_\infty$ we do the following decomposition in three terms, and bound each term:

$$\begin{aligned} \|H_n - H\|_\infty &= \|Z_n^* S_n (C_n + \lambda_1 I)^{-1} - Z^* S C^\dagger\|_\infty \\ &\leq \underbrace{\|(Z_n^* S_n - Z^* S)(C_n + \lambda_1 I)^{-1}\|_\infty}_{(A)} + \underbrace{\|Z^* S((C_n + \lambda_1 I)^{-1} - (C + \lambda_1 I)^{-1})\|_\infty}_{(B)} \\ &+ \underbrace{\|Z^* S((C + \lambda_1 I)^{-1} - C^\dagger)\|_\infty}_{(C)} \end{aligned}$$

Bound (A). We have:

$$(A) = \|(Z_n^* S_n - Z^* S)(C_n + \lambda_1 I)^{-1}\|_\infty \leq \frac{1}{\lambda_1} \|Z_n^* S_n - Z^* S\|_{\text{HS}}$$

From Ciliberto et al. (2016) (proof of lemma 18.), with probability $1 - \delta$: $(A) \leq \frac{4Q\kappa \log \frac{2}{\delta}}{\lambda_1 \sqrt{n}}$.

Bound (B). We have:

$$\begin{aligned} (B) &= \|Z^* S((C + \lambda_1 I)^{-1} - (C_n + \lambda_1 I)^{-1})\|_\infty \\ &= \|Z^* S((C + \lambda_1 I)^{-1}(C_n - C)(C_n + \lambda_1 I)^{-1})\|_\infty \\ &\leq \|Z^* S(C + \lambda_1 I)^{-1}\|_\infty \|(C_n - C)\|_\infty \|(C_n + \lambda_1 I)^{-1}\|_\infty \\ &\leq \frac{1}{\lambda_1} \|h_\psi^*\|_{\mathcal{H}} \|(C_n - C)\|_\infty \end{aligned}$$

where we used the fact that for two invertible operators A, B: $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, and noting that $\|Z^* S(C + \lambda_1 I)^{-1}\|_\infty \leq \|Z^* S(C + \lambda_1 I)^{-1}\|_{\text{HS}} \leq \|H\|_{\text{HS}} = \|h_\psi^*\|_{\mathcal{H}}$. From Ciliberto et al. (2016), with probability $1 - \delta$: $(B) \leq \frac{4\|h_\psi^*\|_{\mathcal{H}} \kappa^2 \log \frac{2}{\delta}}{\lambda_1 \sqrt{n}}$.

Bound (C). We have:

$$\begin{aligned} (C) &= \|Z^* S((C + \lambda_1 I)^{-1} - C^\dagger)\|_\infty \\ &= \|H S^* S((C + \lambda_1 I)^{-1} - C^\dagger)\|_\infty \\ &= \|H(C(C + \lambda_1 I)^{-1} - I)\|_\infty \\ &= \lambda_1 \|H(C + \lambda_1 I)^{-1}\|_\infty \\ &\leq \|h_\psi^*\|_{\mathcal{H}} \end{aligned}$$

We conclude by union bound, with probability at least $1 - 2\delta$:

$$\|H_n - H\|_\infty \leq \frac{4Q\kappa \log \frac{2}{\delta}}{\lambda_1 \sqrt{n}} + \frac{4\|h_\psi^*\|_{\mathcal{H}} \kappa^2 \log \frac{2}{\delta}}{\lambda_1 \sqrt{n}} + \|h_\psi^*\|_{\mathcal{H}}$$

Notice that if we choose $\lambda_1 = (c_8 \log^8(\frac{8}{\delta}))^{-\frac{1-\beta}{2}} n^{-\frac{\beta}{\beta+1}}$ as chosen in Lemma 9, we obtain

$$\|H_n - H\|_\infty \leq (4Q\kappa + R\kappa^2) \log \frac{2}{\delta} \times a \times n^{\frac{\beta}{1+\beta} - \frac{1}{2}} + R \tag{137}$$

with $a = c_8 \log^8(\frac{8}{\delta})^{\frac{1}{2}}$, such that $\|H_n - H\|_\infty \leq 2R\log^9(\frac{8}{\delta})^{\frac{1}{2}}$ when $n \geq N$ with $N > 0$ a constant independent of δ.

