Journal of Machine Learning Research 1 (2000) 1–48 Submitted 4/00; Published 10/00

Random Fourier Features for Operator-Valued Kernels

Brault Romain ROMAIN.BRAULT@TELECOM-PARISTECH.FR
LTCI
Télécom ParisTech
Paris, 46 rue Barrault, France
Université Paris-Saclay

Florence d'Alché-Buc FLORENCE.DALCHE@TELECOM-PARISTECH.FR
LTCI
Télécom ParisTech
Paris, 46 rue Barrault, France
Université Paris-Saclay

Editor: Francis Bach

Abstract

Many problems in Machine Learning can be cast into vector-valued functions approximation. Operator-Valued Kernels *s* and vector-valued Reproducing Kernel Hilbert Spaces provide a theoretical and practical framework to address that issue, extending nicely the well-known setting of scalar-valued kernels. However large scale applications are usually not affordable with these tools that require an important computational power along with a large memory capacity. In this paper, we propose and study scalable methods to perform regression with *s*. To achieve this goal, we extend Random Fourier Features, an approximation technique originally introduced for scalar-valued kernels, to *s*. The idea is to take advantage of an approximated operator-valued feature map in order to come up with a linear model in a finite-dimensional space.

Keywords: Random Fourier Feature, Operator-Valued Kernel

1. Introduction

This paper is dedicated to the definition of a general and flexible approach to learn vector-valued functions together with an efficient implementation of the learning algorithms. To achieve this goal, we study shallow architectures, namely the product of a (nonlinear) operator-valued feature $\widetilde{\phi}(x)$ and a parameter vector θ such that $\widetilde{f}(x) = \widetilde{\phi}(x)^* \theta$, and combine two appealing methodologies: Operator-Valued Kernel Regression and Random Fourier Features.

Operator-Valued Kernels (Micchelli and Pontil, 2005; Carmeli et al., 2010; Kadri et al., 2010; Brouard et al., 2011; Álvarez et al., 2012) extend the classic scalar-valued kernels to functions with values in some *output* Hilbert space. As in the scalar case, (index˙cmd) are used to build Reproducing Kernel Hilbert Spaces (index˙cmd) in which representer theorems apply as for ridge regression or other appropriate loss functional. In these cases, learning a model in the index˙cmd boils down to learning a function of the form $f(x) =$

$\sum_{i=1}^{N} K(x, x_i)\alpha_i$ where $x_1, \ldots, x_N$ are the training input data and each $\alpha_i, i = 1, \ldots, N$ is a vector of the output space $\mathcal{Y}$, and each $K(x, x_i)$ is an operator on $\mathcal{Y}$.

However, index˙cmd suffer from the same drawbacks as classic (scalar-valued) kernel machines: they scale poorly to large datasets because they are exceedingly demanding in terms of memory and computations. We propose to approximate OVKs by extending a methodology called (index˙cmd) (Rahimi and Recht, 2007; Le et al., 2013; Yang et al., 2015b; Sriperumbudur and Szabo, 2015; Bach, 2015; Sutherland and Schneider, 2015; Rudi et al., 2016) so far developed to speed up scalar-valued kernel machines. The index˙cmd approach linearizes a shift-invariant kernel model by generating explicitly an approximated feature map $\tilde{\varphi}$. index˙cmd has been shown to be efficient on large datasets (Rudi et al., 2016) and has been further improved by efficient matrix computations such as (Le et al., 2013, "FastFood") and (Felix et al., 2016, "SORF"), which are considered as the best large scale implementations of kernel methods, along with Nyström approaches proposed in Drineas and Mahoney (2005). Moreover thanks to index˙cmd, kernel methods have been proved to be competitive with deep architectures (Lu et al., 2014; Dai et al., 2014; Yang et al., 2015a).

1.1 Outline and contributions

The paper is structured as follow. In Section 2 we recall briefly how to obtain index˙cmd for scalar-valued kernels and list the state of the art implementation of index˙cmd for large scale kernel learning. Then we define properly , give some important theorems and properties used throughout this paper before given a non exhaustive list of problem tackled with index˙cmd.

Then we move on to our contributions. In Section 3 we propose an index˙cmd construction from $\mathcal{Y}$-Mercer shift invariant index˙cmd that we call (index˙cmd). Then we study the structure of a random feature corresponding to an index˙cmd (without having to specify the target kernel). Eventually we use the framework used to construct index˙cmd to study the regularization properties of index˙cmd in terms of .

In Section 4 we assess theoretically the quality of our index˙cmd: we show that the stochastic index˙cmd estimator converges with high probability toward the target kernel and derive convergence rates. We also give a bound on the variance of the approximated index˙cmd constructed from the corresponding index˙cmd.

In Section 5 we focus on Ridge regression with index˙cmd. First we study the relationship between finding a minimizer in the index˙cmd induce by a given index˙cmd and the feature induced by the corresponding index˙cmd. Then we define a gradient based algorithm to tackle Ridge regression with index˙cmd, show how to obtain an efficient implementation and study its complexity.

Eventually we end this paper by some numerical experiments in Section 6 on toy and real datasets before giving a general conlusion in Section 7.

2. Background

Notations used throughout this paper are summarized in Table 1.

Table 1: Mathematical symbols and their signification (part 1).

Symbol	Meaning
$e \in \mathcal{X}$	The neutral element of the group $\mathcal{X}$.
δ_{ij}	Kronecker delta function. $\delta_{ij} = 0$ if $i \neq j$, 1 otherwise.
$\widehat{\mathcal{X}}$	The Pontryagin dual of $\mathcal{X}$ when $\mathcal{X}$ is a index˙cmd group.
$\langle \cdot, \cdot \rangle_{\mathcal{Y}}$	The canonical inner product of the Hilbert space $\mathcal{Y}$.
$\|\cdot\|_{\mathcal{Y}}$	The canonical norm induced by the inner product of the Hilbert space $\mathcal{Y}$.
$\mathcal{F}(\mathcal{X}; \mathcal{Y})$	Topological vector space of functions from $\mathcal{X}$ to $\mathcal{Y}$.
$\mathcal{C}(\mathcal{X}; \mathcal{Y})$	The topological vector subspace of $\mathcal{F}$ of continuous functions from $\mathcal{X}$ to $\mathcal{Y}$.
$\mathcal{L}(\mathcal{H}; \mathcal{Y})$	The space bounded linear operator from a Hilbert space $\mathcal{H}$ to a Hilbert space $\mathcal{Y}$.
$\|\cdot\|_{\mathcal{Y},\mathcal{Y}'}$	The operator norm $\|\Gamma\|_{\mathcal{Y},\mathcal{Y}'} = \sup_{\|y\|_{\mathcal{Y}}=1} \|\Gamma y\|_{\mathcal{Y}'}$ for all $\Gamma \in \mathcal{L}(\mathcal{Y}, \mathcal{Y}')$
$\mathcal{M}_{m,n}(\mathbb{K})$	The space of matrices of size (m, n).
$\mathcal{L}(\mathcal{Y})$	The space of bounded linear operator from a Hilbert space $\mathcal{Y}$ to itself.
$\mathcal{L}_+(\mathcal{Y})$	The space of non-negative bounded linear operator from a Hilbert space $\mathcal{H}$ to itself.
$\mathcal{B}(\mathcal{X})$	Borel σ-algebra on a topological space $\mathcal{X}$.
$\mathbf{Leb}(\mathcal{X})$	The Lebesgue measure of $\mathcal{X}$.
$\mathbf{Haar}(\mathcal{X})$	A Haar measure of $\mathcal{X}$.
$\mathbf{Pr}_{\mu,\rho}(\mathcal{X})$	A probability measure of $\mathcal{X}$ whose Radon-Nikodym derivative (density) with respect to the measure μ is ρ.
$\mathcal{F}[\cdot]$	The operator.
$L^p(\mathcal{X}, \mu; \mathcal{Y})$	The Banach space of $\|\cdot\|_{\mathcal{Y}}^p$ (Bochner)-integrable function from $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mu)$ to $\mathcal{Y}$ for $p \in \mathbb{R}_+$. $L^p(\mathcal{X}, \mu, \mathbb{R}) := L^p(\mathcal{X}, \mu)$ and $L^p(\mathcal{X}, \mu, \mathbb{R}) = L^p(\mathcal{X}, \mu)$.
$\bigoplus_{j=1}^D x_i$	The direct sum of $D \in \mathbb{N}$ vectors x_i's in the Hilbert spaces $\mathcal{H}_i$. By definition $\langle \bigoplus_{j=1}^D x_j, \bigoplus_{j=1}^D z_j \rangle = \sum_{j=1}^D \langle x_j, z_j \rangle_{\mathcal{H}_i}$.
$\|\cdot\|_p$	The $L^p(\mathcal{X}, \mu, \mathcal{Y})$ norm. $\|f\|_p^p := \int_{\mathcal{X}} \|f(x)\|_{\mathcal{Y}}^p d\mu(x)$. When $\mathcal{X} = \mathbb{N}^*$, $\mathcal{Y} \subseteq \mathbb{R}$ and μ is the counting measure and $p = 2$ it coincide with the Euclidean norm $\|\cdot\|_2$ for finite dimensional vectors.
$\|\cdot\|_\infty$	The uniform norm $\|f\|_\infty = \operatorname{ess\,sup} \{ \|f(x)\|_{\mathcal{Y}} \mid x \in \mathcal{X} \} = \lim_{p \to \infty} \|f\|_p$.
$\|\Gamma\|$	The absolute value of the linear operator $\Gamma \in \mathcal{L}(\mathcal{Y})$, index˙cmd $\|\Gamma\|^2 = \Gamma^* \Gamma$.
$\mathbf{Tr}[\Gamma]$	The trace of a linear operator $\Gamma \in \mathcal{L}(\mathcal{Y})$.
$\|\cdot\|_{\sigma,p}$	The Schatten p-norm, $\|\Gamma\|_{\sigma,p}^p = \mathbf{Tr}[|\Gamma|^p]$ for $\Gamma \in \mathcal{L}(\mathcal{Y})$, where $\mathcal{Y}$ is a Hilbert space. Note that $\|\Gamma\|_{\sigma,\infty} = \rho(\Gamma) \leq \|\Gamma\|_{\mathcal{Y},\mathcal{Y}}$.
$\succcurlyeq$	"Greater than" in the Loewner partial order of operators. $\Gamma_1 \succcurlyeq \Gamma_2$ if $\sigma(\Gamma_1 - \Gamma_2) \subseteq \mathbb{R}_+$.
$\cong$	Given two sets $\mathcal{X}$ and $\mathcal{Y}$, $\mathcal{X} \cong \mathcal{Y}$ if there exists an isomorphism $\varphi : \mathcal{X} \to \mathcal{Y}$.

2.1 Random Fourier Feature maps

methodology introduced by Rahimi and Recht (2007) provides a way to scale up kernel methods when kernels are Mercer and *translation-invariant*. We view the input space $\mathcal{X}$ as a group endowed with the addition. Extensions to other group laws such as Li et al. (2010) are described in Subsection 3.2.2 within the general framework of operator-valued kernels.

Denote $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a positive definite kernel (Aronszajn, 1950) on $\mathbb{R}^d$. A kernel k is said to be *shift-invariant* or *translation-invariant* for the addition if for for all $(x, z, t) \in \left(\mathbb{R}^d\right)^3$ we have $k(x+t, z+t) = k(x, z)$. Then, we define $k_0 : \mathbb{R}^d \to \mathbb{R}$ the function such that $k(x, z) = k_0(x - z)$. k_0 is called the *signature* of kernel k. If k_0 is a continuous function we call the kernel "Mercer". Then, Bochner's theorem (Folland, 1994) is the theoretical result that leads to the Random Fourier Features.

Theorem 1 (Bochner's theorem)

Any continuous positive-definite function (index˙cmd a Mercer kernel) is the of a bounded non-negative Borel measure.

It implies that any positive-definite, continuous and shift-invariant kernel k, have a continuous and positive-definite signature k_0, which is the $\mathcal{F}$ of a non-negative measure μ. We therefore have the $k(x, z) = k_0(x - z) = \int_{\mathbb{R}^d} \exp(-\mathrm{i}\langle\omega, x - z\rangle) d\mu(\omega) = \mathcal{F}[k_0](\omega)$. Moreover $\mu = \mathcal{F}^{-1}[k_0]$. Without loss of generality, we assume that μ is a probability measure, index˙cmd $\int_{\mathbb{R}^d} d\mu(\omega) = 1$ by renormalizing the kernel since $\int_{\mathbb{R}^d} d\mu(\omega) = \int_{\mathbb{R}^d} \exp(-\mathrm{i}\langle\omega, 0\rangle) d\mu(\omega) = k_0(0)$. and we can write the kernel as an expectation over a probability measure μ. For all x, $z \in \mathbb{R}^d$

$$k_0(x - z) = \mathbf{E}_{\omega\sim\mu}\left[\exp(-\mathrm{i}\langle\omega, x - z\rangle)\right].$$

Eventually, if k is real valued we only write the real part, $k(x, z) = \mathbf{E}_{\omega\sim\mu}[\cos\langle\omega, x - z\rangle] = \mathbf{E}_{\omega\sim\mu}[\cos\langle\omega, z\rangle \cos\langle\omega, x\rangle + \sin\langle\omega, z\rangle \sin\langle\omega, x\rangle]$. Let $\bigoplus_{j=1}^{D} x_j$ denote the Dd-length column vector obtained by stacking vectors $x_j \in \mathbb{R}^d$. The feature map $\widetilde{\varphi} : \mathbb{R}^d \to \mathbb{R}^{2D}$ defined as

$$\widetilde{\varphi}(x) = \frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} \begin{pmatrix} \cos\langle x, \omega_j\rangle \\ \sin\langle x, \omega_j\rangle \end{pmatrix}, \quad \omega_j \sim \mathcal{F}^{-1}[k_0] \text{ index˙cmd} \tag{1}$$

is called a *Random Fourier Feature* (map). Each $\omega_j, j = 1, \ldots, D$ is independently and identically sampled from the inverse Fourier transform μ of k_0. This Random Fourier Feature map provides the following Monte-Carlo estimator of the kernel: $\widetilde{k}(x, z) = \widetilde{\varphi}(x)^* \widetilde{\varphi}(z)$. Using trigonometric identities, Rahimi and Recht (2007) showed that the same feature map can also be written

$$\tilde{\varphi}(x) = \sqrt{\frac{2}{D}} \bigoplus_{j=1}^{D} \Big(\cos(\langle x, \omega_j\rangle + b_j)\Big), \tag{2}$$

where $\omega_j \sim \mathcal{F}^{-1}[k_0]$, $b_j \sim \mathcal{U}(0, 2\pi)$ index˙cmd. The feature map defined by Equation 1 and Equation 2 have been compared in Sutherland and Schneider (2015) where they give the condition under wich Equation 1 has lower variance than Equation 2. For instance for the Gaussian kernel, Equation 1 has always lower variance. In practice, Equation 2 is easier to program. In this paper we focus on random Fourier feature of the form given in Equation 1.

The dimension D governs the precision of this approximation, whose uniform convergence towards the target kernel can be found in Rahimi and Recht (2007) and in more recent papers with some refinements proposed in Sutherland and Schneider (2015) and Sriperumbudur and Szabo (2015). Finally, it is important to notice that Random Fourier Feature approach *only* requires two steps before the application of a learning algorithm: (1) define the inverse Fourier transform of the given shift-invariant kernel, (2) compute the randomized feature map using the spectral distribution μ. Rahimi and Recht (2007) show that for the Gaussian kernel $k_0(x-z) = \exp(-\gamma\|x-z\|_2^2)$, the spectral distribution μ is a Gaussian distribution. For the Laplacian kernel $k_0(x-z) = \exp(-\gamma\|x-z\|_1)$, the spectral distribution is a Cauchy distribution.

2.1.1 Extensions of the RFF method

The seminal idea of Rahimi and Recht (2007) has opened a large literature on random features. Nowadays, many classes of kernels other than translation invariant are now proved to have an efficient random feature representation. Kar and Karnick (2012) proposed random feature maps for dot product kernels (rotation invariant) and Hamid et al. (2014) improved the rate of convergence of the approximation error for such kernels by noticing that feature maps for dot product kernels are usually low rank and may not utilize the capacity of the projected feature space efficiently. Pham and Pagh (2013) proposed fast random feature maps for polynomial kernels.

Li et al. (2010) generalized the original index˙cmd of Rahimi and Recht (2007). Instead of computing feature maps for shift-invariant kernels on the additive group $(\mathbb{R}^d, +)$, they used the generalized Fourier transform on any locally compact abelian group to derive random features on the multiplicative group $(\mathbb{R}^d, *)$. In the same spirit Yang et al. (2014b) noticed that an theorem equivalent to Bochner's theorem exists on the semi-group $(\mathbb{R}^d_{>0}, +)$. From this they derived "Random Laplace" features and used them to approximate kernels adapted to learn on histograms.

To speed-up the convergence rate of the random features approximation, Yang et al. (2014a) proposed to sample the random variable from a quasi Monte-Carlo sequence instead of index˙cmd random variables. Le et al. (2013) proposed the "Fastfood" algorithm to reduce the complexity of computing a index˙cmd –using structured matrices and a fast Walsh-Hadarmard transform– from $O_t(Dd)$ to $O_t(D\log(d))$. More recently Felix et al. (2016) proposed also an algorithm "SORF" to compute Gaussian index˙cmd in $O_t(D\log(d))$ but with better convergence rates than "Fastfood" (Le et al., 2013). Mukuta and Harada (2016) proposed a data dependent feature map (comparable to the Nystrom̈ method) by estimating the distribution of the input data, and then finding the eigenfunction decomposition of Mercer's integral operator associated to the kernel.

In the context of large scale learning and deep learning, Lu et al. (2014) showed that index˙cmd can achieve performances comparable to deep-learning methods by combining multiple kernel learning and composition of kernels along with a scalable parallel implementation. Dai et al. (2014) and Xie et al. (2015) combined index˙cmd and stochastic gradient descent to define an online learning algorithm called "Doubly stochastic gradient descent" adapted to large scale learning. Yang et al. (2015a) proposed and studied the idea of re-

placing the last fully interconnected layer of a deep convolutional neural network (LeCun et al., 1995) by the "Fastfood" implementation of index˙cmd.

Eventually Yang et al. (2015b) introduced the algorithm "À la Carte", based on "Fastfood" which is able to learn the spectral distribution

2.2 On Operator-Valued Kernels

We now introduce the theory of (index˙cmd) that provides a flexible framework to study and learn vector-valued functions. The fundations of the general theory of scalar kernels is mostly due to Aronszajn (1950) and provides a unifying point of view for the study of an important class of Hilbert spaces of real or complex valued functions. It has been first applied in the theory of partial differential equation. The theory of (index˙cmd) which extends the scalar-valued kernel was first developped by Pedrick (1957) in his Ph. D Thesis. Since then it has been successfully applied to machine learning by many authors. In particular we introduce the notion of following the propositions of Micchelli and Pontil (2005); Carmeli et al. (2006, 2010).

2.3 Definitions and properties

In machine learning the goal is often to find a function f belonging to a class of functions $\mathcal{F}(\mathcal{X};\mathcal{Y})$ that minimizes a criterion called the true risk. The class of functions we consider are functions living in a Hilbert space $\mathcal{H} \subset \mathcal{F}(\mathcal{X};\mathcal{Y})$. The completeness allows to consider sequences of functions $f_n \in \mathcal{H}$ where the limit $f_n \to f$ is in $\mathcal{H}$. Moreover the existence of an inner product gives rise to a norm and also makes $\mathcal{H}$ a metric space.

Among all these functions $f \in \mathcal{H}$, we consider a subset of functions $f \in \mathcal{H}_K \subset \mathcal{H}$ such that the evaluation map $\mathrm{ev}_x : f \mapsto f(x)$ is bounded for all x. index˙cmd such that $\|\mathrm{ev}_x\|_{\mathcal{H}_K} \leq C_x \in \mathbb{R}$ for all x. For scalar valued kernels the evaluation map is a linear functional. Thus by Riesz's representation theorem there is an isomorphism between evaluating a function at a point and an inner product: $f(x) = \mathrm{ev}_x f = \langle K_x, f\rangle_K$. From this we deduce the reproducing property $K(x,z) = \langle K_x, K_z\rangle_K$ which is the cornerstone of many proofs in machine learning and functional analysis. When dealing with vector-valued functions, the evaluation map ev_x is no longer a linear functional, since it is vector-valued. However, inspired by the theory of scalar valued kernel, many authors showed that if the evaluation map of functions with values in a Hilbert space $\mathcal{Y}$ is bounded, a similar reproducing property can be obtained; namely $\langle y', K(x,z)y\rangle = \langle K_x y', K_z y\rangle_K$ for all y, $y' \in \mathcal{Y}$. This motivates the following definition of a (index˙cmd).

Definition 2 ((Carmeli et al., 2006; Micchelli and Pontil, 2005))

Let $\mathcal{Y}$ be a (real or complex) Hilbert space. A on a locally compact second countable topological space $\mathcal{X}$ is a Hilbert space $\mathcal{H}$ such that

1. *the elements of $\mathcal{H}$ are functions from $\mathcal{X}$ to $\mathcal{Y}$ (index˙cmd $\mathcal{H} \subset \mathcal{F}(\mathcal{X},\mathcal{Y})$);*
2. *for all $x \in \mathcal{X}$, there exists a positive constant C_x such that for all $f \in \mathcal{H}$ $\|f(x)\|_{\mathcal{Y}} \leq C_x \|f\|_{\mathcal{H}}$.*

Throughout this section we show that a index˙cmd defines a unique positive-definite function called (index˙cmd) and conversely an index˙cmd uniquely defines a index˙cmd. The bijection between index˙cmd and index˙cmd has been first proved by Senkene and Tempel'man (1973) in 1973. In this introduction to index˙cmd we follow the definitions and most recent proofs of Carmeli et al. (2010).

Definition 3 (Positive-definite)

Given $\mathcal{X}$ a locally compact second countable topological space and $\mathcal{Y}$ a real Hilbert Space, a map $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ is called a positive-definite kernel if $K(x, z) = K(z, x)^*$ and

$$\sum_{i,j=1}^{N} \langle K(x_i, x_j)y_j, y_i\rangle_{\mathcal{Y}} \geq 0, \tag{3}$$

for all $N \in \mathbb{N}$, for all sequences of points $(x_i)_{i=1}^N$ in $\mathcal{X}^N$, and all sequences of points $(y_i)_{i=1}^N$ in $\mathcal{Y}^N$.

As in the scalar case any defines a unique positive-definite and conversely a positive-definite defines a unique .

Proposition 4 ((Carmeli et al., 2006))

Given a there is a unique positive-definite $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$.

iven $x \in \mathcal{X}$, $K_x : \mathcal{Y} \to \mathcal{F}(\mathcal{X}; \mathcal{Y})$ denotes the linear operator whose action on a vector y is the function $K_x y \in \mathcal{F}(\mathcal{X}; \mathcal{Y})$ defined for all $z \in \mathcal{X}$ by $K_x = \mathrm{ev}_x^*$. As a consequence we have that

$$K(x, z)y = \mathrm{ev}_x \mathrm{ev}_z^* y = K_x^* K_z y = (K_z y)(x). \tag{4}$$

Some direct consequences follow from the definition.

1. The kernel reproduces the value of a function $f \in \mathcal{H}$ at a point $x \in \mathcal{X}$ since for all $y \in \mathcal{Y}$ and $x \in \mathcal{X}$, $\mathrm{ev}_x^* y = K_x y = K(\cdot, x)y$ such that $\langle f(x), y\rangle_{\mathcal{Y}} = \langle f, K(\cdot, x)y\rangle_{\mathcal{H}} = \langle K_x^* f, y\rangle_{\mathcal{Y}}$.
2. For all $x \in \mathcal{X}$ and all $f \in \mathcal{H}$, $\|f(x)\|_{\mathcal{Y}} \leq \sqrt{\|K(x, x)\|_{\mathcal{Y},\mathcal{Y}}}\|f\|_{\mathcal{H}}$. This comes from the fact that $\|K_x\|_{\mathcal{Y},\mathcal{H}} = \|K_x^*\|_{\mathcal{H},\mathcal{Y}} = \sqrt{\|K(x, x)\|_{\mathcal{Y},\mathcal{Y}}}$ and the operator norm is sub-multiplicative.

Additionally given a positive-definite , it defines a unique index˙cmd.

Proposition 5 ((Carmeli et al., 2006))

Given a positive-definite $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$, there is a unique $\mathcal{H}$ on $\mathcal{X}$ with reproducing kernel K.

Since an positive-definite defines a unique (index˙cmd) and conversely a index˙cmd defines a unique , we denote the Hilbert space $\mathcal{H}$ endowed with the scalar product $\langle \cdot, \cdot \rangle$ respectively $\mathcal{H}_K$ and $\langle \cdot, \cdot \rangle_K$. From now we refer to positive-definite or reproducing as . As a consequence, given K an , define $K_x = K(\cdot, x)$ we have

$$K(x, z) = K_x^* K_z \;\; \forall x, z \in \mathcal{X}\,, \tag{5a}$$

$$\mathcal{H}_K = \overline{\mathrm{span}}\; \{\, K_x y \mid \forall x \in \mathcal{X},\;\; \forall y \in \mathcal{Y}\,\}\,. \tag{5b}$$

Where $\overline{\text{span}}$ is the closed span of a given set. Another way to describe functions of $\mathcal{H}_K$ consists in using a suitable feature map.

Proposition 6 (Feature Operator (Carmeli et al., 2010))

Let $\mathcal{H}$ be any Hilbert space and $\phi : \mathcal{X} \to \mathcal{L}(\mathcal{Y}; \mathcal{H})$, with $\phi_x := \phi(x)$. Then the operator $W : \mathcal{H} \to \mathcal{F}(\mathcal{X}; \mathcal{Y})$ defined for all $g \in \mathcal{H}$, and for all $x \in \mathcal{X}$ by $(Wg)(x) = \phi_x^* g$ is a partial isometry from $\mathcal{H}$ onto the index˙cmd $\mathcal{H}_K$ with reproducing kernel $K(x, z) = \phi_x^* \phi_z$, $\forall x, z \in \mathcal{X}$. W^*W is the orthogonal projection onto $(Ker\ W)^\perp = \overline{\text{span}}\ \{\ \phi_x y \mid \forall x \in \mathcal{X},\ \forall y \in \mathcal{Y}\ \}$. Then $\|f\|_K = \inf \{\ \|g\|_{\mathcal{H}} \mid \forall g \in \mathcal{H},\ Wg = f\ \}$.

In this work we mainly focus on the kernel functions inducing a index˙cmd of continuous functions. Such kernel are named $\mathcal{Y}$-Mercer kernels and generalize Mercer kernels.

Definition 7 ($\mathcal{Y}$-Mercer kernel Carmeli et al. (2010))

A positive definite index˙cmd $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ is called $\mathcal{Y}$-Mercer if the associated index˙cmd $\mathcal{H}_K$ is a subspace of the space of continuous functions $\mathcal{C}(\mathcal{X}; \mathcal{Y})$.

2.4 Examples of

In this subsection we list some (index˙cmd) that have been used successfully in the litterature. We do not recall the proof that the following kernels are well defined are refer the interrested reader to the respective authors original work.

index˙cmd have been first introduced in Machine Learning to solve multi-task regression problems. Multi-task regression is encountered in many fields such as structured classification when classes belong to a hierarchy for instance. Instead of solving independently p single output regression task, one would like to take advantage of the relationships between output variables when learning and making a decision.

Proposition 8 (Decomposable kernel (Micheli and Glaunes, 2013))

Let Γ be a non-negative operator of $\mathcal{L}_+(\mathcal{Y})$. K is said to be a decomposable *kernel*[1] *if for all $(x, z) \in \mathcal{X}^2$, $K(x, z) := k(x, z)\Gamma$, where k is a* scalar *kernel.*

When $\mathcal{Y} = \mathbb{R}^p$, the operator Γ can be represented by a matrix which can be interpreted as encoding the relationships between the outputs coordinates. If a graph coding for the proximity between tasks is known, then it is shown in Evgeniou et al. (2005); Baldassarre et al. (2010); Álvarez et al. (2012) that Γ can be chosen equal to the pseudo inverse $L^\dagger$ of the graph Laplacian such that the norm in $\mathcal{H}_K$ is a graph-regularizing penalty for the outputs (tasks). When no prior knowledge is available, Γ can be set to the empirical covariance of the output training data or learned with one of the algorithms proposed in the literature (Dinuzzo et al., 2011; Sindhwani et al., 2013; Lim et al., 2015a). Another interesting property of the decomposable kernel is its universality (a kernel which may approximate an arbitrary continuous target function uniformly on any compact subset of the input space). A reproducing kernel K is said *universal* if the associated index˙cmd $\mathcal{H}_K$ is *dense* in the space of continuous functions $\mathcal{C}(\mathcal{X}, \mathcal{Y})$. The conditions for a kernel to be universal have been discussed in Caponnetto et al. (2008); Carmeli et al. (2010). In

1. Some authors also refer to as *separable* kernels.

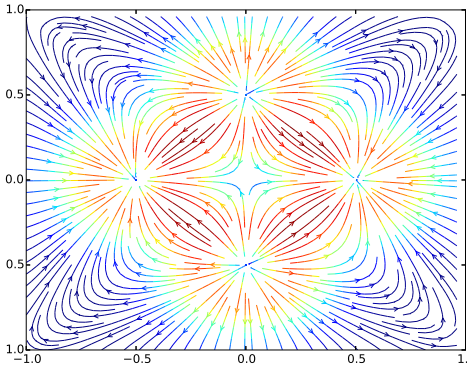

Figure 1: Synthetic $2D$ curl-free field .

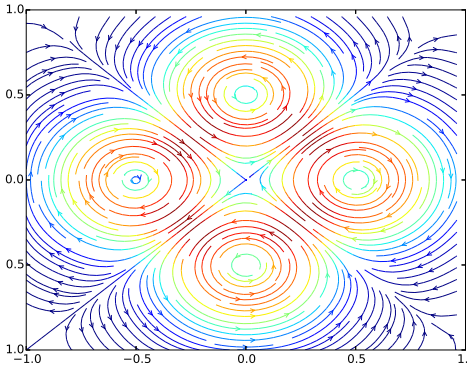

Figure 2: Synthetic $2D$ divergence-free field .

particular they show that a decomposable kernel is universal provided that the scalar kernel k is universal and the operator Γ is injective. Given $(e_k)_{k=1}^p$ a basis of $\mathcal{Y}$, we recall here how the matrix Γ act as a regularizer between the components of the outputs $f_k = \langle f(\cdot), e_k \rangle_{\mathcal{Y}}$ of a function $f \in \mathcal{H}_K$. We prove a generalized version of Proposition 9 to any in Subsection 3.6.

Proposition 9 (Kernels and Regularizers (Álvarez et al., 2012))

Let $K(x, z) := k(x, z)\Gamma$ for all x, $z \in \mathcal{X}$ be a decomposable kernel where Γ is a matrix of size $p \times p$. Then for all $f \in \mathcal{H}_K$, $\|f\|_K = \sum_{i,j=1}^p \left(\Gamma^\dagger\right)_{ij} \langle f_i, f_j \rangle_k$ where $f_i = \langle f, e_i \rangle$ (resp $f_j = \langle f, e_j \rangle$), denotes the i-th (resp j-th) component of f.

Curl-free and divergence-free kernels provide an interesting application of operator-valued kernels (Macedo and Castro, 2008; Baldassarre et al., 2012; Micheli and Glaunes, 2013) to *vector field* learning, for which input and output spaces have the same dimensions ($d = p$). Applications cover shape deformation analysis (Micheli and Glaunes, 2013) and magnetic fields approximations (Wahlström et al., 2013). These kernels discussed in (Fuselier, 2006) allow encoding input-dependent similarities between vector-fields. An illustration of a synthetic $2D$ curl-free and divergence free fields are given respectively in Figure 1 and Figure 2. To obain the curl-free field we took the gradient of a mixture of five two dimensional Gaussians (since the gradient of a potential is always curl-free). We generated the divergence-free field by taking the orthogonal of the curl-free field.

Proposition 10 (Curl-free and Div-free kernel (Macedo and Castro, 2008))

Assume $\mathcal{X} = (\mathbb{R}^d, +)$ and $\mathcal{Y} = \mathbb{R}^p$ with $d = p$. The divergence-free *kernel is defined as $K^{div}(x, z) = K_0^{div}(\delta) = (\nabla\nabla^\mathsf{T} - \Delta I)k_0(\delta)$ and the* curl-free *kernel as $K^{curl}(x, z) = K_0^{curl}(\delta) = -\nabla\nabla^\mathsf{T} k_0(\delta)$, where ∇ is the gradient operator , $\nabla\nabla^\mathsf{T}$ is the Hessian operator and Δ is the Laplacian operator.*

2.5 Shift-Invariant index˙cmd on index˙cmd groups

The main subjects of interest of the present paper are shift-invariant . When referring to a shift-invariant index˙cmd $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ we assume that $\mathcal{X}$ is a locally compact second countable topological group with identity e.

Definition 11 (Shift-invariant index˙cmd)

A reproducing $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ *is called shift-invariant if for all* x, z, $t \in \mathcal{X}$, $K(x \star t, z \star t) = K(x, z)$.

A shift-invariant kernel can be characterized by a function of one variable K_e called the signature of K. Here e denotes the neutral element of the index˙cmd group $\mathcal{X}$ endowed with the binary group operation $\star$.

Proposition 12 (Kernel signature (Carmeli et al., 2010))

Let $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ *be a reproducing kernel. The following conditions are equivalents.*

1. K *is a positive-definite shift-invariant .*
2. *There is a positive-definite function* $K_e : \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ *such that* $K(x, z) = K_e(z^{-1} \star x)$.

If one of the above conditions is satisfied, then $\|K(x, x)\|_{\mathcal{Y}} = \|K_e(e)\|_{\mathcal{Y}}$, $\forall x \in \mathcal{X}$.

