DATA ARE NOT REAL!

ROMAIN BRAULT*

UNDER SUPERVISION OF:
Professor (Prof.) Florence d'Alché-Buc†

Large-scale learning on structured input-output data with operator-valued kernels

Engineer (Eng.)
Computer Science
IBISC
Université d'Évry val d'Essonne

Septembre 2016 – version 0.1

* Email: romain.brault@ibisc.fr
† Email: florence.dalche@telecom-paristech.fr

SUPERVISOR:
Professor (Prof.) Florence d'Alché-Buc

LOCATION:
15, Rue Plumet, 75015 - Paris, France

ABSTRACT

Short summary of the contents... a great guide by Kent Beck how to write good abstracts can be found here:

https://plg.uwaterloo.ca/~migod/research/beckOOPSLA.html

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

Put your publications from the thesis here. The packages `multibib` or `bibtopic` etc. can be used to handle multiple different bibliographies in your document.

We have seen that computer programming is an art,
because it applies accumulated knowledge to the world,
because it requires skill and ingenuity, and especially
because it produces objects of beauty.

ACKNOWLEDGEMENTS

Put your acknowledgements here.

Many thanks to everybody who already sent me a postcard!

Regarding the typography and other help, many thanks go to Marco Kuhlmann, Philipp Lehman, Lothar Schlesier, Jim Young, Lorenzo Pantieri and Enrico Gregorio[1], Jörg Sommer, Joachim Köstler, Daniel Gottschlag, Denis Aydin, Paride Legovini, Steffen Prochnow, Nicolas Repp, Hinrich Harms, Roland Winkler, and the whole LaTeX-community for support, ideas and some great software.

Regarding LyX: The LyX port was initially done by *Nicholas Mariette* in March 2009 and continued by *Ivo Pletikosić* in 2011. Thank you very much for your work and the contributions to the original style.

1 Members of GuIT (Gruppo Italiano Utilizzatori di TeX e LaTeX)

CONTENTS

LIST OF TABLES

ACRONYMS

OVK Operator-Valued Kernel.

ORFF Operator-valued Random Fourier Feature.

RKHS Reproducing Kernel Hilbert Space.

vv-RKHS vector-valued Reproducing Kernel Hilbert Space.

LCA Locally Compact Abelian.

FT Fourier transform.

IFT inverse Fourier transform.

Part I

INTRODUCTION

You can put some informational part preamble text here. Illo principalmente su nos. Non message *occidental* anglo-romanic da. Debitas effortio simplificate sia se, auxiliar summarios da que, se avantiate publicationes via. Pan in terra summarios, capital interlingua se que. Al via multo esser specimen, campo responder que da. Le usate medical addresses pro, europa origine sanctificate nos se.

1 MOTIVATIONS

2 BACKGROUND

2.1 NOTATIONS

The euclidean inner product in $\mathbb{R}^d$ is denoted $\langle\cdot,\cdot\rangle$ and the euclidean norm is denoted $\|\cdot\|$. The unit pure imaginary number $\sqrt{-1}$ is denoted i. $\mathcal{B}(\mathbb{R}^d)$ is the Borel σ-algebra on $\mathbb{R}^d$. If $\mathcal{X}$ and $\mathcal{Y}$ are two vector spaces, we denote by $\mathcal{F}(\mathcal{X};\mathcal{Y})$ the vector space of functions $f : \mathcal{X} \to \mathcal{Y}$ and $\mathcal{C}(\mathcal{X};\mathcal{Y}) \subset \mathcal{F}(\mathcal{X};\mathcal{Y})$ the subspace of continuous functions. If $\mathcal{H}$ is an Hilbert space we denote its scalar product by $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and its norm by $\|\cdot\|_{\mathcal{H}}$. We set $\mathcal{L}(\mathcal{H}) = \mathcal{L}(\mathcal{H};\mathcal{H})$ to be the space of linear operators from $\mathcal{H}$ to itself. If $W \in \mathcal{L}(\mathcal{H})$, Ker W denotes the nullspace, Im W the image and $W^* \in \mathcal{L}(\mathcal{H})$ the adjoint operator (transpose when W is a real matrix). All these notations are summarized in table 1.

2.2 ABOUT STATISTICAL LEARNING

2.3 ON LARGE-SCALE LEARNING

2.4 ELEMENTS OF ABSTRACT HARMONIC ANALYSIS

2.4.1 *Locally compact Abelian groups*

Definition 1. *Locally Compact Abelian group. A group $(\mathcal{X},\star)$ is said to be Locally Compact Abelian if it is a topological* commutative *group $\mathcal{X}$ for which every point has a compact neighborhood and is Hausdorff.*

Locally Compact Abelian (LCA) groups are central to the general definition of Fourier Transform which is related to the concept of Pontryagin duality [8]. Let $(\mathcal{X},\star)$ be a LCA 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* $\hat{\mathcal{X}}$. The dual group of an LCA group is an LCA group and 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 $\hat{\hat{\mathcal{X}}} \cong \mathcal{X}$. I.e. there is a canonical isomorphism between any LCA group and its double dual. To emphasize this duality the following notation is usually adopted: $\omega(x) = (x,\omega) = (\omega,x)$, where $x \in \mathcal{X}$, $\omega \in \hat{\mathcal{X}}$. Another important property involves the complex conjugate of the pairing which is defined as $\overline{(x,\omega)} = (x^{-1},\omega)$.