■

Lemma 16 (Properties of Löwners's partial ordering $\preceq$) *Let A, B be positive semidefinite linear operators on $\mathcal{Y}$ such that $A \preceq B$, and M a bounded linear operator on $\mathcal{Y}$, then*

1. *If A, B are random variables then $\mathbb{E}[A] \preceq \mathbb{E}[B]$.*
2. *$MAM^* \preceq MBM^*$.*

Proof

1) For any $u \in \mathcal{Y}$, we have $\langle u, \mathbb{E}[A]u\rangle_\mathcal{Y} = \mathbb{E}[\langle u, Au\rangle_\mathcal{Y}] \leq \mathbb{E}[\langle u, Bu\rangle_\mathcal{Y}] = \langle u, \mathbb{E}[B]u\rangle_\mathcal{Y}$.

2) From Lemma B.2 in Rudi et al. (2013).

■

A.7 About the Independence of $\phi(x)$ and ϵ

In this section, we discuss the assumption that the random variables $\phi(x)$ and ϵ are independent.

In this work, this assumption allows to obtain shorter and lighter derivations, and an easier reading of the proofs. Nevertheless, such assumption is not exploited by the proposed method, and similar results can be proven without this assumption. More precisely, one can prove bounds with the same dependencies in the parameters of the learning setting, leading to the same conclusions. We discuss how below.

How to obtain similar bounds without this assumption? The independence of $\phi(x)$ and ϵ allow simpler derivations when bounding expectations involving products of these two random variables using $\mathbb{E}[f(\phi(x))g(\epsilon)] = \mathbb{E}[f(\phi(x)) \times \mathbb{E}[g(\epsilon)]$. This is used multiple times from Equations (38) to (48) to prove the Lemma 6, and only there.

We carried out derivations below in order to bound the same quantities but we do not make use of the assumption. Then, we will check that the dependencies in the parameters of the learning setting are similar.

Sketch of the proof (Bound $\|(Z_n^* S_n - HC_n)(C + \lambda_2 I)^{-1/2}\|_\infty$ without the independence assumption). We define

$$\xi_i = \epsilon_i \otimes \phi(x_i)(C + \lambda_2 I)^{-1/2} \tag{138}$$

with $\epsilon_i = y_i - h^*(x_i)$. In this way,

$$\|(Z_n^* S_n - HC_n)(C + \lambda_2 I)^{-1/2}\|_\infty = \|\frac{1}{n}\sum_{i=1}^n \xi_i - \mathbb{E}[\xi]\|_\infty. \tag{139}$$

We aim at applying the Bernstein inequality given in Theorem 14 to the random linear operator $u := \xi - \mathbb{E}[\xi]$. So, we define

$$T := 4\kappa Q \lambda_2^{-1/2} \geq \|u\|_\infty, \tag{140}$$

$$\sigma^2 := \max(\|\mathbb{E}[uu^*]\|_\infty, \|\mathbb{E}[u^*u]\|_\infty), \tag{141}$$

$$d := \operatorname{Tr}(\mathbb{E}[u^*u] + \mathbb{E}[uu^*])/\sigma^2. \tag{142}$$