The notation K_e for the function of completely positive type associated with the reproducing kernel K is consistent with the definition given by Equation 4 since for all $x \in \mathcal{X}$ and all $y \in \mathcal{Y}$, $(K_e y)(x) = K_e(x)y$. We recall that if K is a $\mathcal{Y}$-Mercer kernel, there is a function K_e such that for all x and $z \in \mathcal{X}$, $K(x, z) = K_e(x \star z^{-1})$. Then an index˙cmd K is $\mathcal{Y}$-Mercer if and only if for all $y \in \mathcal{Y}$, $K_e(\cdot)y \in \mathcal{C}(\mathcal{X}; \mathcal{Y})$. In other words a $\mathcal{Y}$-Mercer kernel is nothing but a functions whose signature is continuous and positive definite (Carmeli et al., 2010), which fulfil the conditions required for the "operator-valued" Bochner theorem to apply. Note that if K is a shift invariant $\mathcal{Y}$-Mercer kernel, then $\mathcal{H}_K$ contains continuous *bounded* functions (Carmeli et al., 2010).

2.6 Some applications of Operator-valued kernels

We give here a non exhaustive list of works concerning . A good review of has been conducted in Álvarez et al. (2012). For a theoretical introduction to index˙cmd the interested reader can refer to the papers Carmeli et al. (2006); Caponnetto et al. (2008); Carmeli et al. (2010). Generalization bounds for index˙cmd have been studied in Sindhwani et al. (2013); Kadri et al. (2015); Sangnier et al. (2016); Maurer (2016). Operator-valued Kernel Regression has first been studied in the context of Ridge Regression and Multi-task learning by Micchelli and Pontil (2005). Multi-task regression (Micchelli and Pontil, 2004) and structured multi-class classification (Dinuzzo et al., 2011; Minh et al., 2013b; Mroueh et al., 2012) are undoubtedly the first target applications for working in . have been shown useful to provide a general framework for structured output prediction (Brouard et al., 2011, 2016a) with a link to Output Kernel Regression (Kadri et al., 2013). Beyond structured classification, other various applications such as link prediction, drug activity prediction or recently metabolite identification (Brouard et al., 2016b) and image colorization (Ha Quang et al., 2010) have been developed.

Macedo and Castro (2008); Baldassarre et al. (2012) showed the interest of spectral algorithms in Ridge regression and introduced vector field learning as a new multiple output task in Machine Learning community. Wahlström et al. (2013) applied vector field learning with index˙cmd-based Gaussian processes to the reconstruction of magnetic fields (which are

curl-free). The works of Kadri et al. (2010, 2015) have been the precursors of regression with functional values, opening a new avenue of applications. Appropriate algorithms devoted to on-line learning have been also derived by Audiffren and Kadri (2015). Kernel learning was addressed at least in two ways: first with using Multiple Kernel Learning in Kadri et al. (2012) and second, using various penalties, smooth ones in Dinuzzo et al. (2011); Ciliberto et al. (2015) for decomposable kernels and non smooth ones in Lim et al. (2015b) using proximal methods in the case of decomposable and transformable kernels. Dynamical modeling was tackled in the context of multivariate time series modelling in Lim et al. (2013); Sindhwani et al. (2013); Lim et al. (2015b) and as a generalization of Recursive Least Square Algorithm in Amblard and Kadri (2015). Sangnier et al. (2016) recently explored the minimization of a pinball loss under regularizing constraints induced by a well chosen decomposable kernel in order to handle joint quantile regression.

3. Main contribution: Operator Random Fourier Features

We present in this section a construction methodology devoted to shift-invariant $\mathcal{Y}$-Mercer operator-valued kernels defined on any (index˙cmd) group, noted $(\mathcal{X}, \star)$, for some operation noted $\star$. This allows us to use the general context of Pontryagin duality for of functions on index˙cmd groups. Building upon a generalization of the celebrated Bochner's theorem for operator-valued measures, an operator-valued kernel is seen as the of an operator-valued positive measure. From that result, we extend the principle of index˙cmd for scalar-valued kernels and derive a general methodology to build (index˙cmd) when operator-valued kernels are shift-invariant according to the chosen group operation. Elements of this paper have been developped in Brault et al. (2016).

We present a construction of feature maps called (index˙cmd), such that $f : x \mapsto \widetilde{\phi}(x)^* \theta$ is a continuous function that maps an arbitrary index˙cmd group $\mathcal{X}$ as input space to an arbitrary output Hilbert space $\mathcal{Y}$. First we define a functional *Fourier feature map*, and then propose a Monte-Carlo sampling from this feature map to construct an approximation of a shift-invariant $\mathcal{Y}$-Mercer kernel. Then, we prove the convergence of the kernel approximation $\tilde{K}(x, z) = \widetilde{\phi}(x)^* \widetilde{\phi}(z)$ with high probability on *compact* subsets of the index˙cmd $\mathcal{X}$. Eventually we conclude with some numerical experiments.

3.1 Theoretical study

The following proposition of Zhang et al. (2012); Neeb (1998) extends Bochner's theorem to any shift-invariant $\mathcal{Y}$-Mercer kernel. We give a short intoduction to index˙cmd groups and abstract harmonic analysis in appendix A. For the sake of simplicity, the reader can take $\overline{(x, \omega)} = \overline{\exp(\mathrm{i}\langle x, \omega \rangle_2)} = \exp(-\mathrm{i}\langle x, \omega \rangle_2)$, when $x \in \mathcal{X} = (\mathbb{R}^d, +)$.

Proposition 13 (Operator-valued Bochner's theorem (Zhang et al., 2012; Neeb, 1998))

If a function K from $\mathcal{X} \times \mathcal{X}$ to $\mathcal{Y}$ is a shift-invariant $\mathcal{Y}$-Mercer kernel on $\mathcal{X}$, then there exists a unique positive projection-valued measure $\widehat{Q} : \mathcal{B}(\mathcal{X}) \to \mathcal{L}_+(\mathcal{Y})$ such that for all x, $z \in \mathcal{X}$,

$$K(x, z) = \int_{\widehat{\mathcal{X}}} \overline{(x \star z^{-1}, \omega)} d\widehat{Q}(\omega), \tag{6}$$

where $\widehat{Q}$ belongs to the set of all the projection-valued measures of bounded variation on the σ-algebra of Borel subsets of $\widehat{\mathcal{X}}$. Conversely, from any positive operator-valued measure $\widehat{Q}$, a shift-invariant kernel K can be defined by Equation 6.

Although this theorem is central to the spectral decomposition of shift-invariant $\mathcal{Y}$-Mercer index˙cmd, the following results proved by Carmeli et al. (2010) provides insights about this decomposition that are more relevant in practice. It first gives the necessary conditions to build shift-invariant $\mathcal{Y}$-Mercer kernel with a pair $(A, \widehat{\mu})$ where A is an operator-valued function on $\widehat{\mathcal{X}}$ and $\widehat{\mu}$ is a real-valued positive measure on $\widehat{\mathcal{X}}$. Note that obviously such a pair is not unique and the choice of this paper may have an impact on theoretical properties as well as practical computations. Secondly it also states that any index˙cmd have such a spectral decomposition when $\mathcal{Y}$ is finite dimensional or $\mathcal{X}$.

Proposition 14 (Carmeli et al. (2010))
Let $\widehat{\mu}$ be a positive measure on $\mathcal{B}(\widehat{\mathcal{X}})$ and $A : \widehat{\mathcal{X}} \to \mathcal{L}(\mathcal{Y})$ such that $\langle A(\cdot)y, y' \rangle \in L^1(\mathcal{X}, \widehat{\mu})$ for all $y, y' \in \mathcal{Y}$ and $A(\omega) \succcurlyeq 0$ for $\widehat{\mu}$-almost all $\omega \in \widehat{\mathcal{X}}$. Then, for all $\delta \in \mathcal{X}$,

$$K_e(\delta) = \int_{\widehat{\mathcal{X}}} \overline{(\delta, \omega)} A(\omega) d\widehat{\mu}(\omega) \tag{7}$$

is the kernel signature of a shift-invariant $\mathcal{Y}$-Mercer kernel K such that $K(x, z) = K_e(x \star z^{-1})$. The index˙cmd $\mathcal{H}_K$ is embed in $L^2(\widehat{\mathcal{X}}, \widehat{\mu}; \mathcal{Y}')$ by means of the feature operator

$$(Wg)(x) = \int_{\widehat{\mathcal{X}}} \overline{(x, \omega)} B(\omega) g(\omega) d\widehat{\mu}(\omega), \tag{8}$$

Where $B(\omega)B(\omega)^* = A(\omega)$ and both integrals converge in the weak sense. If $\mathcal{Y}$ is finite dimensional or $\mathcal{X}$ is compact, any shift-invariant kernel is of the above form for some pair $(A, \widehat{\mu})$.

When $p = 1$ one can always assume A is reduced to the scalar 1, $\widehat{\mu}$ is still a bounded positive measure and we retrieve the Bochner theorem applied to the scalar case (Theorem 1).

Proposition 14 shows that a pair $(A, \widehat{\mu})$ entirely characterizes an index˙cmd. Namely a given measure $\widehat{\mu}$ and a function A such that $\langle y', A(.)y \rangle \in L^1(\mathcal{X}, \widehat{\mu})$ for all y, $y' \in \mathcal{Y}$ and $A(\omega) \succcurlyeq 0$ for $\widehat{\mu}$-almost all ω, give rise to an index˙cmd. Since $(A, \widehat{\mu})$ determine a unique kernel we can write $\mathcal{H}_{(A,\widehat{\mu})} \Longrightarrow \mathcal{H}_K$ where K is defined as in Equation 7. However the converse is not true: Given a $\mathcal{Y}$-Mercer shift invariant , there exist infinitely many pairs $(A, \widehat{\mu})$ that characterize an index˙cmd.

The main difference between Equation 6 and Equation 7 is that the first one characterizes an index˙cmd by a unique (index˙cmd), while the second one shows that the index˙cmd that uniquely characterizes a $\mathcal{Y}$-Mercer index˙cmd has an operator-valued density with respect to a *scalar* measure $\widehat{\mu}$; and that this operator-valued density is not unique.

Finally Proposition 14 does not provide any *constructive* way to obtain the pair $(A, \widehat{\mu})$ that characterizes an index˙cmd. The following Subsection 3.1.1 is based on another proposition of Carmeli et al. and shows that if the kernel signature $K_e(\delta)$ of an $index_cmd$ is in L^1 then it is possible to construct *explicitly* a pair $(C, \widehat{\mathbf{Haar}})$ from it. Additionally, we show

that we can always extract a scalar-valued *probability* density function from C such that we obtain a pair $(A, \mathbf{Pr}_{\widehat{\mu},\rho})$ where $\mathbf{Pr}_{\widehat{\mu},\rho}$ is a *probability* distribution absolutely continuous with respect to $\widehat{\mu}$ and with associated (index˙cmd) ρ. Thus for all $\mathcal{Z} \subset \mathcal{B}(\widehat{\mathcal{X}})$, $\mathbf{Pr}_{\widehat{\mu},\rho}(\mathcal{Z}) = \int_{\mathcal{Z}} \rho(\omega) d\widehat{\mu}(\omega)$. When the reference measure $\widehat{\mu}$ is the Lebesgue measure, we note $\mathbf{Pr}_{\widehat{\mu},\rho} = \mathbf{Pr}_{\rho}$. For any function $f : \mathcal{X} \times \widehat{\mathcal{X}} \times \mathcal{Y} \to \mathbb{R}$, we also use the notation $\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}[f(x,\omega,y)] = \mathbf{E}_{\omega \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}}[f(x,\omega,y)] = \int_{\widehat{\mathcal{X}}} f(x,\omega,y) d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}(\omega) = \int_{\widehat{\mathcal{X}}} f(x,\omega,y)\rho(\omega) d\widehat{\mathbf{Haar}}(\omega)$. where the two last equalities hold by the "law of the unconscious statistician" (change of variable formula) and the fact that $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ has density ρ.

3.1.1 Sufficient conditions of existence

While Proposition 14 gives some insights on how to build an approximation of a $\mathcal{Y}$-Mercer kernel, we need a theorem that provides an explicit construction of the pair $(A, \mathbf{Pr}_{\widehat{\mu},\rho})$ from the kernel signature K_e. Proposition 14 in Carmeli et al. (2010) gives the solution, and also provides a sufficient condition for Proposition 14 to apply.

Proposition 15 (Carmeli et al. (2010))

Let K be a shift-invariant $\mathcal{Y}$-Mercer kernel of signature K_e. Suppose that for all $z \in \mathcal{X}$ and for all y, $y' \in \mathcal{Y}$, the function $\langle K_e(.)y, y' \rangle_{\mathcal{Y}} \in L^1(\mathcal{X}, \mathbf{Haar})$ where $\mathcal{X}$ is endowed with the group law $\star$. Denote $C : \widehat{X} \to \mathcal{L}(Y)$, the function defined for all $\omega \in \widehat{\mathcal{X}}$ that satisfies for all y, y' in $\mathcal{Y}$:

$$\langle y', C(\omega)y \rangle_{\mathcal{Y}} = \int_{\mathcal{X}} (\delta, \omega) \langle y', K_e(\delta)y \rangle_{\mathcal{Y}} d\mathbf{Haar}(\delta) = \mathcal{F}^{-1}\left[\langle y', K_e(\cdot)y \rangle\right]_{\mathcal{Y}}(\omega). \tag{9}$$

Then

1. *$C(\omega)$ is a bounded non-negative operator for all $\omega \in \widehat{\mathcal{X}}$,*
2. *$\langle y, C(\cdot)y' \rangle_{\mathcal{Y}} \in L^1\left(\widehat{\mathcal{X}}, \widehat{\mathbf{Haar}}\right)$ for all $y, y' \in \mathcal{X}$,*
3. *for all $\delta \in \mathcal{X}$ and for all y, y' in $\mathcal{Y}$, $\langle y', K_e(\delta)y \rangle_{\mathcal{Y}} = \int_{\widehat{\mathcal{X}}} \overline{(\delta, \omega)} \langle y', C(\omega)y \rangle_{\mathcal{Y}} d\widehat{\mathbf{Haar}}(\omega) = \mathcal{F}\left[\langle y', C(\cdot)y \rangle_{\mathcal{Y}}\right](\delta)$.*

We found some confusion in the literature whether a kernel is the or of a measure. However Lemma 16 clarifies the relation between the and for a translation invariant . Notice that in the real scalar case the and of a shift-invariant kernel are the same, while the difference is significant for index˙cmd. The following lemma is a direct consequence of the definition of $C(\omega)$ as the of the adjoint of K_e and also helps to simplify the definition of index˙cmd.

Lemma 16

Let K_e be the signature of a shift-invariant $\mathcal{Y}$-Mercer kernel such that for all y, $y' \in \mathcal{Y}$, $\langle y', K_e(\cdot)y \rangle_{\mathcal{Y}} \in L^1(\mathcal{X}, \mathbf{Haar})$ and let $\langle y', C(\cdot)y \rangle_{\mathcal{Y}} = \mathcal{F}^{-1}\left[\langle y', K_e(\cdot)y \rangle_{\mathcal{Y}}\right]$. Then

1. *$C(\omega)$ is self-adjoint and C is even.*
2. *$\mathcal{F}^{-1}\left[\langle y', K_e(\cdot)y \rangle_{\mathcal{Y}}\right] = \mathcal{F}\left[\langle y', K_e(\cdot)y \rangle_{\mathcal{Y}}\right]$.*

3. *$K_e(\delta)$ is self-adjoint and K_e is even.*

While Proposition 15 gives an explicit form of the operator $C(\omega)$ defined as the of the kernel K, it is not really convenient to work with the Haar measure $\widehat{\mathbf{Haar}}$ on $\mathcal{B}(\widehat{\mathcal{X}})$. However it is easily possible to turn $\widehat{\mathbf{Haar}}$ into a probability measure to allow efficient integration over an infinite domain.

The following proposition allows to build a spectral decomposition of a shift-invariant $\mathcal{Y}$-Mercer kernel on a index˙cmd group $\mathcal{X}$ endowed with the group law $\star$ with respect to a scalar probability measure, by extracting a scalar probability density function from C.

Proposition 17 (Shift-invariant $\mathcal{Y}$-Mercer kernel spectral decomposition)
Let K_e be the signature of a shift-invariant $\mathcal{Y}$-Mercer kernel. If for all y, $y' \in \mathcal{Y}$, $\langle K_e(.)y, y'\rangle \in L^1(\mathcal{X}, \mathbf{Haar})$ then there exists a positive probability measure $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ and an operator-valued function A an such that for all y, $y' \in \mathcal{Y}$,

$$\langle y', K_e(\delta)y\rangle = \mathbf{E}_{\widehat{\mathbf{Haar}},\rho}\left[\overline{(\delta,\omega)}\langle y', A(\omega)y\rangle\right], \tag{10}$$

with $\langle y', A(\omega)y\rangle\rho(\omega) = \mathcal{F}\left[\langle y', K_e(\cdot)y\rangle\right](\omega)$. Moreover

1. *for all y, $y' \in \mathcal{Y}$, $\langle A(.)y, y'\rangle \in L^1(\widehat{\mathcal{X}}, \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho})$,*
2. *$A(\omega)$ is non-negative for $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$-almost all $\omega \in \widehat{\mathcal{X}}$,*
3. *$A(\cdot)$ and $\rho(\cdot)$ are even functions.*

3.2 Examples of spectral decomposition

In this section we give examples of spectral decomposition for various $\mathcal{Y}$-Mercer kernels, based on Proposition 17.

3.2.1 Gaussian decomposable kernel

Recall that a decomposable $\mathbb{R}^p$-Mercer introduced in Proposition 8 has the form $K(x,z) = k(x,z)\Gamma$, where $k(x,z)$ is a scalar Mercer kernel and $\Gamma \in \mathcal{L}(\mathbb{R}^p)$ is a non-negative operator. Let us focus on $K_e^{dec,gauss}(\cdot) = k_e^{gauss}(\cdot)\Gamma$, the Gaussian decomposable kernel where $K_e^{dec,gauss}$ and k_e^{gauss} are respectively the signature of K and k on the additive group $\mathcal{X} = (\mathbb{R}^d, +)$ – $index_cmd$ $\delta = x - z$ and $e = 0$. The well known Gaussian kernel is defined for all $\delta \in \mathbb{R}^d$ as follows $k_0^{\text{gauss}}(\delta) = \exp\left(-\sigma^{-2}\|\delta\|_2^2\right)/2$ where $\sigma \in \mathbb{R}_{>0}$ is an hyperparameter corresponding to the bandwith of the kernel. The –Pontryagin– dual group of $\mathcal{X} = (\mathbb{R}^d, +)$ is $\widehat{\mathcal{X}} \cong (\mathbb{R}^d, +)$ with the pairing $(\delta, \omega) = \exp(\mathrm{i}\langle\delta, \omega\rangle)$ where δ and $\omega \in \mathbb{R}^d$. In this case the Haar measures on $\mathcal{X}$ and $\widehat{\mathcal{X}}$ are in both cases the Lebesgue measure. However in order to have the property that $\mathcal{F}^{-1}\left[\mathcal{F}\left[f\right]\right] = f$ and $\mathcal{F}^{-1}\left[f\right] = \mathcal{R}\mathcal{F}\left[f\right]$ one must normalize both measures by $\sqrt{2\pi}^{-d}$, index˙cmd for all $\mathcal{Z} \in \mathcal{B}\left(\mathbb{R}^d\right)$, $\sqrt{2\pi}^d\mathbf{Haar}(\mathcal{Z}) = \mathbf{Leb}(\mathcal{Z})$ and $\sqrt{2\pi}^d\widehat{\mathbf{Haar}}(\mathcal{Z}) = \mathbf{Leb}(\mathcal{Z})$. Then the on $(\mathbb{R}^d, +)$ is

$$\mathcal{F}\left[f\right](\omega) = \int_{\mathbb{R}^d} \exp\left(-\mathrm{i}\langle\delta,\omega\rangle\right) f(\delta) d\mathbf{Haar}(\delta) = \int_{\mathbb{R}^d} \exp\left(-\mathrm{i}\langle\delta,\omega\rangle\right) f(\delta) \frac{d\mathbf{Leb}(\delta)}{\sqrt{2\pi}^d}.$$

Since $k_0^{\text{gauss}} \in L^1$ and Γ is bounded, it is possible to apply Proposition 17, and obtain for all y and $y' \in \mathcal{Y}$,

$$\left\langle y', C^{dec,gauss}(\omega) y \right\rangle = \mathcal{F}\left[\left\langle y', K_0^{dec,gauss}(\cdot) y \right\rangle\right](\omega) = \mathcal{F}\left[k_0^{gauss}\right](\omega) \langle y', \Gamma y \rangle.$$

Thus

$$C^{dec,gauss}(\omega) = \int_{\mathbb{R}^d} \exp\left(-\mathrm{i}\langle \omega, \delta \rangle - \frac{\|\delta\|_2^2}{2\sigma^2}\right) \frac{d\mathbf{Leb}(\delta)}{\sqrt{2\pi}^d} \Gamma.$$

Hence

$$C^{dec,gauss}(\omega) = \underbrace{\frac{1}{\sqrt{2\pi \frac{1}{\sigma^2}}^d} \exp\left(-\frac{\sigma^2}{2}\|\omega\|_2^2\right) \sqrt{2\pi}^d}_{\rho(\cdot) = \mathcal{N}(0, \sigma^{-2} I_d)\sqrt{2\pi}^d} \underbrace{\Gamma}_{A(\cdot)=\Gamma}.$$

Therefore the canonical decomposition of $C^{dec,gauss}$ is $A^{dec,gauss}(\omega) = \Gamma$ and $\rho^{dec,gauss} = \mathcal{N}(0, \sigma^{-2} I_d)\sqrt{2\pi}^d$, where $\mathcal{N}$ is the Gaussian probability distribution. Note that this decomposition is done with respect to the *normalized* Lebesgue measure $\widehat{\mathbf{Haar}}$, meaning that for all $\mathcal{Z} \in \mathcal{B}(\widehat{\mathcal{X}})$,

$$\begin{aligned}\mathbf{Pr}_{\widehat{\mathbf{Haar}}, \mathcal{N}(0,\sigma^{-2}I_d)\sqrt{2\pi}^d}(\mathcal{Z}) &= \int_{\mathcal{Z}} \mathcal{N}(0, \sigma^{-2} I_d)\sqrt{2\pi}^d d\widehat{\mathbf{Haar}}(\omega) \\ &= \int_{\widehat{\mathcal{X}}} \mathcal{N}(0, \sigma^{-2} I_d) d\mathbf{Leb}(\omega) = \mathbf{Pr}_{\mathcal{N}(0,\sigma^{-2}I_d)}(\mathcal{Z}).\end{aligned}$$

Thus, the same decomposition with respect to the usual –non-normalized– Lebesgue measure **Leb** yields

$$A^{dec,gauss}(\cdot) = \Gamma \tag{11a}$$

$$\rho^{dec,gauss} = \mathcal{N}(0, \sigma^{-2} I_d). \tag{11b}$$

3.2.2 Skewed-χ^2 decomposable kernel

The skewed-χ^2 scalar kernel (Li et al., 2010), useful for image processing, is defined on the index cmd group $\mathcal{X} = (-c_k; +\infty)_{k=1}^d$, with $c_k \in \mathbb{R}_{>0}$ and endowed with the group operation $\odot$. Let $(e_k)_{k=1}^d$ be the standard basis of $\mathcal{X}$. The operator $\odot : \mathcal{X} \times \mathcal{X} \to \mathcal{X}$ is defined by $x \odot z = ((x_k + c_k)(z_k + c_k) - c_k)_{k=1}^d$. The identity element e is $(1 - c_k)_{k=1}^d$ since $(1-c) \odot x = x$. Thus the inverse element x^{-1} is $((x_k + c_k)^{-1} - c_k)_{k=1}^d$. The skewed-$\chi^2$ scalar kernel reads

$$k_{1-c}^{skewed}(\delta) = \prod_{k=1}^d \frac{2}{\sqrt{\delta_k + c_k} + \sqrt{\frac{1}{\delta_k + c_k}}}. \tag{12}$$

The dual of $\mathcal{X}$ is $\widehat{\mathcal{X}} \cong \mathbb{R}^d$ with the pairing $(\delta, \omega) = \prod_{k=1}^d \exp\left(\mathrm{i} \log(\delta_k + c_k)\omega_k\right)$. The Haar measure are defined for all $\mathcal{Z} \in \mathcal{B}((-c; +\infty)^d)$ and all $\widehat{\mathcal{Z}} \in \mathcal{B}(\mathbb{R}^d)$ by $\sqrt{2\pi}^d \mathbf{Haar}(\mathcal{Z}) = \int_{\mathcal{Z}} \prod_{k=1}^d \frac{1}{z_k + c_k} d\mathbf{Leb}(z)$ and $\sqrt{2\pi}^d \widehat{\mathbf{Haar}}(\widehat{\mathcal{Z}}) = \mathbf{Leb}(\widehat{\mathcal{Z}})$. Thus the is

$$\mathcal{F}[f](\omega) = \int_{(-c;+\infty)^d} \prod_{k=1}^d \frac{\exp\left(-\mathrm{i} \log(\delta_k + c_k)\omega_k\right)}{\delta_k + c_k} f(\delta) \frac{d\mathbf{Leb}(\delta)}{\sqrt{2\pi}^d}.$$

Then, applying Fubini's theorem over product space, and the fact that each dimension is independent

$$\mathcal{F}\left[k_0^{skewed}\right](\omega) = \prod_{k=1}^{d} \int_{-c_k}^{+\infty} \frac{2\exp\left(-\mathrm{i}\log(\delta_k + c_k)\omega_k\right)}{(\delta_k + c_k)\left(\sqrt{\delta_k + c_k} + \sqrt{\frac{1}{\delta_k + c_k}}\right)} \frac{d\mathbf{Leb}(\delta_k)}{\sqrt{2\pi}^d}.$$

Making the change of variable $t_k = (\delta_k + c_k)^{-1}$ yields

$$\mathcal{F}\left[k_0^{skewed}\right](\omega) = \prod_{k=1}^{d} \int_{-\infty}^{+\infty} \frac{2\exp\left(-\mathrm{i}t_k\omega_k\right)}{\exp\left(\frac{1}{2}t_k\right) + \exp\left(-\frac{1}{2}t_k\right)} \frac{d\mathbf{Leb}(t_k)}{\sqrt{2\pi}^d} = \sqrt{2\pi}^d \prod_{k=1}^{d} \operatorname{sech}(\pi\omega_k).$$

Since $k_{1-c}^{\text{skewed}} \in L^1$ and Γ is bounded, it is possible to apply Proposition 17, and obtain

$$C^{dec,skewed}(\omega) = \mathcal{F}\left[k_{1-c}^{skewed}\right](\omega)\Gamma = \underbrace{\sqrt{2\pi}^d \prod_{k=1}^{d} \operatorname{sech}(\pi\omega_k)}_{\rho(\cdot)=\mathcal{S}(0,2^{-1})^d\sqrt{2\pi}^d} \underbrace{\Gamma}_{A(\cdot)}.$$

Hence the decomposition with respect to the usual –non-normalized– Lebesgue measure **Leb** yields

$$A^{dec,skewed}(\cdot) = \Gamma \tag{13a}$$

$$\rho^{dec,skewed} = \mathcal{S}\left(0, 2^{-1}\right)^d. \tag{13b}$$

3.2.3 Curl-free Gaussian kernel

The curl-free Gaussian kernel is defined as $K_0^{curl,gauss} = -\nabla\nabla^\mathsf{T} k_0^{gauss}$. Here $\mathcal{X} = (\mathbb{R}^d, +)$ so the setting is the same than Subsection 3.2.1.

$$\begin{aligned} C^{curl,gauss}(\omega)_{ij} &= \mathcal{F}\left[K_{1-c}^{curl,gauss}(\cdot)_{ij}\right](\omega) \\ &= \sqrt{2\pi\frac{1}{\sigma^2}}^d \exp\left(-\frac{\sigma^2}{2}\|\omega\|_2^2\right)\sqrt{2\pi}^d \omega_i\omega_j. \end{aligned}$$

Hence

$$C^{curl,gauss}(\omega) = \underbrace{\frac{1}{\sqrt{2\pi\frac{1}{\sigma^2}}^d} \exp\left(-\frac{\sigma^2}{2}\|\omega\|_2^2\right)\sqrt{2\pi}^d}_{\mu(\cdot)=\mathcal{N}(0,\sigma^{-2}I_d)\sqrt{2\pi}^d} \underbrace{\omega\omega^\mathsf{T}}_{A(\omega)=\omega\omega^\mathsf{T}}.$$

Here a canonical decomposition is $A^{curl,gauss}(\omega) = \omega\omega^\mathsf{T}$ for all $\omega \in \mathbb{R}^d$ and $\mu^{curl,gauss} = \mathcal{N}(0, \sigma^{-2}I_d)\sqrt{2\pi}^d$ with respect to the normalized Lebesgue measure $d\omega$. Again the decomposition with respect to the usual –non-normalized– Lebesgue measure is for all $\omega \in \mathbb{R}^d$

$$A^{curl,gauss}(\omega) = \omega\omega^\mathsf{T} \tag{14a}$$

$$\mu^{curl,gauss} = \mathcal{N}(0, \sigma^{-2}I_d). \tag{14b}$$

3.2.4 Divergence-free kernel

The divergence-free Gaussian kernel is defined as $K_0^{div,gauss} = (\nabla\nabla^\mathsf{T} - \Delta)k_0^{gauss}$ on the group $\mathcal{X} = (\mathbb{R}^d, +)$. The setting is the same than Subsection 3.2.1. Hence

$$C^{div,gauss}(\omega)_{ij} = \mathcal{F}\left[K_0^{div,gauss}(\cdot)_{ij}\right](\omega) = \left(\delta_{i=j}\sum_{k=1}^{d}\omega_k^2 - \omega_i\omega_j\right)\mathcal{F}\left[k_0^{gauss}\right](\omega).$$

Hence

$$C^{div,gauss}(\omega) = \underbrace{\frac{1}{\sqrt{2\pi\frac{1}{\sigma^2}}^d}\exp\left(-\frac{\sigma^2}{2}\|\omega\|_2^2\right)\sqrt{2\pi}^d}_{\rho(\cdot)=\mathcal{N}(0,\sigma^{-2}I_d)\sqrt{2\pi}^d}\underbrace{\left(I_d\|\omega\|_2^2 - \omega\omega^\mathsf{T}\right)}_{A(\omega)=I_d\|\omega\|_2^2-\omega\omega^\mathsf{T}}.$$

Thus the canonical decomposition with respect to the normalized Lebesgue measure is $A^{div,gauss}(\omega) = I_d\|\omega\|_2^2 - \omega\omega^\mathsf{T}$ and the measure $\rho^{div,gauss} = \mathcal{N}(0,\sigma^{-2}I_d)\sqrt{2\pi}^d$. The canonical decomposition with respect to the usual Lebesgue measure is

$$A^{div,gauss}(\omega) = I_d\|\omega\|_2^2 - \omega\omega^\mathsf{T} \tag{15a}$$
$$\rho^{div,gauss} = \mathcal{N}(0,\sigma^{-2}I_d). \tag{15b}$$

3.3 Operator-valued Random Fourier Features (ORFF)

3.3.1 Building Operator-valued Random Fourier Features

As shown in Proposition 17 it is always possible to find a pair $(A, \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho})$ from a shift invariant $\mathcal{Y}$-Mercer K_e such that $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ is a probability measure, index˙cmd $\int_{\widehat{\mathcal{X}}}\rho d\widehat{\mathbf{Haar}} = 1$ where ρ is the density of $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ and $K_e(\delta) = \mathbf{E}_{\widehat{\mathbf{Haar}},\rho}\overline{(\delta,\omega)}A(\omega)$. In order to obtain an approximation of K from a decomposition $(A, \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho})$ we turn our attention to a Monte-Carlo estimation of the expectation in Equation 10 characterizing a $\mathcal{Y}$-Mercer shift-invariant .

Proposition 18

Let $K(x,z)$ be a shift-invariant $\mathcal{Y}$-Mercer kernel with signature K_e such that for all y, $y' \in \mathcal{Y}$, $\langle y', K_e(\cdot)y\rangle \in L^1(\mathcal{X}, \mathbf{Haar})$. Then one can find a pair $(A, \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho})$ that satisfies Proposition 17. index˙cmd for $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$-almost all ω, and all $y, y' \in \mathcal{Y}$, $\langle y, A(\omega)y'\rangle\rho(\omega) = \mathcal{F}\left[\langle y', K_e(\cdot)y\rangle\right](\omega)$. If $(\omega_j)_{j=1}^D$ be a sequence of $D \in \mathbb{N}^*$ index˙cmd random variables following the law $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ then the operator-valued function $\tilde{K}$ defined for $(x,z) \in \mathcal{X}\times\mathcal{X}$ as

$$\tilde{K}(x,z) = \frac{1}{D}\sum_{j=1}^{D}\overline{(x\star z^{-1},\omega_j)}A(\omega_j)$$

is an approximation of K. index˙cmd it satisfies for all x, $z \in \mathcal{X}$, $\tilde{K}(x,z) \xrightarrow[D\to\infty]{index˙cmd} K(x,z)$ in the weak operator topology, where K is a $\mathcal{Y}$-Mercer index˙cmd.

Algorithm 1: Construction of index˙cmd from index˙cmd

Input : $K(x,z) = K_e(\delta)$ a shift-invariant $\mathcal{Y}$-Mercer kernel on $(\mathcal{X}, \star)$ such that $\forall y, y' \in \mathcal{Y}$, $\langle y', K_e(\cdot)y\rangle \in L^1(\mathbb{R}^d, \mathbf{Haar})$ and D the number of features.

Output : A random feature $\widetilde{\phi}(x)$ such that $\widetilde{\phi}(x)^*\widetilde{\phi}(z) \approx K(x,z)$

1 Define the pairing (x, ω) from the index˙cmd group $(\mathcal{X}, \star)$;

2 Find a decomposition $(A, \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho})$ and B such that $B(\omega)B(\omega)^*\rho(\omega) = A(\omega)\rho(\omega) = \mathcal{F}^{-1}[K_e](\omega)$;

4 Draw D index˙cmd realizations $(\omega_j)_{j=1}^D$ from the probability distribution $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$;

6 **return** $\begin{cases} \widetilde{\phi}(x) \in \mathcal{L}(\mathcal{Y}, \widetilde{\mathcal{H}}) & : y \mapsto \frac{1}{\sqrt{D}} \bigoplus_{j=1}^D (x, \omega_j) B(\omega_j)^* y \\ \widetilde{\phi}(x)^* \in \mathcal{L}(\widetilde{\mathcal{H}}, \mathcal{Y}) & : \theta \mapsto \frac{1}{\sqrt{D}} \sum_{j=1}^D (x, \omega_j) B(\omega_j) \theta_j \end{cases}$;

Now, for efficient computations as motivated in the introduction, we are interested in finding an approximated *feature map* instead of a kernel approximation. Indeed, an approximated feature map will allow to build linear models in regression tasks. The following proposition deals with the feature map construction.

