JMLR: Workshop and Conference Proceedings 63:110–125, 2016 ACML 2016

Random Fourier Features For Operator-Valued Kernels

Romain Brault ROMAIN.BRAULT@TELECOM-PARISTECH.FR
IBISC, Université d'Évry val d'Essonne
LTCI, CNRS, Télécom ParisTech
46 rue Barrault, Paris, 75684 cedex 13, France

Markus Heinonen MARKUS.O.HEINONEN@AALTO.FI
Department of Information and Computer Science, Aalto University
FI-00076 Aalto, PO Box 15400, Finland

Florence d'Alché-Buc FLORENCE.DALCHE@TELECOM-PARISTECH.FR
LTCI, CNRS, Télécom ParisTech
Université Paris-Saclay
46 rue Barrault, Paris, 75684 cedex 13, France

Editors: Robert J. Durrant and Kee-Eung Kim

Abstract

To scale up operator-valued kernel-based regression devoted to multi-task and structured output learning, we extend the celebrated Random Fourier Feature methodology to get an approximation of operator-valued kernels. We propose a general principle for Operator-valued Random Fourier Feature construction relying on a generalization of Bochner's theorem for shift-invariant operator-valued Mercer kernels. We prove the uniform convergence of the kernel approximation for bounded and unbounded operator random Fourier features using appropriate Bernstein matrix concentration inequality. Numerical experiments show the quality of the approximation and the efficiency of the corresponding linear models on multiclass and regression problems.

Keywords: Operator-valued kernel, Random Fourier Features, Concentration inequalities.

1. Introduction

Multi-task regression (Micchelli and Pontil, 2005), structured classification (Dinuzzo et al., 2011), vector field learning (Baldassarre et al., 2012) and vector autoregression (Sindhwani et al., 2013; Lim et al., 2015) are all learning problems that boil down to learning a vector while taking into account an appropriate output structure. In this paper we are interested in a general and flexible approach to efficiently implement and learn vector-valued functions, while allowing couplings between the outputs. To achieve this goal, we turn to shallow architectures, namely the product of a (nonlinear) feature matrix $\tilde{\Phi}(x)$ and a parameter vector θ such that $\tilde{f}(x) = \tilde{\Phi}(x)^*\theta$, and combine two appealing methodologies: Operator-Valued Kernel Regression and Random Fourier Features.

Operator-Valued Kernels (Micchelli and Pontil, 2005; Carmeli et al., 2010; Álvarez et al., 2012) extend the classic scalar-valued kernels to vector-valued functions. As in the scalar case, operator-valued kernels (OVKs) are used to build Reproducing Kernel Hilbert Spaces (RKHS) in which representer theorems apply as for ridge regression or other appropriate

loss functional. In these cases, learning a model in the RKHS 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 $K(x, x_i)$, an operator on vectors of $\mathcal{Y}$. However, OVKs suffer from the same drawback as classic kernel machines: they scale poorly to very large datasets because they are very demanding in terms of memory and computation. Therefore, focusing on the case $\mathcal{Y} = \mathbb{R}^p$, we propose to approximate OVKs by extending a methodology called Random Fourier Features (RFFs) (Rahimi and Recht, 2007; Le et al., 2013; Yang et al., 2015; Sriperumbudur and Szabo, 2015; Bach, 2015; Sutherland and Schneider, 2015) so far developed to speed up scalar-valued kernel machines. The RFF approach linearizes a shift-invariant kernel model by generating explicitly an approximated feature map $\tilde{\phi}$. RFFs has been shown to be efficient on large datasets and has been further improved by efficient matrix computations (Le et al., 2013, "FastFood"), and is considered as one of the best large scale implementations of kernel methods, along with Nÿstrom approaches proposed in Yang et al. (2012).

In this paper, we propose general Random Fourier Features for functions in vector-valued RKHS. After recalling the background of this study, we present the following contributions: (1) we define a general form of Operator Random Fourier Feature (ORFF) map for shift-invariant operator-valued kernels, (2) we construct explicit operator feature maps for a simple bounded kernel, the decomposable kernel, and more complex unbounded kernels curl-free and divergence-free[1] kernels, (3) we show the corresponding kernel approximation uniformly converges with high probability towards the target kernel and (4) we discuss appropriate learning algorithms to benefit from ORFF and illustrate the theoretical approach by a few numerical results.

2. Background

2.1. Random Fourier Features

We consider scalar-valued functions. Denote $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a positive definite kernel on $\mathbb{R}^d$. A kernel k is said to be *shift-invariant* for the addition if for any $a \in \mathbb{R}^d$, $\forall (x, x') \in \mathbb{R}^d \times \mathbb{R}^d, k(x - a, z - a) = k(x, z)$. 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. Bochner theorem is the theoretical result that leads to the Random Fourier Features.

Theorem 1 (Bochner's theorem) *Every positive definite complex function is the Fourier transform of a non-negative measure. It implies that any positive definite, continuous and shift-invariant kernel k is the Fourier transform of a non-negative measure μ:*

$$k(x, z) = k_0(x - z) = \int_{\mathbb{R}^d} e^{-i\langle \omega, x - z\rangle} d\mu(\omega). \tag{1}$$

Without loss of generality, we assume that μ is a probability measure, i.e. $\int_{\mathbb{R}^d} d\mu(\omega) = 1$. Then we can write eq. (1) as an expectation over μ: $k_0(x - z) = \mathbb{E}_\mu \left[e^{-i\langle \omega, x - z\rangle} \right]$. If k is real valued we thus only write the real part: $k(x, z) = \mathbb{E}_\mu[\cos\langle \omega, x - z\rangle] = \mathbb{E}_\mu[\cos\langle \omega, z\rangle \cos\langle \omega, x\rangle +$

1. Also referred to as div-free kernel.

$\sin\langle\omega, z\rangle \sin\langle\omega, x\rangle]$. Let $\bigoplus_{j=1}^{D} x_j$ denote the Dm-length column vector obtained by stacking vectors $x_j \in \mathbb{R}^m$. The feature map $\tilde{\phi} : \mathbb{R}^d \to \mathbb{R}^{2D}$ defined as

$$\tilde{\phi}(x) = \frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} \begin{pmatrix} \cos\langle x, \omega_j\rangle \\ \sin\langle x, \omega_j\rangle \end{pmatrix}, \quad \omega_j \sim \mu \tag{2}$$

is called a *Random Fourier Feature* (map). Each $\omega_j, j = 1, \ldots, D$ is independently sampled from the inverse Fourier transform μ of k_0. This Random Fourier Feature map provides the following Monte-Carlo estimator of the kernel: $\tilde{k}(x, z) = \tilde{\phi}(x)^* \tilde{\phi}(z)$. The dimension D governs the precision of this approximation, whose uniform convergence towards the target kernel (as defined in eq. (1)) 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 learning: (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 $k(x - z) = exp(-\gamma\|x - z\|^2)$, the spectral distribution $\mu(\omega)$ is Gaussian.