Note that $\|\epsilon\| \leq \|y\|_\mathcal{Y} + \|h^*(x)\|_\mathcal{Y} \leq 2Q$, and $\|\phi(x)\| \leq \kappa$. Then, we have

$$\begin{aligned}
\mathbb{E}[uu^*] &= \mathbb{E}[(\epsilon \otimes \phi(x) - \mathbb{E}[\epsilon \otimes \phi(x)])(C + \lambda_2 I)^{-1}(\epsilon \otimes \phi(x) - \mathbb{E}[\epsilon \otimes \phi(x)])^*] && (143)\\
&\preceq \lambda_2^{-1}\mathbb{E}[(\epsilon \otimes \phi(x) - \mathbb{E}[\epsilon \otimes \phi(x)])(\epsilon \otimes \phi(x) - \mathbb{E}[\epsilon \otimes \phi(x)])^*] && (144)\\
&= \lambda_2^{-1}\mathbb{E}[(\epsilon \otimes \phi(x)(\epsilon \otimes \phi(x))^*] - \mathbb{E}[\epsilon \otimes \phi(x)]\mathbb{E}[\epsilon \otimes \phi(x)]^* && (145)\\
&\preceq \lambda_2^{-1}\mathbb{E}[\epsilon \otimes \epsilon\|\phi(x)\|^2] && (146)\\
&\preceq \lambda_2^{-1}\kappa^2\mathbb{E}[\epsilon \otimes \epsilon] = \lambda_2^{-1}\kappa^2 E && (147)
\end{aligned}$$

where $\preceq$ denotes the Löwner's partial ordering of positive semidefinite operators. We used properties of Löwner's partial ordering (cf. Lemma 16). So, we have

$$\|\mathbb{E}[uu^*]\|_\infty \leq \lambda_2^{-1}\kappa^2\|E\|_\infty. \tag{148}$$

Then, similarly, we have

$$\begin{aligned}
\mathbb{E}[u^*u] &= (C + \lambda_2 I)^{-1/2}\mathbb{E}[(\epsilon \otimes \phi(x) - \mathbb{E}[\epsilon \otimes \phi(x))^*(\epsilon \otimes \phi(x) - \mathbb{E}[\epsilon \otimes \phi(x))](C + \lambda_2 I)^{-1/2} && (149)\\
&= (C + \lambda_2 I)^{-1/2}\left(\mathbb{E}[\phi(x) \otimes \phi(x)\|\epsilon\|^2] - \mathbb{E}[\phi(x) \otimes \epsilon]\mathbb{E}[\phi(x) \otimes \epsilon]^*\right)(C + \lambda_2 I)^{-1/2} && (150)\\
&\preceq (C + \lambda_2 I)^{-1/2}4Q^2C(C + \lambda_2 I)^{-1/2} && (151)\\
&\preceq 4Q^2 I_\mathcal{Y}. && (152)
\end{aligned}$$

So, we have

$$\|\mathbb{E}[u^*u]\|_\infty \leq 4Q^2. \tag{153}$$

Now, from previous derivations, if $\lambda_2 < \|C\|_\infty$, we also have

$$\operatorname{Tr}(\mathbb{E}[uu^*]) \leq \lambda_2^{-1}\operatorname{Tr}(E)\kappa^2, \tag{154}$$

$$\operatorname{Tr}(\mathbb{E}[u^*u]) \leq 4Q^2\lambda_2^{-1}\operatorname{Tr}(C), \tag{155}$$

$$\|\mathbb{E}[uu^*]\|_\infty \geq \frac{\|\operatorname{Var}(\epsilon \otimes \phi(x))\|_\infty}{2\|C\|_\infty}. \tag{156}$$

by defining $\operatorname{Var}(\epsilon \otimes \phi(x)) = \mathbb{E}[(\epsilon \otimes \phi(x) - \mathbb{E}[\epsilon \otimes \phi(x)])(\epsilon \otimes \phi(x) - \mathbb{E}[\epsilon \otimes \phi(x)])^*]$. So, we have

$$\begin{aligned}
d &\leq \frac{\operatorname{Tr}(\mathbb{E}[u^*u]) + \operatorname{Tr}(\mathbb{E}[uu^*])}{\|\mathbb{E}[uu^*]\|_\infty} && (157)\\
&\leq \lambda_2^{-1}\frac{2(\operatorname{Tr}(E)\kappa^2 + 4Q^2\operatorname{Tr}(C))\|C\|_\infty}{\|\operatorname{Var}(\epsilon \otimes \phi(x))\|_\infty}. && (158)
\end{aligned}$$