Proposition 19

Assume the same conditions as Proposition 18. Moreover, if one can define $B : \widehat{\mathcal{X}} \to \mathcal{L}(\mathcal{Y}', \mathcal{Y})$ such that for $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$-almost all ω, and all $y, y' \in \mathcal{Y}$, $\langle y, B(\omega)B(\omega)^*y'\rangle\rho(\omega) = \langle y, A(\omega)y'\rangle\rho(\omega) = \mathcal{F}[\langle y, K_e(\cdot)y'\rangle](\omega)$, then the function $\widetilde{\phi} : \widehat{\mathcal{X}} \to \mathcal{L}(\mathcal{Y}, \bigoplus_{j=1}^D \mathcal{Y}')$ defined for all $y \in \mathcal{Y}$ as follows:

$$\widetilde{\phi}(x)y = \frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} (x, \omega_j) B(\omega_j)^* y, \quad \omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho} \text{ index˙cmd,}$$

is an approximated feature map[2] *for the kernel K.*

Remark 20

We find a decomposition such that $A(\omega_j) = B(\omega_j)B(\omega_j)^*$ for all $j \in \mathbb{N}_D^*$ either by exhibiting a closed-form or using a numerical decomposition. Such a decomposition always exists since $A(\omega)$ is positive semi-definite for all $\omega \in \widehat{\mathcal{X}}$.

Notice that an index˙cmd map as defined in Proposition 19 is also the Monte-Carlo sampling of the corresponding functional Fourier feature map $\phi_x : \mathcal{Y} \to L^2(\widehat{\mathcal{X}}, \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}; \mathcal{Y}')$ as defined in Proposition 21. Indeed, for all $y \in \mathcal{Y}$ and all $x \in \mathcal{X}$, $\widetilde{\phi}(x)y = \bigoplus_{j=1}^D (\phi_x y)(\omega_j)$, $\omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ index˙cmd Proposition 19 allows us to define Algorithm 1 for constructing index˙cmd from an operator valued kernel.

2. index˙cmd it satisfies for all $x, z \in \mathcal{X}$, $\widetilde{\phi}(x)^*\widetilde{\phi}(z) \xrightarrow[D\to\infty]{\text{index˙cmd}} K(x,z)$ in the weak operator topology, where K is a $\mathcal{Y}$-Mercer index˙cmd.

Figure 3: Approximation of a function in a VV-RKHS using different realizations of Operator Random Fourier Feature. Top row and bottom row correspond to two different realizations of $\widetilde{K}$, which are *different* . However when D tends to infinity, the different realizations of $\widetilde{K}$ yield the same index˙cmd.

We give a numerical illustration of different $\tilde{K}$ built from different index˙cmd realization $(\omega_j)_{j=1}^D$, $\omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$. In Figure 3, we represent the approximation of a reference function (black line) defined as $(y_1, y_2)^\mathsf{T} = f(x_i) = \sum_{j=1}^{250} \mathbf{K}_{ij} u_j$ where $u_j \sim \mathcal{N}(0, I_2)$ and K is a Gaussian decomposable kernel. We took $\Gamma = .5I_2 + .51_2$ such that the outputs y_1 and y_2 share some similarities. We generated 250 points equally separated on the segment $(-1; 1)$. The Gram matrix is then $\mathbf{K}_{ij} = \exp\left(-\frac{(x_i - x_j)^2}{2(0.1)^2}\right)\Gamma$, for $i, j \in \mathbb{N}^*_{250}$. We took $\Gamma = .5I_2 + .51_2$ such that the outputs y_1 and y_2 share some similarities. We generated 250 points equally separated on the segment $(-1; 1)$. Then we computed an approximate kernel matrix $\tilde{\mathbf{K}} \approx \mathbf{K}$ for 25 increasing values of D ranging from 1 to 10^4. The two graphs in Figure 3 on the top row shows that the more the number of features increases the closer the model $\widetilde{f}(x_i) = \sum_{j=1}^{250} \tilde{\mathbf{K}}_{ij} u_j$ is to f. The bottom row shows the same experiment but for a different realization of $\tilde{\mathbf{K}}$. When D is small the curves of the bottom and top rows are very dissimilar –and sine wave like– while they both converge to f when D increase. We introduce a *functional* feature map, we call *Fourier Feature map*, defined by the following proposition as a direct consequence of Proposition 14.

Proposition 21 (Functional Fourier feature map)

Let $\mathcal{Y}$ and $\mathcal{Y}'$ be two Hilbert spaces. If there exists an operator-valued function $B : \widehat{\mathcal{X}} \to \mathcal{L}(\mathcal{Y}, \mathcal{Y}')$ such that for all y, $y' \in \mathcal{Y}$, $\langle y, B(\omega)B(\omega)^* y' \rangle_{\mathcal{Y}} = \langle y', A(\omega)y \rangle_{\mathcal{Y}}$ $\widehat{\mu}$-almost everywhere and $\langle y', A(\cdot)y \rangle \in L^1(\widehat{\mathcal{X}}, \widehat{\mu})$ then the operator ϕ_x defined for all y in $\mathcal{Y}$ by $(\phi_x y)(\omega) = (x, \omega)B(\omega)^* y$, is a feature map[3] *of some shift-invariant $\mathcal{Y}$-Mercer kernel K.*

With this notation we have $\phi : \mathcal{X} \to \mathcal{L}(\mathcal{Y}; L^2(\widehat{\mathcal{X}}, \widehat{\mu}; \mathcal{Y}'))$ such that $\phi_x \in \mathcal{L}(\mathcal{Y}; L^2(\widehat{\mathcal{X}}, \widehat{\mu}; \mathcal{Y}'))$ where $\phi_x := \phi(x)$. Notice that an index˙cmd map as defined in Proposition 19 is also the Monte-Carlo sampling of the corresponding functional Fourier feature map $\phi_x : \mathcal{Y} \to L^2(\widehat{\mathcal{X}}, \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}; \mathcal{Y}')$ as defined in Proposition 21. Indeed, for all $y \in \mathcal{Y}$ and all $x \in \mathcal{X}$,

$$\widetilde{\phi}(x)y = \bigoplus_{j=1}^{D} (\phi_x y)(\omega_j), \quad \omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho} \text{ index˙cmd}$$

3.4 From Operator Random Fourier Feature maps to OVKs

It is also interesting to notice that we can go the other way and define from the general form of an , an operator-valued kernel.

Proposition 22 (Operator Random Fourier Feature map)

3. index˙cmd it satisfies for all x, $z \in \mathcal{X}$, $\phi_x^* \phi_z = K(x, z)$ where K is a $\mathcal{Y}$-Mercer index˙cmd.

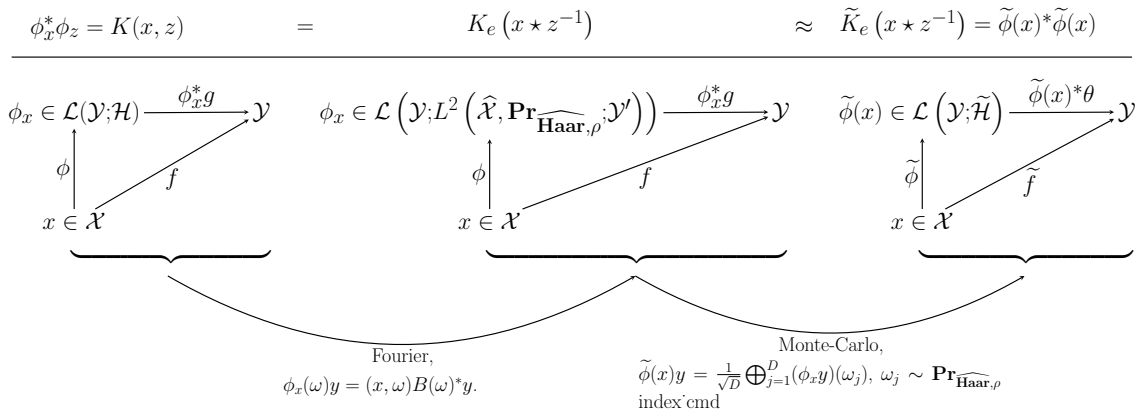

Figure 4: Relationships between feature-maps. For any realization of $\omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ index˙cmd, $\widetilde{\mathcal{H}} = \bigoplus_{j=1}^{D}\mathcal{Y}'$.

Let $\mathcal{Y}$ and $\mathcal{Y}'$ be two Hilbert spaces. If one defines an operator-valued function on the dual of a LCA group $\mathcal{X}$, $B : \widehat{\mathcal{X}} \to \mathcal{L}(\mathcal{Y},\mathcal{Y}')$, and a probability measure $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ on $\mathcal{B}(\widehat{\mathcal{X}})$, such that for all $y \in \mathcal{Y}$ and all $y' \in \mathcal{Y}'$, $\langle y, B(\cdot)y'\rangle \in L^2(\widehat{\mathcal{X}}, \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho})$, then the operator-valued function $\widetilde{\phi} : \mathcal{X} \to \mathcal{L}\left(\mathcal{Y}, \bigoplus_{j=1}^{D}\mathcal{Y}'\right)$ defined for all $x \in \mathcal{X}$ and for all $y \in \mathcal{Y}$ by

$$\widetilde{\phi}(x)y = \frac{1}{\sqrt{D}}\bigoplus_{j=1}^{D}(x,\omega_j)B(\omega_j)^*y, \quad \omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho},\ \textit{index˙cmd}, \tag{16}$$

is an approximated feature map of some $\mathcal{Y}$-Mercer operator-valued kernel[4]*.*

The difference between Proposition 22 and Proposition 19 is that in Proposition 22 we do not assume that $A(\omega)$ and $\mathbf{Pr}_{\widehat{Haar},\rho}$ have been obtained from Proposition 17. We conclude by showing that any realization of an approximate feature map gives a proper operator valued kernel. Hence we can always view $\widetilde{K}(x,z) = \widetilde{\phi}(x)^*\widetilde{\phi}(z)$ —where $\widetilde{\phi}$ is defined as in Proposition 18 (construction from an index˙cmd) or Proposition 22— as a $\mathcal{Y}$-Mercer and thus apply all the classic results of the theory on $\widetilde{K}$.

Proposition 23
Let $\boldsymbol{\omega} \in \widehat{\mathcal{X}}^D$. If for all $y,\ y' \in \mathcal{Y}$ $\langle y', \widetilde{K}_e\left(x \star z^{-1}\right)y\rangle_{\mathcal{Y}} = \langle\widetilde{\phi}(x)y', \widetilde{\phi}(z)y\rangle_{\widetilde{\mathcal{H}}} = \left\langle y', \frac{1}{D}\sum_{j=1}^{D}\overline{(x \star z^{-1},\omega_j)}B(\omega_j)B(\omega_j)^*y\right\rangle_{\mathcal{Y}}$, for all $x,\ z \in \mathcal{X}$, then $\widetilde{K}$ is a shift-invariant $\mathcal{Y}$-Mercer .

Note that the above theorem does not consider the ω_j's as random variables and therefore does not shows the convergence of the kernel $\widetilde{K}$ to some target kernel K. However is shows that any realization of $\widetilde{K}$ when ω_j's are random variables yields a valid $\mathcal{Y}$-Mercer operator-valued kernel. Note that the above theorem does not considers the ω_j's as random variables

4. index˙cmd it satisfies $\widetilde{\phi}(x)^*\widetilde{\phi}(z) \xrightarrow[D\to\infty]{\text{index˙cmd}} K(x,z)$ in the weak operator topology, where K is a $\mathcal{Y}$-Mercer index˙cmd

and therefore does not shows the convergence of the kernel $\widetilde{K}$ to some target kernel K. However is shows that any realization of $\widetilde{K}$ when ω_j's are random variables yields a valid $\mathcal{Y}$-Mercer operator-valued kernel. Indeed, as a result of Proposition 23, in the same way we defined an index˙cmd, we can define an approximate feature operator $\widetilde{W}$ which maps $\widetilde{\mathcal{H}}$ onto $\mathcal{H}_{\widetilde{K}}$, where $\widetilde{K}(x,z) = \widetilde{\phi}(x)^*\widetilde{\phi}(z)$, for all x, $z \in \mathcal{X}$.

Definition 24 (Random Fourier feature operator)
Let $\boldsymbol{\omega} = (\omega_j)_{j=1}^D \in \widehat{\mathcal{X}}^D$ and let $\widetilde{K}_e = \frac{1}{D}\sum_{j=1}^D \overline{(\cdot,\omega_j)}B(\omega_j)B(\omega_j)^*$. We call random Fourier feature operator the linear application $\widetilde{W} : \widetilde{\mathcal{H}} \to \mathcal{H}_{\widetilde{K}}$ defined as

$$\left(\widetilde{W}\theta\right)(x) := \widetilde{\phi}(x)^*\theta = \frac{1}{\sqrt{D}}\sum_{j=1}^D \overline{(x,\omega_j)}B(\omega_j)\theta_j$$

where $\theta = \bigoplus_{j=1}^D \theta_j \in \widetilde{\mathcal{H}}$. Then from Proposition 6, $\left(Ker\,\widetilde{W}\right)^\perp = \overline{\text{span}}\left\{\widetilde{\phi}(x)y \;\middle|\; \forall x \in \mathcal{X},\ \forall y \in \mathcal{Y}\right\} \subseteq \widetilde{\mathcal{H}}$.

The random Fourier feature operator is useful to show the relations between the random Fourier feature map with the functional feature map defined in Proposition 21. The relationship between the generic feature map (defined for all) the functional feature map (defining a shift-invariant $\mathcal{Y}$-Mercer) and the random Fourier feature map is presented in Figure 4.

Proposition 25
For any $g \in \mathcal{H} = L^2(\widehat{\mathcal{X}}, \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}; \mathcal{Y}')$, let $\theta := \frac{1}{\sqrt{D}}\bigoplus_{j=1}^D g(\omega_j)$, $\omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ index˙cmd Then

1. $\left(\widetilde{W}\theta\right)(x) = \widetilde{\phi}(x)^*\theta \xrightarrow[D\to\infty]{index˙cmd} \phi_x^* g = (Wg)(x)$,
2. $\|\theta\|^2_{\widetilde{\mathcal{H}}} \xrightarrow[D\to\infty]{index˙cmd} \|g\|^2_{\mathcal{H}}$,

We write $\widetilde{\phi}(x)^*\widetilde{\phi}(x) \approx K(x,z)$ when $\widetilde{\phi}(x)^*\widetilde{\phi}(x) \xrightarrow{\text{index˙cmd}} K(x,z)$ in the weak operator topology when D tends to infinity. With mild abuse of notation we say that $\widetilde{\phi}(x)$ is an approximate feature map of the functional feature map ϕ_x index˙cmd $\widetilde{\phi}(x) \approx \phi_x$, when for all y', $y \in \mathcal{Y}$,

$$\langle y, K(x,z)y'\rangle_{\mathcal{Y}} = \langle \phi_x y, \phi_z y'\rangle_{L^2(\widehat{\mathcal{X}},\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho};\mathcal{Y}')} \approx \langle \widetilde{\phi}(x)y, \widetilde{\phi}(x)y'\rangle_{\widetilde{\mathcal{H}}} := \langle y, \widetilde{K}(x,z)y'\rangle_{\mathcal{Y}}$$

where ϕ_x is defined in the sense of Proposition 21.

3.5 Examples of Operator Random Fourier Feature maps

We now give two examples of operator-valued random Fourier feature map. First we introduce the general form of an approximated feature map for a matrix-valued kernel on the additive group $(\mathbb{R}^d, +)$.

Example 1 (Matrix-valued kernel on the additive group)
In the following let $K(x,z) = K_0(x-z)$ be a $\mathcal{Y}$-Mercer matrix-valued kernel on $\mathcal{X} = \mathbb{R}^d$, invariant index˙cmd the group operation $+$. Then the function $\widetilde{\phi}$ defined as follow is an of K_0. For all $y \in \mathcal{Y}$,

$$\widetilde{\phi}(x)y = \frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} \begin{pmatrix} \cos\langle x, \omega_j \rangle_2 B(\omega_j)^* y \\ \sin\langle x, \omega_j \rangle_2 B(\omega_j)^* y \end{pmatrix}, \quad \omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho} \text{ index˙cmd.}$$

In particular we deduce the following features maps for the kernels proposed in Subsection 3.2.

- For the decomposable Gaussian kernel $K_0^{dec,gauss}(\delta) = k_0^{gauss}(\delta)\Gamma$ for all $\delta \in \mathbb{R}^d$, let $BB^* = \Gamma$. A bounded –and unbounded– index˙cmd map is

$$\widetilde{\phi}(x)y = \frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} \begin{pmatrix} \cos\langle x, \omega_j \rangle_2 B^* y \\ \sin\langle x, \omega_j \rangle_2 B^* y \end{pmatrix} = (\widetilde{\varphi}(x) \otimes B^*)y,$$

  where $\omega_j \sim \mathbf{Pr}_{\mathcal{N}(0,\sigma^{-2}I_d)}$ index˙cmd and $\widetilde{\varphi}(x) = \frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} \begin{pmatrix} \cos\langle x, \omega_j \rangle_2 \\ \sin\langle x, \omega_j \rangle_2 \end{pmatrix}$ is a scalar index˙cmd map (Rahimi and Recht, 2007).

- For the curl-free Gaussian kernel, $K_0^{curl,gauss} = -\nabla\nabla^\mathsf{T} k_0^{gauss}$ an unbounded index˙cmd map is

$$\widetilde{\phi}(x)y = \frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} \begin{pmatrix} g\cos\langle x, \omega_j \rangle_2 \omega_j^\mathsf{T} y \\ \sin\langle x, \omega_j \rangle_2 \omega_j^\mathsf{T} y \end{pmatrix}, \tag{17}$$

  $\omega_j \sim \mathbf{Pr}_{\mathcal{N}(0,\sigma^{-2}I_d)}$ index˙cmd and a bounded index˙cmd map is

$$\widetilde{\phi}(x)y = \frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} \begin{pmatrix} \cos\langle x, \omega_j \rangle_2 \frac{\omega_j^\mathsf{T}}{\|\omega_j\|} y \\ \sin\langle x, \omega_j \rangle_2 \frac{\omega_j^\mathsf{T}}{\|\omega_j\|} y \end{pmatrix}, \quad \omega_j \sim \mathbf{Pr}_\rho \text{ index˙cmd.}$$

  where $\rho(\omega) = \frac{\sigma^2\|\omega\|^2}{d}\mathcal{N}(0,\sigma^{-2}I_d)(\omega)$ for all $\omega \in \mathbb{R}^d$.

- For the divergence-free Gaussian kernel $K_0^{div,gauss}(x,z) = (\nabla\nabla^\mathsf{T} - \Delta I_d)k_0^{gauss}(x,z)$ an unbounded index˙cmd map is

$$\widetilde{\phi}(x)y = \frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} \begin{pmatrix} \cos\langle x, \omega_j \rangle_2 B(\omega_j)^\mathsf{T} y \\ \sin\langle x, \omega_j \rangle_2 B(\omega_j)^\mathsf{T} y \end{pmatrix} \tag{18}$$

  where $\omega_j \sim \mathbf{Pr}_\rho$ index˙cmd and $B(\omega) = \left(\|\omega\| I_d - \omega\omega^\mathsf{T}\right)$ and $\rho = \mathcal{N}(0,\sigma^{-2}I_d)$ for all $\omega \in \mathbb{R}^d$. A bounded index˙cmd map is

$$\widetilde{\phi}(x)y = \frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} g \begin{pmatrix} \cos\langle x, \omega_j \rangle_2 B(\omega_j)^\mathsf{T} y \\ \sin\langle x, \omega_j \rangle_2 B(\omega_j)^\mathsf{T} y \end{pmatrix}, \quad \omega_j \sim \mathbf{Pr}_\rho \text{ index˙cmd},$$

where $B(\omega) = \left(I_d - \frac{\omega\omega^\intercal}{\|\omega\|^2}\right)$ and $\rho(\omega) = \frac{\sigma^2\|\omega\|^2}{d}\mathcal{N}(0, \sigma^{-2}I_d)$ for allg $\omega \in \mathbb{R}^d$.

The second example extends scalar-valued Random Fourier Features on the skewed multiplicative group –described in appendix A.3 and Subsection 3.2.2– to the operator-valued case.

Example 2 (Matrix-valued kernel on the skewed multiplicative group)

In the following, $K(x,z) = K_{1-c}(x \odot z^{-1})$ is a $\mathcal{Y}$-Mercer matrix-valued kernel on $\mathcal{X} = (-c; +\infty)^d$ invariant index˙cmd the group operation[5] $\odot$. Then the function $\widetilde{\phi}$ defined as follow is an of K_{1-c}. For all $y \in \mathcal{Y}$,

$$\widetilde{\phi}(x)y = \frac{1}{\sqrt{D}}\bigoplus_{j=1}^{D}\begin{pmatrix}\cos\langle\log(x+c), \omega_j\rangle_2 B(\omega_j)^* y \\ \sin\langle\log(x+c), \omega_j\rangle_2 B(\omega_j)^* y\end{pmatrix}, \quad \omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho} \text{ index˙cmd.}$$

3.6 Regularization property

We have shown so far that it is always possible to construct a feature map that allows to approximate a shift-invariant $\mathcal{Y}$-Mercer kernel. However we could also propose a construction of such map by studying the regularization induced with respect to the of a target function $f \in \mathcal{H}_K$. In other words, what is the norm in $L^2(\widehat{\mathcal{X}}, \widehat{\mathbf{Haar}}; \mathcal{Y}')$ induced by $\|\cdot\|_K$?

Proposition 26

Let K be a shift-invariant $\mathcal{Y}$-Mercer Kernel such that for all y, y' in $\mathcal{Y}$, $\langle y', K_e(\cdot)y\rangle_{\mathcal{Y}} \in L^1(\mathcal{X}, \mathbf{Haar})$. Then for all $f \in \mathcal{H}_K$

$$\|f\|_K^2 = \int_{\widehat{\mathcal{X}}} \frac{\left\langle \mathcal{F}[f](\omega), A(\omega)^\dagger \mathcal{F}[f](\omega)\right\rangle_{\mathcal{Y}}}{\rho(\omega)} d\widehat{\mathbf{Haar}}(\omega). \tag{19}$$

where $\langle y', A(\omega)y\rangle\rho(\omega) := \mathcal{F}[\langle y', K_e(\cdot)y\rangle](\omega)$.

Note that if $K(x,z) = k(x,z)$ is a scalar kernel then for all ω in $\widehat{\mathcal{X}}$, $A(\omega) = 1$. Therefore we recover the well known result for kernels that is for any $f \in \mathcal{H}_k$ we have $\|f\|_k = \int_{\widehat{\mathcal{X}}} \mathcal{F}[k_e](\omega)^{-1}\mathcal{F}[f](\omega)^2 d\widehat{\mathbf{Haar}}(\omega)$ (Yang et al., 2012; Vert; Smola et al., 1998). Eventually from this last equation we also recover Proposition 9 for decomposable kernels. If $A(\omega) = \Gamma \in \mathcal{L}_+(\mathbb{R}^p)$, $\|f\|_K = \sum_{i,j=1}^p \left(\Gamma^\dagger\right)_{ij}\langle f_i, f_j\rangle_k$ We also note that the regularization property in $\mathcal{H}_K$ does not depends (as expected) on the decomposition of $A(\omega)$ into $B(\omega)B(\omega)^*$. Therefore the decomposition should be chosen such that it optimizes the computation cost. For instance if $A(\omega) \in \mathcal{L}(\mathbb{R}^p)$ has rank r, one could find an operator $B(\omega) \in \mathcal{L}(\mathbb{R}^p, \mathbb{R}^r)$ such that $A(\omega) = B(\omega)B(\omega)^*$. Moreover, in light of Equation 19 the regularization property of the kernel with respect to the , it is also possible to define an approximate feature map of an from its regularization properties in the index˙cmd as proposed in Algorithm 2.

5. The group operation $\odot$ is defined in Subsection 3.2.2.

Algorithm 2: Construction of index˙cmd

Input :

- The pairing (x, ω) of the index˙cmd group $(\mathcal{X}, \star)$.
- A probability measure $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ with density ρ index˙cmd the haar measure $\widehat{\mathbf{Haar}}$ on $\widehat{\mathcal{X}}$.
- An operator-valued function $B : \widehat{\mathcal{X}} \to \mathcal{L}(\mathcal{Y}, \mathcal{Y}')$ such that for all $y\ y' \in \mathcal{Y}$, $\langle y', B(\cdot)B(\cdot)^* y\rangle \in L^1(\widehat{\mathcal{X}}, \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho})$.
- D the number of features.

Output : A random feature $\widetilde{\phi}(x)$ such that $\widetilde{\phi}(x)^*\widetilde{\phi}(z) \approx K(x, z)$.

Draw D random vectors $(\omega_j)_{j=1}^D$ index˙cmd from the probability law $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$;

return $\begin{cases} \widetilde{\phi}(x) \in \mathcal{L}(\mathcal{Y}, \widetilde{\mathcal{H}}) & : y \mapsto \frac{1}{\sqrt{D}} \bigoplus_{j=1}^D (x, \omega_j) B(\omega_j)^* y \\ \widetilde{\phi}(x)^* \in \mathcal{L}(\widetilde{\mathcal{H}}, \mathcal{Y}) & : \theta \mapsto \frac{1}{\sqrt{D}} \sum_{j=1}^D (x, \omega_j) B(\omega_j) \theta_j \end{cases}$;

4. Main contribution: convergence with high probability of the estimator

We are now interested in a non-asymptotic analysis of the index˙cmd approximation of shift-invariant $\mathcal{Y}$-Mercer kernels on index˙cmd group $\mathcal{X}$ endowed with the operation group $\star$ where $\mathcal{X}$ is a Banach space (The more general case where $\mathcal{X}$ is a Polish space is discussed in the appendix appendix B.2). For a given D, we study how close is the approximation $\tilde{K}(x, z) = \widetilde{\phi}(x)^*\widetilde{\phi}(z)$ to the target kernel $K(x, z)$ for any x, z in $\mathcal{X}$.

If $A \in \mathcal{L}_+(\mathcal{Y})$ we denote $\|A\|_{\mathcal{Y},\mathcal{Y}}$ its operator norm (the induced norm). For x and z in some non-empty compact $\mathcal{C} \subset \mathbb{R}^d$, we consider: $F(x \star z^{-1}) = \tilde{K}(x, z) - K(x, z)$ and study how the uniform norm $\left\|\tilde{K} - K\right\|_{\mathcal{C}\times\mathcal{C}} := \sup_{(x,z)\in\mathcal{C}\times\mathcal{C}} \left\|\tilde{K}(x, z) - K(x, z)\right\|_{\mathcal{Y},\mathcal{Y}}$ behaves according to D. All along this document we denote $\delta = x \star z^{-1}$ for all x and $z \in \mathcal{X}$. Figure 5 empirically shows convergence of three different index˙cmd approximations for x, z sampled from the compact $[-1, 1]^4$ and using an increasing number of sample points D. The log-log plot shows that all three kernels have the same convergence rate, up to a multiplicative factor.

A typical application is the study of the deviation of the empirical mean of random variables to their expectation. This means that given an error ϵ between the kernel approximation $\widetilde{K}$ and the true kernel K, if we are given enough samples to construct $\widetilde{K}$, the probability of measuring an error greater than ϵ is essentially zero (it drops at an exponential rate with respect to the number of samples D). To measure the error between the kernel approximation and the true kernel at a given point many metrics are possible. index˙cmd any matrix norm such as the Hilbert-Schmidt norm, trace norm, the operator norm or Schatten norms. In this work we focus on measuring the error in terms of operator

norm. For all x, $z \in \mathcal{X}$ we look for a bound on

$$\begin{aligned}
&\mathbf{Pr}_\rho \left\{ (\omega_j)_{j=1}^D \;\middle|\; \left\| \widetilde{K}(x,z) - K(x,z) \right\|_{\mathcal{Y},\mathcal{Y}} \geq \epsilon \right\} \\
&= \mathbf{Pr}_\rho \left\{ (\omega_j)_{j=1}^D \;\middle|\; \sup_{0 \neq y \in \mathcal{Y}} \frac{\left\| (\widetilde{K}(x,z) - K(x,z))y \right\|_{\mathcal{Y}}}{\|y\|_{\mathcal{Y}}} \geq \epsilon \right\}
\end{aligned}$$

In other words, given any vector $y \in \mathcal{Y}$ we study how the residual operator $\widetilde{K} - K$ is able to send y to zero. We believe that this way of measuring the "error" to be more intuitive. Moreover, on contrary to an error measure with the Hilbert-Schmidt norm, the operator norm error does not grows linearly with the dimension of the output space as the Hilbert-Schmidt norm does. On the other hand the Hilbert-schmidt norm makes the studied random variables Hilbert space valued, for which it is much easier to derive concentration inequalities (Smale and Zhou, 2007; Pinelis, 1994; Naor, 2012). Note that in the scalar case ($A(\omega) = 1$) the Hilbert-Schmidt norm error and the operator norm are the same and measure the deviation between $\widetilde{K}$ and K as the absolute value of their difference.

A raw concentration inequality of the kernel estimator gives the error on one point. If one is interesting in bounding the maximum error over N points, applying a union bound on all the point would yield a bound that grows linearly with N. This would suggest that when the number of points increase, even if all of them are concentrated in a small subset of $\mathcal{X}$, we should draw increasingly more features to have an error below ϵ with high probability. However if we restrict ourselves to study the error on a compact subset of $\mathcal{X}$ (and in practice data points lies often in a closed bounded subset of $\mathbb{R}^d$), we can cover this compact subset by a finite number of closed balls and apply the concentration inequality and the union bound only on the center of each ball. Then if the function $\left\| \widetilde{K}_e - K_e \right\|$ is smooth enough on each ball (index˙cmd Lipschitz) we can guarantee with high probability that the error between the centers of the balls will not be too high. Eventually we obtain a bound in the worst case scenario on all the points in a subset $\mathcal{C}$ of $\mathcal{X}$. This bound depends on the covering number $\mathcal{N}(\mathcal{C}, r)$ of $\mathcal{X}$ with ball of radius r. When $\mathcal{X}$ is a Banach space, the covering number is proportional to the diameter of the diameter of $\mathcal{C} \subseteq \mathcal{X}$.

Prior to the presentation of general results, we briefly recall the uniform convergence of index˙cmd approximation for a scalar shift invariant kernel on the additive index˙cmd group $\mathbb{R}^d$ and introduce a direct corollary about decomposable shift-invariant index˙cmd on the index˙cmd group $(\mathbb{R}^d, +)$.

4.1 Random Fourier Features in the scalar case and decomposable OVK

Rahimi and Recht (2007) proved the uniform convergence of (index˙cmd) approximation for a scalar shift-invariant kernel on the index˙cmd group $\mathbb{R}^d$ endowed with the group operation $\star = +$. In the case of the shift-invariant decomposable index˙cmd, an upper bound on the error can be obtained as a direct consequence of the result in the scalar case obtained by Rahimi and Recht (2007) and other authors (Sutherland and Schneider, 2015; Sriperumbudur and Szabo, 2015).

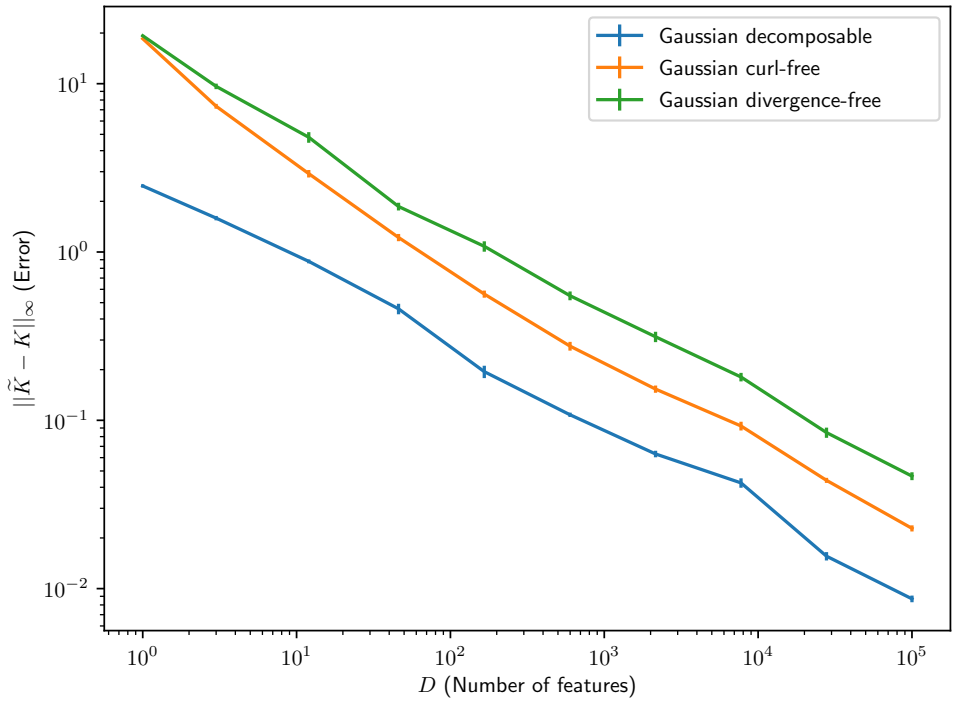

Figure 5: Error reconstructing the target operator-valued kernel K with index˙cmd approximation $\tilde{K}$ for the decomposable, curl-free and divergence-free kernel.