2.2. Operator-Valued Kernels (OVK)

We now turn to vector-valued functions and consider vector-valued Reproducing Kernel Hilbert spaces (vv-RKHS) theory. The definitions are given for input space $\mathcal{X} \subset \mathbb{C}^d$ and output space $\mathcal{Y} \subset \mathbb{C}^p$. We will define operator-valued kernel as reproducing kernels. Given $\mathcal{X}$ and $\mathcal{Y}$, a map $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ is called a $\mathcal{Y}$-reproducing kernel if

$$\sum\nolimits_{i,j=1}^{N} \langle K(x_i, x_j) y_j, y_i\rangle \geq 0,$$

for all $x_1, \ldots, x_N$ in $\mathcal{X}$, all $y_1, \ldots, y_N$ in $\mathcal{Y}$ and $N \geq 1$. 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 $\forall z \in \mathcal{X}$ by $(K_x y)(z) = K(z, x)y$. 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 $K_x \in \mathcal{L}(\mathcal{Y}; \mathcal{H}_K)$, $\forall x \in \mathcal{X}$ and $f(x) = K_x^* f$, $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}_K$, 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 (Carmeli et al., 2010) we have

$$K(x, z) = K_x^* K_z \ \ \forall x, z \in \mathcal{X} \text{ and } \mathcal{H}_K = \overline{\text{span}} \ \{K_x y \mid \forall x \in \mathcal{X}, \ \forall y \in \mathcal{Y}\}$$

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

Proposition 2 (Feature map) *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 by $(Wg)(x) = \Phi_x^* g$, $\forall g \in \mathcal{H}, \forall x \in \mathcal{X}$ is a partial isometry from $\mathcal{H}$ onto the reproducing kernel Hilbert space $\mathcal{H}_K$ with reproducing kernel*

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

We call Φ a *feature map*. In this paper, we are interested on finding feature maps of this form for shift-invariant $\mathbb{R}^p$-Mercer kernels using the following definitions. A reproducing kernel K on $\mathbb{R}^d$ is a $\mathbb{R}^p$-Mercer provided that $\mathcal{H}_K$ is a subspace of $\mathcal{C}(\mathbb{R}^d;\mathbb{R}^p)$. It is said to be a *shift-invariant kernel*[2] for the addition if $K(x+a,z+a)=K(x,z),\forall(x,z,a)\in\mathcal{X}^3$. It is characterized by a function $K_0:\mathcal{X}\to\mathcal{L}(\mathcal{Y})$ of completely positive type such that $K(x,z)=K_0(\delta)$, with $\delta=x-z$.

3. Operator-valued Random Fourier Features

3.1. Spectral representation of shift-invariant vector-valued Mercer kernels

The goal of this work is to build approximated matrix-valued feature map for shift-invariant $\mathbb{R}^p$-Mercer kernels, denoted with K, such that any function $f\in\mathcal{H}_K$ can be approximated by a function $\tilde{f}$ defined by: $\tilde{f}(x)=\tilde{\Phi}(x)^*\theta$, where $\tilde{\Phi}(x)$ is a matrix of size $(m\times p)$ and θ is an m-dimensional vector. For this purpose, we use results of Carmeli et al. (2010) and Zhang et al. (2012) to define the Fourier transform of shift-invariant Operator-Valued Mercer. In this work, we focus on the finite real case $\mathcal{X}=\mathbb{R}^d$ and $\mathcal{Y}=\mathbb{R}^p$. However the whole framework stands for Hilbert spaces of infinite dimension. The following proposition of Zhang et al. (2012) extends Bochner's theorem to any shift-invariant $\mathbb{R}^p$-Mercer kernel.

Proposition 3 (Operator-valued Bochner's theorem) *A continuous function K from $\mathbb{R}^d\times\mathbb{R}^d$ to $\mathcal{L}(\mathbb{R}^p)$ is a shift-invariant reproducing kernel if and only if $\forall x,z\in\mathbb{R}^d$, it is the Fourier transform of a positive operator-valued measure $\mathcal{M}:\mathcal{B}(\mathbb{R}^d)\to\mathcal{L}_+(\mathbb{R}^p)$ with $K(x,z)=\int_{\mathbb{R}^d}e^{-i\langle x-z,\omega\rangle}d\mathcal{M}(\omega)$, where $\mathcal{M}$ belongs to the set of all the $\mathcal{L}_+(\mathbb{R}^p)$-valued measures of bounded variation on the σ-algebra of Borel subsets of $\mathbb{R}^d$.*

However it is much more convenient to use a more explicit result that involves real-valued (positive) measures. The following proposition instantiates the proposition 13 in Carmeli et al. (2010) to matrix-valued operators.

Proposition 4 (Spectral decomposition of shift-invariant OVK) *Let μ be a positive measure on $\mathbb{R}^d$ and $A:\mathbb{R}^d\to\mathcal{L}(\mathbb{R}^p)$ such that $A(\cdot)_{\ell m}\in L^1(\mathbb{R}^d,d\mu)$ for all $\ell,m'\in\{1,\dots,p\}$ and $A(\omega)\geq 0$ for μ-almost all ω. Then, for all $\delta\in\mathbb{R}^d$, for all $\ell,m\in\{1,\dots,p\}$,*

$$K_0(\delta)_{\ell m}=\int_{\mathbb{R}^d}e^{-i\langle\delta,\omega\rangle}A(\omega)_{\ell m}d\mu(\omega) \tag{3}$$

is the kernel signature of a shift-invariant $\mathbb{R}^p$-Mercer kernel K such that $K(x,z)=K_0(x-z)$. In other terms, each real-valued function $K_0(\cdot)_{\ell m}$ is the Fourier transform of $A(\cdot)_{\ell m}p_\mu(\cdot)$ where $p_\mu(\omega)=\frac{d\mu}{d\omega}$ is the Radon-Nikodym derivative (density) of the measure μ. Any shift-invariant kernel is of the above form for some pair $(A(\omega),\mu(\omega))$.

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 (theorem 1). Now we introduce the following proposition that is a direct consequence of proposition 4.