Conclusion. Then, one can bound $\|(Z_n^* S_n - HC_n)(C + \lambda_2 I)^{-1/2}\|_\infty$ as in the proof of Lemma 6 by applying the Bernstein inequality given in Theorem 14.

The dependencies in the learning setting's parameters of the resulting bound will depend on the dependencies in the learning setting's parameters of the obtained bounds on $\|\mathbb{E}[u^* u]\|_\infty$, $\|\mathbb{E}[uu^*]\|_\infty$, and d.

Notice that the bounds on $\|\mathbb{E}[u^* u]\|_\infty$, $\|\mathbb{E}[uu^*]\|_\infty$ have the same dependencies in the learning setting's parameters than the ones obtained in Lemma 6 on $\|\mathbb{E}[\xi^* \xi]\|_\infty$, $\|\mathbb{E}[\xi \xi^*]\|_\infty$.

The bound on d obtained above without the independence assumption has poorer dependencies in the learning setting's parameters than the one obtained in Lemma 6. More precisely, d has poorer dependencies in t_1 and λ_2. Nevertheless, it remains polynomial dependencies in t_1^{-1} and λ_2^{-1}, such that the resulting $\eta = \log(\frac{4d}{\delta})$, in the proof of Lemma 7, has similar dependencies in the learning setting's parameters than the one obtained in Lemma 7.

We conclude that, without the independence assumption of $\phi(x)$ and ϵ, one can prove bounds similar to Theorem 1, namely with the same dependencies in the parameters of the learning setting.

Appendix B. Additional Experimental Details

In this section, we give an additional synthetic experiment (Section B.1) that aims at discussing the difference between the output source condition (Assumption 3) and the standard source condition (Ciliberto et al., 2020). We also give additional details on the experiments for the sake of reproducibility (Sections B.2, B.3).

B.1 Difference Between Standard Source Condition and Assumption 3.

From Assumption 1 we have $M = HCH^*$. Hence, Assumption 3 measures the alignment between HCH^* and HH^*. Notice that it's a different assumption than requiring the alignment of C and H^*H (source condition). Indeed, in general strong Assumption 3 doesn't imply strong source condition. For instance, when H is finite rank (e.g. $H = y_0 \otimes h_0$ with $y_0 \in \mathcal{Y}, h_0 \in \mathcal{H}_x$), Assumption 3 is verified with $\beta = 0$ (best case), while the source condition can be arbitrarily bad (e.g. if $\langle h_0 | C^{-(1-v)} h_0 \rangle_{\mathcal{H}_x} = +\infty$ with $v > 0$, then the source condition can't be verified for $r \leq v$). Source condition is verified with $r = 1 - 2u$ by operators of the form $H = H_0 C^u$ with $H_0 \in \mathcal{Y} \otimes \mathcal{H}_x$, $\|H_0\|_{\text{HS}} < +\infty$, $u \in [0, \frac{1}{2}]$. Similarly, Assumption 3 is verified with $\beta = \frac{1}{2u+1}$ by operators of the form $H = (H_0 C H_0^*)^u H_0$ with $\|H_0\|_\infty < +\infty$, $u \in [0, +\infty[$.
We illustrate this empirically. For $d = 200$, $\mathcal{X} = \mathcal{H}_x = \mathcal{Y} = \mathbb{R}^d$, we choose $\mu_p(C) = \frac{1}{p^2}$ and draw randomly the eigenvector associated to each eigenvalue. We draw $H_0 \in \mathbb{R}^{d \times d}$ with independently drawn coefficients from the standard normal distribution. Notice that β and r can be measured as the increasing rates, when $t, \lambda \to 0$, in $t^{-\beta}$ and λ^{-r} of the quantities $\|(M+t)^{-\frac{1}{2}} H\|_\infty^2$ and $\|H(C+\lambda)^{-\frac{1}{2}}\|_\infty^2$. Hence, we compute and plot on Figure 5 $\|H(C+\lambda)^{-\frac{1}{2}}\|_\infty^2$ w.r.t λ (left), and $\|(M+t)^{-\frac{1}{2}} H\|_\infty^2$ w.r.t t (right), with $H = (H_0 C H_0^*)^\gamma H_0$ for various $\gamma \in [0, 1.5]$. We also plot in Figure 5 (right) the slopes $\beta = \frac{1}{2\gamma+1}$. Firstly, we see that Assumption 3 indeed improved when γ increases, while the source condition is low and does not change. Then, as explained $H = (H_0 C H_0^*)^\gamma H_0$ verifies Assumption 3 with at least $\beta = \frac{1}{2\gamma+1}$, but depending on H_0 it might be verified for $\beta \ll \frac{1}{2\gamma+1}$. Nonetheless, notice that with our generated H_0, $\beta = \frac{1}{2\gamma+1}$ are sharp for $H = (H_0 C H_0^*)^\gamma H_0$.