Theorem 27 (Uniform error bound for index˙cmd, Rahimi and Recht (2007))

Let $\mathcal{C}$ be a compact of subset of $\mathbb{R}^d$ of diameter $|\mathcal{C}|$. Let k be a shift invariant kernel, differentiable with a bounded second derivative and $\mathbf{Pr}_\rho$ its normalized such that it defines a probability measure. Let $\widetilde{k} = \sum_{j=1}^D \cos\langle \cdot, \omega_j \rangle \approx k(x,z)$ and $\sigma^2 = \mathbf{E}_\rho \|\omega\|_2^2$. Then we have

$$\mathbf{Pr}_\rho \left\{ (\omega_j)_{j=1}^D \;\middle|\; \left\| \tilde{k} - k \right\|_{\mathcal{C}\times\mathcal{C}} \geq \epsilon \right\} \leq 2^8 \left(\frac{\sigma|\mathcal{C}|}{\epsilon} \right)^2 \exp\left(-\frac{\epsilon^2 D}{4(d+2)} \right)$$

From Theorem 27, we can deduce the following corollary about the uniform convergence of the index˙cmd approximation of the decomposable kernel. We recall that for a given pair x, z in $\mathcal{C}$, $\tilde{K}(x,z) = \widetilde{\phi}(x)^* \widetilde{\phi}(z) = \Gamma \tilde{k}(x,z)$ and $K_0(x-z) = \Gamma \mathbf{E}_{\widehat{\mathbf{Haar}},\rho}[\tilde{k}(x,z)]$.

Corollary 28 (Uniform error bound for decomposable index˙cmd)

Let $\mathcal{C}$ be a compact of subset of $\mathbb{R}^d$ of diameter $|\mathcal{C}|$. Let K be a decomposable kernel built from a positive operator self-adjoint Γ, and k a shift invariant kernel with bounded second derivative such that $\widetilde{K} = \sum_{j=1}^D \cos\langle \cdot, \omega_j \rangle \Gamma \approx K$ and $\sigma^2 = \mathbf{E}_\rho \|\omega\|_2^2$. Then

$$\mathbf{Pr}_\rho \left\{ (\omega_j)_{j=1}^D \;\middle|\; \left\| \widetilde{K} - K \right\|_{\mathcal{C}\times\mathcal{C}} \geq \epsilon \right\} \leq 2^8 \left(\frac{\sigma \|\Gamma\|_{\mathcal{Y},\mathcal{Y}} |\mathcal{C}|}{\epsilon} \right)^2 \exp\left(-\frac{\epsilon^2 D}{4\|\Gamma\|_2^2 (d+2)} \right)$$

Note that a similar corollary could have been obtained for the recent result of Sutherland and Schneider (2015) who refined the bound proposed by Rahimi and Recht by using a Bernstein

concentration inequality instead of the Hoeffding inequality. More recently Sriperumbudur and Szabo (2015) showed an optimal bound for . The improvement of Sriperumbudur and Szabo (2015) is mainly in the constant factors where the bound does not depend linearly on the diameter $|\mathcal{C}|$ of $\mathcal{C}$ but exhibit a logarithmic dependency $\log(|\mathcal{C}|)$, hence requiring significantly less random features to reach a desired uniform error with high probability. Moreover, Sutherland and Schneider (2015) also considered a bound on the expected max error $\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}\left\|\widetilde{K}-K\right\|_\infty$, which is obtained using Dudley's entropy integral (Dudley, 1967; Boucheron et al., 2013) as a bound on the supremum of an empirical process by the covering number of the indexing set. This useful theorem is also part of the proof of Sriperumbudur and Szabo (2015).

4.2 Uniform convergence of approximation on groups

Before introducing the new theorem, we give the definition of the Orlicz norm which gives a proxy-bound on the norm of subexponential random variables.

Definition 29 (Orlicz norm (Van Der Vaart and Wellner, 1996))

Let $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-decreasing convex function with $\psi(0) = 0$. For a random variable X on a measured space $(\Omega, \mathcal{T}(\Omega), \mu)$, the quantity $\|X\|_\psi = \inf\{\, C > 0 \mid \mathbf{E}_\mu[\psi(|X|/C)] \leq 1 \,\}$. is called the Orlicz norm of X.

Here, the function ψ is chosen as $\psi(u) = \psi_\alpha(u)$ where $\psi_\alpha(u) := e^{u^\alpha} - 1$. When $\alpha = 1$, a random variable with finite Orlicz norm is called a *subexponential variable* because its tails decrease at an exponential rate. Let X be a self-adjoint random operator. Given a scalar-valued measure μ, we call *variance* of an operator X the quantity $\mathbf{Var}_\mu[X] = \mathbf{E}_\mu[X - \mathbf{E}_\mu[X]]^2$. Among the possible concentration inequalities adapted to random operators (Tropp et al., 2015; Minsker, 2011; Ledoux and Talagrand, 2013; Pinelis, 1994; Koltchinskii et al., 2013), we focus on the results of Tropp et al. (2015); Minsker (2011), for their robustness to high or potentially infinite dimension of the output space $\mathcal{Y}$. To guarantee a good scaling with the dimension of $\mathcal{Y}$ we introduce the notion of intrinsic dimension (or effective rank) of an operator.

Definition 30

Let A be a trace class operator acting on a Hilbert space $\mathcal{Y}$. We call intrinsic dimension the quantity: $\mathrm{IntDim}(A) = \|A\|_{\mathcal{Y},\mathcal{Y}}^{-1} \mathbf{Tr}[A]$.

Indeed the bound proposed in our first publication at index˙cmd (Brault et al., 2016) based on Koltchinskii et al. (2013) depends on p while the present bound depends on the intrinsic dimension of the variance of $A(\omega)$ which is always smaller than p when the operator $A(\omega)$ is Hilbert-Schmidt ($p \leq \infty$).

Corollary 31

Let $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ be a shift-invariant $\mathcal{Y}$-Mercer kernel, where $\mathcal{Y}$ is a finite dimensional Hilbert space of dimension p and $\mathcal{X}$ a finite dimensional Banach space of dimension d. Moreover, let $\mathcal{C}$ be a closed ball of $\mathcal{X}$ centred at the origin of diameter $|\mathcal{C}|$, $A : \widehat{\mathcal{X}} \to \mathcal{L}(\mathcal{Y})$

and $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ *a pair such that*

$$\tilde{K}_e = \sum_{j=1}^{D} \cos(\cdot, \omega_j) A(\omega_j) \approx K_e, \quad \omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho} \text{ index˙cmd..}$$

Let $\mathcal{D}_\mathcal{C} = \mathcal{C} \star \mathcal{C}^{-1}$ *and* $V(\delta) \succcurlyeq \mathbf{Var}_{\widehat{\mathbf{Haar}},\rho} \tilde{K}_e(\delta)$ *for all* $\delta \in \mathcal{D}_\mathcal{C}$ *and* H_ω *be the Lipschitz constant of the function* $h : x \mapsto (x, \omega)$*. If the three following constants exist*

$$m \geq \int_{\widehat{\mathcal{X}}} H_\omega \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}(\omega) < \infty$$

and

$$u \geq 4 \left(\left\| \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} \right\|_{\psi_1} + \sup_{\delta \in \mathcal{D}_\mathcal{C}} \|K_e(\delta)\|_{\mathcal{Y},\mathcal{Y}} \right) < \infty$$

and

$$v \geq \sup_{\delta \in \mathcal{D}_\mathcal{C}} D \|V(\delta)\|_{\mathcal{Y},\mathcal{Y}} < \infty.$$

Define $p_{int} \geq \sup_{\delta \in \mathcal{D}_\mathcal{C}} \mathrm{IntDim}(V(\delta))$*, then for all* $0 < \epsilon \leq m|C|$*,*

$$\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho} \left\{ (\omega_j)_{j=1}^D \;\middle|\; \left\| \tilde{K} - K \right\|_{\mathcal{C}\times\mathcal{C}} \geq \epsilon \right\}$$
$$\leq 8\sqrt{2} \left(\frac{m|\mathcal{C}|}{\epsilon} \right) \left(p_{int} r_{v/D}(\epsilon) \right)^{\frac{1}{d+1}} \begin{cases} \exp\left(-D \frac{\epsilon^2}{8v(d+1)\left(1+\frac{1}{p}\right)} \right), & \epsilon \leq \frac{v}{u} \frac{1+1/p}{K(v,p)} \\ \exp\left(-D \frac{\epsilon}{8u(d+1)K(v,p)} \right), & \text{otherwise,} \end{cases}$$

where $K(v,p) = \log\left(16\sqrt{2}p\right) + \log\left(\frac{u^2}{v}\right)$ *and* $r_{v/D}(\epsilon) = 1 + \frac{3}{\epsilon^2 \log^2(1+D\epsilon/v)}$.

We give a comprehensive full proof of the theorem in appendix B.2. It follows the usual scheme derived in Rahimi and Recht (2007) and Sutherland and Schneider (2015) and involves Bernstein concentration inequality for unbounded symmetric matrices (Theorem 45).

4.3 Dealing with infinite dimensional operators

We studied the concentration of index˙cmd under the assumption that $\mathcal{Y}$ is finite dimensional. Indeed a d term characterizing the dimension of the input space $\mathcal{X}$ appears in the bound proposed in Corollary 31, and when d tends to infinity, the exponential part goes to zero so that the probability is bounded by a constant greater than one. Unfortunately, considering unbounded random operators Minsker (2011) doesn't give any tighter solution.

In our first bound presented at index˙cmd, we presented a bound based on a matrix concentration inequality for unbounded random variable. Compared to this previous bound, Corollary 31 does not depend on the dimensionality p of the output space $\mathcal{Y}$ but on the intrinsic dimension of the operator $A(\omega)$. However to remove the dependency in p in the exponential part, we must turn our attention to operator concentration inequalities for bounded random variable. To the best of our knowledge we are not aware of concentration inequalities working for "unbounded" operator- valued random variables. Following the same proof than Corollary 31 we obtain Corollary 32.

Corollary 32

Let $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ be a shift-invariant $\mathcal{Y}$-Mercer kernel, where $\mathcal{Y}$ is a Hilbert space and $\mathcal{X}$ a finite dimensional Banach space of dimension D. Moreover, let $\mathcal{C}$ be a closed ball of $\mathcal{X}$ centered at the origin of diameter $|\mathcal{C}|$, subset of $\mathcal{X}$, $A : \widehat{\mathcal{X}} \to \mathcal{L}(\mathcal{Y})$ and $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ a pair such that

$$\tilde{K}_e = \sum_{j=1}^{D} \cos(\cdot, \omega_j) A(\omega_j) \approx K_e, \quad \omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho} \text{ index˙cmd.}$$

where $A(\omega_j)$ is a Hilbert-Schmidt operator for all $j \in \mathbb{N}_D^*$. Let $\mathcal{D}_\mathcal{C} = \mathcal{C} \star \mathcal{C}^{-1}$ and $V(\delta) \succcurlyeq \mathbf{Var}_{\widehat{\mathbf{Haar}},\rho} \tilde{K}_e(\delta)$ for all $\delta \in \mathcal{D}_\mathcal{C}$ and H_ω be the Lipschitz constant of the function $h : x \mapsto (x, \omega)$. If the three following constants exists

$$m \geq \int_{\widehat{\mathcal{X}}} H_\omega \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}(\omega) < \infty$$

and

$$u \geq \operatorname*{ess\,sup}_{\omega \in \widehat{\mathcal{X}}} \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} + \sup_{\delta \in \mathcal{D}_\mathcal{C}} \|K_e(\delta)\|_{\mathcal{Y},\mathcal{Y}} < \infty$$

and

$$v \geq \sup_{\delta \in \mathcal{D}_\mathcal{C}} D \|V(\delta)\|_{\mathcal{Y},\mathcal{Y}} < \infty.$$

define $p_{int} \geq \sup_{\delta \in \mathcal{D}_\mathcal{C}} \operatorname{IntDim}(V(\delta))$ then for all $\sqrt{\frac{v}{D}} + \frac{u}{3D} < \epsilon < m|\mathcal{C}|$,

$$\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}\left\{ (\omega_j)_{j=1}^D \;\middle|\; \sup_{\delta \in \mathcal{D}_\mathcal{C}} \|F(\delta)\|_{\mathcal{Y},\mathcal{Y}} \geq \epsilon \right\} \leq \; 8\sqrt{2} \left(\frac{m|\mathcal{C}|}{\epsilon}\right) p_{int}^{\frac{1}{d+1}} \exp\left(-D\psi_{v,d,u}(\epsilon)\right)$$

where $\psi_{v,d,u}(\epsilon) = \frac{\epsilon^2}{2(d+1)(v+u\epsilon/3)}$.

Again a full comprehensive proof is given in appendix B.2 of the appendix. Notice that in this result, The dimension $p = \dim \mathcal{Y}$ does not appear. Only the intrinsic dimension of the variance of the estimator. Moreover when d is large, the term $p_{int}^{\frac{1}{d+1}}$ goes to one, so that the impact of the intrinsic dimension on the bound vanish when the dimension of the input space is large. subsectionVariance of the approximation We now provide a bound on the norm of the variance of $\tilde{K}$, required to apply Corollaries 31 and 32. This is an extension of the proof of Sutherland and Schneider (2015) to the operator-valued case, and we recover their results in the scalar case when $A(\omega) = 1$. An illustration of the bound is provided in Figure 6 for the decomposable and the curl-free index˙cmd.

Proposition 33 (Bounding the *variance* of $\tilde{\boldsymbol{K}}_{\boldsymbol{e}}$)

Let K be a shift invariant $\mathcal{Y}$-Mercer kernel on a second countable index˙cmd topological space $\mathcal{X}$. Let $A : \widehat{\mathcal{X}} \to \mathcal{L}(\mathcal{Y})$ and $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ a pair such that $\tilde{K}_e = \sum_{j=1}^{D} \cos(\cdot, \omega_j) A(\omega_j) \approx K_e$, $\omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ index˙cmd Then,

$$\mathbf{Var}_{\widehat{\mathbf{Haar}},\rho}\left[\tilde{K}_e(\delta)\right] \preccurlyeq \frac{1}{2D}\left((K_e(2\delta) + K_e(e)) \, \mathbf{E}_{\widehat{\mathbf{Haar}},\rho}[A(\omega)] - 2K_e(\delta)^2 + \mathbf{Var}_{\widehat{\mathbf{Haar}},\rho}[A(\omega)] \right)$$

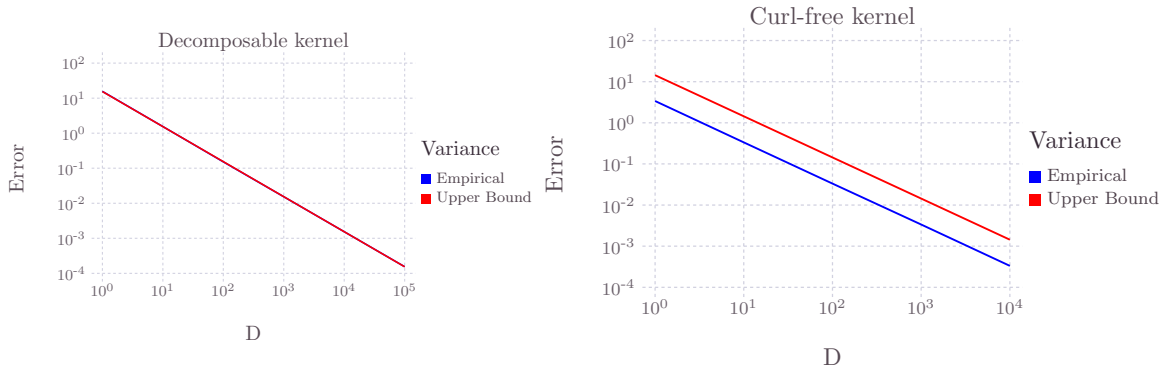

Figure 6: Comparison between an empirical bound on the norm of the variance of the decomposable (left) and curl-free (right) ORFF obtained and the theoretical bound proposed in Proposition 33 versus D.

4.4 Application on decomposable, curl-free and divergence-free

First, the two following examples discuss the form of H_ω for the additive group and the skewed-multiplicative group. Here we view $\mathcal{X} = \mathbb{R}^d$ as a Banach space endowed with the Euclidean norm. Thus the Lipschitz constant H_ω is bounded by the supremum of the norm of the gradient of h_ω.

Example 3 (Additive group)

On the additive group, $h_\omega(\delta) = \langle \omega, \delta \rangle$. *Hence* $H_\omega = \|\omega\|_2$.

Example 4 (Skewed-multiplicative group)

On the skewed multiplicative group, $h_\omega(\delta) = \langle \omega, \log(\delta + c) \rangle$. *Therefore* $\sup_{\delta \in \mathcal{C}} \|\nabla h_\omega(\delta)\|_2 = \sup_{\delta \in \mathcal{C}} \|\omega/(\delta + c)\|_2$. *Eventually* $\mathcal{C}$ *is compact subset of* $\mathcal{X}$ *and finite dimensional thus* $\mathcal{C}$ *is closed and bounded. Thus* $H_\omega = \|\omega\|_2 / (\min_{\delta \in \mathcal{C}} \|\delta\|_2 + c)$.

Now we compute upper bounds on the norm of the variance and Orlicz norm of the three index˙cmd we took as examples.

4.4.1 Decomposable kernel

notice that in the case of the Gaussian decomposable kernel, index˙cmd $A(\omega) = A$, $e = 0$, $K_0(\delta) = Ak_0(\delta)$, $k_0(\delta) \geq 0$ and $k_0(\delta) = 1$, then we have

$$D \left\| \mathbf{Var}_\mu \left[\tilde{K}_0(\delta) \right] \right\|_{\mathcal{Y},\mathcal{Y}} \leq (1 + k_0(2\delta)) \|A\|_{\mathcal{Y},\mathcal{Y}}/2 + k_0(\delta)^2.$$

4.4.2 Curl-free and divergence-free kernels:

recall that in this case $p = d$. For the (Gaussian) curl-free kernel, $A(\omega) = \omega\omega^*$ where $\omega \in \mathbb{R}^d \sim \mathcal{N}(0, \sigma^{-2} I_d)$ thus $\mathbf{E}_\mu[A(\omega)] = I_d/\sigma^2$ and $\mathbf{Var}_\mu[A(\omega)] = (d+1) I_d/\sigma^4$. Hence,

$$D \left\| \mathbf{Var}_\mu \left[\tilde{K}_0(\delta) \right] \right\|_{\mathcal{Y},\mathcal{Y}} \leq \frac{1}{2} \left\| \frac{1}{\sigma^2} K_0(2\delta) - 2K_0(\delta)^2 \right\|_{\mathcal{Y},\mathcal{Y}} + \frac{(d+1)}{\sigma^4}.$$

This bound is illustrated by Figure 5 B, for a given datapoint. Eventually for the Gaussian divergence-free kernel, $A(\omega) = I\|\omega\|_2^2 - \omega\omega^*$, thus $\mathbf{E}_\mu[A(\omega)] = I_d(d-1)/\sigma^2$ and $\mathbf{Var}_\mu[A(\omega)] = d(4d-3)I_d/\sigma^4$. Hence,

$$D\left\|\mathbf{Var}_\mu\left[\tilde{K}_0(\delta)\right]\right\|_{\mathcal{Y},\mathcal{Y}} \leq \frac{1}{2}\left\|\frac{(d-1)}{\sigma^2}K_0(2\delta) - 2K_0(\delta)^2\right\|_{\mathcal{Y},\mathcal{Y}} + \frac{d(4d-3)}{\sigma^4}.$$

To conclude, we ensure that the random variable $\|A(\omega)\|_{\mathcal{Y},\mathcal{Y}}$ has a finite Orlicz norm with $\psi = \psi_1$ in these three cases.

4.4.3 Computing the Orlicz norm

for a random variable with strictly monotonic moment generating function (MGF), one can characterize its inverse ψ_1 Orlicz norm by taking the functional inverse of the MGF evaluated at 2 (see Lemma 43 of the appendix). In other words $\|X\|_{\psi_1}^{-1} = \mathrm{MGF}(x)_X^{-1}(2)$. For the Gaussian curl-free and divergence-free kernel,

$$\left\|A^{div}(\omega)\right\|_{\mathcal{Y},\mathcal{Y}} = \left\|A^{curl}(\omega)\right\|_{\mathcal{Y},\mathcal{Y}} = \|\omega\|_2^2,$$

where $\omega \sim \mathcal{N}(0, I_d/\sigma^2)$, hence $\|A(\omega)\|_2 \sim \Gamma(p/2, 2/\sigma^2)$. The MGF of this gamma distribution is $\mathrm{MGF}(x)(t) = (1 - 2t/\sigma^2)^{-(p/2)}$. Eventually

$$\left\|\left\|A^{div}(\omega)\right\|_{\mathcal{Y},\mathcal{Y}}\right\|_{\psi_1}^{-1} = \left\|\left\|A^{curl}(\omega)\right\|_{\mathcal{Y},\mathcal{Y}}\right\|_{\psi_1}^{-1} = \frac{\sigma^2}{2}\left(1 - 4^{-\frac{1}{p}}\right).$$

5. Learning with index˙cmd

Before focusing on learning function with an ORFF model, we briefly review the context of supervised learning in index˙cmd. model.

5.1 Supervised learning within index˙cmd

Let $\boldsymbol{s} = (x_i, y_i)_{i=1}^N \in (\mathcal{X} \times \mathcal{Y})^N$ be a sequence of training samples. Given a local loss function $L : \mathcal{X} \times \mathcal{F} \times \mathcal{Y} \to \overline{\mathbb{R}}$ such that L is proper, convex and lower semi-continuous in $\mathcal{F}$, we are interested in finding a *vector-valued function* $f_{\boldsymbol{s}} : \mathcal{X} \to \mathcal{Y}$, that lives in a index˙cmd and minimize a tradeoff between a data fitting term L and a regularization term to prevent from overfitting. Namely finding $f_{\boldsymbol{s}} \in \mathcal{H}_K$ such that

$$f_{\boldsymbol{s}} = \operatorname*{arg\,min}_{f \in \mathcal{H}_K} \frac{1}{N}\sum_{i=1}^N L(x_i, f, y_i) + \frac{\lambda}{2}\|f\|_K^2 \tag{20}$$

where $\lambda \in \mathbb{R}_+$ is a (Tychonov) regularization hyperparameter. We call the quantity

$$\mathcal{R}_\lambda(f, \boldsymbol{s}) = \frac{1}{N}\sum_{i=1}^N L(x_i, f, y_i) + \frac{\lambda}{2}\|f\|_K^2, \quad \forall f \in \mathcal{H}_K, \forall \boldsymbol{s} \in (\mathcal{X} \times \mathcal{Y})^N.$$

the (Tychonov) regularized risk of the model $f \in \mathcal{H}_K$ according the local loss L. A common choice for L is the squared error loss $L : (x, f, y) \mapsto \|f(x) - y\|_{\mathcal{Y}}^2$ which yields the vector-valued ridge regression problem.

5.1.1 Representer theorem and Feature equivalence

Regression in has been well studied (Álvarez et al., 2012; Argyriou et al., 2009; Minh et al., 2013a, 2016; Sangnier et al., 2016; Kadri et al., 2015; Micchelli and Pontil, 2005; Brouard et al., 2016a), and a cornerstone of learning in index˙cmd is the representer theorem[6], which allows to replace the search of a minimizer in a infinite dimensional index˙cmd by a finite number of parameters $(u_i)_{i=1}^N$, $u_i \in \mathcal{Y}$.

In the following we suppose we are given a cost function $c : \mathcal{Y} \times \mathcal{Y} \to \overline{\mathbb{R}}$, such that $c(f(x), y)$ returns the error of the prediction $f(x)$ index˙cmd the ground truth y. A loss function of a model f with respect to an example $(x, y) \in \mathcal{X} \times \mathcal{Y}$ can be naturally defined from a cost function as $L(x, f, y) = c(f(x), y)$. Conceptually the function c evaluates the quality of the prediction versus its ground truth $y \in \mathcal{Y}$ while the loss function L evaluates the quality of the model f at a training point $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

Theorem 34 (Representer theorem)

Let K be a $\mathcal{Y}$-Mercer and $\mathcal{H}_K$ its corresponding $\mathcal{Y}$-Reproducing Kernel Hilbert space. Let $c : \mathcal{Y} \times \mathcal{Y} \to \overline{\mathbb{R}}$ be a cost function such that $L(x, f, y) = c(Vf(x), y)$ is a proper convex lower semi-continuous function in f for all $x \in \mathcal{X}$ and all $y \in \mathcal{Y}$. Eventually let $\lambda \in \mathbb{R}_{>0}$ be the Tychonov regularization hyperparameters The solution $f_{\boldsymbol{s}} \in \mathcal{H}_K$ of the regularized optimization problem

$$f_{\boldsymbol{s}} = \underset{f \in \mathcal{H}_K}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} c(f(x_i), y_i) + \frac{\lambda}{2} \|f\|_K^2 \tag{21}$$

has the form $f_{\boldsymbol{s}} = \sum_{j=1}^N K(\cdot, x_j) u_{\boldsymbol{s},j}$ where $u_{\boldsymbol{s},j} \in \mathcal{Y}$ and

$$u_{\boldsymbol{s}} = \underset{u \in \bigoplus_{i=1}^N \mathcal{Y}}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} c\left(\sum_{k=1}^{N} K(x_i, x_j) u_j, y_i \right) + \frac{\lambda}{2} \sum_{k=1}^{N} u_i^* K(x_i, x_k) u_k. \tag{22}$$

The first representer theorem was introduced by Wahba (1990) in the case where $\mathcal{Y} = \mathbb{R}$. The extension to an arbitrary Hilbert space $\mathcal{Y}$ has been proved by many authors in different forms (Brouard et al., 2011; Kadri et al., 2015; Micchelli and Pontil, 2005). The idea behind the representer theorem is that even though we minimize over the whole space $\mathcal{H}_K$, when $\lambda > 0$, the solution of Equation 21 falls inevitably into the set $\mathcal{H}_{K,\boldsymbol{s}} = \left\{ \sum_{j=1}^N K_{x_j} u_j \;\middle|\; \forall (u_i)_{i=1}^N \in \mathcal{Y}^N \right\}$. Therefore the result can be expressed as a finite linear combination of basis functions of the form $K(\cdot, x_k)$. Notice that we can perform the kernel expansion of $f_{\boldsymbol{s}} = \sum_{j=1}^N K(\cdot, x_j) u_{\boldsymbol{s},j}$ even though $\lambda = 0$. However $f_{\boldsymbol{s}}$ is no longer the solution of Equation 21 over the whole space $\mathcal{H}_K$ but a projection on the subspace $\mathcal{H}_{K,\boldsymbol{s}}$. The representer theorem show that minimizing a functional in a index˙cmd yields a solution which depends on all the points in the training set. Assuming that for all x_i and $x \in \mathcal{X}$ and for all $u_i \in \mathcal{Y}$ it takes time $O(P)$ to compute $K(x_i, x)u_i$, making a prediction using the representer theorem takes $O(NP)$. Obviously If $\mathcal{Y} = \mathbb{R}^p$, Then $P = O(p^2)$ thus making a prediction cost $O(Np^2)$ operations.

6. Sometimes referred to as minimal norm interpolation theorem.

5.2 Learning with Operator Random Fourier Feature maps

Instead of learning a model f that depends on all the points of the training set, we would like to learn a parametric model of the form $\widetilde{f}(x) = \widetilde{\phi}(x)^*\theta$, where θ lives in some space $\widetilde{\mathcal{H}}$. We are interested in finding a parameter vector $\theta_{\boldsymbol{s}}$ such that

$$\theta_{\boldsymbol{s}} = \underset{\theta\in\widetilde{\mathcal{H}}}{\arg\min}\, \mathfrak{R}_\lambda(\theta, \boldsymbol{s}) = \underset{\theta\in\widetilde{\mathcal{H}}}{\arg\min}\, \frac{1}{N}\sum_{i=1}^{N} c\left(\widetilde{\phi}(x_i)^*\theta, y_i\right) + \frac{\lambda}{2}\|\theta\|^2_{\widetilde{\mathcal{H}}} \tag{23}$$

The following theorem states that when $\lambda > 0$ then learning with a feature map is equivalent to learn with a kernel. Moreover if $f_{\boldsymbol{s}} \in \mathcal{H}_K$ is a solution of Equation 21 and $\theta_{\boldsymbol{s}} \in \mathcal{H}$ is the solution of Equation 24, then $f_{\boldsymbol{s}} = \phi(\cdot)^*\theta_{\boldsymbol{s}}$. This equivalence could have been obtained by means of Lagrange duality. However in this proof we do not use such tool: we only focus on the representer theorem and the fact that there exists a partial isometry W between the index˙cmd and a feature space $\mathcal{H}$. We show that if $\theta_{\boldsymbol{s}}$ is a solution of *Equation* 23, then theta belongs to $(\text{Ker } W)^\perp$, thus there is an isometry between $\theta_{\boldsymbol{s}} \in \widetilde{\mathcal{H}}$ and $\mathcal{H}_{\widetilde{K}}$: namely W.

Theorem 35 (Feature equivalence)
Let $\widetilde{K}$ be an such that for all x, $z \in \mathcal{X}$, $\widetilde{\phi}(x)^*\widetilde{\phi}(z) = \widetilde{K}(x, z)$ where $\widetilde{K}$ is a $\mathcal{Y}$-Mercer index˙cmd and $\mathcal{H}_{\widetilde{K}}$ its corresponding $\mathcal{Y}$-Reproducing kernel Hilbert space. Let $c : \mathcal{Y} \times \mathcal{Y} \to \overline{\mathbb{R}}$ be a cost function such that $L\left(x, \widetilde{f}, y\right) = c\left(\widetilde{f}(x), y\right)$ is a proper convex lower semi-continuous function in $\widetilde{f} \in \mathcal{H}_{\widetilde{K}}$ for all $x \in \mathcal{X}$ and all $y \in \mathcal{Y}$. Eventually let $\lambda \in \mathbb{R}_{>0}\mathbb{R}_+$ be the Tychonov regularization hyperparameter. The solution $f_{\boldsymbol{s}} \in \mathcal{H}_{\widetilde{K}}$ of the regularized optimization problem

$$\widetilde{f}_{\boldsymbol{s}} = \underset{\widetilde{f}\in\mathcal{H}_{\widetilde{K}}}{\arg\min}\, \frac{1}{N}\sum_{i=1}^{N} c\left(\widetilde{f}(x_i), y_i\right) + \frac{\lambda}{2}\left\|\widetilde{f}\right\|^2_{\widetilde{K}} \tag{24}$$

has the form $\widetilde{f}_{\boldsymbol{s}} = \widetilde{\phi}(\cdot)^*\theta_{\boldsymbol{s}}$, where $\theta_{\boldsymbol{s}} \in (Ker\, \widetilde{W})^\perp$ and

$$\theta_{\boldsymbol{s}} = \underset{\theta\in\widetilde{\mathcal{H}}}{\arg\min}\, \frac{1}{N}\sum_{i=1}^{N} c\left(\widetilde{\phi}(x_i)^*\theta, y_i\right) + \frac{\lambda}{2}\|\theta\|^2_{\widetilde{\mathcal{H}}} \tag{25}$$

In the aforementioned theorem, we use the notation $\widetilde{K}$ and $\widetilde{\phi}$ because our main subject of interest is the index˙cmd map. However this theorem works for *any* feature maps $\phi(x) \in \mathcal{L}(\mathcal{Y}, \mathcal{H})$ even when $\mathcal{H}$ is infinite dimensional.[7]. This shows that when $\lambda > 0$ the solution of Equation 22 with the approximated kernel $K(x, z) \approx \widetilde{K}(x, z) = \widetilde{\phi}(x)^*\widetilde{\phi}(z)$ is the same than the solution of Equation 25 up to an isometric isomorphm (see appendix B.3.2). Namely, if $u_{\boldsymbol{s}}$ is the solution of Equation 22, $\theta_{\boldsymbol{s}}$ is the solution of Equation 25 and $\lambda > 0$ we have

$$\theta_{\boldsymbol{s}} = \sum_{i=1}^{N} \widetilde{\phi}(x_i)(u_{\boldsymbol{s}})_i \in (\text{Ker } W)^\perp \subseteq \widetilde{\mathcal{H}}.$$

7. If $\phi(x) : \mathcal{L}(\mathcal{Y}, \mathcal{H})$ and $\dim(\mathcal{H}) = \infty$, the decomposition $\mathcal{H} = (\text{Ker } W) \oplus (\text{Ker } W)^\perp$ holds since $\mathcal{H}$ is a Hilbert space and W is a bounded operator.

If $\lambda_K = 0$ we can still find a solution $u_{\boldsymbol{s}}$ of Equation 22. By construction of the kernel expansion, we have $u_{\boldsymbol{s}} \in (\text{Ker } W)^\perp$. However looking at the proof of Theorem 35 we see that $\theta_{\boldsymbol{s}}$ might *not* belong to $(\text{Ker } W)^\perp$. We can compute a residual vector $r_{\boldsymbol{s}} = \sum_{i=1}^N \widetilde{\phi}(x_i)(u_{\boldsymbol{s}})_i - \theta_{\boldsymbol{s}}$. Since $\sum_{j=1}^N \widetilde{\phi}(x_j) \in (\text{Ker } W)^\perp$ by construction, if $r_{\boldsymbol{s}} = 0$, it means that λ_K is large enough for both representer theorem and index˙cmd representer theorem to apply. If $r_{\boldsymbol{s}} \neq 0$ but $\widetilde{\phi}(\cdot)^* r_{\boldsymbol{s}} = 0$ it means that both $\theta_{\boldsymbol{s}}$ and $\sum_{j=1}^N \widetilde{\phi}(x_j) u_{\boldsymbol{s}}$ are in $(\text{Ker } W)^\perp$, thus the representer theorem fails to find the "true" solution over the whole space $\mathcal{H}_{\widetilde{K}}$ but returns a projection onto $\mathcal{H}_{\widetilde{K},\boldsymbol{s}}$ of the solution. If $r_{\boldsymbol{s}} \neq 0$ and $\widetilde{\phi}(\cdot)^* r_{\boldsymbol{s}} \neq 0$ means that $\theta_{\boldsymbol{s}}$ is *not* in $(\text{Ker } W)^\perp$, thus the feature equivalence theorem fails to apply. Since $r_{\boldsymbol{s}} = \sum_{i=1}^N \widetilde{\phi}(x_i)(u_{\boldsymbol{s}})_i - \theta_{\boldsymbol{s}}^\perp - \theta_{\boldsymbol{s}}^\parallel$ and $\sum_{i=1}^N \widetilde{\phi}(x_i)(u_{\boldsymbol{s}})_i$ is in $(\text{Ker } W)^\perp$, with mild abuse of notation we write $r_{\boldsymbol{s}} = \theta^\parallel$. This remark is illustrated in Figure 7.