2. Also referred to as *translation-invariant kernel*.

Proposition 5 (Fourier feature map) *Given the conditions of proposition 4, we define $B(\omega)$ such that $A(\omega) = B(\omega)B(\omega)^*$. Then the function $\Phi_x : \mathbb{R}^p \to L^2(\mathbb{R}^d, \mu; \mathbb{R}^p)$ defined for all $x \in \mathbb{R}^p$ by*

$$\forall y \in \mathbb{R}^p, \quad (\Phi_x y)(\omega) = e^{i\langle x,\omega\rangle} B(\omega)^* y \tag{4}$$

is a feature map of the shift-invariant kernel K: i.e. for all x, z in $\mathbb{R}^d$, $\Phi_x^* \Phi_z = K(x, z)$.

Proof. For all $y, y' \in \mathbb{R}^p$,

$$(\Phi_x y)(\cdot)^*(\Phi_z y')(\cdot) = \int_{\mathbb{R}^d} e^{i\langle x,\omega\rangle} y^* B(\omega) e^{-i\langle z,\omega\rangle} B(\omega)^* y' d\mu(\omega) = \int_{\mathbb{R}^d} e^{i\langle x-z,\omega\rangle} \langle y, A(\omega)y'\rangle d\mu(\omega),$$

Taking $y = e_\ell$ and $y' = e_m$, where e_ℓ's are basis vectors of $\mathbb{R}^p$ yields from proposition 4 $(\Phi_x e_\ell)(\cdot)^*(\Phi_z e_m)(\cdot) = (\Phi_x^* \Phi_z)_{\ell m} = \int_{\mathbb{R}^d} e^{-i\langle \delta,\omega\rangle} A(\omega)_{\ell m} d\mu(\omega) = K_0(\delta)_{\ell m}$. ■

To define an approximation of a given operator-valued kernel, we need an inversion theorem that provides an explicit construction of the pair $A(\omega), \mu(\omega)$ from the kernel signature. We use Carmeli et al. (2010, Prop. 14.) instantiated to $\mathbb{R}^p$-Mercer kernel to find such a pair.

Proposition 6 (Carmeli et al. (2010)) *Let K be a shift-invariant $\mathbb{R}^p$-Mercer kernel. Suppose that $\forall \ell, m \in \{1, \ldots, p\}$, $K_0(\cdot)_{\ell m} \in L^1(\mathbb{R}^d, dx)$ where dx denotes the Lebesgue measure. Define C such that for all $\omega \in \mathbb{R}^d$, and for all $\ell, m \in \{1, \ldots, p\}$,*

$$C(\omega)_{\ell m} = \int_{\mathbb{R}^d} e^{i\langle \delta,\omega\rangle} K_0(\delta)_{\ell m} d\delta. \text{ Then} \tag{5}$$

i) $C(\omega)$ is a non-negative matrix for all $\omega \in \mathbb{R}^d$,

ii) $\forall \ell, m \in \{1, \ldots, p\}$, $C(\cdot) \in L^1(\mathbb{R}^d, d\omega)$,

iii) $\forall \delta \in \mathbb{R}^d$, $\forall \ell, m \in \{1, \ldots, p\}$, $K_0(\delta)_{\ell m} = \int_{\mathbb{R}^d} e^{-i\langle \delta,\omega\rangle} C(\omega)_{\ell m} d\omega$.

From eq. (3) and eq. (5), we can write the following equality concerning the matrix-valued kernel signature K_0, coefficient-wise: $\forall \delta \in \mathbb{R}^d$, $\forall \ell, m \in \{1, \ldots, p\}$, $\int_{\mathbb{R}^d} e^{-i\langle \delta,\omega\rangle} C(\omega)_{\ell m} d\omega = \int_{\mathbb{R}^d} e^{-i\langle \delta,\omega\rangle} A(\omega)_{\ell m} d\mu(\omega)$. We then conclude that the following equality holds almost everywhere for $\omega \in \mathbb{R}^d$: $C(\omega)_{\ell m} = A(\omega)_{\ell m} p_\mu(\omega)$ where $p_\mu(\omega) = \frac{d\mu}{d\omega}$. Without loss of generality we assume that $\int_{\mathbb{R}^d} d\mu(\omega) = 1$ and thus, μ is a probability distribution. Note that this is always possible through an appropriate normalization of the kernel. We note p_μ is the density of μ. Eventually proposition 4 results in an expectation: $K_0(x - z) = \mathbb{E}_\mu[e^{-i\langle x-z,\omega\rangle} A(\omega)]$.

3.2. Construction of Operator Random Fourier Feature

Given a $\mathbb{R}^p$-Mercer shift-invariant kernel K on $\mathbb{R}^d$, we build an Operator-Valued Random Fourier Feature (ORFF) map in three steps presented in algorithm 1. It relies on a Monte-Carlo approximation of the spectral representation of K presented in eqs. (3) and (4).

Algorithm 1: Construction of ORFF

Input : $K(x,z) = K_0(\delta)$ a $\mathbb{R}^p$-shift-invariant Mercer kernel such that $K_0(\delta)_{\ell m} \in L^1(\mathbb{R}^d, dx)$.

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

1 Compute $C : \mathbb{R}^d \to \mathcal{L}(\mathbb{R}^p)$ from eq. (5) by using the inverse Fourier transform of K_0, the signature of K;

2 Find $B(\omega)$, $p_\mu(\omega)$ such that $B(\omega)B(\omega)^* p_\mu(\omega) = C(\omega)$;

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

4 **return** $\tilde{\Phi}(x) = \frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} e^{-i\langle x, \omega_j \rangle} B(\omega_j)^*$;

3.3. Monte-Carlo approximation

Let $\bigoplus_{j=1}^{D} X_j$ denote the block matrix of size $rD \times s$ obtained by stacking D matrices $X_1, \ldots, X_D$ of size $r \times s$. Assuming steps 1 and 2 have been performed, for all $j = 1, \ldots, n$, we find a decomposition $A(\omega_j) = B(\omega_j)B(\omega_j)^*$ either by exhibiting a general analytical closed-form or using a numerical decomposition. Denote $p \times p'$ the dimension of the matrix $B(\omega)$. Based on proposition 5, we propose a randomized matrix-valued feature map: $\forall x \in \mathbb{R}^d$,

$$\tilde{\Phi}(x) = \frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} \Phi_x(\omega_j) = \frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} e^{-i\langle x, \omega_j \rangle} B(\omega_j)^*, \tag{6}$$