Table 1: Mathematical symbols used throughout the parper and their signification.

Symbol	Meaning
i	Unit pure imaginary number $\sqrt{-1}$.
e	Euler constant.
$\langle \cdot, \cdot \rangle$	Euclidean inner product.
$\lVert \cdot \rVert$	Euclidean norm.
$\mathcal{X}$	Input space ().
$\hat{\mathcal{X}}$	The Pontryagin dual of $\mathcal{X}$.
$\mathcal{Y}$	Output space (Hilbert space).
$\mathcal{H}$	Feature space (Hilbert space).
$\langle \cdot, \cdot \rangle_{\mathcal{Y}}$	The canonical inner product of the Hilbert space $\mathcal{Y}$.
$\lVert \cdot \rVert_{\mathcal{Y}}$	The canonical norm induced by the inner product of the Hilbert space $\mathcal{Y}$.
$\mathcal{F}(\mathcal{X};\mathcal{Y})$	Vector space of function from $\mathcal{X}$ to $\mathcal{Y}$.
$\mathcal{C}(\mathcal{X};\mathcal{Y})$	The vector subspace of $\mathcal{F}$ of continuous function from $\mathcal{X}$ to $\mathcal{Y}$.
$\mathcal{L}(\mathcal{H};\mathcal{Y})$	The set of bounded linear operator from a Hilbert space $\mathcal{H}$ to a Hilbert space $\mathcal{Y}$.
$\mathcal{L}(\mathcal{Y})$	The set of bounded linear operator from a Hilbert space $\mathcal{H}$ to itself.
$\mathcal{L}_+(\mathcal{Y})$	The set of non-negative bounded linear operator from a Hilbert space $\mathcal{H}$ to itself.
$\mathcal{B}(\mathcal{X})$	Borel σ-algebra on $\mathcal{X}$.
$\mu(\mathcal{X})$	A scalar positive measure of $\mathcal{X}$.
$p_\mu(x)$	The Radon-Nikodym derivative of μ w. r. t. the Lebesgue measure.
dx, $d\omega$	The canonical Haar measure of the LCA group $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. (resp. $(\hat{\mathcal{X}}, \mathcal{B}(\hat{\mathcal{X}}))$.
$L^p(\mathcal{X}, dx)$	The Banach space of $\lvert\cdot\rvert^p$-integrable function from $(\mathcal{X}, \mathcal{B}(\mathcal{X}, dx))$ to $\mathbb{C}$.
$L^p(\mathcal{X}, dx; \mathcal{Y})$	The Banach space of $\lVert\cdot\rVert_{\mathcal{Y}^p}$ (Bochner)-integrable function from $(\mathcal{X}, \mathcal{B}(\mathcal{X}), dx)$ to $\mathcal{Y}$.

We notice that for any pairing depending of ω, there exists a function $h_\omega : \mathcal{X} \to \mathbb{R}$ such that: $(x, \omega) = \exp(-ih_\omega(x))$ since any pairing maps into $\mathbb{U}$. Moreover,

$$\begin{aligned}
(x \star z^{-1}, \omega) = \omega(x)\omega(z^{-1}) &= \exp(-ih_\omega(x)) \exp(-ih_\omega(z^{-1})) \\
&= \exp(-ih_\omega(x)) \exp(+ih_\omega(z)).
\end{aligned}$$

Table 2: Classification of Fourier transforms in terms of their domain and transform domain.

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

Table 2 provide an explicit list of pairings for various groups based on $\mathbb{R}^d$ or its subsets. We especially mention the duality pairing associated to the skewed multiplicative LCA group $\mathcal{X} = ((-c; +\infty)^d, \odot)$ (x_k+c)(z_k+c) - c, Hence $h_\omega(x) = \sum_{k=1}^d \omega_k \log(x_k + c)$. This group together with the operation $\odot$ has been proposed by [10] to handle histograms features especially useful in image recognition applications.

2.4.2 *The Fourier transform*

For a function with values in a separable Hilbert space $f \in L^1(\mathcal{X}, dx; \mathcal{Y})$, where dx is the Haar measure on $\mathcal{X}$, we denote $\mathcal{F}[f]$ its Fourier transform (FT) which is defined by

$$\forall \omega \in \hat{\mathcal{X}}, \quad \mathcal{F}[f](\omega) = \int_{\hat{\mathcal{X}}} \overline{(x, \omega)} f(x) dx.$$

For a measure defined on $\mathcal{X}$, there exists a unique suitably normalized measure $d\omega$ on $\hat{\mathcal{X}}$ such that $\forall f \in L^1(\mathcal{X}, dx; \mathcal{Y})$ and if $\mathcal{F}[f] \in L^1(\hat{\mathcal{X}}, d\omega, \mathcal{Y})$ we have

$$\forall x \in \mathcal{X}, \quad f(x) = \int_{\hat{\mathcal{X}}} \mathcal{F}[f](\omega)(x, \omega) d\omega. \tag{1}$$