In Figure 7, we generated the data from a since wave to which we add some Gaussian noise. We learned a Gaussian kernel based index˙cmd model (blue curve) and a kernel model (yellow curve) where the kernel is obtained from the index˙cmd map. The left column represents the fit of the model to the points for four different valued of λ (top to bottom: 10^{-2}, 10^{-5}, $10e^{10}$, 0). The middle column shows if the index˙cmd solution $\theta_{\boldsymbol{s}}$ is in $(\text{Ker } \tilde{W})^\perp$. This is true for all values of λ. The right column shows that even though $\theta_{\boldsymbol{s}}$ is in $(\text{Ker } \tilde{W})^\perp$, when $\lambda \to 0$ learning with index˙cmd is different from learning with the kernel constructed from the index˙cmd maps since the coefficients of $\theta^\parallel$ are all different from 0.

5.3 Solving ORFF-based regression

In order to find a solution to Equation 23, we turn our attention to gradient descent methods. We define an algorithm (Algorithm 3) to find efficiently a solution to Equation 23 when $c(y, y') = \|y = y'\|_{\mathcal{Y}}^2$ and study its complexity.

5.3.1 Gradient methods

Since the solution of Equation 23 is unique when $\lambda > 0$, a sufficient and necessary condition is that the gradient of $\mathfrak{R}_\lambda$ at the minimizer $\theta_{\boldsymbol{s}}$ is zero. We use the Frechet derivative, the strongest notion of derivative in Banach spaces (Conway, 2013; Kurdila and Zabarankin, 2006) which directly generalizes the notion of gradient to Banach spaces. The chain rule is valid in this context (Kurdila and Zabarankin, 2006, theorem 4.1.1 page 140). Hence

$$\nabla_\theta c\left(\widetilde{\phi}(x_i)^*\theta, y_i\right) = \widetilde{\phi}(x_i)\left(\frac{\partial}{\partial y} c\left(y, y_i\right)\Big|_{y=\widetilde{\phi}(x_i)^*\theta}\right)^*, \text{ and } \nabla_\theta \|\theta\|_{\widetilde{\mathcal{H}}}^2 = 2\theta.$$

Provided that $c(y, y_i)$ is Frechet differentiable index˙cmd y, for all y and $y_i \in \mathcal{Y}$ we have $\nabla_\theta \mathfrak{R}_\lambda(\theta, \boldsymbol{s}) \in \widetilde{\mathcal{H}}$ and

$$\nabla_\theta \mathfrak{R}_\lambda(\theta, \boldsymbol{s}) = \frac{1}{N}\sum_{i=1}^N \widetilde{\phi}(x_i)\left(\frac{\partial}{\partial y} c\left(y, y_i\right)\Big|_{y=\widetilde{\phi}(x_i)^*\theta}\right)^* + \lambda\theta \tag{26}$$

Example 5 (Naive closed form for the squared error cost)

Figure 7: index˙cmd equivalence theorem.

Algorithm 3: Naive closed form for the squared error cost.

Input :

- $\boldsymbol{s} = (x_i, y_i)_{i=1}^N \in (\mathcal{X} \times \mathbb{R}^p)^N$ a sequence of supervised training points,
- $\widetilde{\phi}(x_i) \in \mathcal{L}(\mathbb{R}^p, \mathbb{R}^r)$ a feature map defined for all $x_i \in \mathcal{X}$,
- $\lambda \in \mathbb{R}_{>0}$ the Tychonov regularization term,

Output : A model $h : \mathcal{X} \to \mathbb{R}^p$, $h(x) = \widetilde{\phi}(x)^\mathsf{T}\theta_{\boldsymbol{s}}$. such that $\theta_{\boldsymbol{s}}$ minimize Equation 23, where $c(y, y') = \|y - y'\|_2^2$ and $\mathbb{R}^r$ and $\mathbb{R}^p$

$\mathbf{P} \leftarrow \frac{1}{N}\sum_{i=1}^N \widetilde{\phi}(x_i)\widetilde{\phi}(x_i)^\mathsf{T} \in \mathcal{L}(\mathbb{R}^r, \mathbb{R}^r)$;

$\mathbf{Y} \leftarrow \frac{1}{N}\sum_{i=1}^N \widetilde{\phi}(x_i)y_i \in \mathbb{R}^r$;

$\theta_{\boldsymbol{s}} \leftarrow \mathrm{solve}_\theta\left((\mathbf{P} + \lambda I_r)\theta = \mathbf{Y}\right)$;

return $h : x \mapsto \widetilde{\phi}(x)^\mathsf{T}\theta_{\boldsymbol{s}}$;

Consider the cost function defined for all $y,\ y' \in \mathcal{Y}$ *by* $c(y, y') = \frac{1}{2}\|y - y\|_{\mathcal{Y}}^2$. *Then* $\left(\frac{\partial}{\partial y}c\left(y, y_i\right)\Big|_{y=\widetilde{\phi}(x_i)^*\theta}\right)^* = \left(\widetilde{\phi}(x_i)^*\theta - y_i\right)$. *Thus, since the optimal solution* $\theta_{\boldsymbol{s}}$ *verifies* $\nabla_{\theta_{\boldsymbol{s}}}\mathfrak{R}_\lambda(\theta_{\boldsymbol{s}}, \boldsymbol{s}) = 0$ *we have* $\frac{1}{N}\sum_{i=1}^N \widetilde{\phi}(x_i)\left(\widetilde{\phi}(x_i)^*\theta_{\boldsymbol{s}} - y_i\right) + \lambda\theta_{\boldsymbol{s}} = 0$. *Therefore,*

$$\left(\frac{1}{N}\sum_{i=1}^N \widetilde{\phi}(x_i)\widetilde{\phi}(x_i)^* + \lambda I_{\widetilde{\mathcal{H}}}\right)\theta_{\boldsymbol{s}} = \frac{1}{N}\sum_{i=1}^N \widetilde{\phi}(x_i)y_i. \tag{27}$$

Suppose that $\mathcal{Y} \subseteq \mathbb{R}^p$, *and for all* $x \in \mathcal{X}$, $\widetilde{\phi}(x) : \mathbb{R}^r \to \mathbb{R}^p$ *where all spaces are endowed with the Euclidean inner product. From this we can derive Algorithm 3 which returns the closed form solution of Equation* 23 *for* $c(y, y') = \frac{1}{2}\|y - y'\|_2^2$.

5.3.2 Complexity analysis

Algorithm 3 constitutes our first step toward large-scale learning with . We can easily compute the time complexity of Algorithm 3 when all the operators act on finite dimensional Hilbert spaces. Suppose that $p = \dim(\mathcal{Y}) < \infty$ and for all $x \in \mathcal{X}$, $\widetilde{\phi}(x) : \mathcal{Y} \to \widetilde{\mathcal{H}}$ where $r = \dim(\widetilde{\mathcal{H}}) < \infty$ is the dimension of the redescription space $\widetilde{\mathcal{H}} = \mathbb{R}^r$. Since p and $r < \infty$, we view the operators $\widetilde{\phi}(x)$ and $I_{\widetilde{\mathcal{H}}}$ as matrices. Step 1 costs $O_t(Nr^2p)$. Steps 2 costs $O_t(Nrp)$. For step 3, the naive inversion of the operator costs $O_t(r^3)$. Eventually the overall complexity of Algorithm 3 is $O_t\left(r^2(Np + r)\right)$, while the space complexity is $O_s(r^2)$.

This complexity is to compare with the kernelized solution. Let

$$\mathbf{K} : \begin{cases} \mathcal{Y}^N \to \mathcal{Y}^N \\ u \mapsto \bigoplus_{i=1}^{N+U}\sum_{j=1}^{N+U} K(x_i, x_j)u_j \end{cases}$$

When $\mathcal{Y} = \mathbb{R}$,

$$\mathbf{K} = \begin{pmatrix} K(x_1, x_1) & \dots & K(x_1, x_{N+U}) \\ \vdots & \ddots & \vdots \\ K(x_{N+U}, x_1) & \dots & K(x_{N+U}, x_{N+U}) \end{pmatrix}$$

is called the Gram matrix of K. When $\mathcal{Y} = \mathbb{R}^p$, $\mathbf{K}$ is a matrix-valued Gram matrix of size $pN \times pN$ where each entry $\mathbf{K}_{ij} \in \mathcal{M}_{p,p}(\mathbb{R})$. Then the equivalent kernelized solution $u_{\boldsymbol{s}}$ of Theorem 34 is

$$\left(\frac{1}{N}\mathbf{K} + \lambda I_{\bigoplus_{i=1}^N \mathcal{Y}}\right) u_{\boldsymbol{s}} = \bigoplus_{i=1}^N y_i.$$

which has time complexity $O_t\left(N^3p^3\right)$ and space complexity $O_s\left(N^2p^2\right)$. Suppose we are given a generic index˙cmd map (see Subsection 3.5). Then $r = 2Dp$, where D is the number of samples. Hence Algorithm 3 is better that its kernelized counterpart when $r = 2Dp$ is small compared to Np. Thus, roughly speaking it is better to use Algorithm 3 when the number of features, r, required is small compared to the number of training points. Notice that Algorithm 3 has a linear complexity with respect to the number of supervised training points N so it is better suited to large scale learning provided that D does not grows linearly with N. Yet naive learning with Algorithm 3 by viewing all the operators as matrices is still problematic. Indeed learning p independent models with scalar Random Fourier Features would cost $O_t\left(D^2p^3(N+D)\right)$ since $r = 2Dp$. This Means that learning vector-valued function has increased the (expected) complexity from p to p^3. However in some cases we can drastically reduce the complexity by viewing the feature-maps as linear operators rather than matrices.

5.4 Efficient learning with ORFF

When developping Algorithm 3 we considered that the feature map $\widetilde{\phi}(x)$ was a matrix from $\mathbb{R}^p$ to $\mathbb{R}^r$ for all $x \in \mathcal{X}$, and therefore that computing $\widetilde{\phi}(x)\widetilde{\phi}(z)^\mathsf{T}$ has a time complexity of $O(r^2p)$. While this holds true in the most generic senario, in many cases the feature maps present some structure or sparsity allowing to reduce the computational cost of evaluating the feature map. We focus on the given by Algorithm 1, developped in Subsection 3.3 and Subsection 3.5 and treat the decomposable kernel, the curl-free kernel and the divergence-free kernel as an example. We recall that if $\mathcal{Y}' = \mathbb{R}^{p'}$ and $\mathcal{Y} = \mathbb{R}^p$, then $\widetilde{\mathcal{H}} = \mathbb{R}^{2Dp'}$ thus the given in Section 3 have the form

$$\begin{cases} \widetilde{\phi}(x) \in \mathcal{L}\left(\mathbb{R}^p, \mathbb{R}^{2Dp'}\right) & : y \mapsto \frac{1}{\sqrt{D}} \bigoplus_{j=1}^D (x, \omega_j) B(\omega_j)^\mathsf{T} y \\ \widetilde{\phi}(x)^\mathsf{T} \in \mathcal{L}\left(\mathbb{R}^{2Dp'}, \mathbb{R}^p\right) & : \theta \mapsto \frac{1}{\sqrt{D}} \sum_{j=1}^D (x, \omega_j) B(\omega_j) \theta_j \end{cases},$$

where $\omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ index˙cmd and $B(\omega_j) \in \mathcal{L}\left(\mathbb{R}^p, \mathbb{R}^{p'}\right)$ for all $\omega_j \in \widehat{\mathcal{X}}$. Hence the can be seen as the block matrix $\in \mathcal{M}_{2Dp',p}(\mathbb{R})$

$$\widetilde{\phi}(x) = \begin{pmatrix} \cos\langle x, \omega_1\rangle B(\omega_1) & \sin\langle x, \omega_1\rangle B(\omega_1) & \dots & \cos\langle x, \omega_D\rangle B(\omega_D) & \sin\langle x, \omega_D\rangle B(\omega_D) \end{pmatrix}^\mathsf{T} \tag{28}$$

5.4.1 Case of study: the decomposable kernel

Throughout this section we show how the mathematical formulation relates to a concrete (Python) implementation. We propose a Python implementation based on NumPy (Oliphant, 2006), SciPy (Jones et al., 2014) and Scikit-learn (Pedregosa et al., 2011). Following Equation 28, the feature map associated to the decomposable kernel would be

$$\begin{aligned}\widetilde{\phi}(x) &= \frac{1}{\sqrt{D}}\begin{pmatrix}\cos\langle x,\omega_1\rangle B & \sin\langle x,\omega_1\rangle B & \dots & \cos\langle x,\omega_D\rangle B & \sin\langle x,\omega_D\rangle B\end{pmatrix}^\mathsf{T} \\ &= \underbrace{\frac{1}{\sqrt{D}}\begin{pmatrix}\cos\langle x,\omega_1\rangle & \sin\langle x,\omega_1\rangle & \dots & \cos\langle x,\omega_D\rangle & \sin\langle x,\omega_D\rangle\end{pmatrix}^\mathsf{T}}_{\widetilde{\varphi}(x)} \otimes B^\mathsf{T},\end{aligned}$$

$\omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ index˙cmd, which would lead to the following naive python implementation for the Gaussian (RBF) kernel of parameter γ, whose associated spectral distribution is $\mathbf{Pr}_\rho = \mathcal{N}(0, 2\gamma)$. Let $\theta \in \mathbb{R}^{2Dp'}$ and $y \in \mathbb{R}^{\text{ı}}$. With such implementation evaluating a matrix vector product such as $\widetilde{\phi}(x)^\mathsf{T}\theta$ or $\widetilde{\phi}(x)y$ have $O_t(2Dp'p)$ time complexity and $O_s(2Dp'p)$ of space complexity, which is utterly inefficient. Indeed, recall that if $B \in \mathcal{M}_{p,p'}\left(\mathbb{R}^{p'}\right)$ is matrix, the operator $\widetilde{\phi}(x)$ corresponding to the decomposable kernel is

$$\widetilde{\phi}(x)y = \frac{1}{\sqrt{D}}\bigoplus_{j=1}^{D}\begin{pmatrix}\cos\langle x,\omega_j\rangle B^\mathsf{T}y \\ \sin\langle x,\omega_j\rangle B^\mathsf{T}y\end{pmatrix} = \left(\frac{1}{\sqrt{D}}\bigoplus_{j=1}^{D}\begin{pmatrix}\cos\langle x,\omega_j\rangle \\ \sin\langle x,\omega_j\rangle\end{pmatrix}\right) \otimes (B^\mathsf{T}y) \tag{29}$$

and

$$\widetilde{\phi}(x)^\mathsf{T}\theta = \frac{1}{\sqrt{D}}\sum_{j=1}^{D}\cos\langle x,\omega_j\rangle B\theta_j + \sin\langle x,\omega_j\rangle B\theta_j = B\left(\frac{1}{\sqrt{D}}\sum_{j=1}^{D}\left(\cos\langle x,\omega_j\rangle + \sin\langle x,\omega_j\rangle\right)\theta_j\right). \tag{30}$$

Which requires only evaluation of B on y and can be implemented easily in Python thanks to SciPy's LinearOperator. Note that the computation of these expressions can be fully vectorized[8] using the vectorization property of the Kronecker product. In the following we consider $\Theta \in \mathcal{M}_{2D,u'}(\mathbb{R})$ and the operator $\mathbf{vec} : \mathcal{M}_{p',2D}(\mathbb{R}) \to \mathbb{R}^{2Dp'}$ which turns a matrix into a vector (index˙cmd $\theta_{p'i+j} = \mathbf{vec}(\Theta_{ij})$, $i \in \mathbb{N}_{(2D-1)}$ and $j \in \mathbb{N}^*_{p'}$). Then $\left(\widetilde{\varphi}(x) \otimes B^\mathsf{T}\right)^\mathsf{T}\theta = \left(\widetilde{\varphi}(x)^\mathsf{T} \otimes B\right)\mathbf{vec}(\Theta) = \mathbf{vec}\left(B\Theta\widetilde{\varphi}(x)\right)$. with this trick, many authors (Sindhwani et al., 2013; Brault et al., 2016; Rosasco et al., 2010; Carmeli et al., 2010) notice that the decomposable kernel usually yields a Stein equation (Penzl, 1998). Indeed rewriting step 3 of Algorithm 3 gives a system to solve of the form

$$\widetilde{\varphi}(X)\widetilde{\varphi}(X)^\mathsf{T}\Theta B^\mathsf{T}B + \lambda\Theta - Y = 0 \Leftrightarrow \left(\widetilde{\varphi}(X)\widetilde{\varphi}(X)^\mathsf{T} \otimes B^\mathsf{T}B + \lambda I_{2Dp'}\right)\theta - Y = 0$$

Many solvers exists to solve efficiently this kind of systems[9], but most of them share the particularity that they are not just restricted to handle Stein equations. Broadly speaking,

8. See Walt et al. (2011).
9. For instance Sleijpen et al. (2010).

Figure 8: Efficient decomposable Gaussian ORFF (lower is better).

Figure 9: Efficient curl-free Gaussian ORFF (lower is better).

iterative solvers (or matrix free solvers) are designed to solve any systems of equation of ther form $PX = C$, where P is a linear operator (not a matrix). This is exacly our case where $\widetilde{\varphi}(x) \otimes B^T$ is the matrix form of the operator $\Theta \mapsto \mathbf{vec}(B\Theta\widetilde{\varphi}X)$.

This leads us to the following (more efficient) Python implementation of the Decomposable index˙cmd "operator" to be feed to a matrix-free solvers. **?? PythonTeX ??**It is worth mentioning that the same strategy can be applied in many different language. For instance in C++, the library Eigen (Guennebaud et al., 2010) allows to wrap a sparse matrix with a custom type, where the user overloads the transpose and dot product operator (as in Python). Then the custom user operator behaves as a (sparse) matrix –see `https://eigen.tuxfamily.org/dox/group__MatrixfreeSolverExample.html`. With this implementation the time complexity of $\widetilde{\phi}(x)^\mathsf{T}\theta$ and $\widetilde{\phi}(x)y$ falls down to $O_t((D+p)p')$ and the same holds for space complexity.

A quick experiment shows the advantage of seeing the decomposable kernel as a linear operator rather than a matrix. We draw $N = 100$ points $(x_i)_{i=1}^N$ in the interval $(0,1)^{20}$ and use a decomposable kernel with matrix $\Gamma = BB^\mathsf{T} \in \mathcal{M}_{p,p}(\mathbb{R})$ where $B \in \mathcal{M}_{p,p}(\mathbb{R})$ is a random matrix with coefficients drawn uniformly in $(0,1)$. We compute $\widetilde{\phi}(x)^\mathsf{T}\theta$ for all x_i's, where $\theta \in \mathcal{M}_{2D,1}(\mathbb{R})$, $D = 100$, with the implementation `EfficientDecomposableGaussianORFF`, Equation 30. The coefficients of θ were drawn at random uniformly in $(0,1)$. We report the execution time in Figure 8 for different values of p, $1 \leq p \leq 100$. The left plot reports the execution time in seconds of the construction of the feature. The middle plot reports the execution time of $\widetilde{\phi}(x)^\mathsf{T}\theta$, and the right plot the memory used in bytes to store $\widetilde{\phi}(x)$ for all x_i's. We averaged the results over ten runs.

Curl-free kernel. We use the unbounded index˙cmd map presented in Equation 17. We draw $N = 1000$ points $(x_i)_{i=1}^N$ in the interval $(0,1)^p$ and use a curl-free kernel. We compute $\widetilde{\phi}(x)^\mathsf{T}\theta$ for all x_i's, where $\theta \in \mathcal{M}_{2D,1}(\mathbb{R})$, $D = 500$, with the matrix implementation and the `LinearOperator` implementation. The coefficients of θ were drawn at random uniformly in $(0,1)$. We report the execution time in Figure 9 for different values of p, $1 \leq p \leq 100$. The left plot reports the execution time in seconds of the construction of the features. The middle plot reports the execution time of $\widetilde{\phi}(x)^\mathsf{T}\theta$, and the right plot the memory used in bytes to store $\widetilde{\phi}(x)$ for all x_i's. We averaged the results over fifty runs. As we can see the linear-operator implementation is one order of magnitude slower than its matrix counterpart. However it uses considerably less memory.

Divergence-free kernel. We use the unbounded index˙cmd map presented in Equation 18. We draw $N = 100$ points $(x_i)_{i=1}^N$ in the interval $(0,1)^p$ and use a curl-free kernel. We compute $\widetilde{\phi}(x)^\mathsf{T}\theta$ for all x_i's, where $\theta \in \mathcal{M}_{2Dp,1}(\mathbb{R})$, $D = 100$, with the matrix implementation and the `LinearOperator` implementation. The coefficients of θ were drawn at random uniformly in $(0,1)$. We report the execution time in Figure 9 for different values of

Figure 10: Efficient divergence-free Gaussian ORFF (lower is better).

p, $1 \leq p \leq 100$. The left plot reports the execution time in seconds of the construction of the feature. The middle plot reports the execution time of $\widetilde{\phi}(x)^\mathsf{T}\theta$, and the right plot the memory used in bytes to store $\widetilde{\phi}(x)$ for all x_i's. We averaged the results over ten runs. We draw the same conclusions as the curl-free kernel.

6. Numerical experiments

We present a set of experiments to complete the theoretical contribution and illustrate the behavior of ORFF-regression. First we study how well the ORFF regression recover the result of operator-valued kernel regression. Second we show the advantages of ORFF regression over independent RFF regression. A code implementing ORFF is available at `https://github.com/operalib/operalib` a framework for OVK Learning.

6.1 Learning with ORFF vs learning with OVK

6.1.1 Datasets

The *first dataset* considered is the handwritten digits recognition dataset MNIST[10]. We select a training set of $12,000$ images and a test set of $10,000$ images. The inputs are images represented as a vector $x_i \in [0, 255]^{784}$ and the targets $y_i \in \mathbb{N}_9$ are integers between 0 and 9. First we scaled the inputs such that they take values in $[-1, 1]^{784}$. Then we binarize the targets such that each number is represented by a unique binary vector of dimension 10. The vector y_i is zero everywhere except on the dimension corresponding to the class where it is one. For instance the class 4 is encoded $\begin{pmatrix}0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0\end{pmatrix}^\mathsf{T}$. To predict classes, we use the simplex coding method presented in Mroueh et al. (2012). The intuition behind simplex coding is to project the binarized labels of dimension p onto the most separated vectors on the hypersphere of dimension $p-1$. For ORFF we can encode directly this projection in the B matrix of the decomposable kernel $K_0(\delta) = BB^* k_0(\delta)$ where k_0 is a Gaussian kernel. The matrix B is computed via the recursion

$$B_{p+1} = \begin{pmatrix} 1 & u^T \\ 0_{p-1} & \sqrt{1-p^{-2}}B_p \end{pmatrix}, \qquad B_2 = \begin{pmatrix} 1 & -1 \end{pmatrix},$$

where $u = \begin{pmatrix} -p^{-2} & \dots & -p^{-2} \end{pmatrix}^T \in \mathbb{R}^{p-1}$ and $0_{p-1} = \begin{pmatrix} 0 & \dots & 0 \end{pmatrix}^T \in \mathbb{R}^{p-1}$. For we project the binarized targets on the simplex as a preprocessing step, before learning with the decomposable $K_0(\delta) = I_p k_0(\delta)$, where k_0 is a scalar Gaussian kernel.

The *second dataset* is a simulated five dimensional ($5D$) vector field with structure. We generate a scalar field as a random function $f : [-1, 1]^5 \to \mathbb{R}$, where $\widetilde{f}(x) = \widetilde{\varphi}(x)^* \theta$ where θ is a random matrix with each entry following a standard normal distribution, $\widetilde{\varphi}$ is a scalar Gaussian RFF with bandwidth $\sigma = 0.4$. The input data x are generated from a uniform probability distribution. We take the gradient of $\widetilde{f}$ to generate the curl-free $5D$ vector field.

10. available at `http://yann.lecun.com/exdb/mnist`

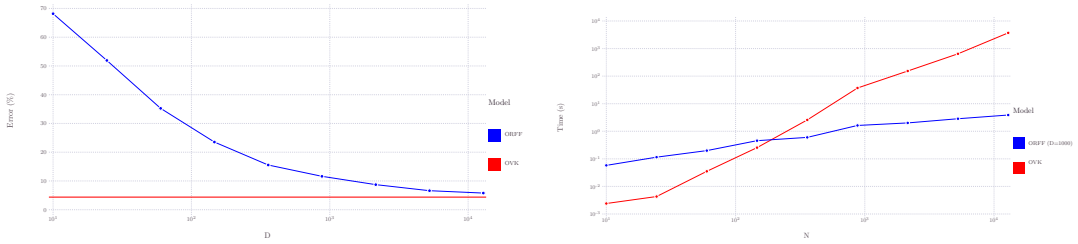

Figure 11: Empirical comparison of ORFF and OVK regression on MNIST dataset and empirical behavior of ORFF regression versus D and N.

The *third dataset* is a synthetic of data from $\mathbb{R}^{20} \to \mathbb{R}^4$ as described in Audiffren and Kadri (2015). In this dataset, inputs $(x_1, \ldots, x_{20})$ are generated independently and uniformly over $[0, 1]$ and the different outputs are computed as follows. Let $\varphi(x) = (x_1^2, x_4^2, x_1x_2, x_3x_5, x_2, x_4, 1)$ and (w_i) denotes the index˙cmd copies of a seven dimensional Gaussian distribution with zero mean and covariance $\Sigma \in \mathcal{M}_{7,7}(\mathbb{R})$ such that $\begin{pmatrix} 0.5 & 0.25 & 0.1 & 0.05 & 0.15 & 0.1 & 0.15 \end{pmatrix}$ Then, the outputs of the different tasks are generated as $y_i = w_i\varphi(x)$. We use this dataset with $p = 4$, 10^5 instances and for the train set and also 10^5 instances for the test set.

6.1.2 Results

Performance of ORFF regression on the first dataset. We trained both index˙cmd and index˙cmd models on MNIST dataset with a decomposable Gaussian kernel with signature $K_0(\delta) = \exp\left(-\|\delta\|/(2\sigma^2)\right)\Gamma$. To apply *Algorithm* 3 after noticing that in the case of the decomposable kernel with $\lambda_M = 0$, it boilds down to a Stein equation (Brault et al., 2016, section 5.1), we use an off-the-shelf solver[11] able to handle Stein's equation. For both methods we choose $\sigma = 20$ and use a 2-fold cross validation on the training set to select the optimal λ. First, Figure 11 compares the running time between OVK and ORFF models using $D = 1000$ Fourier features against the number of datapoints N. The log-log plot shows ORFF scaling better than the OVK index˙cmd the number of points. Second, Figure 11 shows the test prediction error versus the number of ORFFs D, when using $N = 1000$ training points. As expected, the ORFF model converges toward the OVK model when the number of features increases.

Performance of ORFF regression on the second dataset. We perform a similar experiment on the second dataset (5D-vector field with structure). We use a Gaussian curl-free kernel with bandwidth equal to the median of the pairwise distances and tune the hyperparameter λ on a grid. Here we optimize Equation 23, where c is the squared error cost, using Scipy's index˙cmd (Byrd et al., 1995) solver[12] with the gradients given in Equation 26 and the efficient linear operator described in Subsection 5.4 (index˙cmd Equations 29 and 30). Figure 12 (bottom row) reports the R^2 (coefficient of determination) score on the test set versus the number of curl-index˙cmd D with a comparison with curl-index˙cmd. In this

11. Available at `http://ta.twi.tudelft.nl/nw/users/gijzen/IDR.html`
12. Available at `http://docs.scipy.org/doc/scipy/reference/optimize.html`

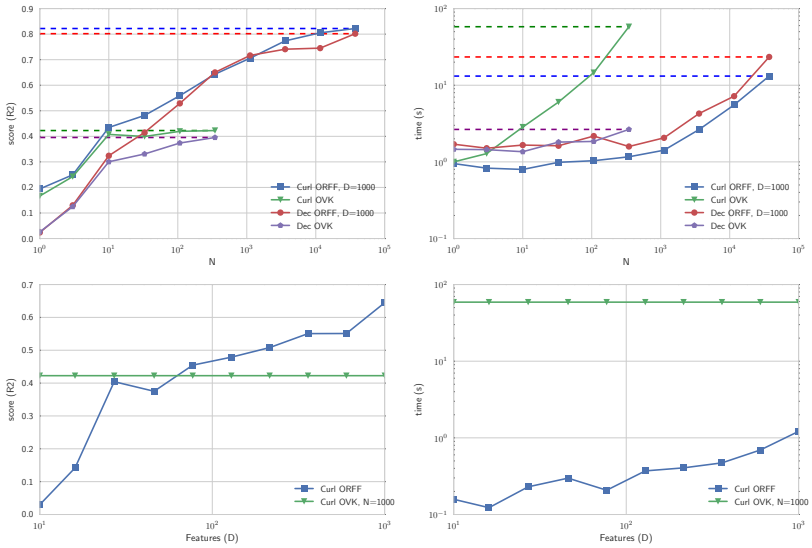

Figure 12: Empirical comparison between curl-free ORFF, curl-free OVK, independent ORFF, independent OVK on a synthetic vector field regression task.

experiment, we see that curl-index˙cmd can even be better than curl-index˙cmd, suggesting that index˙cmd might play an additional regularizing role. It also shows the computation time of curl-index˙cmd and curl-index˙cmd. We see that index˙cmd regression does not scale with large datasets, while index˙cmd regression does. When $N > 10^4$, index˙cmd regression exceeds memory capacity.

Structured prediction vs Independent (RFF) prediction. On the second dataset, Figure 12 (top row) compares R^2 score and time of index˙cmd regression using the trivial identity decomposable kernel, index˙cmd independent index˙cmd, to curl-free index˙cmd regression. Curl-free index˙cmd outperforms independent index˙cmd, as expected, since the dataset involves structured outputs.

Impact of the number of random features (D). In this setting we solved the optimisation problem for both index˙cmd and index˙cmd using a index˙cmd. Figure 13 top row shows that for a fixed number of instance in the train set, index˙cmd performs better than index˙cmd in terms of accuracy (R^2). However index˙cmd scales better than index˙cmd index˙cmd the number of data. index˙cmd is able to process more data than index˙cmd in the same time and thus reach a better accuracy for a given amount of time. Bottom row shows that index˙cmd tends to reach index˙cmd's accuracy for a fixed number of data when the number of features increase.

Multitask learning. In this experiment we are interested in multitask learning with operator-valued random Fourier features, and see whether the approximation of a joint index˙cmd performs better than an independent index˙cmd. In this setting we assume that

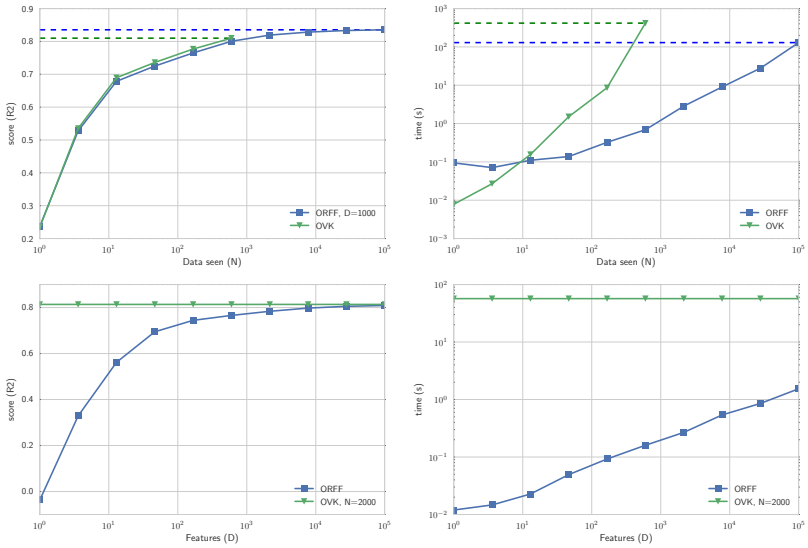

Figure 13: Decomposable kernel on the third dataset: R^2 score vs number of data in the train set (N)

for each entry $x_i \in \mathbb{R}^d$ we only have access to one observation $y_i \in \mathbb{R}$ corresponding to a task t_i. We used the SARCOS dataset, taken from `http://www.gaussianprocess.org/gpml/data/` website. This is an inverse dynamics problem, index˙cmd we have to predict the 7 joint torques given the joint positions, velocities and accelerations. Hence, we have to solve a regression problem with 21 inputs and 7 outputs which is a very nonlinear function. It has $45K$ inputs data. Suppose that we are given a collection of inputs data $x_1, \ldots, x_N \in \mathbb{R}^{21}$ and a collection of output data $((y_1, t_1) \ldots, (y_N, t_N)) \in (\mathbb{R} \times \mathbb{N}_T)^N$ where T is the number of tasks. We consider the following multitask loss function $L(h(x), (y, t)) = \frac{1}{2} \left(\langle h(x), e_t \rangle_2 - y \right)^2$, This loss function is adapted to datasets where the number of data per tasks is unbalanced (index˙cmd for one input data we observe the value of only one task and not all the tasks.). We optimise the regularized risk

$$\frac{1}{N} \sum_{i=1}^{N} L\left(h(x_i), (y_i, t_i)\right) + \frac{\lambda}{2N} ||h||^2_{\mathcal{H}} = \frac{1}{2N} \sum_{i=1}^{N} \left(\langle h(x_i), e_{t_i} \rangle - y_i \right)^2 + \frac{\lambda}{2N} ||h||^2_{\mathcal{H}}$$

We used a model h based on the decomposable kernel $h(x) = (\varphi(x)^T \otimes B)\theta$ we chose B such that $BB^T = A$, where A is the inverse graph Laplacian L of the similarities between the tasks, parametrized by an hyperparameter $\gamma \in \mathbb{R}_+$. $L_{kl} = \exp\left(-\gamma \sqrt{\sum_{i=1}^{N} \left(y_i^k - y_i^l \right)^2} \right)$. We draw N data randomly for each task, hence creating a dataset of $N \times 7$ data and computed the nMSE on the proposed test set (4.5K points). We repeated the experiments 80 times to avoid randomness. We choose $D = \frac{\max(N, 500)}{2}$ features, and optimized the problem with a second order batch gradient. Table 3 shows that using the index˙cmd approximation of an

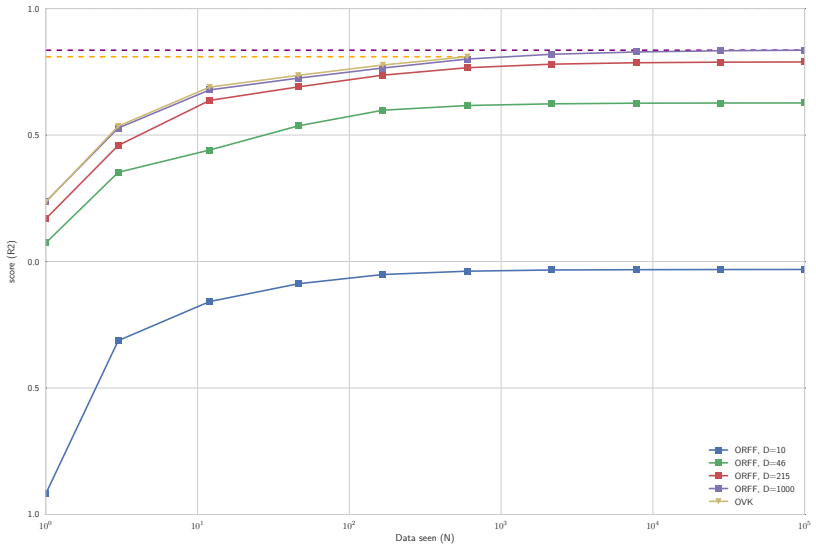

Figure 14: Decomposable kernel on the third dataset: R^2 score vs number of data in the train set (N) for different number for different number of random samples (D).