Where $\forall j \in \{1, \ldots, D\}$, ω_j are independent identically distributed (i.i.d.) random vectors following the probability law μ. The corresponding approximation for the kernel is then for all $x, z \in \mathbb{R}^d$,

$$\tilde{K}(x,z) = \tilde{\Phi}(x)^*\tilde{\Phi}(z) = \sum\nolimits_{j=1}^{D} \frac{\Phi_x(\omega_j)^*\Phi_z(\omega_j)}{D} = \sum\nolimits_{j=1}^{D} \frac{e^{-i\langle x-z, \omega_j \rangle}}{D} A(\omega_j). \tag{7}$$

From the weak law of large numbers, one can verify that the Monte-Carlo estimator $\tilde{\Phi}(x)^*\tilde{\Phi}(z)$ converges in probability in the weak operator topology to the target kernel $K(x,z)$ when D tends to infinity. Namely,

$$\tilde{K}(x,z) = \tilde{\Phi}(x)^*\tilde{\Phi}(z) \xrightarrow[D\to\infty]{p.} \mathbb{E}_\mu \left[e^{-i\langle x-z, \omega \rangle} A(\omega) \right] = K(x,z)$$

We also use the notation $\tilde{K}^j(\delta) = \Phi_x(\omega_j)^*\Phi_z(\omega_j)$ such that $\sum_{j=1}^{D} \tilde{K}^j(\delta)/D = \tilde{K}(x,z)$ and $\tilde{K}_0(\delta) = \tilde{K}(x,z)$. As for the scalar-valued kernel, a real-valued matrix-valued function has a real matrix-valued Fourier transform if $A(\omega)$ is even with respect to ω. Taking this point into account, we define the feature map of a real matrix-valued kernel as

$$\tilde{\Phi}(x) = \frac{1}{\sqrt{D}} \bigoplus_{j=1}^{D} \begin{pmatrix} \cos\langle x, \omega_j \rangle B(\omega_j)^* \\ \sin\langle x, \omega_j \rangle B(\omega_j)^* \end{pmatrix}, \quad \omega_j \sim \mu.$$

The kernel approximation becomes

$$\tilde{K}(x,z) = \sum_{j=1}^{D} \frac{\cos\langle x, \omega_j \rangle \cos\langle z, \omega_j \rangle + \sin\langle x, \omega_j \rangle \sin\langle z, \omega_j \rangle}{D} A(\omega_j) = \sum_{j=1}^{D} \frac{\cos\langle x - z, \omega_j \rangle}{D} A(\omega_j).$$

Algorithm 1 summarizes the construction of ORFF. In the following, we give an explicit construction of ORFFs for three well-known $\mathbb{R}^p$-Mercer and shift-invariant kernels: the *decomposable kernel* introduced in Micchelli and Pontil (2005) for multi-task regression and the *curl-free* and the *divergence-free* kernels studied in Macedo and Castro (2008) and Baldassarre et al. (2012) for vector field learning. All these kernels are defined using a scalar-valued shift-invariant Mercer kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ whose signature is denoted k_0. A usual choice is to choose k as a Gaussian kernel with $k_0(\delta) = \exp\left(-\frac{\|\delta\|^2}{2\sigma^2}\right)$, which gives $\mu = \mathcal{N}(0, \sigma^{-2}I)$ as its inverse Fourier transform.

Definition 7 (Decomposable kernel) *Let A be a $(p \times p)$ positive semi-definite matrix. If $\forall x, z \in \mathbb{R}^d, K^{dec}(x, z) = k_0(x - z)A$, then K is a $\mathbb{R}^p$-Mercer shift-invariant reproducing kernel.*

The matrix A encodes 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) that A can be chosen equal to the pseudo-inverse $L^\dagger$ of the graph Laplacian, and then the ℓ_2 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 (Dinuzzo et al., 2011; Sindhwani et al., 2013; Lim et al., 2015). In the following the Fourier transform is referred to as $\mathcal{F}[\cdot]$ and the inverse Fourier transform as $\mathcal{F}^{-1}[\cdot]$.

Example 1 (ORFF for decomposable kernel)

$$C^{dec}(\omega)_{\ell m} = \int_{\mathcal{X}} e^{i\langle \delta, \omega \rangle} k_0(\delta) A_{\ell m} d\delta = A_{\ell m} \mathcal{F}^{-1}[k_0](\omega)$$

Hence, $A^{dec}(\omega) = A$ and $p^{dec}_\mu(\omega) = \mathcal{F}^{-1}[k_0](\omega)$.

ORFF for curl-free and div-free kernels: Curl-free and divergence-free kernels provide an interesting application of operator-valued kernels 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 also discussed in Fuselier (2006) allow encoding input-dependent similarities between vector-fields.

Definition 8 (Curl-free and Div-free kernel) *We have $d = p$. The divergence-free kernel is defined as $K^{div}(x, z) = K^{div}_0(\delta) = (\nabla\nabla^* - \Delta I)k_0(\delta)$ and the curl-free kernel as $K^{curl}(x, z) = K^{curl}_0(\delta) = -\nabla\nabla^* k_0(\delta)$, where $\nabla\nabla^*$ is the Hessian operator and Δ 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 (Macedo and Castro, 2008; Baldassarre et al., 2012). For curl-free kernel we use the differentiation properties of the Fourier transform.

Example 2 (ORFF for curl-free kernel) $\forall \ell, m \in \{1, \ldots, p\}$,

$$C^{curl}(\omega)_{\ell m} = -\mathcal{F}^{-1}\left[\frac{\partial}{\partial \delta_\ell}\frac{\partial}{\partial \delta_m} k_0\right](\omega) = \omega_\ell \omega_m \mathcal{F}^{-1}[k_0](\omega)$$

Hence, $A^{curl}(\omega) = \omega\omega^*$ *and* $p_\mu^{curl}(\omega) = \mathcal{F}^{-1}[k_0](\omega)$. *We can obtain directly:* $B^{curl}(\omega) = \omega$.

For the divergence-free kernel we first compute the Fourier transform of the Laplacian of a scalar kernel using differentiation and linearity properties of the Fourier transform. We denote $\delta_{\{\ell=m\}}$ as the Kronecker delta which is 1 if $\ell = m$ and zero otherwise.