Moreover if $d\omega$ is normalized, $\mathcal{F}$ extends to a unitary operator from $L^2(\mathcal{X}, dx, \mathcal{Y})$ onto $L^2(\hat{\mathcal{X}}, d\omega, \mathcal{Y})$ Then the inverse Fourier transform (IFT) of a function $g \in L^1(\hat{\mathcal{X}}, d\omega, \mathcal{Y})$ (where $d\omega$ is a Haar measure on $\hat{\mathcal{X}}$ suitably normalize w. r. t. the Haar measure dx) is noted $\mathcal{F}^{-1}[g]$ defined by

$$\forall x \in \mathcal{X}, \quad \mathcal{F}^{-1}[g](x) = \int_{\hat{\mathcal{X}}} (x, \omega) g(\omega) d\omega,$$

Section 2.4.1 gives some examples of real Abelian groups with their associated dual and pairing. The interested reader can refer to Folland [8] for a more detailed construction of LCA, Pontryagin duality and Fourier transforms on LCA. For the familiar case of a scalar-valued function f on the LCA group $(\mathbb{R}^d, +)$, we have:

$$\forall \omega \in \hat{\mathcal{X}}, \quad \mathcal{F}[f](\omega) = \int_{\mathbb{R}^d} e^{-i\langle \omega, x - z\rangle} f(x) dx, \tag{2}$$

the Haar measure being here the Lebesgue measure.

2.5 ON OPERATOR-VALUED KERNELS

We now introduce the theory of vector-valued Reproducing Kernel Hilbert Space (vv-RKHS) that provides a flexible framework to study and learn vector-valued functions.

2.5.1 *Definitions and properties*

An operator-valued kernel is defined here as a $\mathcal{Y}$-reproducing kernel Carmeli et al. [5].

Definition 2. *Given $\mathcal{X}$, a Polish space and $\mathcal{Y}$, a Hilbert Space, a map $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ is called a $\mathcal{Y}$-reproducing kernel if*

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

for all $x_1, \dots, x_N$ in $\mathcal{X}$, all $y_1, \dots, y_N$ in $\mathcal{Y}$ and $N \geqslant 1$. Given $x \in \mathcal{X}$, $K_x : \mathcal{Y} \to \mathcal{F}(\mathcal{X}; \mathcal{Y})$ denotes the linear operator whose action on a vector y is the function $K_x y \in \mathcal{F}(\mathcal{X}; \mathcal{Y})$ defined by $(K_x y)(z) = K(z, x)y$, for all $z \in \mathcal{X}$.

Additionally, given a $\mathcal{Y}$-reproducing kernel K, there is a unique Hilbert space $\mathcal{H}_K \subset \mathcal{F}(\mathcal{X}; \mathcal{Y})$ satisfying $K_x \in \mathcal{L}(\mathcal{Y}; \mathcal{H}_K)$, for all $x \in \mathcal{X}$ and $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}_K,\ f(x) = K_x^* f$, where $K_x^* : \mathcal{H}_K \to \mathcal{Y}$ is the adjoint of K_x. The space $\mathcal{H}_K$ is called the *vector-valued Reproducing Kernel Hilbert Space* associated with K. The corresponding product and norm are denoted by $\langle .,. \rangle_K$ and $\|.\|_K$, respectively. As a consequence [5] we have:

$$\begin{aligned} K(x, z) &= K_x^* K_z \quad \forall x, z \in \mathcal{X}, \\ \mathcal{H}_K &= \overline{\operatorname{span}} \{ K_x y \mid \forall x \in \mathcal{X}, \ \ \forall y \in \mathcal{Y} \}. \end{aligned}$$

Another way to describe functions of $\mathcal{H}_K$ consists in using a suitable feature map.

Proposition 3 (Feature Operator Carmeli et al. [5])**.** *Let $\mathcal{H}$ be a 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 vv-RKHS $\mathcal{H}_K$ with reproducing kernel*

$$K(x, z) = \phi_x^* \phi_z, \ \ \forall x, z \in \mathcal{X}.$$

$W^* W$ is the orthogonal projection onto

$$\mathit{Ker}\ W^{\perp} = \overline{\operatorname{span}} \{ \phi_x y \mid \forall x \in \mathcal{X}, \ \ \forall y \in \mathcal{Y} \}.$$

Then $\|f\|_K = \inf\{ \|g\|_{\mathcal{H}} \mid \forall g \in \mathcal{H}, \ \ Wg = f \}$.

We call ϕ a *feature map*, W a *feature operator* and $\mathcal{H}$ a *feature space*.

2.5.2 *Examples of operator-valued kernels*

Operator-valued kernels 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.

Definition 4 (Decomposable kernel)**.** *Let A be a positive semi-definite operator of $\mathcal{L}(\mathcal{Y})$. K is said to be a $\mathcal{Y}$-Mercer* decomposable *kernel if for all $(x, z) \in \mathcal{X}^2$,*

Some authors also refer to as separable *kernels.*

$$K(x, z) = k(x, z)A,$$

where k is a scalar *Mercer kernel.*

When $\mathcal{Y} = \mathbb{R}^p$, the matrix A is 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 Álvarez, Rosasco, and Lawrence [1], Baldassarre et al. [2], and Evgeniou, Micchelli, and Pontil [7] that A 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, A can be set to the empirical covariance of the output training data or learned with one of the algorithms proposed in the literature [6, 11, 14]. 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 vv-RKHS $\mathcal{H}_K$ is dense in the space $\mathcal{C}(\mathcal{X}, \mathcal{Y})$. The conditions for a kernel to be universal have been discussed in Caponnetto et al. [4] and Carmeli et al. [5]. In particular they show that a decomposable kernel is universal provided that the scalar kernel k is universal and the operator A is injective.