B.2 Image Reconstruction

Link to downloadable data set `https://web.stanford.edu/~hastie/StatLearnSparsity_files/DATA/zipcode.html`

SPEN USPS experiments' details. We used an implementation of SPEN in python with PyTorch by Philippe Beardsell and Chih-Chao Hsu (cf. https://github.com/philqc/deep-value-networks-pytorch). Small changes have been made. SPEN was trained using standard architecture from Belanger and McCallum (2016), that is a simple 2-hidden layers neural network for the feature network with equal layer size $n_h = 110$, and a single-hidden layer neural network for the structure learning network with size $n_s = 50$. The size of the two hidden layers $n_h \in [10, 30, 50, 70, 90, 110, 130]$ was selected during the pre-training of the feature network using 5 repeated random sub-sampling validation (80%/20%) selecting the best mean validation MSE (cf. Figure 6 for convergence of this phase). $n_s \in [5, 10, 20, 50, 70]$ was selected during the training phase of the SPEN network (training of the structure learn-

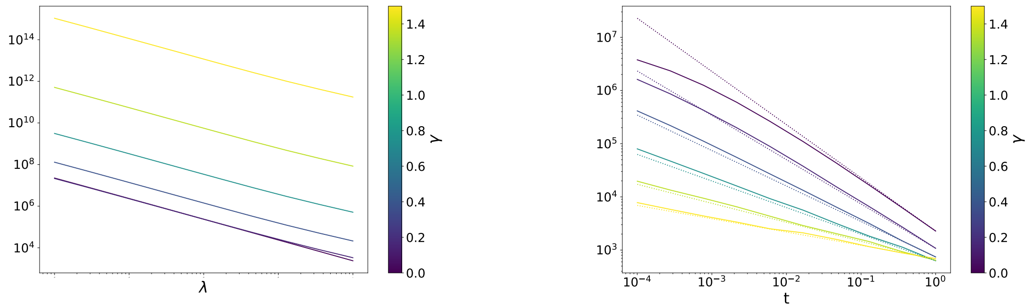

Figure 5: Source condition $\|H(C+\lambda)^{-\frac{1}{2}}\|_\infty^2$ w.r.t λ (left) and output source condition $\|(M+t)^{-\frac{1}{2}}H\|_\infty^2$ w.r.t t (right) in log-log scale for $H = (H_0CH_0^*)^\gamma H_0$ and various $\gamma \in \{0, 0.1, 0.25, 0.5, 0.9, 1.5\}$.

ing network plus the last layer of the feature network) doing approximate loss-augmented inference (cf. Figure 6 for inferences' convergences), and minimizing the SSVM loss, using 5 repeated random sub-sampling validation (80%/20%) selecting the best mean validation MSE (cf. Figure 6 for convergence of this phase).