Example 3 (ORFF for divergence-free kernel)

$$\begin{aligned} C^{div}(\omega)_{\ell m} &= \mathcal{F}^{-1}\left[\frac{\partial}{\partial\delta_\ell}\frac{\partial}{\partial\delta_m}k_0 - \delta_{\{\ell=m\}}\Delta k_0\right] = \mathcal{F}^{-1}\left[\frac{\partial}{\partial\delta_\ell}\frac{\partial}{\partial\delta_m}k_0\right] - \delta_{\{\ell=m\}}\mathcal{F}^{-1}[\Delta k_0] \\ &= (\delta_{\{\ell=m\}} - \omega_\ell\omega_m)\|\omega\|_2^2\mathcal{F}^{-1}[k_0], \end{aligned}$$

since $\mathcal{F}^{-1}[\Delta k_0(\delta)] = \sum_{k=1}^p \mathcal{F}^{-1}\left[\frac{\partial}{\partial\delta_k}k_0\right] = -\|\omega\|_2^2\mathcal{F}^{-1}[k_0]$. *Hence* $A^{div}(\omega) = I\|\omega\|_2^2 - \omega\omega^*$ *and* $p_\mu^{div}(\omega) = \mathcal{F}^{-1}[k_0](\omega)$. *Here,* $B^{div}(\omega) = I\|\omega\| - \omega\omega^*/\|\omega\|$.

4. Theoretical guaranties for the ORFF approximation error

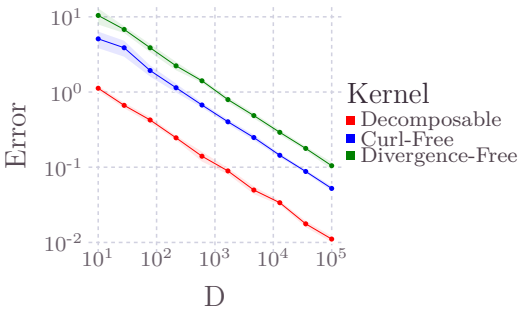
4.1. Uniform error bound

We are now interested on measuring how close ORFF approximation $\tilde{K}$ is to K given D. If A is a real matrix, we denote $\|A\|_2$ its spectral norm, defined as the square root of the largest eigenvalue of A. If F is an operator-valued function we use the shortcut notation $\|F\|_\infty = \sup_x \|F(x)\|_2$. For x and z in $\mathcal{C} \subset \mathbb{R}^d$, we study how,

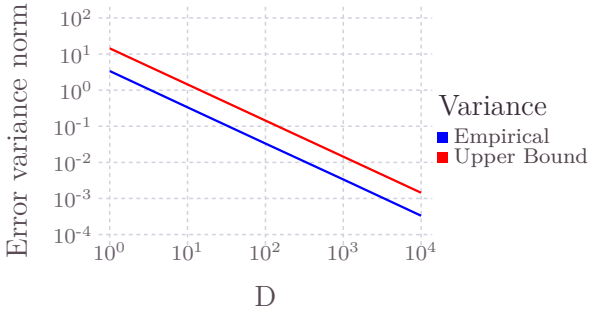
$$\left\|\tilde{K} - K\right\|_\infty = \sup_{x,z\in\mathcal{C}}\left\|\tilde{K}(x,z) - K(x,z)\right\|_2 \tag{8}$$

behaves according to D, that is the the maximal approximation error of the largest eigenvalue across the domain of K. Figure 1 A empirically shows convergence of three different OVK approximations for 1000 data uniformly drawn from the compact $[-1,1]^4$ 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.

To bound the approximation error, we turn to concentration inequalities devoted to random matrices (Boucheron et al., 2013). For decomposable kernel, the error bound can be directly obtained as a consequence of the uniform convergence of RFFs in the scalar case proved in Rahimi and Recht (2007); Sutherland and Schneider (2015); Sriperumbudur and Szabo (2015); since in this case $\|\tilde{K}(x,z) - K(x,z)\|_2 = \|A\|_2|\tilde{k}(x,z) - k(x,z)|$. This theorem and its proof are presented in corollary 2 of the supplementary material. More interestingly, we propose a new bound for Operator Random Fourier Feature approximation in the general case. It relies on two main ideas: (i) Matrix-Bernstein concentration inequality for random matrices need to be used instead of concentration inequality for scalar random variables, (ii) a general theorem valid for random matrices with bounded norms (case for decomposable kernel ORFF approximation) as well as unbounded norms (curl and divergence-free kernels) that behave as subexponential random variables. 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.



A. Empirical Approximation Error versus D for different OVKs.



B. Comparison between an empirical bound on the norm of the variance of the curl-free ORFF obtained and the theoretical bound proposed in proposition 12 versus D.

Figure 1: Empirical approximation error and bounds on the norm of the variance of $\tilde{K}$

Definition 9 (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 $(\Omega, \mathcal{T}(\Omega), \mu)$, $\|X\|_\psi := \inf\{C > 0 \mid \mathbb{E}[\psi(|X|/C)] \leq 1\}$.*

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 random matrix of size $p \times p$. We call *variance* of X and use the notation $\mathbb{V}[X] = \mathbb{E}[X - \mathbb{E}[X]]^2$. With this convention, $\mathbb{V}[X]_{\ell m} = \sum_{r=1}^{p} \mathrm{Cov}[X_{\ell r}, X_{rm}]$.

Theorem 10 *Assume K is a shift-invariant $\mathbb{R}^p$-Mercer kernel on $\mathcal{C}$, a compact subset of $\mathbb{R}^d$ of diameter $|\mathcal{C}|$. Let $\tilde{K}$ be the ORFF approximation of K depending on D (as defined in eq. (7)), K_0 be the kernel signature of K and $p_\mu(\cdot)A(\cdot)$ be the inverse Fourier transform of the kernel's signature (in the sense of proposition 6). Let us define the constants b, σ_p^2, $m \in \mathbb{R}_+$ as*

$$b = D\left\|\mathbb{V}_\mu\left[\tilde{K}\right]\right\|_\infty \text{ and } \sigma_p^2 = \mathbb{E}_\mu\left[\|\omega\|_2^2\|A(\omega)\|_2^2\right] \text{ and } m = 4\left(\left\|\|A(\omega)\|_2\right\|_{\psi_1} + \|K\|_\infty\right)$$

where $\omega \sim \mu$, then for all ϵ in $\mathbb{R}_+$,

$$\mathbb{P}\left\{\left\|\tilde{K} - K\right\|_\infty \geq \epsilon\right\} \leq C_{d,p}\left(\frac{\sigma_p|\mathcal{C}|}{\epsilon}\right)^{\frac{2}{1+2/d}} \begin{cases} \exp\left(-\frac{\epsilon^2 D}{8(d+2)\left(b+\frac{\epsilon\bar{u}}{6}\right)}\right) & \text{if } \epsilon\bar{u} \leq 2(e-1)b \\ \exp\left(-\frac{\epsilon D}{(d+2)(e-1)\bar{u}}\right) & \text{otherwise,} \end{cases}$$

where $\bar{u} = 2m\log\left(2^{\frac{3}{2}}\left(\frac{m}{b}\right)^2\right)$ and $C_{d,p} = p\left(\left(\frac{d}{2}\right)^{\frac{-d}{d+2}} + \left(\frac{d}{2}\right)^{\frac{2}{d+2}}\right)2^{\frac{6d+2}{d+2}}$.