Curl-free and divergence-free kernels provide an interesting application of operator-valued kernels [3, 12, 13] to *vector field* learning, for which input and output spaces have the same dimensions ($d = p$). Applications cover shape deformation analysis [13] and magnetic fields approximations [15]. These kernels discussed in [9] allow encoding input-dependent similarities between vector-fields.

Definition 5 (Curl-free and Div-free kernel)**.** *Assume $\mathcal{X} = (\mathbb{R}^d, +)$ and $\mathcal{Y} = \mathbb{R}^p$ with $d = p$. The* divergence-free *kernel is defined as*

$$K^{div}(x, z) = K_0^{div}(\delta) = (\nabla \nabla^T - \acute{} I) k_0(\delta)$$

and the curl-free *kernel as*

$$K^{curl}(x, z) = K_0^{curl}(\delta) = -\nabla\nabla^T k_0(\delta),$$

where $\nabla\nabla^T$ *is the Hessian operator and ´ is the Laplacian operator.*

Although taken separately these kernels are not universal, a convex combination of the curl-free and divergence-free kernels allows to learn any vector field that satisfies the Helmholtz decomposition theorem [3, 12].

Part II

CONTRIBUTIONS

You can put some informational part preamble text here. Illo principalmente su nos. Non message *occidental* anglo-romanic da. Debitas effortio simplificate sia se, auxiliar summarios da que, se avantiate publicationes via. Pan in terra summarios, capital interlingua se que. Al via multo esser specimen, campo responder que da. Le usate medical addresses pro, europa origine sanctificate nos se.

3 OPERATOR-VALUED RANDOM FOURIER FEATURES

3.1 MOTIVATIONS

Random Fourier Features have been proved useful to implement efficiently kernel methods in the scalar case, allowing to learn a linear model based on an approximated feature map. In this work, we are interested to construct approximated operator-valued feature maps to learn vector-valued functions. With an explicit (approximated) feature map, one converts the problem of learning a function f in the vector-valued Reproducing Kernel Hilbert Space $\mathcal{H}_K$ into the learning of a linear model $\tilde{f}$ defined by:

$$\tilde{f}(x) = \tilde{\phi}(x)^*\theta,$$

where $\phi : \mathcal{X} \to \mathcal{L}(\mathcal{H}, \mathcal{Y})$ and $\theta \in \mathcal{H}$. The methodology we propose works for operator-valued kernels defined on any Locally Compact Abelian (LCA) group, noted $(\mathcal{X}, \star)$, for some operation noted $\star$. This allows us to use the general context of Pontryagin duality for Fourier transform of functions on LCA groups. Building upon a generalization of Bochner's theorem for operator-valued measures, an operator-valued kernel is seen as the *Fourier transform* of an operator-valued positive measure. From that result, we extend the principle of Random Fourier Feature for scalar-valued kernels and derive a general methodology to build Operator Random Fourier Feature when operator-valued kernels are shift-invariant according to the chosen group operation.

3.2 CONSTRUCTION

We present a construction of Operator-valued Random Fourier Feature (ORFF) such that $f : x \mapsto \tilde{\phi}(x)^*\theta$ is a continuous function that maps an arbitrary LCA 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 $\tilde{K}(x, z) = \tilde{\phi}(x)^*\tilde{\phi}(z)$ with high probability on *compact* subsets of the LCA $\mathcal{X}$, when $\mathcal{Y}$ is *finite dimensional*. Eventually we conclude with some numerical experiments.

3.2.1 *Theoretical study*

The following proposition of Carmeli et al. [5] and Zhang, Xu, and Zhang [16] extends Bochner's theorem to any shift-invariant $\mathcal{Y}$-Mercer kernel.

Proposition 6 (Operator-valued Bochner's theorem [16])**.** *If a continuous 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 operator-valued measure* $M : \mathcal{B}(\mathcal{X}) \to \mathcal{L}_+(\mathcal{Y})$ *such that for all* $x, z \in \mathcal{X}$*,*

$$K(x, z) = \int_{\hat{\mathcal{X}}} \overline{(x \star z^{-1}, \omega)} dM(\omega), \tag{3}$$

where M *belongs to the set of all the* $\mathcal{L}_+(\mathcal{Y})$*-valued measures of bounded variation on the* σ*-algebra of Borel subsets of* $\hat{\mathcal{X}}$*. Conversely, from any positive operator-valued measure* M*, a shift-invariant kernel* K *can be defined by theorem 6.*

Although this theorem is central to the spectral decomposition of shift-invariant $\mathcal{Y}$-Mercer OVK, the following results proved by Carmeli et al. [5] provides insights about this decomposition that are more relevant in practise. It first shows how to build shift-invariant $\mathcal{Y}$-Mercer kernel but more importantly, also states that any operator-valued spectral decomposition of such OVKs when $\mathcal{Y}$ is finite dimensional or $\mathcal{X}$ is compact can be written using a pair (A, μ) where A is an operator-valued function on $\hat{\mathcal{X}}$ and μ is a real-valued positive measure on $\hat{\mathcal{X}}$. Note that obviously such a pair is not unique ad the choice of this paper may have an impact on theoretical properties as well as practical computations.