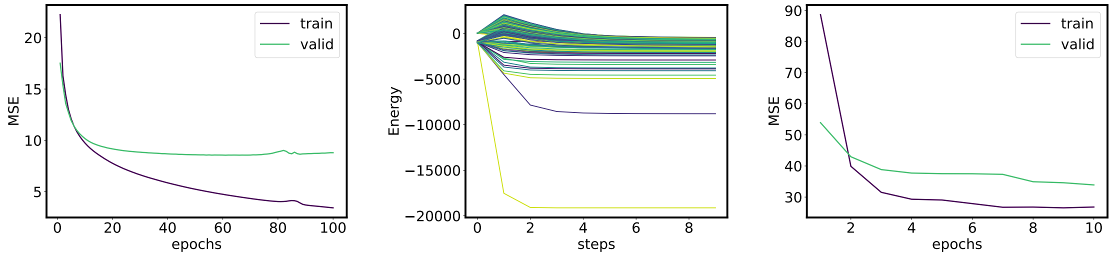

Figure 6: Left: Convergence of train/validation MSE when pre-training the feature network. / Center: approximate loss-augmented inferences' convergences. / Right: Convergence of train/validation SSVM loss when training the SPEN network.

B.3 Multi-label Classification

Link to downloadable data set `http://mulan.sourceforge.net/datasets-mlc.html`

References

Mauricio A. Álvarez, Lorenzo Rosasco, and Neil D. Lawrence. Kernels for vector-valued functions: A review. *Found. Trends Mach. Learn.*, 4(3):195–266, 2012. ISSN 1935-8237.

Theodore Wilbur Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, pages 327–351, 1951.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Luca Baldassarre, Lorenzo Rosasco, Annalisa Barla, and Alessandro Verri. Multi-output learning via spectral filtering. *Machine learning*, 87(3):259–301, 2012.

David Belanger and Andrew McCallum. Structured prediction energy networks. In *International Conference on Machine Learning*, pages 983–992. PMLR, 2016.

Céline Brouard, Florence d'Alché-Buc, and Marie Szafranski. Semi-supervised penalized output kernel regression for link prediction. In *Proceedings of the 28th International Conference on Machine Learning*, pages 593–600, 2011.

Céline Brouard, Huibin Shen, Kai Dührkop, Florence d'Alché Buc, Sebastian Böcker, and Juho Rousu. Fast metabolite identification with input output kernel regression. *Bioinformatics*, 32(12):i28–i36, 2016a.

Céline Brouard, Marie Szafranski, and Florence d'Alché-Buc. Input output kernel regression: Supervised and semi-supervised structured output prediction with operator-valued kernels. *The Journal of Machine Learning Research*, 17(1):6105–6152, 2016b.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. In *Advances in neural information processing systems*, pages 4412–4420, 2016.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A general framework for consistent structured prediction with implicit loss embeddings. *J. Mach. Learn. Res.*, 21(98):1–67, 2020.

Rina Foygel, Michael Horrell, Mathias Drton, and John Lafferty. Nonparametric reduced rank regression. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Thomas Gärtner. A survey of kernels for structured data. *ACM SIGKDD explorations newsletter*, 5(1):49–58, 2003.

Pierre Geurts, Louis Wehenkel, and Florence d'Alché Buc. Kernelizing the output of tree-based methods. In *Proceedings of the 23rd international conference on Machine learning*, pages 345–352, 2006.

Michael Gygli, Mohammad Norouzi, and Anelia Angelova. Deep value networks learn to evaluate and iteratively refine structured outputs. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1341–1351, 2017.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

Hachem Kadri, Mohammad Ghavamzadeh, and Philippe Preux. A generalized kernel approach to structured output learning. In *International Conference on Machine Learning*, pages 471–479. PMLR, 2013.

Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, Alain Rakotomamonjy, and Julien Audiffren. Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 17(20):1–54, 2016.

Ioannis Katakis, Grigorios Tsoumakas, and Ioannis Vlahavas. Multilabel text classification for automated tag suggestion. *ECML PKDD Discovery Challenge 2008*, page 75, 2008.

Anna Korba, Alexandre Garcia, and Florence d'Alché Buc. A structured prediction approach for label ranking. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Pierre Laforgue, Alex Lambert, Luc Brogat-Motte, and Florence d'Alché Buc. Duality in rkhss with infinite dimensional outputs: Application to robust losses. In *International Conference on Machine Learning*, pages 5598–5607. PMLR, 2020.

Néhémy Lim, Florence d'Alché Buc, Cédric Auliac, and George Michailidis. Operator-valued kernel-based vector autoregressive models for network inference. *Machine learning*, 99(3): 489–513, 2015.

Xi Victoria Lin, Sameer Singh, Luheng He, Ben Taskar, and Luke Zettlemoyer. Multi-label learning with posterior regularization. In *NIPS Workshop on Modern Machine Learning and Natural Language Processing*, 2014.

Giulia Luise, Dimitrios Stamos, Massimiliano Pontil, and Carlo Ciliberto. Leveraging low-rank relations between surrogate tasks in structured prediction. In *International Conference on Machine Learning*, pages 4193–4202. PMLR, 2019.

Helmut Lütkepohl. Vector autoregressive models. In *Handbook of research methods and applications in empirical macroeconomics*. Edward Elgar Publishing, 2013.

Charles A Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural computation*, 17(1):177–204, 2005.

Stanislav Minsker. On some extensions of bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 127:111–119, 2017.

Ashin Mukherjee and Ji Zhu. Reduced rank ridge regression and its kernel extensions. *Statistical analysis and data mining: the ASA data science journal*, 4(6):612–622, 2011.

Guillaume Rabusseau and Hachem Kadri. Low-rank regression with tensor responses. *Advances in Neural Information Processing Systems*, 29:1867–1875, 2016.

Bernardino Romera-Paredes, Hane Aung, Nadia Bianchi-Berthouze, and Massimiliano Pontil. Multilinear multitask learning. In *International Conference on Machine Learning*, pages 1444–1452. PMLR, 2013.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *NIPS*, pages 3215–3225, 2017.

Alessandro Rudi, Guillermo D. Cañas, and Lorenzo Rosasco. On the sample complexity of subspace learning. In *Advances in Neural Information Processing Systems*, pages 2067–2075, 2013.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *NIPS*, pages 1657–1665, 2015.

Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

E. Senkene and Arkady Tempel'man. Hilbert spaces of operator-valued functions. *Mathematical transactions of the Academy of Sciences of the Lithuanian SSR*, 13(4):665–670, 1973.

Nicholas Sterge, Bharath Sriperumbudur, Lorenzo Rosasco, and Alessandro Rudi. Gain with no pain: Efficiency of kernel-pca by nyström sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 3642–3652. PMLR, 2020.

Joel A. Tropp. User-friendly tools for random matrices: An introduction. Technical report, California Institute of Technology Division of Engineering and Applied Science, 2012.

Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of sgd for least-squares in the interpolation regime. *Advances in Neural Information Processing Systems*, 34:21581–21591, 2021.

Raja Velu and Gregory C Reinsel. *Multivariate reduced-rank regression: theory and applications*, volume 136. Springer Science & Business Media, 2013.

G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in. Applied Mathematics*. Philadelphia: SIAM, 1990.

Jason Weston, Olivier Chapelle, Vladimir Vapnik, André Elisseeff, and Bernhard Schölkopf. Kernel dependency estimation. In *Advances in neural information processing systems*, pages 897–904, 2003.