We give here a sketch of the proof here and a complete comprehensive proof of 10 is given in section B of the supplementary material.

Sketch of proof. In the following, let $F(\delta) = F(x-z) = \tilde{K}(x,z) - K(x,z)$. As in Rahimi and Recht (2007) let $\mathcal{D}_\mathcal{C} = \{x - z | x, z \in \mathcal{C}\}$ with diameter at most $2l$ where l is the diameter

of $\mathcal{C}$. Since $\mathcal{C}$ is supposed compact, so is $\mathcal{D}_\mathcal{C}$. It is then possible to find an ϵ-net covering $\mathcal{D}_\mathcal{C}$ with at most $T = (4|\mathcal{C}|/r)^d$ balls of radius r. Let us call $\delta_i, i = 1, \dots, T$ the center of the i-th ball, called *anchors* of the ϵ-net. Denote L_F the Lipschitz constant of F. We introduce the following lemma proved in the supplements:

Lemma 11 *If (H1): $L_F \leq \frac{\epsilon}{2r}$ and (H2) $\|F(\delta_i)\|_2 \leq \frac{\epsilon}{2}$, for all $0 < i < T$, then $\forall \delta \in \mathcal{D}_\mathcal{C}$, $\|F(\delta)\|_2 \leq \epsilon$.*

To apply the lemma, we must check assumptions (H1) and (H2).

Sketch of proof of (H1). We bound the Lipschitz constant by noticing that F is differentiable, so $L_F = \left\|\frac{\partial F}{\partial \delta}(\delta^*)\right\|_2$ where $\delta^* = \arg\max_{\delta \in \mathcal{D}_\mathcal{C}} \left\|\frac{\partial F}{\partial \delta}(\delta)\right\|_2$. Using Jensen's inequality and applying Markov's inequality yields

$$\mathbb{P}\left\{L_F \geq \frac{\epsilon}{2r}\right\} = \mathbb{P}\left\{L_F^2 \geq \left(\frac{\epsilon}{2r}\right)^2\right\} \leq \mathbb{E}_\mu\left[\|\omega\|_2^2 \|A(\omega)\|_2^2\right]\left(\frac{2r}{\epsilon}\right)^2. \tag{9}$$

We set $\sigma_p^2 = \mathbb{E}_\mu\left[\|\omega\|_2^2\|A(\omega)\|_2^2\right]$ and suppose its existence.

Sketch of proof of (H2). To obtain a bound on the anchors we apply Koltchinskii et al. (2013, theorem 4). We suppose that the two constants $b = \sup_{\delta \in \mathcal{D}_\mathcal{C}} D\left\|\mathbb{V}_\mu\left[\tilde{K}_0(\delta)\right]\right\|_2$ and $\bar{u} = \log\left(2\left(\frac{m}{b}\right)^2 + 1\right)$, where $m = 4\left(\left\|\|A(\omega)\|_2\right\|_{\psi_1} + \sup_{\delta \in \mathcal{D}_\mathcal{C}}\|K_0(\delta)\|_2\right)$ and $\omega \sim \mu$, exists. Then,

$$\forall i = 1, \dots, T, \;\; \mathbb{P}\left\{\|F(\delta_i)\|_2 \geq \epsilon\right\} \leq 2p \begin{cases} \exp\left(-\frac{D\epsilon^2}{4b + 2\epsilon\bar{u}/3}\right) & \text{if } \epsilon\bar{u} \leq 2(e-1)b \\ \exp\left(-\frac{D\epsilon}{(e-1)\bar{u}}\right) & \text{otherwise.} \end{cases} \tag{10}$$

Combining (H1) and (H2). Now applying the lemma and taking the union bound over the centers of the ϵ-net yields $\mathbb{P}\left\{\sup_{\delta \in \mathcal{D}_\mathcal{C}}\|F(\delta)\|_2 \leq \epsilon\right\} \geq 1 - \kappa_1 r^{-d} - \kappa_2 r^2$, with $\kappa_2 = 4\sigma_p^2\epsilon^{-2}$ and $\kappa_1 = 2p(4|\mathcal{C}|)^d \begin{cases} \exp\left(-\frac{\epsilon^2 D}{16\left(b + \frac{\epsilon}{6}\bar{u}\right)}\right) & \text{if } \epsilon\bar{u} \leq 2(e-1)b \\ \exp\left(-\frac{\epsilon D}{2(e-1)\bar{u}}\right) & \text{otherwise} \end{cases}$. We choose r such that $d\kappa_1 r^{-d-1} - 2\kappa_2 r = 0$, i.e. $r = \left(\frac{d\kappa_1}{2\kappa_2}\right)^{\frac{1}{d+2}}$. The bound becomes

$$\mathbb{P}\left\{\sup_{\delta \in \mathcal{D}_\mathcal{C}}\|F(\delta)\| \geq \epsilon\right\} \leq pC_d' 2^{\frac{6d+2}{d+2}}\left(\frac{\sigma_p|\mathcal{C}|}{\epsilon}\right)^{\frac{2}{1+2/d}} \begin{cases} \exp\left(-\frac{\epsilon^2}{8(d+2)\left(b + \frac{\epsilon}{6}\bar{u}\right)}\right) & \text{if } \epsilon\bar{u} \leq 2(e-1)b \\ \exp\left(-\frac{\epsilon}{(d+2)(e-1)\bar{u}}\right) & \text{otherwise.} \end{cases}$$

where $C_d' = \left(\left(\frac{d}{2}\right)^{\frac{-d}{d+2}} + \left(\frac{d}{2}\right)^{\frac{2}{d+2}}\right)$. Conclude by taking $C_{d,p} = pC_d' 2^{\frac{6d+2}{d+2}}$. ∎

4.2. Variance of the ORFF approximation

We now provide a bound on the norm of the variance of $\tilde{K}$, required to apply theorem 10.