Proposition 7 (Carmeli et al. [5])**.** *Let* μ *be a positive measure on* $\mathcal{B}(\hat{\mathcal{X}})$ *and* $A : \hat{\mathcal{X}} \to \mathcal{L}(\mathcal{Y})$ *such that* $\langle A(.)y, y' \rangle \in L^1(\mathcal{X}, d\mu)$ *for all* $y, y' \in \mathcal{Y}$ *and* $A(\omega) \succcurlyeq 0$ *for* μ*-almost all* ω*. Then, for all* $\delta \in \mathcal{X}$ *and for all* $y, y' \in \mathcal{Y}$*,*

$$\langle y, K_e(\delta)y \rangle_{\mathcal{Y}} = \int_{\hat{\mathcal{X}}} \overline{(\delta, \omega)} \langle y, A(\omega)y' \rangle_{\mathcal{Y}} d\mu(\omega) \tag{4}$$

is the kernel signature of a shift-invariant $\mathcal{Y}$*-Mercer kernel* K *such that* $K(x, z) = K_e(x \star z^{-1})$*, where the integral converges in the weak operator topology. If* $\mathcal{Y}$ *is finite dimensional or* $\mathcal{X}$ *is compact, any shift-invariant kernel is of the above form for some pair* $(A(\omega), \mu(\omega))$*.*

This theorem is more interesting than theorem 6 in the sense that it shows that we are certain of the existence of a *scalar* measure μ and a positive operator $A(\omega)$, provided that $\mathcal{X}$ is compact or $\mathcal{Y}$ is finite dimensional. When $p = 1$ one can always assume A is reduced to the scalar 1, μ is still a bounded positive measure and we retrieve the Bochner theorem applied to the scalar case (**??**).

3.2.2 *Functional Fourier feature map*

Let us introduce a functional feature map, we call here *Fourier Feature map*, defined by the following proposition as a direct consequence of theorem 7.

Proposition 8 (Fourier feature map)**.** *If there exist an operator-valued function* $B : \hat{\mathcal{X}} \to \mathcal{L}(\mathcal{Y}, \mathcal{Y}')$ *such that for all* $y, y' \in \mathcal{Y}$*,* $\langle y, B(\omega)B(\omega)^* y' \rangle =$

$\langle y, A(\omega)y'\rangle$ μ-almost everywhere and $\langle y, A(\omega)y'\rangle \in L^1(\hat{\mathcal{X}}, d\mu)$ then the operator ϕ_x defined for all y in $\mathcal{Y}$ by

$$(\phi_x y)(\omega) = (x, \omega)B(\omega)^* y, \tag{5}$$

is a feature map[1] *of some shift-invariant kernel* K.

[1] *I. e. it satisfies for all $x, z \in \mathcal{X}$, $\phi_x^* \phi_z = K(x, z)$ where K is an operator-valued kernel.*

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

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

which defines a $\mathcal{Y}$-Mercer according to theorem 7 of Carmeli et al. [5]. With this notation notice that $\phi : \mathcal{X} \to \mathcal{L}(\mathcal{Y}; L^2(\hat{\mathcal{X}}, \mu; \mathcal{Y}'))$ such that $\phi_x \in \mathcal{L}(\mathcal{Y}; L^2(\hat{\mathcal{X}}, \mu; \mathcal{Y}'))$ where $\phi_x := \phi(x)$. □

3.2.3 *Sufficient conditions of existence*

While theorem 7 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 $A(\omega), \mu(\omega)$ from the kernel signature. Proposition 14 in Carmeli et al. [5] gives the solution, and also provide a sufficient condition for theorem 7 to apply.

Proposition 9 (Carmeli et al. [5])**.** *Let K be a shift-invariant $\mathcal{Y}$-Mercer kernel. Suppose that $\forall z \in \mathcal{X}$ and $\forall y, y' \in \mathcal{Y}$, $\langle K_e(.)y, y'\rangle \in L^1(\mathcal{X}, dx)$ where dx denotes the Haar measure on $(\mathcal{X}, \star)$. Define C such that for all $\omega \in \hat{\mathcal{X}}$ and for all y, y' in $\mathcal{Y}$,*

$$\begin{aligned}\langle y, C(\omega)y'\rangle &= \int_{\mathcal{X}} (\delta, \omega)\langle y, K_e(\delta)y'\rangle d\delta \\ &= \mathcal{F}^{-1}\left[\langle y, K_e(\cdot)y'\rangle\right](\omega)\end{aligned} \tag{6}$$

Then

i) $C(\omega)$ *is a bounded non-negative operator for all* $\omega \in \hat{\mathcal{X}}$,

ii) $\langle y, C(.)y'\rangle \in L^1(\hat{\mathcal{X}}, d\omega)$ *for all* $y, y' \in \mathcal{X}$,

iii) for all $\delta \in \mathcal{X}$ *and for all* y, y' *in* $\mathcal{Y}$,

$$\langle y, K_e(\delta)y'\rangle = \int_{\hat{\mathcal{X}}} \overline{(\delta, \omega)} \langle y, C(\omega)y'\rangle d\omega.$$

Gathering the two propositions, we present now the following property that allows to build a spectral decomposition of a shift-invariant $\mathcal{Y}$-Mercer kernel on a LCA group $(\mathcal{X}, \star)$.