N	Independant (%)	Laplacian (%)	p-value	T
50×7	23.138 ± 0.577	22.254 ± 0.536	2.68%	4(s)
100×7	16.191 ± 0.221	15.568 ± 0.187	$< 0.1\%$	16(s)
200×7	12.713 ± 0.0978	12.554 ± 0.0838	1.52%	12(s)
400×7	10.785 ± 0.0579	10.651 ± 0.0466	$< 0.1\%$	10(s)
800×7	7.512 ± 0.0344	7.512 ± 0.0344	100%	15(s)
3200×7	5.658 ± 0.0187	5.658 ± 0.0187	100%	20(s)

Table 3: Error (% of nMSE) on SARCOS dataset.

operator-valued kernel with a good prior on the data improves the performances index˙cmd the independent index˙cmd. However the advantage seems to be less important the more data are available.

7. Conclusion

index˙cmd naturally extend the celebrated kernel method used to learn scalar-valued functions, to the case of learning vector-valued functions. Although index˙cmd are appealing from a theoretical aspect, these methods scale poorly in terms of computation time when the number of data is high. Indeed, to evaluate the value of function with an , it requires to evaluate an on all the point in the given dataset. Hence naive learning with kernels usu-

ally scales cubicly in time with the number of data. In the context of large-scale learning such scaling is not acceptable. Through this work we propose a methodology to tackle this difficulty.

Enlightened by the literature on large-scale learning with *scalar*-valued kernel, in particular the work of Rahimi and Recht (Rahimi and Recht, 2007), we propose to replace an index˙cmd by a random feature map that we called . Our contribution start with the formal mathematical construction of this feature from an index˙cmd. Then we show that it is also possible to obtain a kernel from an index˙cmd. Eventually we analyse the regularization properties in terms of of $\mathcal{Y}$-Mercer kernels. Then we moved on giving a bound on the error due to the random approximation of the index˙cmd with high probability. We showed that it is possible to bound the error even though the index˙cmd estimator of an index˙cmd is not a bounded random variable. Moreover we also give a bound when the dimension of the output data infinite.

After ensuring that an index˙cmd is a good approximation of a kernel, we moved on giving a framework for supervised learnin with . We showed that learning with a feature map is equivalent to learn with the reconstructed index˙cmd under some mild conditions. Then we focused on an efficient implementation of index˙cmd by viewing them as linear operators rather than matrices and using matrix-free (iterative) solvers and concluded with some numerical experiments.

Following Rahimi and Recht a generalization bound for index˙cmd kernel ridge would probably suggest that the number of feature to draw is proportional to the number of data. However new results of Rudi et al. (2016) suggest that the number of feature should be proportional to the *square root* of the number of data. In a future work, we shall investigate this results and extend it to index˙cmd.

Since the contruction of index˙cmd is valid for infinite dimensional Hilbert spaces such as function spaces, we would also like to investigate learning function valued functions in an efficient manner.

Acknowledgments

The authors are grateful to Maxime Sangnier (UPMC, France) for the insightful discussions and Markus Heinonen (Aalto University, Sweden) for the preliminary experiments.

Appendix A. Reminder on Abstract Harmonic Analysis

A.1 Locally compact Abelian groups

Definition 36 ((index˙cmd) group.)

A group $\mathcal{X}$ endowed with a binary operation $\star$ is said to be a Locally Compact Abelian group if $\mathcal{X}$ is a topological commutative *group index˙cmd $\star$ for which every point has a compact neighborhood and is Hausdorff (T2).*

Moreover given a element z of a index˙cmd group $\mathcal{X}$, we define the set $z \star \mathcal{X} = \mathcal{X} \star z = \{ z \star x \mid \forall x \in \mathcal{X} \}$ and the set $\mathcal{X}^{-1} = \left\{ x^{-1} \;\middle|\; \forall x \in \mathcal{X} \right\}$. We also note e the neutral element of $\mathcal{X}$ such that $x \star e = e \star x = e$ for all $x \in \mathcal{X}$. Throughout this paper we focus on positive definite function. Let $\mathcal{Y}$ be a complex separable Hilbert space. A function $f : \mathcal{X} \to \mathcal{Y}$ is positive definite if for all $N \in \mathbb{N}$ and all $y \in \mathcal{Y}$,

$$\sum_{i,j\,=1}^{N} \left\langle y_i, f\left(x_j^{-1} \star x_i\right) y_j \right\rangle_{\mathcal{Y}} \geq 0 \tag{31}$$

for all sequences $(y_i)_{i \in \mathbb{N}_N^*} \in \mathcal{Y}^N$ and all sequences $(x_i)_{i \in \mathbb{N}_N^*} \in \mathcal{X}^N$. If $\mathcal{Y}$ is real we add the assumption that $f(x^{-1}) = f(x)^*$ for all $x \in \mathcal{X}$

A.2 Even and odd functions

Let $\mathcal{X}$ be a index˙cmd group and $\mathbb{K}$ be a field viewed as an additive group. We say that a function $f : \mathcal{X} \to \mathbb{K}$ is even if for all $x \in \mathcal{X}$, $f(x) = f\left(x^{-1}\right)$ and odd if $f(x) = -f\left(x^{-1}\right)$. The definition can be extended to operator-valued functions.

Definition 37 (Even and odd operator-valued function on a index˙cmd group)

Let $\mathcal{X}$ be a measured index˙cmd group and $\mathcal{Y}$ be a Hilbert space, and $\mathcal{L}(\mathcal{Y})$ the space of bounded linear operators from $\mathcal{Y}$ to itself viewed as an additive group. A function $f : \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ is (weakly) even if for all $x \in \mathcal{X}$ and all y, $y' \in \mathcal{Y}$, $\langle y, f\left(x^{-1}\right) y' \rangle_{\mathcal{Y}} = \langle y, f(x) y' \rangle_{\mathcal{Y}}$ and (weakly) odd if $\langle y, f\left(x^{-1}\right) y' \rangle_{\mathcal{Y}} = -\langle y, f(x) y' \rangle_{\mathcal{Y}}$.

It is easy to check that if f is odd then $\int_{\mathcal{X}} \langle y, f(x) y' \rangle_{\mathcal{Y}} d\mathbf{Haar}(x) = 0$. Besides the product of an even and an odd function is odd. Indeed for all f, $g \in \mathcal{F}(\mathcal{X}; \mathcal{L}(\mathcal{Y}))$, where f is even and g odd. Define $h(x) = \langle y, f(x) g(x) y' \rangle$. Then we have $h\left(x^{-1}\right) = \langle y, f\left(x^{-1}\right) g\left(x^{-1}\right) y' \rangle_{\mathcal{Y}} = \langle y, f(x) \left(-g(x)\right) y' \rangle_{\mathcal{Y}} = -h(x)$.

A.3 Characters

index˙cmd groups are central to the general definition of Fourier Transform which is related to the concept of Pontryagin duality (Folland, 1994). Let $(\mathcal{X}, \star)$ be a index˙cmd group with e its neutral element and the notation, x^{-1}, for the inverse of $x \in \mathcal{X}$. A *character* is a complex continuous homomorphism $\omega : \mathcal{X} \to \mathbb{U}$ from $\mathcal{X}$ to the set of complex numbers of unit module $\mathbb{U}$. The set of all characters of $\mathcal{X}$ forms the Pontryagin *dual group* $\widehat{\mathcal{X}}$. The dual group of an index˙cmd group is an index˙cmd group so that we can endow $\widehat{\mathcal{X}}$ with a "dual" Haar measure noted $\widehat{\mathbf{Haar}}$. Then the dual group operation is defined by $(\omega_1 \star' \omega_2)(x) = \omega_1(x)\omega_2(x) \in \mathbb{U}$. The Pontryagin duality theorem states that $\widehat{\widehat{\mathcal{X}}} \cong \mathcal{X}$. index˙cmd there is a canonical isomorphism between any index˙cmd group and its double dual. To emphasize this duality the following notation is usually adopted: $\omega(x) = (x, \omega) = (\omega, x) = x(\omega)$, where $x \in \mathcal{X} \cong \widehat{\widehat{\mathcal{X}}}$ and $\omega \in \widehat{\mathcal{X}}$. The form $(\cdot, \cdot)$ defined in appendix A.3 is called (duality) pairing. Another important property involves the complex conjugate of the pairing which is defined as $\overline{(x, \omega)} = \left(x^{-1}, \omega\right) = \left(x, \omega^{-1}\right)$. We notice that for any pairing depending of ω, there exists a function $h_\omega : \mathcal{X} \to \mathbb{R}$ such that $(x, \omega) = \exp(\mathrm{i} h_\omega(x))$ since any pairing maps into $\mathbb{U}$. Moreover, $\left(x \star z^{-1}, \omega\right) = \omega(x)\omega\left(z^{-1}\right) = \exp\left(+\mathrm{i} h_\omega\left(x\right)\right) \exp\left(+\mathrm{i} h_\omega\left(z^{-1}\right)\right) =$

Table 4: Classification of s in terms of their domain and transform domain.

$\mathcal{X} =$	$\widehat{\mathcal{X}} \cong$	Operation	Pairing
$\mathbb{R}^d$	$\mathbb{R}^d$	$+$	$(x,\omega) = \exp\left(\mathrm{i}\langle x,\omega\rangle_2\right)$
$\mathbb{R}^d_{*,+}$	$\mathbb{R}^d$	$\cdot$	$(x,\omega) = \exp\left(\mathrm{i}\langle \log(x),\omega\rangle_2\right)$
$(-c;+\infty)^d$	$\mathbb{R}^d$	$\odot$	$(x,\omega) = \exp\left(\mathrm{i}\langle \log(x+c),\omega\rangle_2\right)$

$\exp\left(+\mathrm{i}h_\omega\left(x\right)\right)\exp\left(-\mathrm{i}h_\omega\left(z\right)\right)$. Table 4 provides an explicit list of pairings for various groups based on $\mathbb{R}^d$ or its subsets. The interested reader can refer to Folland (1994) for a more detailed construction of index˙cmd, Pontryagin duality and s on index˙cmd.

A.4 The

For a function with values in a separable Hilbert space, $f \in L^1(\mathcal{X}, \mathbf{Haar}; \mathcal{Y})$, we denote $\mathcal{F}\left[f\right]$ its (index˙cmd) which is defined by

$$\forall \omega \in \widehat{\mathcal{X}},\ \ \mathcal{F}\left[f\right](\omega) = \int_{\mathcal{X}} \overline{(x,\omega)} f(x) d\mathbf{Haar}(x).$$

The (index˙cmd) of a function $g \in L^1(\widehat{\mathcal{X}}, \widehat{\mathbf{Haar}}; \mathcal{Y})$ is noted $\mathcal{F}^{-1}\left[g\right]$ defined by $\forall x \in \mathcal{X},\ \ \mathcal{F}^{-1}\left[g\right](x) = \int_{\widehat{\mathcal{X}}} (x,\omega) g(\omega) d\widehat{\mathbf{Haar}}(\omega)$, We also define the flip operator $\mathcal{R}$ by $(\mathcal{R}f)(x) := f\left(x^{-1}\right)$.

Theorem 38 (Fourier inversion)

Given a measure **Haar** *defined on* $\mathcal{X}$*, there exists a unique suitably normalized dual measure* $\widehat{\mathbf{Haar}}$ *on* $\widehat{\mathcal{X}}$ *such that for all* $f \in L^1(\mathcal{X}, \mathbf{Haar}; \mathcal{Y})$ *and if* $\mathcal{F}\left[f\right] \in L^1(\widehat{\mathcal{X}}, \widehat{\mathbf{Haar}}; \mathcal{Y})$ *we have*

$$f(x) = \int_{\widehat{\mathcal{X}}} (x,\omega) \mathcal{F}\left[f\right](\omega) d\widehat{\mathbf{Haar}}(\omega), \quad \textit{for } \mathbf{Haar}\textit{-almost all } x \in \mathcal{X}. \tag{32}$$

index˙cmd such that $(\mathcal{R}\mathcal{F}\mathcal{F}\left[f\right])(x) = \mathcal{F}^{-1}\mathcal{F}\left[f\right](x) = f(x)$ *for* **Haar***-almost all* $x \in \mathcal{X}$*. If* f *is continuous this relation holds for all* $x \in \mathcal{X}$*.*

Thus when a Haar measure **Haar** on $\mathcal{X}$ is given, the measure on $\widehat{\mathcal{X}}$ that makes Theorem 38 true is called the dual measure of **Haar**, noted $\widehat{\mathbf{Haar}}$. Let $c \in \mathbb{R}_*$ If $c\mathbf{Haar}$ is the measure on $\mathcal{X}$, then $c^{-1}\widehat{\mathbf{Haar}}$ is the dual measure on $\widehat{\mathcal{X}}$. Hence one must replace $\widehat{\mathbf{Haar}}$ by $c^{-1}\widehat{\mathbf{Haar}}$ in the inversion formula to compensate. Whenever $\widehat{\mathbf{Haar}} = \mathbf{Haar}$ we say that the Haar measure is self-dual. For the familiar case of a scalar-valued function f on the index˙cmd group $(\mathbb{R}^d, +)$, we have for all $\omega \in \widehat{\mathcal{X}} = \mathbb{R}^d$

$$\mathcal{F}\left[f\right](\omega) = \int_{\mathcal{X}} \overline{(x,\omega)} f(x) d\mathbf{Haar}(x) = \int_{\mathbb{R}^d} \exp(-\mathrm{i}\langle x,\omega\rangle_2) f(x) d\mathbf{Leb}(x), \tag{33}$$

the Haar measure being here the Lebesgue measure. Notice that the normalization factor of $\widehat{\mathbf{Haar}}$ on $\widehat{\mathcal{X}}$ depends on the measure **Haar** on $\mathcal{X}$ *and* the duality pairing. For instance

let $\mathcal{X} = (\mathbb{R}^d, +)$. If one endow $\mathcal{X}$ with the Lebesgue measure as the Haar measure, the Haar measure on the dual is defined for all $\mathcal{Z} \in \mathcal{B}(\mathbb{R}^d)$ by

$$\mathbf{Haar}(\mathcal{Z}) = \mathbf{Leb}(\mathcal{Z}), \quad \text{and} \quad \widehat{\mathbf{Haar}}(\mathcal{Z}) = \frac{1}{(2\pi)^d}\mathbf{Leb}(\mathcal{Z}),$$

in order to have $\mathcal{F}^{-1}\mathcal{F}[f] = f$. If one use the cleaner equivalent pairing $(x, \omega) = \exp(2\mathrm{i}\pi\langle x, \omega\rangle_2)$ rather than $(x, \omega) = \exp(\mathrm{i}\langle x, \omega\rangle_2)$, then $\widehat{\mathbf{Haar}}(\mathcal{Z}) = \mathbf{Leb}(\mathcal{Z})$. The pairing $(x, \omega) = \exp(2\mathrm{i}\pi\langle x, \omega\rangle_2)$ looks more attractive in theory since it limits the messy factor outside the integral sign and make the Haar measure self-dual. However it is of lesser use in practice since it yields additional unnecessary computation when evaluating the pairing. Hence for symmetry reason on $(\mathbb{R}^d, +)$ and reduce computations we settle with the Haar measure on $\mathbb{R}^d$ groups (additive and multiplicative) defined as $\widehat{\mathbf{Haar}}(\mathcal{Z}) = \mathbf{Haar}(\mathcal{Z}) = \sqrt{2\pi}^{-d}\mathbf{Leb}(\mathcal{Z})$. We conclude this subsection by recalling the injectevity property of the .

Corollary 39 (injectivity)

Given μ and ν two measures, if $\mathcal{F}[\mu] = \mathcal{F}[\nu]$ then $\mu = \nu$. Moreover given two functions f and $g \in L^1(\mathcal{X}, \mathbf{Haar}; \mathcal{Y})$ if $\mathcal{F}[f] = \mathcal{F}[g]$ then $f = g$

Appendix B. Proofs

In this section we give the proofs of our contributions stated in the main body of the paper.

B.1 Construction

B.1.1 Proof of Lemma 16

Proof For any function f on $(\mathcal{X}, \star)$ define the flip operator $\mathcal{R}$ by $(\mathcal{R}f)(x) := f\left(x^{-1}\right)$. For any shift invariant $\mathcal{Y}$-Mercer kernel and for all $\delta \in \mathcal{X}$, $K_e(\delta) = K_e\left(\delta^{-1}\right)^*$. Indeed from the definition of a shift-invariant kernel, $K_e\left(\delta^{-1}\right) = K\left(\delta^{-1}, e\right) = K(e, \delta) = K(\delta, e)^* = K_e(\delta)^*$.

Item 1: taking the yields, $\langle y', C(\omega)y\rangle_{\mathcal{Y}} = \mathcal{F}^{-1}\left[\langle y', K_e(\cdot)y\rangle_{\mathcal{Y}}\right](\omega) = \mathcal{R}\mathcal{F}^{-1}\left[\langle K_e(\cdot)y', y\rangle_{\mathcal{Y}}\right](\omega) = \mathcal{R}\langle C(\cdot)y', y\rangle_{\mathcal{Y}}(\omega) = \left\langle y', C\left(\omega^{-1}\right)^* y\right\rangle_{\mathcal{Y}}$. Hence $C(\omega) = C\left(\omega^{-1}\right)^*$. Suppose that $\mathcal{Y}$ is a complex Hilbert space. Since for all $\omega \in \widehat{\mathcal{X}}$, $C(\omega)$ is bounded and non-negative so $C(\omega)$ is self-adjoint. Besides we have $C(\omega) = C\left(\omega^{-1}\right)^*$ so C must be even. Suppose that $\mathcal{Y}$ is a real Hilbert space. The of a real valued function obeys $\mathcal{F}[f](\omega) = \overline{\mathcal{F}[f](\omega^{-1})}$. Therefore since $C(\omega)$ is non-negative for all $\omega \in \widehat{\mathcal{X}}$, $\langle y', C(\omega)y\rangle = \overline{\langle y', C\left(\omega^{-1}\right)y\rangle} = \langle y, C\left(\omega^{-1}\right)^* y'\rangle = \langle y, C(\omega)y'\rangle$. Hence $C(\omega)$ is self-adjoint and thus C is even.

Item 2: simply, for all $y, y' \in \mathcal{Y}$, $\langle y, C(\omega^{-1})y'\rangle = \langle y', C(\omega)y\rangle$ thus $\mathcal{F}^{-1}\left[\langle y', K_e(\cdot)y\rangle_{\mathcal{Y}}\right](\omega) = \langle y', C(\omega)y\rangle = \mathcal{R}\langle y', C(\cdot)y\rangle(\omega) = \mathcal{R}\mathcal{F}^{-1}\left[\langle y', K_e(\cdot)y\rangle_{\mathcal{Y}}\right](\omega) = \mathcal{F}\left[\langle y', K_e(\cdot)y\rangle_{\mathcal{Y}}\right](\omega)$.

Item 3: from Item 2 we have $\mathcal{F}^{-1}\left[\langle y', K_e(\cdot)y\rangle\right] = \mathcal{F}^{-1}\mathcal{R}\langle y', K_e(\cdot)y\rangle$. By injectivity of the , K_e is even. Since $K_e(\delta) = K_e(\delta^{-1})^*$, we must have $K_e(\delta) = K_e(\delta)^*$. ■

B.1.2 Proof of Proposition 17

Proof This is a simple consequence of Proposition 15 and Lemma 16. By taking $\langle y', C(\omega)y\rangle = \mathcal{F}^{-1}\left[\langle y', K_e(\cdot)y\rangle\right](\omega) = \mathcal{F}\left[\langle y', K_e(\cdot)y\rangle\right](\omega)$ we can write the following equality concerning the index˙cmd signature K_e: $\langle y', K_e(\delta)y\rangle(\omega) = \int_{\widehat{\mathcal{X}}} \overline{(\delta,\omega)}\langle y', C(\omega)y\rangle d\widehat{\mathbf{Haar}}(\omega) = \int_{\widehat{\mathcal{X}}} \overline{(\delta,\omega)}\left\langle y', \frac{1}{\rho(\omega)}C(\omega)y\right\rangle \rho(\omega)d\widehat{\mathbf{Haar}}(\omega)$. It is always possible to choose $\rho(\omega)$ such that $\int_{\widehat{\mathcal{X}}}\rho(\omega)d\widehat{\mathbf{Haar}}(\omega) = 1$. For instance choose

$$\rho(\omega) = \frac{\|C(\omega)\|_{\mathcal{Y},\mathcal{Y}}}{\int_{\widehat{\mathcal{X}}}\|C(\omega)\|_{\mathcal{Y},\mathcal{Y}}d\widehat{\mathbf{Haar}}(\omega)}$$

Since for all $y,\ y' \in \mathcal{Y}$, $\langle y', C(\cdot)y\rangle \in L^1(\widehat{\mathcal{X}}, \widehat{\mathbf{Haar}})$ and $\mathcal{Y}$ is a separable Hilbert space, by Pettis measurability theorem, $\int_{\widehat{\mathcal{X}}}\|C(\omega)\|_{\mathcal{Y},\mathcal{Y}}d\widehat{\mathbf{Haar}}(\omega)$ is finite and so is $\|C(\omega)\|_{\mathcal{Y},\mathcal{Y}}$ for all $\omega \in \widehat{\mathcal{X}}$. Therefore $\rho(\omega)$ is the density of a probability measure $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$, index˙cmd conclude by taking $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}(\mathcal{Z}) = \int_{\mathcal{Z}}\rho(\omega)d\widehat{\mathbf{Haar}}(\omega)$, for all $\mathcal{Z} \in \mathcal{B}(\widehat{\mathcal{X}})$. ■

B.1.3 Proof of Proposition 18

Proof Suppose that for all $y,\ y' \in \mathcal{Y}$, $\langle y', A(\omega)y\rangle\rho(\omega) = \mathcal{F}^{-1}\left[\langle y', K_e(\cdot)y\rangle\right](\omega)$ where ρ is a probability distribution (see Proposition 17). From the strong law of large numbers $\frac{1}{D}\sum_{j=1}^{D}\overline{(x\star z^{-1},\omega_j)}A(\omega_j) \xrightarrow[D\to\infty]{\text{index˙cmd}} \mathbf{E}_{\widehat{\mathbf{Haar}},\rho}[\overline{(x\star z^{-1},\omega_j)}A(\omega)]$ where the integral converges in the weak operator topology. Then by Proposition 17 we recover K_e when $D \to \infty$ since, $\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}[\overline{(x\star z^{-1},\omega_j)}A(\omega)] = K_e(x\star z^{-1})$. ■

B.1.4 Proof of Proposition 19

Proof Let $(\omega_j)_{j=1}^{D}$ be a sequence of $D \in \mathbb{N}^*$ index˙cmd random variables following the law $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$. For all $x,\ z \in \mathcal{X}$ and all $y,\ y' \in \mathcal{Y}$,

$$\left\langle \widetilde{\phi}(x)y, \widetilde{\phi}(z)y'\right\rangle_{\bigoplus_{j=1}^{D}\mathcal{Y}'} = \frac{1}{D}\left\langle \bigoplus_{j=1}^{D}\left((x,\omega_j)B(\omega_j)^*y\right), \bigoplus_{j=1}^{D}\left((z,\omega_j)B(\omega_j)^*y'\right)\right\rangle$$

By definition of the inner product in direct sum of Hilbert spaces,

$$\begin{aligned}&\frac{1}{D}\left\langle \bigoplus_{j=1}^{D}\left((x,\omega_j)B(\omega_j)^*y\right), \bigoplus_{j=1}^{D}\left((z,\omega_j)B(\omega_j)^*y'\right)\right\rangle \\ &= \frac{1}{D}\sum_{j=1}^{D}\left\langle y, \overline{(x,\omega_j)}B(\omega_j)(z,\omega_j)B(\omega_j)^*y'\right\rangle_{\mathcal{Y}} = \left\langle y, \left(\frac{1}{D}\sum_{j=1}^{D}\overline{(x\star z^{-1},\omega_j)}A(\omega_j)\right)y'\right\rangle_{\mathcal{Y}},\end{aligned}$$

Eventually apply Proposition 18 to obtain the convergence of the Monte-Carlo plug-in estimator to the true kernel K. ■

B.1.5 Proof of Proposition 21

Proof For all y, $y' \in \mathcal{Y}$ and x, $z \in \mathcal{X}$,

$$\langle y, \phi_x^* \phi_z y' \rangle_{\mathcal{Y}} = \langle \phi_x y, \phi_z y' \rangle_{L^2(\widehat{\mathcal{X}}, \widehat{\mu}; \mathcal{Y}')} = \int_{\widehat{\mathcal{X}}} \overline{(x, \omega)} \langle y, B(\omega)(z, \omega) B(\omega)^* y' \rangle d\widehat{\mu}(\omega)$$
$$= \int_{\widehat{\mathcal{X}}} \overline{(x \star z^{-1}, \omega)} \langle y, B(\omega) B(\omega)^* y' \rangle d\widehat{\mu}(\omega) = \int_{\widehat{\mathcal{X}}} \overline{(x \star z^{-1}, \omega)} \langle y, A(\omega) y' \rangle d\widehat{\mu}(\omega),$$

which defines a $\mathcal{Y}$-Mercer according to Proposition 14 of Carmeli et al. (2010). ■

B.1.6 Proof of Proposition 22

Proof From the strong law of large numbers $\frac{1}{D} \sum_{j=1}^{D} \overline{(x \star z^{-1}, \omega_j)} A(\omega_j) \xrightarrow[D \to \infty]{\text{index·cmd}}$ $\mathbf{E}_{\widehat{\mathbf{Haar}}, \rho}[\overline{(x \star z^{-1}, \omega_j)} A(\omega)]$ where the integral converges in the weak operator topology. Then by Proposition 14, $\mathbf{E}_{\widehat{\mathbf{Haar}}, \rho}[\overline{(x \star z^{-1}, \omega_j)} A(\omega)] = K_e(x \star z^{-1})$. ■

B.1.7 Proof of Proposition 23

Proof Apply Proposition 6 to $\widetilde{\phi}$ considering the Hilbert space $\widetilde{\mathcal{H}}$ to show that $\widetilde{K}$ is an index·cmd. Then Proposition 12 shows that $\widetilde{K}$ is shift-invariant since $\widetilde{K}(x, z) = \widetilde{K}_e\left(x \star z^{-1}\right)$. Since $B(\omega)$ is a bounded operator, $\widetilde{K}$ is $\mathcal{Y}$-Mercer because all the functions in the sum are continuous. ■

B.1.8 Proof of Proposition 25

Proof [of item 1] Since $(\omega_j)_{j=1}^D$ are index·cmd random vectors, for all $y \in \mathcal{Y}$ and for all $y' \in \mathcal{Y}'$, $\langle y, B(\cdot) y' \rangle \in L^2(\widehat{\mathcal{X}}, \mathbf{Pr}_{\widehat{\mathbf{Haar}}, \rho})$ and $g \in L^2(\widehat{\mathcal{X}}, \mathbf{Pr}_{\widehat{\mathbf{Haar}}, \rho}; \mathcal{Y}')$,

$$(\widetilde{W}\theta)(x) = \widetilde{\phi}(x)^* \theta = \frac{1}{D} \sum_{j=1}^{D} \overline{(x, \omega_j)} B(\omega_j) g(\omega_j), \qquad \omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}}, \rho} \text{ index·cmd}$$
$$\xrightarrow[D \to \infty]{\text{index·cmd}} \int_{\widehat{\mathcal{X}}} \overline{(x, \omega)} B(\omega) g(\omega) d\mathbf{Pr}_{\widehat{\mathbf{Haar}}, \rho}(\omega) = (Wg)(x) := \phi_x^* g.$$

from the strong law of large numbers. ■

Proof [of item 2] Again, since $(\omega_j)_{j-1}^D$ are index·cmd random vectors and $g \in L^2(\widehat{\mathcal{X}}, \mathbf{Pr}_{\widehat{\mathbf{Haar}}, \rho}; \mathcal{Y}')$,

$$\|\theta\|_{\widetilde{\mathcal{H}}}^2 = \frac{1}{D} \sum_{j=1}^{D} \|g(\omega_j)\|_{\mathcal{Y}'}^2, \qquad \omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}}, \rho} \text{ index·cmd}$$
$$\xrightarrow[D \to \infty]{\text{index·cmd}} \int_{\widehat{\mathcal{X}}} \|g(\omega)\|_{\mathcal{Y}'}^2 d\mathbf{Pr}_{\widehat{\mathbf{Haar}}, \rho}(\omega) = \|g\|_{L^2\left(\widehat{\mathcal{X}}, \mathbf{Pr}_{\widehat{\mathbf{Haar}}, \rho}; \mathcal{Y}'\right)}^2.$$

from the strong law of large numbers. ■

B.1.9 Proof of Proposition 26

Proof We first show how the relates to the feature operator. Since $\mathcal{H}_K$ is embedded into $\mathcal{H} = L^2(\widehat{\mathcal{X}}, \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}; \mathcal{Y}')$ by means of the feature operator W, we have for all $f \in \mathcal{H}_k$, for all $f \in \mathcal{H}$ and for all $x \in \mathcal{X}$

$$\mathcal{F}\left[\mathcal{F}^{-1}\left[f\right]\right](x) = \int_{\widehat{\mathcal{X}}} \overline{(x,\omega)}\mathcal{F}^{-1}\left[f\right](\omega) d\widehat{\mathbf{Haar}}(\omega) = f(x)$$
$$(Wg)(x) = \int_{\widehat{\mathcal{X}}} \overline{(x,\omega)}\rho(\omega)B(\omega)g(\omega) d\widehat{\mathbf{Haar}}(\omega) = f(x).$$

By injectivity of the , $\mathcal{F}^{-1}\left[f\right](\omega) = \rho(\omega)B(\omega)g(\omega)$. From Proposition 6 we have

$$\begin{aligned}\|f\|_K^2 &= \inf\left\{ \|g\|_{\mathcal{H}}^2 \;\middle|\; \forall g \in \mathcal{H},\ \ Wg = f \right\} \\ &= \inf\left\{ \int_{\widehat{\mathcal{X}}} \|g(\omega)\|_{\mathcal{Y}'}^2 d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}(\omega) \;\middle|\; \forall g \in \mathcal{H},\ \ \mathcal{F}^{-1}\left[f\right] = \rho(\cdot)B(\cdot)g(\cdot) \right\}.\end{aligned}$$

The pseudo inverse of the operator $B(\omega)$ – noted $B(\omega)^\dagger$ – is the unique solution of the system $\mathcal{F}^{-1}\left[f\right](\omega) = \rho(\omega)B(\omega)g(\omega)$ index˙cmd $g(\omega)$ with minimal norm[13]. Eventually, $\|f\|_K^2 = \int_{\widehat{\mathcal{X}}} \frac{\left\|B(\omega)^\dagger \mathcal{F}^{-1}[f](\omega)\right\|_{\mathcal{Y}}^2}{\rho(\omega)^2} d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}(\omega)$ Using the fact that $\mathcal{F}^{-1}\left[\cdot\right] = \mathcal{F}\mathcal{R}[\cdot]$ and $\mathcal{F}^2[\cdot] = \mathcal{R}[\cdot]$,

$$\begin{aligned}\|f\|_K^2 &= \int_{\widehat{\mathcal{X}}} \frac{\left\|\mathcal{R}\left[B(\cdot)^\dagger \rho(\cdot)\right](\omega)\mathcal{F}\left[f\right](\omega)\right\|_{\mathcal{Y}}^2}{\rho(\omega)^2} d\widehat{\mathbf{Haar}}(\omega) \\ &= \int_{\widehat{\mathcal{X}}} \frac{\langle B(\omega)^\dagger \mathcal{F}\left[f\right](\omega), B(\omega)^\dagger \mathcal{F}\left[f\right](\omega)\rangle_{\mathcal{Y}}}{\rho(\omega)} d\widehat{\mathbf{Haar}}(\omega) \\ &= \int_{\widehat{\mathcal{X}}} \frac{\langle \mathcal{F}\left[f\right](\omega), A(\omega)^\dagger \mathcal{F}\left[f\right](\omega)\rangle_{\mathcal{Y}}}{\rho(\omega)} d\widehat{\mathbf{Haar}}(\omega).\end{aligned}$$

■

B.2 Convergence with high probability of the ORFF estimator

We recall the notations $\delta = x \star z^{-1}$, for all $x,\ z \in \mathcal{X}$, $\tilde{K}(x,z) = \widetilde{\phi}(x)^* \widetilde{\phi}(z)$, $\tilde{K}^j(x,z) = \phi_x(\omega_j)^* \phi_z(\omega_j)$, where $\omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ and $K_e(\delta) = K(x,z)$ and $\tilde{K}_e(\delta) = \tilde{K}(x,z)$. For the sake of readabilty, we use throughout the proof the quantities: $F(\delta) := \tilde{K}(x,z) - K(x,z)$ and $F^j(\delta) := \frac{1}{D}\left(\tilde{K}^j(x,z) - K(x,z)\right)$. We also view $\mathcal{X}$ as a metric space endowed with the distance $d_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. Compared to the scalar case, the proof follows the same scheme as the one described in (Rahimi and Recht, 2007; Sutherland and Schneider, 2015),

13. Note that since $B(\omega)$ is bounded the pseudo inverse of $B(\omega)$ is well defined for $\widehat{\mathbf{Haar}}$-almost all ω.

but we consider an operator norm as measure of the error and therefore concentration inequality dealing with these operator norm. The main feature of Proposition 46 is that it covers the case of bounded index˙cmd as well as unbounded index˙cmd. In the case of bounded index˙cmd, a Bernstein inequality for matrix concentration such that the one proved in Mackey et al. (2014, Corollary 5.2) or the formulation of Tropp (2012) recalled in Koltchinskii et al. (2013) is suitable. However some kernels like the curl and the divergence-free kernels do not have obvious bounded $\left\|F^j\right\|_{\mathcal{Y},\mathcal{Y}}$ but exhibit F^j with subexponential tails. Therefore, we use an operator Bernstein concentration inequality adapted for random matrices with subexponential norms.