Proposition 12 (Bounding the *variance* of $\tilde{K}$) *Let K be a shift-invariant $\mathbb{R}^p$-Mercer kernel on $\mathcal{C}$, a compact subset of $\mathbb{R}^d$, $\tilde{K}$ be the ORFF approximation of K (as defined in eq. (7)) and $\mathcal{D}_\mathcal{C} = \{x - z \mid x, z \in \mathcal{C}\}$. Then*

$$\forall \delta \in \mathcal{D}_\mathcal{C}, \quad \left\| \mathbb{V}_\mu \left[\tilde{K}_0(\delta) \right] \right\|_2 \leq \frac{\left\| (K_0(2\delta) + K_0(0)) \mathbb{E}_\mu[A(\omega)] - 2K_0(\delta)^2 \right\|_2 + 2\|\mathbb{V}_\mu[A(\omega)]\|_2}{2D}.$$

Proof. It relies on the i.i.d. property of the random vectors ω_j and trigonometric identities (see the proof in section C of the supplementary material). ■

4.3. Application on decomposable, curl and div-free OVKs

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

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

$$D \left\| \mathbb{V}_\mu \left[\tilde{K}_0(\delta) \right] \right\|_2 \leq (1 + k_0(2\delta)) \|A\|_2 / 2 + k_0(\delta)^2.$$

Curl-free and div-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 $\mathbb{E}_\mu[A(\omega)] = I_d/\sigma^2$ and $\mathbb{V}_\mu[A(\omega)] = (d+1) I_d/\sigma^4$. Hence,

$$D \left\| \mathbb{V}_\mu \left[\tilde{K}_0(\delta) \right] \right\|_2 \leq \frac{1}{2} \left\| \frac{1}{\sigma^2} K_0(2\delta) - 2K_0(\delta)^2 \right\|_2 + \frac{(d+1)}{\sigma^4}.$$

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

$$D \left\| \mathbb{V}_\mu \left[\tilde{K}_0(\delta) \right] \right\|_2 \leq \frac{1}{2} \left\| \frac{(d-1)}{\sigma^2} K_0(2\delta) - 2K_0(\delta)^2 \right\|_2 + \frac{d(4d-3)}{\sigma^4}.$$

Eventually, we ensure that the random variable $\|A(\omega)\|$ has a finite Orlicz norm with $\psi = \psi_1$ in these three cases.

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 7 of the supplementary material). In other words $\|X\|_{\psi_1}^{-1} = \mathrm{MGF}(x)_X^{-1}(2)$. For the Gaussian curl-free and divergence-free kernel, $\left\|A^{div}(\omega)\right\|_2 = \left\|A^{curl}(\omega)\right\|_2 = \|\omega\|_2^2$, 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

$$\left\| \left\| A^{div}(\omega) \right\|_2 \right\|_{\psi_1}^{-1} = \left\| \left\| A^{curl}(\omega) \right\|_2 \right\|_{\psi_1}^{-1} = \frac{\sigma^2}{2} \left(1 - 4^{-\frac{1}{p}} \right).$$

5. Learning with ORFF

While theoretically relevant, the approximation error bounds are too loose to be used to find a safe value for D. In the following, we choose appropriate learning algorithms to use ORFF in vector-valued regression in order to study the empirical behavior of these methods. Code implementing ORFF is available at https://github.com/operalib/operalib/tree/ORFF in the branch ORFF of Operalib, a framework for OVK learning.

5.1. Penalized regression with ORFF

Once we have an approximated feature map, we can use it to provide a feature matrix of size $p'D \times p$ with matrix $B(\omega)$ of size $p \times p'$ such that $A(\omega) = B(\omega)B(\omega)^*$. A function $f \in \mathcal{H}_K$ is then approximated by a linear model $\tilde{f}(x) = \tilde{\Phi}(x)^*\theta$, where $\theta \in \mathbb{R}^{p'D}$. Let $\mathcal{S} = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}^p, i = 1, \ldots, N\}$ be a collection of i.i.d training samples. Given a local loss function $L : \mathcal{S} \to \mathbb{R}^+$ and a ℓ_2 penalty, we minimize

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} L\left(\tilde{\Phi}(x_i)^*\theta, y_i\right) + \lambda \|\theta\|_2^2, \tag{11}$$

instead of minimizing $\mathcal{L}(f) = \frac{1}{N}\sum_{i=1}^{N} L(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}_K}^2$. To find a minimizer of the optimization problem eq. (11) many optimization algorithms are available. For instance, in a large-scale context, a stochastic gradient descent algorithm would be suitable: we can adapt the algorithm to the kind of kernel/problematic. We investigate two optimization algorithms: a Stein equation solver appropriate for decomposable kernels and a (stochastic) gradient descent for non-decomposable kernels (e.g. the curl-free and div-free kernels).

Closed form for the decomposable kernel: for the real decomposable kernel $K_0(\delta) = k(\delta)A$ when $L(y, y') = \|y - y'\|_2^2$ (Kernel Ridge regression in $\mathcal{H}_K$), the learning problem described in eq. (11) can be re-written in terms of matrices to find the unique minimizer Θ_*, where $\mathrm{vec}(\Theta) = \theta$ such that θ is a $p'D$ vector and Θ a $p' \times D$ matrix. We use the notation $X = \bigoplus_{i=1}^{N} x_i$. If $\tilde{\phi}$ is a scalar feature map ($\tilde{\phi}(X) = \bigoplus_{i=1}^{N} \tilde{\phi}(x_i)$ is a matrix of size $D \times N$) for the scalar kernel k_0. Let $\|.\|_F$ be the Frobenius norm. Then

$$\tilde{\Phi}(X)^*\theta = (\tilde{\phi}(X)^* \otimes B)\theta = B\Theta\tilde{\phi}(X) \text{ and } \theta_* = \underset{\Theta \in \mathbb{R}^{p' \times D}}{\arg\min} \left\| B\Theta\tilde{\phi}(X) - Y \right\|_F^2 + \lambda \|\Theta\|_F^2. \tag{12}$$

This is a convex optimization problem and a sufficient condition is $\tilde{\phi}(X)\tilde{\phi}(X)^*\Theta_* B^* B - \tilde{\phi}(X)Y^*B + \lambda\Theta_* = 0$, which is a Stein equation.

Gradient computation for the general case. When it is not possible or desirable to use Stein's equations solver one can apply a (stochastic) gradient descent algorithm. The gradient computation for and ℓ_2-loss applied to ORFF model is briefly recalled in section D.1 of the supplementary material.

5.2. Numerical illustration

We present numerical experiments to illustrate and complete the theoretical contribution with bounded and unbounded ORFFs.

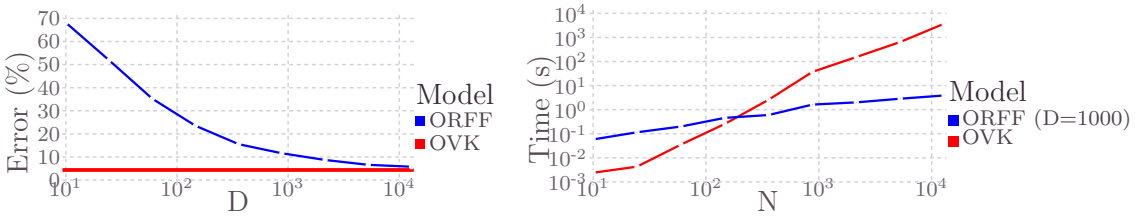


Figure 2: Empirical comparison of ORFF and OVK regression on MNIST dataset and empirical behavior of ORFF regression versus D and N.