Proposition 10 (Sufficient condition for shift-invariant $\mathcal{Y}$-Mercer kernel spectral decomposition)**.** *Let* K_e *be the signature of a shift-invariant* $\mathcal{Y}$*-Mercer kernel on* $(\mathcal{X}, \star)$ *and suppose that for all* $y, y' \in \mathcal{Y}$, $\langle K_e(.)y, y' \rangle \in L^1(\mathcal{X}, dx)$.

If $\mathcal{Y}$ *is of finite dimension or* $\mathcal{X}$ *is compact, then there exists* μ*, a positive measure on* $\mathcal{B}(\hat{\mathcal{X}})$*, and* $A : \hat{\mathcal{X}} \to \mathcal{L}_+(\mathcal{Y})$*, a operator-valued functions such that for all* $y, y' \in \mathcal{Y}$ $\langle A(.)y, y' \rangle \in L^1(\mathcal{X}, d\mu)$ *and for all* y, y' *in* $\mathcal{Y}$,

$$\langle y, K_e(\delta)y' \rangle = \int_{\hat{\mathcal{X}}} \overline{(\delta, \omega)} \langle y, A(\omega)y' \rangle p_\mu(\omega) d\omega.$$

where $\langle y, A(\omega)y' \rangle p_\mu(\omega) = \mathcal{F}^{-1}\left[\langle y, K_e(x \star z^{-1})y' \rangle\right]$.

Proof. From theorem 7 and theorem 9, if $\mathcal{X}$ is compact or $\mathcal{Y}$ is finite dimensional, we can write the following equality concerning the OVK signature K_e. For all $\delta \in \mathcal{X}$ and for all y, y' in $\mathcal{Y}$

$$\int_{\hat{\mathcal{X}}} \overline{(\delta, \omega)} \langle y, C(\omega)y' \rangle d\omega = \int_{\hat{\mathcal{X}}} \overline{(\delta, \omega)} \langle y, A(\omega)y' \rangle d\mu(\omega).$$

Since both sides of the equation define continuous functions, the following equality holds μ-almost everywhere. For all $\omega \in \hat{\mathcal{X}}$ and for all $y, y' \in \mathcal{Y}$,

$$\langle y, C(\omega)y' \rangle = \langle y, A(\omega)y' \rangle p_\mu(\omega) = \mathcal{F}^{-1}\left[\langle y, K_e(\cdot)y' \rangle\right](\omega), \tag{7}$$

where $p_\mu(\omega) = \frac{d\mu}{d\omega}$ is the Radon-Nikodym derivative of the measure μ, i. e. its density. □

In the case where $\mathcal{Y} = \mathbb{R}^p$, we rewrite section 3.2.3 coefficient-wisely by choosing the orthonormal basis of $\mathcal{Y}$, $(e_1, \dots, e_p)$, such that for all $i, j \in \{1, \dots, p\}$,

$$\langle e_i, C(\omega)e_j \rangle = C(\omega)_{ij} = A(\omega)_{ij} p_\mu(\omega) = \mathcal{F}^{-1}\left[K_e(\delta)_{ij}\right]. \tag{8}$$

It follows that for all $i, j \in \{1, \dots, p\}$,

$$K_e(x \star z^{-1})_{ij} = \mathcal{F}\left[A(\cdot)_{ij}\right] \tag{9}$$

Remark 11. *Note that although the inverse Fourier transform of* K_e *yields a unique operator-valued function* $C(\cdot)$*, the decomposition of* $C(\omega)$ *into* $A(\omega)p_\mu(\omega)$ *is not unique. The choice of the decomposition may be justified by the computational cost or by the nature of the constants involved in the uniform convergence of the estimator.*

3.2.4 *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 Fourier transform of a target function $f \in \mathcal{H}_K$. In other words, what is the norm in $L^2(\hat{\mathcal{X}}, d\omega, \mathcal{Y}')$ induced by $\|\cdot\|_K$?

Proposition 12. *Let* K *be a shift-invariant* $\mathcal{Y}$*-Mercer Kernel such that for all* y, y' *in* $\mathcal{Y}$*,* $\langle y, K_e(\cdot)y'\rangle \in L^1(\mathcal{X}, dx)$ *and* $\langle y, A(\omega)y'\rangle p_\mu(\omega) \coloneqq \mathcal{F}^{-1}\left[\langle y, K_e(\cdot)y'\rangle\right](\omega)$*. Finally let* $f \in \mathcal{H}_K$*. Then*

$$\|f\|_K^2 = \int_{\hat{\mathcal{X}}} \frac{\langle \mathcal{F}[f](\omega), A(\omega)^\dagger \mathcal{F}[f](\omega)\rangle_{\mathcal{Y}}}{p_\mu(\omega)} d\omega. \tag{10}$$

Proof. We first show how the Fourier transform relates to the feature operator. Since $\mathcal{Y}$ is embed into $L^2(\hat{\mathcal{X}}, \mu, \mathcal{Y})$ by mean of the feature operator W, we have:

$$\begin{aligned}\mathcal{F}\left[\mathcal{F}^{-1}[g]\right](x) &= \int_{\hat{\mathcal{X}}} \overline{(x,\omega)}\mathcal{F}^{-1}[g](\omega)d\omega = g(x)\\ (Wg)(x) &= \int_{\hat{\mathcal{X}}} \overline{(x,\omega)}p_\mu(\omega)B(\omega)g(\omega)d\omega = g(x).\end{aligned}$$