B.2.1 Epsilon-net

Let $\mathcal{C} \subseteq \mathcal{X}$ be a compact subset of $\mathcal{X}$. Let $\mathcal{D}_{\mathcal{C}} = \left\{ x \star z^{-1} \;\middle|\; x, z \in \mathcal{C} \right\}$ with diameter at most $2|\mathcal{C}|$ where $|\mathcal{C}|$ is the diameter of $\mathcal{C}$. Since $\mathcal{C}$ is supposed compact, so is $\mathcal{D}_{\mathcal{C}}$. Since $\mathcal{D}_{\mathcal{C}}$ is also a metric space it is well known that a compact metric space is totally bounded. Thus it is possible to find a finite ϵ-net covering $\mathcal{D}_{\mathcal{C}}$. We call $T = \mathcal{N}(\mathcal{D}_{\mathcal{C}}, r)$ the number of closed balls of radius r required to cover $\mathcal{D}_{\mathcal{C}}$. For instance if $\mathcal{D}_{\mathcal{C}}$ is a subspace finite dimensional Banach space with diameter at most $2|\mathcal{C}|$ it is possible to cover the space with at most $T = (4|\mathcal{C}|/r)^d$ balls of radius r (see Cucker and Smale (2001, proposition 5)). Let us call $\delta_i, i = 1, \ldots, T$ the center of the i-th ball, also called anchor of the ϵ-net. Denote L_F the Lipschitz constant of F. Let $\|\cdot\|_{\mathcal{Y},\mathcal{Y}}$ be the operator norm on $\mathcal{L}(\mathcal{Y})$ (largest eigenvalue). We introduce the following technical lemma.

Lemma 40

$\forall \delta \in \mathcal{D}_{\mathcal{C}}$, *if*

$$L_F \leq \frac{\epsilon}{2r} \tag{34}$$

and

$$\|F(\delta_i)\|_{\mathcal{Y},\mathcal{Y}} \leq \frac{\epsilon}{2}, \quad \textit{for all } i \in \mathbb{N}_T^* \tag{35}$$

then $\|F(\delta)\|_{\mathcal{Y},\mathcal{Y}} \leq \epsilon$.

Proof $\|F(\delta)\|_{\mathcal{Y},\mathcal{Y}} = \|F(\delta) - F(\delta_i) + F(\delta_i)\|_{\mathcal{Y},\mathcal{Y}} \leq \|F(\delta) - F(\delta_i)\|_{\mathcal{Y},\mathcal{Y}} + \|F(\delta_i)\|_{\mathcal{Y},\mathcal{Y}}$ for all $0 < i < T$. Using the Lipschitz continuity of F we have $\|F(\delta) - F(\delta_i)\|_{\mathcal{Y},\mathcal{Y}} \leq d_{\mathcal{X}}(\delta, \delta_i) L_F \leq rL_F$ hence $\|F(\delta)\|_{\mathcal{Y},\mathcal{Y}} \leq rL_F + \|F(\delta_i)\|_{\mathcal{Y},\mathcal{Y}} = \frac{r\epsilon}{2r} + \frac{\epsilon}{2} = \epsilon$. ■

To apply the lemma, we must bound the Lipschitz constant of the operator-valued function F (Equation 34) and $\|F(\delta_i)\|_{\mathcal{Y},\mathcal{Y}}$, for all $i = 1, \ldots, T$ as well (Equation 35).

B.2.2 Bounding the Lipschitz constant

This proof is a slight generalization of Minh (2016) to arbitrary metric spaces. It differ from our first approach (Brault et al., 2016), based on the proof of Sutherland and Schneider (2015) which was only valid for a finite dimensional input space $\mathcal{X}$ and imposed a twice differentiability condition on the considered kernel.

Lemma 41

Let $H_\omega \in \mathbb{R}_+$ be the Lipschitz constant of $h_\omega(\cdot)$ and assume that $\int_{\widehat{\mathcal{X}}} H_\omega \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}(\omega) < \infty$. Then the operator-valued function $K_e : \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ is Lipschitz with

$$\|K_e(x) - K_e(z)\|_{\mathcal{Y},\mathcal{Y}} \leq d_{\mathcal{X}}(x,z) \int_{\widehat{\mathcal{X}}} H_\omega \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}(\omega). \tag{36}$$

Proof We use the fact that the cosine function is Lipschitz with constant 1 and h_ω Lipschitz with constant H_ω. For all x, $z \in \mathcal{X}$ we have

$$\begin{aligned}
\left\|\tilde{K}_e(x) - K_e(z)\right\|_{\mathcal{Y},\mathcal{Y}} &= \left\| \int_{\widehat{\mathcal{X}}} (\cos h_\omega(x) - \cos h_\omega(z))\, A(\omega) d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho} \right\|_{\mathcal{Y},\mathcal{Y}} \\
&\leq \int_{\widehat{\mathcal{X}}} |\cos h_\omega(x) - \cos h_\omega(z)| \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho} \\
&\leq \int_{\widehat{\mathcal{X}}} |h_\omega(x) - h_\omega(z)| \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho} \\
&\leq d_{\mathcal{X}}(x,z) \int_{\widehat{\mathcal{X}}} H_\omega \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}
\end{aligned}$$

■

In the same way, considering $\tilde{K}_e(\delta) = \frac{1}{D}\sum_{j=1}^{D} \cos h_{\omega_j}(\delta) A(\omega_j)$, where $\omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$, we can show that $\tilde{K}_e$ is Lipschitz with $\left\|\tilde{K}_e(x) - \tilde{K}_e(z)\right\|_{\mathcal{Y},\mathcal{Y}} \leq d_{\mathcal{X}}(x,z) \frac{1}{D}\sum_{j=1}^{D} H_{\omega_j} \|A(\omega_j)\|_{\mathcal{Y},\mathcal{Y}}$. Combining the Lipschitz continuity of $\tilde{K}_e$ and $\tilde{K}$ (Lemma 41) we obtain

$$\begin{aligned}
\|F(x) - F(z)\|_{\mathcal{Y},\mathcal{Y}} &= \left\|\tilde{K}_e(x) - \tilde{K}_e(x) - \tilde{K}_e(z) + K_e(z)\right\|_{\mathcal{Y},\mathcal{Y}} \\
&\leq \left\|\tilde{K}_e(x) - \tilde{K}_e(z)\right\|_{\mathcal{Y},\mathcal{Y}} + \|K_e(x) - K_e(z)\|_{\mathcal{Y},\mathcal{Y}} \\
&\leq d_{\mathcal{X}}(x,z) \left(\int_{\widehat{\mathcal{X}}} H_\omega \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho} + \frac{1}{D}\sum_{j=1}^{D} H_{\omega_j} \|A(\omega_j)\|_{\mathcal{Y},\mathcal{Y}} \right)
\end{aligned}$$

Taking the expectation yields $\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}[L_F] = 2\int_{\widehat{\mathcal{X}}} H_\omega \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}(\omega)$ Thus by Markov's inequality,

$$\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}\{\, (\omega_j)_{j=1}^{D} \mid L_F \geq \epsilon \,\} \leq \frac{\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}[L_F]}{\epsilon} \leq \frac{2}{\epsilon} \int_{\widehat{\mathcal{X}}} H_\omega \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}. \tag{37}$$

B.2.3 Bounding F on a given anchor point δ_i

To bound $\|F(\delta_i)\|_{\mathcal{Y},\mathcal{Y}}$, Hoeffding inequality devoted to matrix concentration (Mackey et al., 2014) can be applied. We prefer here to turn to tighter and refined inequalities such as Matrix Bernstein inequalities (Sutherland and Schneider (2015) also pointed that for the scalar case). The first non-commutative (matrix) concentration inequalities are due to the pioneer work of Ahlswede and Winter (2002), using bound on the moment generating function. This gave rise to many applications Tropp (2012); Oliveira (2009); Koltchinskii et al. (2013) ranging from analysis of randomized optimization algorithm to analysis of

random graphs and generalization bounds usefull in machine learning. The concentration inequatilty of Koltchinskii et al. (2013) we used in our original paper (Brault et al., 2016) has the default to grow linearly with the dimension p of the output space $\mathcal{Y}$. However if the evaluation of the operator-valued kernel at two points yields a low-rank matrix, this bound could be improved since only a few principal dimensions are relevant. Moreover this bound cannot be used when dealing with operator-valued kernel acting on infinite dimensional Hilbert spaces. Recent results of Minsker (2011) consider the notion of intrinsic dimension to avoid this "curse of dimensionality" (see definition 30 for the definition). When A is approximately low-rank (index˙cmd many eigenvalues are small), or go quickly to zero, the intrinsic dimension can be much lower than the dimensionality. Indeed, $1 \leq \mathrm{IntDim}(A) \leq \mathrm{Rank}(A) \leq \dim(A)$.

Theorem 42 (Bounded non-commutative Bernstein with intrinsic dimension (Minsker, 2011; Tropp et al., 2015))

Consider a sequence $(X_j)_{j=1}^D$ of D independent Hilbert-Schmidt self-adjoint random operators acting on a separable Hilbert $\mathcal{Y}$ space that satisfy $\mathbf{E}X_j = 0$ for all $j \in \mathbb{N}_D^*$. Suppose that there exist some constant $U \geq 2\|X_j\|_{\mathcal{Y},\mathcal{Y}}$ almost surely for all $j \in \mathbb{N}_D^*$. Define a semi-definite upper bound for the the operator-valued variance $V \succcurlyeq \sum_{j=1}^D \mathbf{E}X_j^2$. Then for all $\epsilon \geq \sqrt{\|V\|_{\mathcal{Y},\mathcal{Y}}} + U/3$,

$$\mathbf{Pr}\left\{ \left\| \sum_{j=1}^{D} X_j \right\|_{\mathcal{Y},\mathcal{Y}} \geq \epsilon \right\} \leq 4\,\mathrm{IntDim}(V) \exp\left(-\psi_{V,U}(\epsilon)\right)$$

where $\psi_{V,U}(\epsilon) = \frac{\epsilon^2}{2\|V\|_{\mathcal{Y},\mathcal{Y}} + 2U\epsilon/3}$

he concentration inequality is restricted to the case where $\epsilon \geq \sqrt{\|V\|_{\mathcal{Y},\mathcal{Y}}} + U/3$ since the probability is vacuous on the contrary. The assumption that X_j's are Hilbert-Schmidt operators comes from the fact that the product of two such operator yields a trace-class operator, for which the intrinsic dimension is well defined.

However, to cover the general case including unbounded index˙cmd like curl and divergence-free index˙cmd, we choose a version of Bernstein matrix concentration inequality proposed in Koltchinskii et al. (2013) that allows to consider matrices that are not uniformly bounded but have subexponential tails. In the following we use the notion of Orlicz norm to bound random variable by their tail behavior rather than their value (see definition 29). For the sake of simplicity, we now fix $\psi(t) = \psi_1(t) = \exp(t) - 1$. Although the Orlicz norm should be adapted to the tail of the distribution of the random operator we want to quantify to obtain the sharpest bounds. We also introduce two technical lemmas related to Orlicz norm. The first one relates the ψ_1-Orlicz norm to the moment generating function (MGF).

Lemma 43

Let X be a random variable with a strictly monotonic moment-generating function. We have $\|X\|_{\psi_1}^{-1} = MGF_{|X|}^{-1}(2)$.

Proof We have

$$\|X\|_{\psi_1} = \inf\{\, C > 0 \;\mid\; \mathbf{E}[\exp(|X|/C)] \leq 2 \,\} = \frac{1}{\sup\{\, C > 0 \;\mid\; \mathrm{MGF}_{|X|}(C) \leq 2 \,\}}.$$

X has strictly monotonic moment-generating thus $C^{-1} = \mathrm{MGF}^{-1}_{|X|}(2)$. Hence $\|X\|^{-1}_{\psi_1} = \mathrm{MGF}^{-1}_{|X|}(2)$. ■

The second lemma gives the Orlicz norm of a positive constant.

Lemma 44

If $a \in \mathbb{R}_+$ then $\|a\|_{\psi_1} = \frac{a}{\ln(2)} < 2a$.

Proof We consider a as a positive constant random variable, whose (index˙cmd) is $\mathrm{MGF}_a(t) = \exp(at)$. From Lemma 43, $\|a\|_{\psi_1} = \frac{1}{\mathrm{MGF}^{-1}_X(2)}$. Then $\mathrm{MGF}^{-1}_{|a|}(2) = \frac{\ln(2)}{|a|}$, $a \neq 0$. If $a = 0$ then $\|a\|_{\psi_1} = 0$ by definition of a norm. Thus $\|a\|_{\psi_1} = \frac{a}{\ln(2)}$. ■

We now turn our attention to Minsker (2011)'s theorem to for unbounded random variables.

Theorem 45 (Unbounded non-commutative Bernstein with intrinsic dimension)

Consider a sequence $(X_j)_{j=1}^D$ of D independent self-adjoint random operators acting on a finite dimensional Hilbert space $\mathcal{Y}$ of dimension p that satisfy $\mathbf{E}X_j = 0$ for all $j \in \mathbb{N}^*_D$. Suppose that there exist some constant $U \geq \left\| \|X_j\|_{\mathcal{Y},\mathcal{Y}} \right\|_\psi$ for all $j \in \mathbb{N}^*_D$. Define a semi-definite upper bound for the the operator-valued variance $V \succcurlyeq \sum_{j=1}^D \mathbf{E}X_j^2$. Then for all $\epsilon > 0$,

$$\mathbf{Pr}\left\{ \left\| \sum_{j=1}^D X_j \right\|_{\mathcal{Y},\mathcal{Y}} \geq \epsilon \right\} \leq \begin{cases} 2\,\mathrm{IntDim}(V) \exp\left(-\frac{\epsilon^2}{2\|V\|_{\mathcal{Y},\mathcal{Y}}\left(1+\frac{1}{p}\right)} \right) r_V(\epsilon), & \epsilon \leq \frac{\|V\|_{\mathcal{Y},\mathcal{Y}}}{2U} \frac{1+1/p}{K(V,p)} \\ 2\,\mathrm{IntDim}(V) \exp\left(-\frac{\epsilon}{4UK(V,p)} \right) r_V(\epsilon), & \textit{otherwise.} \end{cases}$$

where $K(V,p) = \log\left(16\sqrt{2}p\right) + \log\left(\frac{DU^2}{\|V\|_{\mathcal{Y},\mathcal{Y}}}\right)$ and $r_V(\epsilon) = 1 + \frac{3}{\epsilon^2 \log^2(1+\epsilon/\|V\|_{\mathcal{Y},\mathcal{Y}})}$

Let $\psi = \psi_1$. To use Theorem 45, we set $X_j = F^j(\delta_i)$. We have indeed $\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}[F^j(\delta_i)] = 0$ since $\tilde{K}(\delta_i)$ is the Monte-Carlo approximation of $K_e(\delta_i)$ and the matrices $F^j(\delta_i)$ are self-adjoint. We assume we can bound all the Orlicz norms of the $F^j(\delta_i) = \frac{1}{D}(\tilde{K}^j(\delta_i) - K_e(\delta_i))$. In the following we use constants u_i such that $u_i = DU$. Using Lemma 44 and the sub-additivity of the $\|\cdot\|_{\mathcal{Y},\mathcal{Y}}$ and $\|\cdot\|_{\psi_1}$ norm,

$$\begin{aligned} u_i &= 2D \max_{1\leq j\leq D} \left\| \left\| F^j(\delta_i) \right\|_{\mathcal{Y},\mathcal{Y}} \right\|_{\psi_1} \leq 2 \max_{1\leq j\leq D} \left\| \left\| \tilde{K}^j(\delta_i) \right\|_{\mathcal{Y},\mathcal{Y}} \right\|_{\psi_1} + 2 \left\| \|K_e(\delta_i)\|_{\mathcal{Y},\mathcal{Y}} \right\|_{\psi_1} \\ &< 4 \max_{1\leq j\leq D} \left\| \|A(\omega_j)\|_{\mathcal{Y},\mathcal{Y}} \right\|_{\psi_1} + 4\|K_e(\delta_i)\|_{\mathcal{Y},\mathcal{Y}} = 4\left(\left\| \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} \right\|_{\psi_1} + \|K_e(\delta_i)\|_{\mathcal{Y},\mathcal{Y}} \right) \end{aligned}$$

In the same way we defined the constants $v_i = DV$, $v_i = D\sum_{j=1}^{D} \mathbf{E}_{\widehat{\mathbf{Haar}},\rho} F^j(\delta_i)^2 = D\mathbf{Var}_{\widehat{\mathbf{Haar}},\rho}\left[\tilde{K}(\delta_i)\right]$ Then applying Theorem 45, we get for all $i \in \mathbb{N}^*_{\mathcal{N}(\mathcal{D}_\mathcal{C},r)}$ (i is the index of each anchor)

$$\begin{aligned}
&\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}\left\{ (\omega_j)_{j=1}^D \;\middle|\; \|F(\delta_i)\|_{\mathcal{Y},\mathcal{Y}} \geq \epsilon \right\} \\
&\leq \begin{cases} 4\,\mathrm{IntDim}(v_i)\exp\left(-D\frac{\epsilon^2}{2\|v_i\|_{\mathcal{Y},\mathcal{Y}}\left(1+\frac{1}{p}\right)}\right) r_{v_i/D}(\epsilon), & \epsilon \leq \frac{\|v_i\|_{\mathcal{Y},\mathcal{Y}}}{2u_i}\frac{1+1/p}{K(v_i,p)} \\ 4\,\mathrm{IntDim}(v_i)\exp\left(-D\frac{\epsilon}{4u_iK(v_i,p)}\right) r_{v_i/D}(\epsilon), & \text{otherwise.} \end{cases}
\end{aligned}$$

with

$$K(v_i,p) = \log\left(16\sqrt{2}p\right) + \log\left(\frac{u_i^2}{\|v_i\|_{\mathcal{Y},\mathcal{Y}}}\right)$$

and

$$r_{v_i/D} = 1 + \frac{3}{\epsilon^2\log^2(1 + D\epsilon/\|v_i\|_{\mathcal{Y},\mathcal{Y}})}.$$

To unify the bound on each anchor we define two constant

$$u = 4\left(\left\|\|A(\omega)\|_{\mathcal{Y},\mathcal{Y}}\right\|_{\psi_1} + \sup_{\delta\in\mathcal{D}_\mathcal{C}}\|K_e(\delta)\|_{\mathcal{Y},\mathcal{Y}}\right) \geq \max_{i=1,\dots T} u_i$$

and

$$v = \sup_{\delta\in\mathcal{D}_\mathcal{C}} D\mathbf{Var}_{\widehat{\mathbf{Haar}},\rho}\left[\tilde{K}_e(\delta)\right] \geq \max_{i=1,\dots T} v_i.$$

B.2.4 Union Bound and examples

Taking the union bound over the anchors yields

$$\begin{aligned}
&\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}\left\{ (\omega_j)_{j=1}^D \;\middle|\; \bigcup_{i=1}^{\mathcal{N}(\mathcal{D}_\mathcal{C},r)} \|F(\delta_i)\|_{\mathcal{Y},\mathcal{Y}} \geq \epsilon \right\} \\
&\leq 4\mathcal{N}(\mathcal{D}_\mathcal{C},r) r_{v/D}(\epsilon)\,\mathrm{IntDim}(v) \begin{cases} \exp\left(-D\frac{\epsilon^2}{2\|v\|_{\mathcal{Y},\mathcal{Y}}\left(1+\frac{1}{p}\right)}\right), & \epsilon \leq \frac{\|v\|_{\mathcal{Y},\mathcal{Y}}}{2u}\frac{1+1/p}{K(v,p)} \\ \exp\left(-D\frac{\epsilon}{4uK(v,p)}\right), & \text{otherwise.} \end{cases}
\end{aligned} \tag{38}$$

Hence combining Equation 37 and Equation 38 gives and summing up the hypothesis yields the following proposition

Proposition 46

Let $K : \mathcal{X}\times\mathcal{X} \to \mathcal{L}(\mathcal{Y})$ be a shift-invariant $\mathcal{Y}$-Mercer kernel, where $\mathcal{Y}$ is a finite dimensional Hilbert space of dimension p and $\mathcal{X}$ a metric space. Moreover, let $\mathcal{C}$ be a compact subset of $\mathcal{X}$, $A : \widehat{\mathcal{X}} \to \mathcal{L}(\mathcal{Y})$ and $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ a pair such that $\tilde{K}_e = \sum_{j=1}^{D}\cos(\cdot,\omega_j)A(\omega_j) \approx K_e$

$\omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ *index˙cmd. Let* $V(\delta) \succcurlyeq \mathbf{Var}_{\widehat{\mathbf{Haar}},\rho}\tilde{K}_e(\delta)$, *for all* $\delta \in \mathcal{D}_\mathcal{C}$ *and* H_ω *be the Lipschitz constant of the function* $h : x \mapsto (x, \omega)$. *If the three following constant exists*

$$m \geq \int_{\widehat{\mathcal{X}}} H_\omega \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho} < \infty$$

and

$$u \geq 4\left(\left\|\|A(\omega)\|_{\mathcal{Y},\mathcal{Y}}\right\|_{\psi_1} + \sup_{\delta\in\mathcal{D}_\mathcal{C}} \|K_e(\delta)\|_{\mathcal{Y},\mathcal{Y}}\right) < \infty$$

and

$$v \geq \sup_{\delta\in\mathcal{D}_\mathcal{C}} D\|V(\delta)\|_{\mathcal{Y},\mathcal{Y}} < \infty.$$

Define $p_{int} \geq \sup_{\delta\in\mathcal{D}_\mathcal{C}} \mathrm{IntDim}(V(\delta))$ *then for all* $r \in \mathbb{R}_+^*$ *and all* $\epsilon \in \mathbb{R}_+^*$,

$$\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}\left\{ (\omega_j)_{j=1}^D \;\middle|\; \left\|\tilde{K} - K\right\|_{\mathcal{C}\times\mathcal{C}} \geq \epsilon \right\} \leq 4\left(\frac{rm}{\epsilon} + p_{int}\mathcal{N}(\mathcal{D}_\mathcal{C}, r) r_{v/D}(\epsilon) \begin{cases} \exp\left(-D\frac{\epsilon^2}{8v\left(1+\frac{1}{p}\right)}\right), & \epsilon \leq \frac{v}{u}\frac{1+1/p}{K(v,p)} \\ \exp\left(-D\frac{\epsilon}{8uK(v,p)}\right), & \textit{otherwise.} \end{cases}\right)$$

where

$$K(v,p) = \log\left(16\sqrt{2}p\right) + \log\left(\frac{u^2}{\|v\|_{\mathcal{Y},\mathcal{Y}}}\right)$$

and

$$r_{v/D}(\epsilon) = 1 + \frac{3}{\epsilon^2 \log^2(1 + D\epsilon/\|v\|_{\mathcal{Y},\mathcal{Y}})}.$$

Proof Let $m = \int_{\widehat{\mathcal{X}}} H_\omega \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$. From Lemma 41, $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}\left\{ (\omega_j)_{j=1}^D \;\middle|\; L_F \geq \frac{\epsilon}{2r} \right\} \leq \frac{4rm}{\epsilon}$. Thus from Lemma 40, for all $r \in \mathbb{R}_+^*$,

$$\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}\left\{ (\omega_j)_{j=1}^D \;\middle|\; \sup_{\delta\in\mathcal{D}_\mathcal{C}} \|F(\delta)\|_{\mathcal{Y},\mathcal{Y}} \geq \epsilon \right\} \leq$$

$$\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}\left\{ (\omega_j)_{j=1}^D \;\middle|\; L_F \geq \frac{\epsilon}{2r} \right\} + \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}\left\{ (\omega_j)_{j=1}^D \;\middle|\; \bigcup_{i=1}^{\mathcal{N}(\mathcal{D}_\mathcal{C},r)} \|F(\delta_i)\|_{\mathcal{Y},\mathcal{Y}} \geq \epsilon \right\} = 4\frac{rm}{\epsilon}$$

$$+ 4\mathcal{N}(\mathcal{D}_\mathcal{C}, r) r_{v/D}(\epsilon)\,\mathrm{IntDim}(v) \begin{cases} \exp\left(-D\frac{\epsilon^2}{8\|v\|_{\mathcal{Y},\mathcal{Y}}\left(1+\frac{1}{p}\right)}\right), & \epsilon \leq \frac{\|v\|_{\mathcal{Y},\mathcal{Y}}}{u}\frac{1+1/p}{K(v,p)} \\ \exp\left(-D\frac{\epsilon}{8uK(v,p)}\right), & \text{otherwise.} \end{cases}$$

■

With minor modifications we can obtain a second inequality for the case where the random operators $A(\omega_j)$ are bounded almost surely. This second bound with more restrictions on A has the advantage of working in infinite dimension as long as $A(\omega_j)$ is a Hilbert-Schmidt operator.

Proposition 47

Let $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ be a shift-invariant $\mathcal{Y}$-Mercer kernel, where $\mathcal{Y}$ is a Hilbert space and $\mathcal{X}$ a metric space. Moreover, let $\mathcal{C}$ be a compact subset of $\mathcal{X}$, $A : \widehat{\mathcal{X}} \to \mathcal{L}(\mathcal{Y})$ and $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ a pair such that $\tilde{K}_e = \sum_{j=1}^{D} \cos(\cdot, \omega_j) A(\omega_j) \approx K_e$, $\omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ index˙cmd. where $A(\omega_j)$ is a Hilbert-Schmidt operator for all $j \in \mathbb{N}_D^*$. Let $\mathcal{D}_{\mathcal{C}} = \mathcal{C} \star \mathcal{C}^{-1}$ and $V(\delta) \succcurlyeq \mathbf{Var}_{\widehat{\mathbf{Haar}},\rho} \tilde{K}_e(\delta)$, for all $\delta \in \mathcal{D}_{\mathcal{C}}$ and H_ω be the Lipschitz constant of the function $h : x \mapsto (x, \omega)$. If the three following constant exists

$$m \geq \int_{\widehat{\mathcal{X}}} H_\omega \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho} < \infty$$

and

$$u \geq \operatorname*{ess\,sup}_{\omega \in \widehat{\mathcal{X}}} \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} + \sup_{\delta \in \mathcal{D}_{\mathcal{C}}} \|K_e(\delta)\|_{\mathcal{Y},\mathcal{Y}} < \infty$$

and

$$v \geq \sup_{\delta \in \mathcal{D}_{\mathcal{C}}} D \|V(\delta)\|_{\mathcal{Y},\mathcal{Y}} < \infty.$$

define $p_{int} \geq \sup_{\delta \in \mathcal{D}_{\mathcal{C}}} \operatorname{IntDim}(V(\delta))$ then for all $r \in \mathbb{R}_+^*$ and all $\epsilon > \sqrt{\frac{v}{D}} + \frac{1}{3D} u$,

$$\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho} \left\{ (\omega_j)_{j=1}^{D} \;\middle|\; \sup_{\delta \in \mathcal{D}_{\mathcal{C}}} \|F(\delta)\|_{\mathcal{Y},\mathcal{Y}} \geq \epsilon \right\} \leq \; 4 \left(\frac{rm}{\epsilon} + p_{int} \mathcal{N}(\mathcal{D}_{\mathcal{C}}, r) \exp\left(-D \psi_{v,u}(\epsilon)\right) \right)$$

where $\psi_{v,u}(\epsilon) = \frac{\epsilon^2}{2(v + u\epsilon/3)}$.

When the covering number $\mathcal{N}(\mathcal{D}_{\mathcal{C}}, r)$ of the metric space $\mathcal{D}_{\mathcal{C}}$ has an analytical form, it is possible to optimize the bound over the radius r of the covering balls. As an example, we refine Proposition 46 and Proposition 47 in the case where $\mathcal{C}$ is a finite dimensional Banach space.

Corollary 48

Let $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ be a shift-invariant $\mathcal{Y}$-Mercer kernel, where $\mathcal{Y}$ is a finite dimensional Hilbert space of dimension p and $\mathcal{X}$ a finite dimensional Banach space of dimension d. Moreover, let $\mathcal{C}$ be a closed ball of $\mathcal{X}$ centered at the origin of diameter $|\mathcal{C}|$, $A : \widehat{\mathcal{X}} \to \mathcal{L}(\mathcal{Y})$ and $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ a pair such that $\tilde{K}_e = \sum_{j=1}^{D} \cos(\cdot, \omega_j) A(\omega_j) \approx K_e$, $\omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ index˙cmd. Let $\mathcal{D}_{\mathcal{C}} = \mathcal{C} \star \mathcal{C}^{-1}$ and $V(\delta) \succcurlyeq \mathbf{Var}_{\widehat{\mathbf{Haar}},\rho} \tilde{K}_e(\delta)$, for all $\delta \in \mathcal{D}_{\mathcal{C}}$ Let H_ω be the Lipschitz constant of $h_\omega : x \mapsto (x, \omega)$. If the three following constant exists

$$m \geq \int_{\widehat{\mathcal{X}}} H_\omega \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho} < \infty$$

and

$$u \geq 4\left(\left\|\|A(\omega)\|_{\mathcal{Y},\mathcal{Y}}\right\|_{\psi_1} + \sup_{\delta\in\mathcal{D}_\mathcal{C}}\|K_e(\delta)\|_{\mathcal{Y},\mathcal{Y}}\right) < \infty$$

and

$$v \geq \sup_{\delta\in\mathcal{D}_\mathcal{C}} D\|V(\delta)\|_{\mathcal{Y},\mathcal{Y}} < \infty.$$

Define $p_{int} \geq \sup_{\delta\in\mathcal{D}_\mathcal{C}} \mathrm{IntDim}(V(\delta))$*, then for all* $0 < \epsilon \leq m|C|$*,*

$$\begin{aligned}
&\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}\left\{ (\omega_j)_{j=1}^D \;\middle|\; \left\|\tilde{K} - K\right\|_{\mathcal{C}\times\mathcal{C}} \geq \epsilon \right\} \\
&\leq 8\sqrt{2}\left(\frac{m|\mathcal{C}|}{\epsilon}\right)\left(p_{int} r_{v/D}(\epsilon)\right)^{\frac{1}{d+1}} \begin{cases} \exp\left(-D\frac{\epsilon^2}{8v(d+1)\left(1+\frac{1}{p}\right)}\right), & \epsilon \leq \frac{v}{u}\frac{1+1/p}{K(v,p)} \\ \exp\left(-D\frac{\epsilon}{8u(d+1)K(v,p)}\right), & \textit{otherwise,} \end{cases}
\end{aligned}$$

where $K(v,p) = \log\left(16\sqrt{2}p\right) + \log\left(\frac{u^2}{v}\right)$ *and* $r_{v/D}(\epsilon) = 1 + \frac{3}{\epsilon^2\log^2(1+D\epsilon/v)}$.

Proof As we have seen in appendix B.2.1, suppose that $\mathcal{X}$ is a finite dimensional Banach space. Let $\mathcal{C} \subset \mathcal{X}$ be a closed ball centered at the origin of diameter $|\mathcal{C}| = C$ then the difference ball centered at the origin

$$\mathcal{D}_\mathcal{C} = \mathcal{C} \star \mathcal{C}^{-1} = \left\{ x \star\ z^{-1} \;\middle|\; \|x\|_\mathcal{X} \leq C/2, \|z\|_\mathcal{X} \leq C/2, (x,z) \in \mathcal{X}^2 \right\} \subset \mathcal{X}$$

is closed and bounded, so compact and has diameter $|C| = 2C$. It is possible to cover it with $\log(\mathcal{N}(\mathcal{D}_\mathcal{C}, r)) = d\log\left(\frac{2|C|}{r}\right)$ closed balls of radius r. Pluging back into Equation yields

$$\begin{aligned}
\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}\left\{ (\omega_j)_{j=1}^D \;\middle|\; \left\|\tilde{K} - K\right\|_{\mathcal{C}\times\mathcal{C}} \geq \epsilon \right\} \leq 4 \Bigg(&\frac{rm}{\epsilon} + p_{int}\left(\frac{2|C|}{r}\right)^d r_{v/D}(\epsilon) \\
&\begin{cases} \exp\left(-D\frac{\epsilon^2}{8v\left(1+\frac{1}{p}\right)}\right), & \epsilon \leq \frac{v}{u}\frac{1+1/p}{K(v,p)} \\ \exp\left(-D\frac{\epsilon}{8uK(v,p)}\right), & \text{otherwise.} \end{cases} \Bigg)
\end{aligned}$$

The right hand side of the equation has the form $ar + br^{-d}$ with $a = \frac{m}{\epsilon}$ and

$$b = p_{int}(2|\mathcal{C}|)^d r_{v/D}(\epsilon) \begin{cases} \exp\left(-D\frac{\epsilon^2}{8v\left(1+\frac{1}{p}\right)}\right), & \epsilon \leq \frac{v}{u}\frac{1+1/p}{K(v,p)} \\ \exp\left(-D\frac{\epsilon}{8uK(v,p)}\right), & \text{otherwise.} \end{cases}$$

Following Rahimi and Recht (2007); Sutherland and Schneider (2015); Minh (2016), we optimize over r. It is a convex continuous function on $\mathbb{R}_+$ and achieve minimum at $r =$

$\left(\frac{bd}{a}\right)^{\frac{1}{d+1}}$ and the minimum value is $r_* = a^{\frac{d}{d+1}} b^{\frac{1}{d+1}} \left(d^{\frac{1}{d+1}} + d^{-\frac{d}{d+1}}\right)$, hence

$$\begin{aligned}
&\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}\left\{ (\omega_j)_{j=1}^D \;\middle|\; \left\|\tilde{K} - K\right\|_{\mathcal{C}\times\mathcal{C}} \geq \epsilon \right\} \\
&\leq C_d \left(\frac{2m|\mathcal{C}|}{\epsilon}\right)^{\frac{d}{d+1}} \left(p_{int} r_{v/D}(\epsilon)\right)^{\frac{1}{d+1}} \begin{cases} \exp\left(-D\frac{\epsilon^2}{8v(d+1)\left(1+\frac{1}{p}\right)}\right), & \epsilon \leq \frac{v}{u}\frac{1+1/p}{K(v,p)} \\ \exp\left(-D\frac{\epsilon}{8u(d+1)K(v,p)}\right), & \text{otherwise,} \end{cases} \\
&\leq 8\sqrt{2}\left(\frac{m|\mathcal{C}|}{\epsilon}\right) \left(p_{int} r_{v/D}(\epsilon)\right)^{\frac{1}{d+1}} \begin{cases} \exp\left(-D\frac{\epsilon^2}{8v(d+1)\left(1+\frac{1}{p}\right)}\right), & \epsilon \leq \frac{v}{u}\frac{1+1/p}{K(v,p)} \\ \exp\left(-D\frac{\epsilon}{8u(d+1)K(v,p)}\right), & \text{otherwise,} \end{cases}
\end{aligned}$$

where $C_d = 4\left(d^{\frac{1}{d+1}} + d^{-\frac{d}{d+1}}\right)$. Eventually when $\mathcal{X}$ is a Banach space, the Lipschitz constant of h_ω is the supremum of the gradient $H_\omega = \sup_{\delta\in\mathcal{D}_\mathcal{C}} \|(\nabla h_\omega)(\delta)\|_{\widehat{\mathcal{X}}}$. ■

Following the same proof technique we obtain the second bound for bounded index˙cmd.