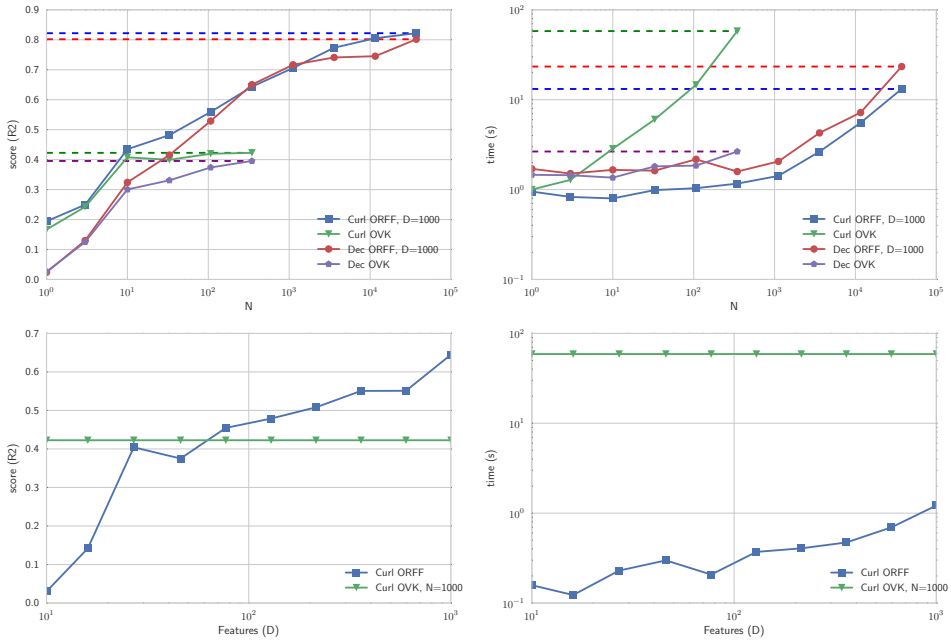


Figure 3: Empirical comparison between curl-free ORFF, curl-free OVK, independent ORFF, independent OVK on a synthetic vector field regression task.

Datasets: the first dataset considered is the handwrittent character recognition set, MNIST[3]. A training (resp. test) set of 12000 (resp. 10000) images were selected. The inputs are images represented as a vector $x_i \in [0, 255]^{784}$ and the targets are integers between 0 and 9. We scale the inputs such that they take values in $[-1, 1]^{784}$. We binarize the targets with a one-hot encoder. To predict classes, we use simplex coding method presented in Mroueh et al. (2012). The intuition behind simplex coding is to project the binarized labels of di-

3. Available at http://yann.lecun.com/exdb/mnist

mension p onto the most separated vectors on the hypersphere of dimension $p-1$. For ORFF we encode this projection in the matrix B 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

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

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 OVK we project the binarized targets on the simplex as a pre-processing step, before learning with the kernel $K_0(\delta) = I_p k_0(\delta)$, where k_0 is a also Gaussian kernel. The *second dataset* is a simulated 5D-vector field with structure. We generate a scalar field as a random function $f : [-1, 1]^5 \to \mathbb{R}$, where $f(x) = \tilde{\phi}(x)^T \theta$ where θ is a random normal matrix, $\tilde{\phi}$ 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 f to generate the 5D-curl-free vector-field. We also report additional results on a *third dataset* with 10^5 data from $\mathbb{R}^{20} \to \mathbb{R}^4$ used in Audiffren and Kadri (2013), in section D.2 of the supplements.

Performance of ORFF regression: we trained both ORFF and OVK models on MNIST dataset with a decomposable Gaussian kernel with signature $K_0(\delta) = \exp(-\|\delta\|/\sigma^2)A$. To find a solution of the optimization problem described in eq. (12), we use off-the-shelf solver[4] 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, fig. 2 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 w.r.t the number of points. Second, fig. 2 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.

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. We optimize eq. (11) using Scipy's L-BFGS-B solver[5]. Figure 3 (bottom row) reports the R2 score on the test set versus the number of curl-ORFF D with a comparison with curl-OVK. In this experiment, we see that curl-OFF can even be better than curl-OVK, suggesting that ORFF might play an additional regularizing role. It also shows the computation time of curl-ORFF and curl-OVK. We see that OVK regression does not scale with large datasets, while ORFF regression does. When $N > 10^4$, OVK regression exceeds memory capacity.

Structured prediction vs Independent (RFF) prediction: on the second dataset, fig. 3 (top row) compares R2 score and time of ORFF regression using the trivial identity decomposable kernel, e.g. independent RFFs, to curl-free ORFF regression. Curl-free ORFF outperforms independent RFFs, as expected, since the dataset involves structured outputs.

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

6. Conclusion

We introduced ORFF, a general and versatile framework for shift-invariant OVK approximation. We proved the uniform convergence of the approximation error for bounded and unbounded ORFFs. The complexity in time of these approximations together with the linear learning algorithm make this implementation scalable with the data size and thus appealing compared to OVK regression as shown in numerical experiments. Further work concerns generalization bounds and consistency for ORFF-regression. Finally this work opens the door to building deeper architectures by stacking vector-valued functions while keeping a kernel view for large datasets.

Acknowledgments

R. Brault was funded by University of Évry (PhD grant numbered 76391). The authors are grateful to Maxime Sangnier for his relevant comments.

References

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.

J. Audiffren and H. Kadri. Online learning with multiple operator-valued kernels. *arXiv preprint arXiv:1311.0222*, 2013.

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. 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.

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.

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.

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

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

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*, pages 213–226. Institute of Mathematical Statistics, 2013.

Q. V. Le, T. Sarlós, and A. J. Smola. Fastfood - computing hilbert space expansions in loglinear time. In *ICML 2013, Atlanta, USA, 16-21*, pages 244–252, 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, 2015.

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.

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.

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

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

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.

B. Sriperumbudur and Z. Szabo. Optimal rates for random fourier features. In C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, and R. Garnett, editors, *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.

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

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

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.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *NIPS 25*, pages 476–484, 2012.

Z. Yang, A. G. Wilson, A. J. Smola, and L. Song. A la carte - learning fast kernels. In *AISTATS* Yang et al. (2015).

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