Hence, $\mathcal{F}^{-1}[f](\omega) = p_\mu(\omega)B(\omega)g(\omega)$ μ-almost everywhere. From theorem 3 we have

$$\begin{aligned}&\|f\|_K^2\\ &\quad= \inf\left\{ \|g\|_{\mathcal{H}}^2 \;\middle|\; \forall g \in \mathcal{H},\; Wg = f \right\}\\ &= \inf\left\{ \int_{\hat{\mathcal{X}}} \|g(\omega)\|_{\mathcal{Y}}^2 d\mu(\omega) \;\middle|\; \forall g \in \mathcal{H},\; \mathcal{F}^{-1}[f](\omega) = p_\mu(\omega)B(\omega)g(\omega) \right\}.\end{aligned}$$

The pseudo inverse of the operator $B(\omega)$ (noted $B(\omega)^\dagger$) is the unique solution of the system $\mathcal{F}^{-1}[f](\omega) = p_\mu(\omega)B(\omega)g(\omega)$ w.r.t. to $g(\omega)$ with minimal norm. Eventually,

$$\begin{aligned}\|f\|_K^2 &= \int_{\hat{\mathcal{X}}} \frac{\left\|B(\omega)^\dagger \mathcal{F}^{-1}[f](\omega)\right\|_{\mathcal{Y}}^2}{p_\mu(\omega)^2} d\mu(\omega)\\ &= \int_{\hat{\mathcal{X}}} \frac{\left\|B(\omega)^\dagger \mathcal{F}[f](\omega)\right\|_{\mathcal{Y}}^2}{p_\mu(\omega)^2} d\mu(\omega)\end{aligned} \tag{11}$$

Conclude the proof by taking $d\mu(\omega) = p_\mu(\omega)d\omega$. □

Note that if $K(x,z) = k(x,z)$ is a scalar kernel then for all ω in $\hat{\mathcal{X}}$, $A(\omega) = 1$. Therefore we recover a well known results for kernels that is for any $f \in \mathcal{H}_k$ we have $\|f\|_k = \int_{\hat{\mathcal{X}}} \mathcal{F}[k_e](\omega)^{-1}\mathcal{F}[f](\omega)^2 d\omega$. 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 $A(\omega) = B(\omega)B(\omega)^*$.

3.2.5 *Building Operator-valued Random Fourier Features*

Without loss of generality we assume that $\int_{\mathcal{X}} d\mu(\omega) = 1$ and thus, μ is a probability distribution and p_μ, a probability density. Note that this is always possible through an appropriate rescaling of the kernel.

Given a $\mathcal{Y}$-Mercer shift-invariant kernel K on $\mathcal{X}$, an approximation of K can be obtained using a decomposition (A, μ) and a plug-in Monte-Carlo estimator instead of the expectation. However, for efficient computations, as motivated in the introduction, we are interested in finding an approximated feature map more than a kernel approximation. Indeed, an approximated feature map will allow to build linear models in regression tasks. The following proposition provides the general form of an Operator Random Fourier Feature map:

Proposition 13. *If one can find* $B : \hat{\mathcal{X}} \to \mathcal{L}(Y)$, *such that for all* y, $y' \in \mathcal{Y}$, $\langle y, A(\omega)y'\rangle = \langle y, B(\omega)B(\omega)^* y'\rangle \in L^1(\hat{\mathcal{X}}, d\mu)$, *then the operator-valued function*

$$\tilde{\phi}(x) = \frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} (x, \omega_j) B(\omega_j)^*, \qquad \omega_j \sim \mu \tag{12}$$

is an approximated feature map of kernel K

Proof. Let $\omega_1, \dots, \omega_D$ be D i. i. d random vectors following the law μ. For all $(x, z) \in \mathcal{X}^2$,

$$\begin{aligned} \tilde{\phi}(x)^* \tilde{\phi}(z) &= \left(\frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} \exp\left(i h_{\omega_j}(x) B(\omega_j)^* \right) \right)^* \\ &\qquad \left(\frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} \exp\left(i h_{\omega_j}(z) B(\omega_j)^* \right) \right) \\ &= \frac{1}{D} \sum_{j=1}^{D} \exp\left(-i(h_{\omega_j}(x) - h_{\omega_j}(z)\right) A(\omega_j) \\ &= \frac{1}{D} \sum_{j=1}^{D} \overline{(x, z)} A(\omega_j) \end{aligned}$$

From the strong law of large numbers, $\frac{1}{D} \sum_{j=1}^{D} \overline{(x, z)} A(\omega_j)$ converges almost-surely in the weak operator topology to $\mathbf{E}_\mu[\overline{(x \star z^{-1}, \omega_j)} A(\omega)]$ when D tends to infinity. □

Remark 14. *We find a decomposition such that for all* $j = 1, \dots, D$, $A(\omega_j) = B(\omega_j)B(\omega_j)^*$ *either by exhibiting an analytic closed-form or using a numerical decomposition.*

This proposition leads to the following construction algorithm.