Corollary 49

Let $K : \mathcal{X}\times\mathcal{X} \to \mathcal{L}(\mathcal{Y})$ be a shift-invariant $\mathcal{Y}$-Mercer kernel, where $\mathcal{Y}$ is a Hilbert space and $\mathcal{X}$ a finite dimensional Banach space of dimension D. Moreover, let $\mathcal{C}$ be a closed ball of $\mathcal{X}$ centered at the origin of diameter $|\mathcal{C}|$, subset of $\mathcal{X}$, $A : \widehat{\mathcal{X}} \to \mathcal{L}(\mathcal{Y})$ and $\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ a pair such that $\tilde{K}_e = \sum_{j=1}^D \cos(\cdot,\omega_j)A(\omega_j) \approx K_e$, $\omega_j \sim \mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}$ index˙cmd. where $A(\omega_j)$ is a Hilbert-Schmidt operator for all $j \in \mathbb{N}_D^*$. Let $\mathcal{D}_\mathcal{C} = \mathcal{C}\star\mathcal{C}^{-1}$ and $V(\delta) \succcurlyeq \mathbf{Var}_{\widehat{\mathbf{Haar}},\rho}\tilde{K}_e(\delta)$ for all $\delta \in \mathcal{D}_\mathcal{C}$ and H_ω be the Lipschitz constant of the function $h : x \mapsto (x,\omega)$. If the three following constant exists

$$m \geq \int_{\widehat{\mathcal{X}}} H_\omega \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} d\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho} < \infty$$

and

$$u \geq \operatorname*{ess\,sup}_{\omega\in\widehat{\mathcal{X}}} \|A(\omega)\|_{\mathcal{Y},\mathcal{Y}} + \sup_{\delta\in\mathcal{D}_\mathcal{C}} \|K_e(\delta)\|_{\mathcal{Y},\mathcal{Y}} < \infty$$

and

$$v \geq \sup_{\delta\in\mathcal{D}_\mathcal{C}} D\|V(\delta)\|_{\mathcal{Y},\mathcal{Y}} < \infty.$$

define $p_{int} \geq \sup_{\delta\in\mathcal{D}_\mathcal{C}} \operatorname{IntDim}(V(\delta))$ then for all $\sqrt{\frac{v}{D}} + \frac{u}{3D} < \epsilon < m|\mathcal{C}|$,

$$\mathbf{Pr}_{\widehat{\mathbf{Haar}},\rho}\left\{ (\omega_j)_{j=1}^D \;\middle|\; \sup_{\delta\in\mathcal{D}_\mathcal{C}} \|F(\delta)\|_{\mathcal{Y},\mathcal{Y}} \geq \epsilon \right\} \leq \; 8\sqrt{2}\left(\frac{m|\mathcal{C}|}{\epsilon}\right) p_{int}^{\frac{1}{d+1}} \exp\left(-D\psi_{v,d,u}(\epsilon)\right)$$

where $\psi_{v,d,u}(\epsilon) = \frac{\epsilon^2}{2(d+1)(v+u\epsilon/3)}$.

B.2.5 Proof of the ORFF estimator variance bound (Proposition 33).

We use the notations $\delta = x \star z^{-1}$ for all $x, z \in \mathcal{X}$, $\tilde{K}(x,z) = \widetilde{\phi}(x)^* \widetilde{\phi}(z)$, $\tilde{K}^j(x,z) = \phi_x(\omega_j)^* \phi_z(\omega_j)$ and $K_e(\delta) = K_e(x,z)$.

Proof Let $\delta \in \mathcal{D}_\mathcal{C}$ be a constant. From the definition of the variance of a random variable and using the fact that the $(\omega_j)_{j=1}^D$ are index˙cmd random variables,

$$\begin{aligned}\mathbf{Var}_{\widehat{\mathbf{Haar}},\rho}\left[\tilde{K}_e(\delta)\right] &= \mathbf{E}_{\widehat{\mathbf{Haar}},\rho}\left[\frac{1}{D}\sum_{j=1}^D \tilde{K}_e^j(\delta) - K_e(\delta)\right]^2 = \frac{1}{D^2}\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}\left[\sum_{j=1}^D \tilde{K}_e^j(\delta) - K_e(\delta)\right]^2 \\ &= \frac{1}{D}\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}\left[\tilde{K}_e^j(\delta)^2 - \tilde{K}_e^j(\delta)K_e(\delta) - K_e(\delta)\tilde{K}_e^j(\delta) + K_e(\delta)^2\right]\end{aligned}$$

From the definition of $\tilde{K}_e^j$, $\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}\tilde{K}_e^j(\delta) = K_e(\delta)$, which leads to

$$\mathbf{Var}_{\widehat{\mathbf{Haar}},\rho}\left[\tilde{K}_e(\delta)\right] = \frac{1}{D}\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}\left[\tilde{K}_e^j(\delta)^2 - K_e(\delta)^2\right]$$

A trigonometric identity gives us $(\cos(\delta,\omega))^2 = \frac{1}{2}\left(\cos(2\delta,\omega) + \cos(e,\omega)\right)$. Thus

$$\mathbf{Var}_{\widehat{\mathbf{Haar}},\rho}\left[\tilde{K}_e(\delta)\right] = \frac{1}{2D}\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}\left[\left(\cos(2\delta,\omega) + \cos(e,\omega)\right)A(\omega)^2 - 2K_e(\delta)^2\right].$$

Also,

$$\begin{aligned}\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}\left[\cos(2\delta,\omega)A(\omega)^2\right] &= \mathbf{E}_{\widehat{\mathbf{Haar}},\rho}\left[\cos(2\delta,\omega)A(\omega)\right]\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}\left[A(\omega)\right] \\ &\quad + \mathbf{Cov}_{\widehat{\mathbf{Haar}},\rho}\left[\cos(2\delta,\omega)A(\omega), A(\omega)\right] \\ &= K_e(2\delta)\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}\left[A(\omega)\right] + \mathbf{Cov}_{\widehat{\mathbf{Haar}},\rho}\left[\cos(2\delta,\omega)A(\omega), A(\omega)\right]\end{aligned}$$

Similarly we obtain

$$\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}\left[\cos(e,\omega)A(\omega)^2\right] = K_e(e)\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}\left[A(\omega)\right] + \mathbf{Cov}_{\widehat{\mathbf{Haar}},\rho}\left[\cos(e,\omega)A(\omega), A(\omega)\right]$$

Therefore

$$\begin{aligned}\mathbf{Var}_{\widehat{\mathbf{Haar}},\rho}\left[\tilde{K}_e(\delta)\right] &= \frac{1}{2D}\Big(\left(K_e(2\delta) + K_e(e)\right)\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}\left[A(\omega)\right] - 2K_e(\delta)^2 \\ &\qquad + \mathbf{Cov}_{\widehat{\mathbf{Haar}},\rho}\left[\left(\cos(2\delta,\omega) + \cos(e,\omega)\right)A(\omega), A(\omega)\right]\Big) \\ &= \frac{1}{2D}\Big(\left(K_e(2\delta) + K_e(e)\right)\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}\left[A(\omega)\right] - 2K_e(\delta)^2 \\ &\qquad + \mathbf{Cov}_{\widehat{\mathbf{Haar}},\rho}\left[\left(\cos(\delta,\omega)\right)^2 A(\omega), A(\omega)\right]\Big) \\ &\preccurlyeq \frac{1}{2D}\Big(\left(K_e(2\delta) + K_e(e)\right)\mathbf{E}_{\widehat{\mathbf{Haar}},\rho}\left[A(\omega)\right] - 2K_e(\delta)^2 + \mathbf{Var}_{\widehat{\mathbf{Haar}},\rho}\left[A(\omega)\right]\Big)\end{aligned}$$

■

B.3 Learning

B.3.1 Proof of Theorem 34

Proof Since $f(x) = K_x^* f$, the optimization problem reads

$$f_{\boldsymbol{s}} = \operatorname*{arg\,min}_{f \in \mathcal{H}_K} \frac{1}{N} \sum_{i=1}^{N} c(K_{x_i}^* f, y_i) + \frac{\lambda}{2} \|f\|_K^2$$

Let $W_{\boldsymbol{s}} : \mathcal{H}_K \to \bigoplus_{i=1}^N \mathcal{Y}$ be the restriction[14] linear operator defined as $W_{\boldsymbol{s}} f = \bigoplus_{i=1}^N K_{x_i}^* f$, with $K_{x_i}^* : \mathcal{H}_K \to \mathcal{Y}$ and $K_{x_i} : \mathcal{Y} \to \mathcal{H}_K$. Let $Y = \bigoplus_{i=1}^N y_i \in \mathcal{Y}^N$. We have $\langle Y, W_{\boldsymbol{s}} f \rangle_{\bigoplus_{i=1}^N \mathcal{Y}} = \sum_{i=1}^N \langle y_i, K_{x_i}^* f \rangle_{\mathcal{Y}} = \sum_{i=1}^N \langle K_{x_i} y_i, f \rangle_{\mathcal{H}_K}$. Thus the adjoint operator $W_{\boldsymbol{s}}^* : \bigoplus_{i=1}^N \mathcal{Y} \to \mathcal{H}_K$ is $W_{\boldsymbol{s}}^* Y = \sum_{i=1}^N K_{x_i} y_i$, and the operator $W_{\boldsymbol{s}}^* W_{\boldsymbol{s}} : \mathcal{H}_K \to \mathcal{H}_K$ is $W_{\boldsymbol{s}}^* W_{\boldsymbol{s}} f = \sum_{i=1}^N K_{x_i} K_{x_i}^* f$. Let $\mathfrak{R}_\lambda(f, \boldsymbol{s}) = \underbrace{\frac{1}{N} \sum_{i=1}^{N} c(f(x_i), y_i)}_{=\mathfrak{R}_c} + \frac{\lambda}{2} \|f\|_K^2$. To ensure that $\mathfrak{R}_\lambda$ has a global minimizer we need the following technical lemma (which is a consequence of the Hahn-Banach theorem for lower-semicontimuous functional, see Kurdila and Zabarankin (2006)).

Lemma 50

Let $\mathfrak{R}$ be a proper, convex, lower semi-continuous functional, defined on a Hilbert space $\mathcal{H}$. If $\mathfrak{R}$ is strongly convex, then $\mathfrak{R}$ is coercive.

Since c is proper, lower semi-continuous and convex by assumption, thus the term $\mathfrak{R}_c$ is also proper, lower semi-continuous and convex. Moreover the term $\frac{\lambda}{2}\|f\|_K^2$ is strongly convex. Thus $\mathfrak{R}_\lambda$ is strongly convex. Apply Lemma 50 to obtain the coercivity of $\mathfrak{R}_\lambda$, and then Mazur-Schauder's theorem (see Górniewicz (1999); Kurdila and Zabarankin (2006)) to show that $\mathfrak{R}_\lambda$ has a unique minimizer and is attained. Then let $\mathcal{H}_{K,\boldsymbol{s}} = \left\{ \sum_{j=1}^N K_{x_j} u_j \;\middle|\; \forall (u_i)_{i=1}^N \in \mathcal{Y}^N \right\}$. For $f \in \mathcal{H}_{K,\boldsymbol{s}}^{\perp}$[15], the operator $W_{\boldsymbol{s}}$ satisfies $\langle Y, W_{\boldsymbol{s}} f \rangle_{\bigoplus_{i=1}^N \mathcal{Y}} = \langle \underbrace{f}_{\in \mathcal{H}_{K,\boldsymbol{s}}^{\perp}}, \underbrace{\sum_{i=1}^{N} K_{x_i} V^* y_i}_{\in \mathcal{H}_{K,\boldsymbol{s}}} \rangle_{\mathcal{H}_K} = 0$ for all sequences $(y_i)_{i=1}^N$, since $y_i \in \mathcal{Y}$. Hence,

$$(f(x_i))_{i=1}^N = 0 \tag{39}$$

In the same way, $\sum_{i=1}^N \langle K_{x_i}^* f, u_i \rangle_{\mathcal{Y}} = \langle \underbrace{f}_{\in \mathcal{H}_{K,\boldsymbol{s}}^{\perp}}, \underbrace{\sum_{j=1}^{N} K_{x_j} u_j}_{\in \mathcal{H}_{K,\boldsymbol{s}}} \rangle_{\mathcal{H}_K} = 0.$ for all sequences $(u_i)_{i=1}^N \in \mathcal{Y}^N$. As a result,

$$(f(x_i))_{i=1}^N = 0. \tag{40}$$

14. $W_{\boldsymbol{s}}$ is sometimes called the sampling or evaluation operator as in Minh et al. (2016). However we prefer calling it "restriction operator" as in Rosasco et al. (2010) since $W_{\boldsymbol{s}} f$ is the restriction of f to the points in $\boldsymbol{s}$.

15. $\mathcal{H}_{K,\boldsymbol{s}}^{\perp} \oplus \mathcal{H}_{K,\boldsymbol{s}} = \mathcal{H}_K$ because $W_{\boldsymbol{s}}$ is bounded.

Now for an arbitrary $f \in \mathcal{H}_\mathcal{K}$, consider the orthogonal decomposition $f = f^\perp + f^\parallel$, where $f^\perp \in \mathcal{H}_{K,\boldsymbol{s}}^\perp$ and $f^\parallel \in \mathcal{H}_{K,\boldsymbol{s}}$. Then since $\left\|f^\perp + f^\parallel\right\|_{\mathcal{H}_K}^2 = \left\|f^\perp\right\|_{\mathcal{H}_K}^2 + \left\|f^\parallel\right\|_{\mathcal{H}_K}^2$, Equation 39 and Equation 40 shows that if $\lambda > 0$, clearly then $\mathfrak{R}_\lambda(f, \boldsymbol{s}) = \mathfrak{R}_\lambda\left(f^\perp + f^\parallel, \boldsymbol{s}\right) \geq \mathfrak{R}_\lambda\left(f^\parallel, \boldsymbol{s}\right)$ The last inequality holds only when $\left\|f^\perp\right\|_{\mathcal{H}_K} = 0$, that is when $f^\perp = 0$. As a result since the minimizer of $\mathfrak{R}_\lambda$is unique and attained, it must lies in $\mathcal{H}_{K,\boldsymbol{s}}$. ■

B.3.2 Proof of Theorem 35

Proof Since $\widetilde{K}$ is an operator-valued kernel, from Theorem 34, Equation 24 has a solution of the form

$$\widetilde{f}_{\boldsymbol{s}} = \sum_{i=1}^N \widetilde{K}(\cdot, x_i)u_i = \sum_{i=1}^N \widetilde{\phi}(\cdot)^*\widetilde{\phi}(x_i)u_i = \widetilde{\phi}(\cdot)^* \underbrace{\left(\sum_{i=1}^N \widetilde{\phi}(x_i)u_i\right)}_{=\theta \in \left(\operatorname{Ker}\widetilde{W}\right)^\perp \subset \widetilde{\mathcal{H}}},$$

where $u_i \in \mathcal{Y}$ and $x_i \in \mathcal{X}$. Let $\theta_{\boldsymbol{s}} = \arg\min_{\theta \in \left(\operatorname{Ker}\widetilde{W}\right)^\perp} \frac{1}{N}\sum_{i=1}^N c\left(\widetilde{\phi}(x_i)^*\theta, y_i\right) + \frac{\lambda}{2}\left\|\widetilde{\phi}(\cdot)^*\theta\right\|_{\widetilde{K}}^2$. Since $\theta \in (\operatorname{Ker}\ \widetilde{W})^\perp$ and W is an isometry from $(\operatorname{Ker}\ \widetilde{W})^\perp \subset \widetilde{\mathcal{H}}$ onto $\mathcal{H}_{\widetilde{K}}$, we have $\left\|\widetilde{\phi}(\cdot)^*\theta\right\|_{\widetilde{K}}^2 = \|\theta\|_{\widetilde{\mathcal{H}}}^2$. Hence $\theta_{\boldsymbol{s}} = \arg\min_{\theta \in \left(\operatorname{Ker}\widetilde{W}\right)^\perp} \frac{1}{N}\sum_{i=1}^N c\left(\widetilde{\phi}(x_i)^*\theta, y_i\right) + \frac{\lambda}{2}\|\theta\|_{\widetilde{\mathcal{H}}}^2$ Finding a minimizer $\theta_{\boldsymbol{s}}$ over $\left(\operatorname{Ker}\widetilde{W}\right)^\perp$ is not the same as finding a minimizer over $\widetilde{\mathcal{H}}$. Although in both cases Mazur-Schauder's theorem guarantees that the respective minimizers are unique, they might not be the same. Since $\widetilde{W}$ is bounded, Ker $\widetilde{W}$ is closed, so that we can perform the decomposition $\widetilde{\mathcal{H}} = \left(\operatorname{Ker}\widetilde{W}\right)^\perp \oplus \left(\operatorname{Ker}\widetilde{W}\right)$. Then clearly by linearity of W and the fact that for all $\theta^\parallel \in \operatorname{Ker}\widetilde{W}$, $\widetilde{W}\theta^\parallel = 0$, if $\lambda > 0$ we have $\theta_{\boldsymbol{s}} = \arg\min_{\theta \in \widetilde{\mathcal{H}}} \frac{1}{N}\sum_{i=1}^N c\left(\widetilde{\phi}(x_i)^*\theta, y_i\right) + \frac{\lambda}{2}\|\theta\|_{\widetilde{\mathcal{H}}}^2$ Thus $\theta_{\boldsymbol{s}} = \arg\min_{\substack{\theta^\perp \in \left(\operatorname{Ker}\widetilde{W}\right)^\perp, \\ \theta^\parallel \in \operatorname{Ker}\widetilde{W}}} \frac{1}{N}\sum_{i=1}^N c\left(\left(\widetilde{W}\theta^\perp\right)(x) + \underbrace{\left(\widetilde{W}\theta^\parallel\right)(x)}_{=0 \text{ for all } \theta^\parallel}, y_i\right) + \frac{\lambda}{2}\left\|\theta^\perp\right\|_{\widetilde{\mathcal{H}}}^2 + \underbrace{\frac{\lambda}{2}\left\|\theta^\parallel\right\|_{\widetilde{\mathcal{H}}}^2}_{=0 \text{ only if } \theta^\parallel = 0}$ Thus $\theta_{\boldsymbol{s}} = \arg\min_{\theta^\perp \in \left(\operatorname{Ker}\widetilde{W}\right)^\perp} \frac{1}{N}\sum_{i=1}^N c\left(\left(\widetilde{W}\theta^\perp\right)(x), y_i\right) + \frac{\lambda}{2}\left\|\theta^\perp\right\|_{\widetilde{\mathcal{H}}}^2$ Hence minimizing over $\left(\operatorname{Ker}\widetilde{W}\right)^\perp$ or $\widetilde{\mathcal{H}}\omega$ is the same when $\lambda > 0$. Eventually, $\theta_{\boldsymbol{s}} = \arg\min_{\theta \in \widetilde{\mathcal{H}}} \frac{1}{N}\sum_{i=1}^N c\left(\widetilde{\phi}(x_i)^*\theta, y_i\right) + \frac{\lambda}{2}\|\theta\|_{\widetilde{\mathcal{H}}}^2$. ■

References

R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory 48(3)*, pages 569–679, 2002.

M. A. Álvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: a review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012.

P.-O. Amblard and H. Kadri. Operator-valued kernel recursive least squares algorithm. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 2376–2380. IEEE, 2015.

A. Argyriou, C. A. Micchelli, and M. Pontil. When is there a representer theorem? Vector vs matrix regularizers. *Journal of Machine Learning Research*, 10:2507–2529, 2009.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, pages 337–404, 1950.

J. Audiffren and H. Kadri. Online learning with operator-valued kernels. In *European symposium on artificial neural networks (ESANN)*, 2015.

F. Bach. On the equivalence between quadrature rules and random features. HAl-report-/hal-01118276, 2015.

L. Baldassarre, L. Rosasco, A. Barla, and A. Verri. Vector field learning via spectral filtering. In J. Balcazar, F. Bonchi, A. Gionis, and M. Sebag, editors, *ECML/PKDD*, volume 6321 of *LNCS*, pages 56–71. Springer Berlin / Heidelberg, 2010.

L. Baldassarre, L. Rosasco, A. Barla, and A. Verri. Multi-output learning via spectral filtering. *Machine Learning*, 87(3):259–301, 2012.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities*. Oxford Press, 2013.

R. Brault, M. Heinonen, and F. d'Alché Buc. Random fourier features for operator-valued kernels. In *Proceedings of The 8th Asian Conference on Machine Learning*, pages 110–125, 2016.

C. Brouard, F. d'Alché-Buc, and M. Szafranski. Semi-supervised penalized output kernel regression for link prediction. In *Proc. of the 28th Int. Conf. on Machine Learning*, 2011.

C. Brouard, F. d'Alché-Buc, and M. Szafranski. Input output kernel regression. *to appear in JMLR*, 2016a.

C. Brouard, H. Shen, K. Dührkop, F. d'Alché Buc, S. Böcker, and J. Rousu. Fast metabolite identification with input output kernel regression. *Bioinformatics*, 32(12):i28–i36, 2016b.

R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

A. Caponnetto, C. A. Micchelli, M. , and Y. Ying. Universal multitask kernels. *Journal of Machine Learning Research*, 9:1615–1646, 2008.

C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4(04):377–408, 2006.

C. Carmeli, E. De Vito, A. Toigo, and V. Umanità. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8:19–61, 2010.

C. Ciliberto, Y. Mroueh, T. Poggio, and L. Rosasco. Convex Learning of Multiple Tasks and their Structure. In *Proc. of the 32nd International Conference on Machine Learning*, 2015.

J. B. Conway. *A course in functional analysis*, volume 96. Springer Science & Business Media, 2013.

F. Cucker and S. Smale. On the mathematical foundation of learning. *AMERICAN MATHEMATICAL SOCIETY*, 39(1):1–49, 2001.

B. Dai, B. Xie, N. He, Y. Liang, A. Raj, M.-F. F. Balcan, and L. Song. Scalable kernel methods via doubly stochastic gradients. In *Advances in Neural Information Processing Systems*, pages 3041–3049, 2014.

F. Dinuzzo, C. Ong, P. Gehler, and G. Pillonetto. Learning output kernels with block coordinate descent. In *Proc. of the 28th Int. Conf. on Machine Learning*, 2011.

P. Drineas and M. W. Mahoney. On the nyström method for approximating a gram matrix for improved kernel-based learning. *journal of machine learning research*, 6(Dec):2153–2175, 2005.

R. M. Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *JMLR*, 6:615–637, 2005.

X. Y. Felix, A. T. Suresh, K. M. Choromanski, D. N. Holtmann-Rice, and S. Kumar. Orthogonal random features. In *Advances in Neural Information Processing Systems*, pages 1975–1983, 2016.

G. B. Folland. *A course in abstract harmonic analysis*. CRC press, 1994.

E. Fuselier. *Refined Error Estimates for Matrix-Valued Radial Basis Functions*. PhD thesis, Texas A&M University, 2006.

L. Górniewicz. *Topological fixed point theory of multivalued mappings*, volume 495. Springer, 1999.

G. Guennebaud, B. Jacob, et al. Eigen v3, 2010. URL `http://eigen.tuxfamily.org`.

M. Ha Quang, S. H. Kang, and T. M. Le. Image and video colorization using vector-valued reproducing kernel hilbert spaces. *Journal of Mathematical Imaging and Vision*, 37(1): 49–65, 2010.

R. Hamid, Y. Xiao, A. Gittens, and D. DeCoste. Compact random feature maps. In *ICML*, pages 19–27, 2014.

E. Jones, T. Oliphant, and P. Peterson. {SciPy}: open source scientific tools for {Python}. 2014.

H. Kadri, E. Duflos, P. Preux, S. Canu, and M. Davy. Nonlinear functional regression: a functional rkhs approach. In *JMLR Proc. of International Conference on Artificial Intelligence and Statistics*, volume 9, 2010.

H. Kadri, A. Rakotomamonjy, P. Preux, and F. R. Bach. Multiple operator-valued kernel learning. In *Advances in NIPS*, pages 2429–2437, 2012.

H. Kadri, M. Ghavamzadeh, and P. Preux. A generalized kernel approach to structured output learning. In *Proc. of the 30th International Conference on Machine Learning*, 2013.

H. Kadri, E. Duflos, P. Preux, S. Canu, A. Rakotomamonjy, and J. Audiffren. Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 16:1–54, 2015.

P. Kar and H. Karnick. Random feature maps for dot product kernels. In *AISTATS*, volume 22, pages 583–591, 2012.

V. Koltchinskii et al. A remark on low rank matrix recovery and noncommutative bernstein type inequalities. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 213–226. Institute of Mathematical Statistics, 2013.

A. J. Kurdila and M. Zabarankin. Convex functional analysis, 2006.

Q. V. Le, T. Sarlós, and A. J. Smola. Fastfood - computing hilbert space expansions in loglinear time. In *Proc. of ICML 2013, Atlanta, USA, 16-21 June 2013*, pages 244–252, 2013.

Y. LeCun, Y. Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

F. Li, C. Ionescu, and C. Sminchisescu. *Pattern Recognition: 32nd DAGM Symposium, Darmstadt, Germany, September 22-24, 2010. Proc.*, chapter Random Fourier Approximations for Skewed Multiplicative Histogram Kernels, pages 262–271. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

N. Lim, Y. Şenbabaoğlu, G. Michailidis, and F. d'Alché Buc. Okvar-boost: a novel boosting algorithm to infer nonlinear dynamics and interactions in gene regulatory networks. *Bioinformatics*, 29(11):1416–1423, 2013.

N. Lim, F. d'Alché-Buc, C. Auliac, and G. Michailidis. Operator-valued kernel-based vector autoregressive models for network inference. *Machine Learning*, 99(3):489–513, 2015a.

N. Lim, F. d'Alché Buc, C. Auliac, and G. Michailidis. Operator-valued kernel-based vector autoregressive models for network inference. *Machine learning*, 99(3):489–513, 2015b.

Z. Lu, A. May, K. Liu, A. B. Garakani, D. Guo, A. Bellet, L. Fan, M. Collins, B. Kingsbury, M. Picheny, et al. How to scale up kernel methods to be as good as deep neural nets. *arXiv preprint arXiv:1411.4000*, 2014.

Y. Macedo and R. Castro. Learning div-free and curl-free vector fields by matrix-valued kernels. Technical report, Preprint A 679/2010 IMPA, 2008.

L. Mackey, M. I. Jordan, R. Chen, B. Farrel, and J. Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42:3:906–945, 2014.

A. Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer, 2016.

C. A. Micchelli and M. Pontil. Kernels for multi–task learning. In *NIPS*, volume 86, page 89, 2004.

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

M. Micheli and J. Glaunes. Matrix-valued kernels for shape deformation analysis. Technical report, Arxiv report, 2013.

H. Q. Minh. Operator-valued bochner theorem, fourier feature maps for operator-valued kernels, and vector-valued learning. *arXiv preprint arXiv:1608.05639*, 2016.

H. Q. Minh, L. Bazzani, and V. Murino. A unifying framework for vector-valued manifold regularization and multi-view learning. In *Proc. of the 30th International Conference on Machine Learning*, 2013a.

H. Q. Minh, L. Bazzani, and V. Murino. A unifying framework for vector-valued manifold regularization and multi-view learning. In *ICML (2)*, pages 100–108, 2013b.

H. Q. Minh, L. Bazzani, and V. Murino. A unifying framework in vector-valued reproducing kernel hilbert spaces for manifold regularization and co-regularized multi-view learning. *Journal of Machine Learning Research*, 17(25):1–72, 2016.

S. Minsker. On some extensions of bernstein's inequality for self-adjoint operators. *arXiv preprint arXiv:1112.5448*, 2011.

Y. Mroueh, T. Poggio, L. Rosasco, and J.-j. Slotine. Multiclass learning with simplex coding. In *Advances in NIPS*, pages 2789–2797, 2012.

Y. Mukuta and T. Harada. Kernel approximation via empirical orthogonal decomposition for unsupervised feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5222–5230, 2016.

A. Naor. On the banach-space-valued azuma inequality and small-set isoperimetry of alon–roichman graphs. *Combinatorics, Probability and Computing*, 21(04):623–634, 2012.

K.-H. Neeb. Operator-valued positive definite kernels on tubes. *Monatshefte für Mathematik*, 126(2):125–160, 1998.

T. E. Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

R. I. Oliveira. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *arXiv preprint arXiv:0911.0600*, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

G. Pedrick. Theory of reproducing kernels for hilbert spaces of vector-valued functions. Technical report, University of Kansas, Department of Mathematics, 1957.

T. Penzl. Numerical solution of generalized lyapunov equations. *Advances in Computational Mathematics*, 8(1):33–48, 1998.

N. Pham and R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD*, pages 239–247. ACM, 2013.

I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NIPS 2007*, pages 1177–1184, 2007.

L. Rosasco, M. Belkin, and E. D. Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11(Feb):905–934, 2010.

A. Rudi, R. Camoriano, and L. Rosasco. Generalization properties of learning with random features. *arXiv preprint arXiv:1602.04474*, 2016.

M. Sangnier, O. Fercoq, and F. d'Alché Buc. Joint quantile regression in vector-valued rkhss. *Neural Information Processing Systems*, 2016.

E. Senkene and A. Tempel'man. Hilbert spaces of operator-valued functions. *Lithuanian Mathematical Journal*, 13(4):665–670, 1973.

V. Sindhwani, H. Q. Minh, and A. Lozano. Scalable matrix-valued kernel learning for high-dimensional nonlinear multivariate regression and granger causality. In *Proc. of UAI'13, Bellevue, WA, USA, August 11-15, 2013*. AUAI Press, Corvallis, Oregon, 2013.

G. L. Sleijpen, P. Sonneveld, and M. B. Van Gijzen. Bi-cgstab as an induced dimension reduction method. *Applied Numerical Mathematics*, 60(11):1100–1114, 2010.

S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

A. J. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural networks*, 11(4):637–649, 1998.

B. Sriperumbudur and Z. Szabo. Optimal rates for random fourier features. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in NIPS 28*, pages 1144–1152, 2015.

D. J. Sutherland and J. G. Schneider. On the error of random fourier features. In *Proc. of UAI 2015, July 12-16, 2015, Amsterdam, The Netherlands*, pages 862–871, 2015.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

J. A. Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

A. W. Van Der Vaart and J. A. Wellner. Weak convergence. In *Weak Convergence and Empirical Processes*, pages 16–28. Springer, 1996.

J.-P. Vert. Regularization of kernel methods by decreasing the bandwidth of the gaussian kernel.

G. Wahba. *Spline model for observational data*. Philadelphia, Society for Industrial and Applied Mathematics, 1990.

N. Wahlström, M. Kok, T. Schön, and F. Gustafsson. Modeling magnetic fields using gaussian processes. In *in Proc. of the 38th ICASSP*, 2013.

S. v. d. Walt, S. C. Colbert, and G. Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

B. Xie, Y. Liang, and L. Song. Scale up nonlinear component analysis with doubly stochastic gradients. In *Advances in Neural Information Processing Systems*, pages 2341–2349, 2015.

J. Yang, V. Sindhwani, H. Avron, and M. Mahoney. Quasi-monte carlo feature maps for shift-invariant kernels. In *Proceedings of The 31st International Conference on Machine Learning (ICML-14)*, pages 485–493, 2014a.

J. Yang, V. Sindhwani, Q. Fan, H. Avron, and M. W. Mahoney. Random laplace feature maps for semigroup kernels on histograms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 971–978, 2014b.

T. Yang, Y.-F. Li, M. Mahdavi, R. Jin, and Z. Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *NIPS 25*, pages 476–484, 2012.

Z. Yang, M. Moczulski, M. Denil, N. de Freitas, A. Smola, L. Song, and Z. Wang. Deep fried convnets. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1476–1483, 2015a.

Z. Yang, A. G. Wilson, A. J. Smola, and L. Song. A la carte - learning fast kernels. In *Proc. of AISTATS 2015, San Diego, California, USA, 2015*, 2015b.

H. Zhang, Y. Xu, and Q. Zhang. Refinement of operator-valued reproducing kernels. *Journal of Machine Learning Research*, 13:91–136, 2012.