Algorithm 1: Construction of ORFF

Input : $K(x,z) = K_e(\delta)$ a $\mathcal{Y}$-shift-invariant Mercer kernel such that $\forall y, y' \in \mathcal{Y}$, $\langle y, K_e(\delta) y' \rangle \in L^1(\mathbb{R}^d, dx)$.

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

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

2 Find a decomposition $(B(\omega), p_\mu(\omega))$ such that $B(\omega)B(\omega)^* p_\mu(\omega) = \mathcal{F}^{-1}\left[K_e\right](\omega)$;

3 Draw D random vectors ω_j, $j = 1, \dots, D$ from the probability distribution μ;

4 **return** $\tilde{\phi}(x) = \frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} (x, \omega_j) B(\omega_j)^*$;

3.3 UNIFORM BOUND ON THE APPROXIMATION

3.4 LEARNING WITH OPERATOR-VALUED RANDOM-FOURIER FEATURES

3.5 CONSISTENCY AND GENERALIZATION BOUNDS

3.6 CONCLUSIONS

4 CONCURRENT METHODS

4.1 BACKGROUND

4.2 THE NYSTRÖM METHOD

4.3 SUB-SAMPLING THE DATA

4.4 CONCLUSIONS

Part III

FINAL WORDS

You can put some informational part preamble text here. Illo principalmente su nos. Non message *occidental* anglo-romanic da. Debitas effortio simplificate sia se, auxiliar summarios da que, se avantiate publicationes via. Pan in terra summarios, capital interlingua se que. Al via multo esser specimen, campo responder que da. Le usate medical addresses pro, europa origine sanctificate nos se.

5 CONLUSIONS

Part IV

APPENDIX

A

OPERATOR-VALUED FUNCTIONS AND INTEGRATION

B PROOFS OF THEOREMS

BIBLIOGRAPHY

[1] M. A. Álvarez, L. Rosasco, and N. D. Lawrence. "Kernels for vector-valued functions: a review." In: *Foundations and Trends in Machine Learning* 4.3 (2012), pp. 195–266.

[2] L. Baldassarre, L. Rosasco, A. Barla, and A. Verri. "Vector Field Learning via Spectral Filtering." In: *ECML/PKDD*. Ed. by J. Balcazar, F. Bonchi, A. Gionis, and M. Sebag. Vol. 6321. LNCS. Springer Berlin / Heidelberg, 2010, pp. 56–71.

[3] L. Baldassarre, L. Rosasco, A. Barla, and A. Verri. "Multi-output learning via spectral filtering." In: *Machine Learning* 87.3 (2012), pp. 259–301.

[4] A. Caponnetto, C. A. Micchelli, M., and Y. Ying. "Universal MultiTask Kernels." In: *Journal of Machine Learning Research* 9 (2008), pp. 1615–1646.

[5] C. Carmeli, E. De Vito, A. Toigo, and V. Umanità. "Vector valued reproducing kernel Hilbert spaces and universality." In: *Analysis and Applications* 8 (2010), pp. 19–61.

[6] F. Dinuzzo, C.S. Ong, P. Gehler, and G. Pillonetto. "Learning Output Kernels with Block Coordinate Descent." In: *Proc. of the 28th Int. Conf. on Machine Learning*. 2011.

[7] T. Evgeniou, C. A. Micchelli, and M. Pontil. "Learning Multiple Tasks with kernel methods." In: *JMLR* 6 (2005), pp. 615–637.

[8] Gerald B Folland. *A course in abstract harmonic analysis*. CRC press, 1994.

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

[10] F. Li, C. Ionescu, and C. Sminchisescu. "Pattern Recognition: 32nd DAGM Symposium, Darmstadt, Germany, September 22-24, 2010. Proc." In: ed. by M. Goesele, S. Roth, A. Kuijper, B. Schiele, and K. Schindler. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. Chap. Random Fourier Approximations for Skewed Multiplicative Histogram Kernels, pp. 262–271. ISBN: 978-3-642-15986-2. DOI: 10.1007/978-3-642-15986-2_27. URL: http://dx.doi.org/10.1007/978-3-642-15986-2_27.

[11] N. Lim, F. d'Alché-Buc, C. Auliac, and G. Michailidis. "Operator-valued kernel-based vector autoregressive models for network inference." In: *Machine Learning* 99.3 (2015), pp. 489–513.

[12] Y. Macedo and R. Castro. *Learning Div-Free and Curl-Free Vector Fields by Matrix-Valued Kernels*. Tech. rep. Preprint A 679/2010 IMPA, 2008.

[13] M. Micheli and J. Glaunes. *Matrix-valued kernels for shape deformation analysis*. Tech. rep. Arxiv report, 2013.

[14] V. Sindhwani, H. Q. Minh, and A.C. 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.

[15] N. Wahlström, M. Kok, T.B. Schön, and Fredrik Gustafsson. “Modeling magnetic fields using Gaussian processes.” In: *in Proc. of the 38th ICASSP*. 2013.

[16] Haizhang Zhang, Yuesheng Xu, and Qinghui Zhang. “Refinement of Operator-valued Reproducing Kernels.” In: *Journal of Machine Learning Research* 13 (2012), pp. 91–136.

DECLARATION

Put your declaration here.

15, Rue Plumet, 75015 - Paris, France, Septembre 2016

Romain Brault

colophon

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

https://bitbucket.org/amiede/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

Final Version as of October 17, 2016 (`classicthesis` version 0.1).