Periodic Kernel Approximation by Index Set Fourier Series Features

Anthony Tompkins
School of Computer Science
The University of Sydney
anthony.tompkins@sydney.edu.au

Fabio Ramos
School of Computer Science
NVIDIA USA, The University of Sydney
fabio.ramos@sydney.edu.au

Abstract

Periodicity is often studied in timeseries modelling with autoregressive methods but is less popular in the kernel literature, particularly for multi-dimensional problems such as in textures, crystallography, quantum mechanics, and robotics. Large datasets often make modelling periodicity untenable for otherwise powerful non-parametric methods like Gaussian Processes (GPs) which typically incur an $\mathcal{O}(N^3)$ computational cost, while approximate feature methods are impeded by their approximate accuracy. We introduce a method that efficiently decomposes multi-dimensional periodic kernels into a set of basis functions by exploiting multivariate Fourier series. Termed *Index Set Fourier Series Features* (ISFSF), we show that our approximation produces significantly less generalisation error than alternative approximations such as those based on random and deterministic Fourier features on regression problems with periodic data.

1 INTRODUCTION

The phenomena of periodicity permeates a vast number of natural and artificial processes [9, 16, 4]. However, it is rare to come across its enquiry in machine learning and more specifically with regard to periodic kernels, kernel methods on manifolds, and their feature-space approximations. Almost all existing work focuses on non-parametric full-kernel methods. Although non-parametric methods [45] are exceptionally flexible methods for statistical modelling, they inherently lack scalability. A particular non-parametric method using *kernel* functions, is the GP [36].

However, the inability to truly scale GP inference to large datasets is a major limitation of such methods. While there have been various efforts to approximate GPs with lower rank solutions based on *inducing points* [40, 17] these methods are still constrained by their data-dependence. Inspired by the applicability of feature-space kernels for scalable GP regression [23] we note the real-world significance of effective periodic kernel representations for tasks such as texture in-painting [48], predictive representations of infinitely periodic crystal lattices and machine-learning aided discovery of materials [32, 37, 11], and in robotics for problems involving periodic systems [30, 10]. Scalable multivariate periodic kernel approximations, unaddressed in the literature, motivates the main contribution of this paper.

Specifically, in our contributions we provide:

- Fourier series approximations of multivariate stationary periodic kernels with an efficient sparse construction; and
- a general bound for the cardinality of the resulting full and sparse feature sets as well as an upper bound to the truncation error for the multivariate feature approximation.

We compare in detail the proposed method against recent state-of-the-art kernel approximations in terms of both the kernel approximation and, more importantly, predictive generalisation error. Empirical results on real datasets and robot simulations further demonstrate that deterministic index set based features provide significantly improved convergence generalisation properties by reducing both the data samples *and* the number of features required for equivalently accurate predictions.

2 RELATED WORK

While much work has been done for data-independent kernel approximations such as RFFs, as opposed to Nyström [46, 14], there is limited work on such approximations of

periodic kernels. Two recent works [41, 42] explore approximations for periodic kernels in univariate timeseries where some response varies periodically with respect to time. However, it is not clear how to tractably generalise such decompositions into multiple dimensions where the response varies periodically as a function of *multiple* inputs.

The work of [33, 34] termed *Random Fourier Features* (RFFs) is the idea of explicit data-*independent* feature maps using ideas from harmonic analysis and sketching theory. By approximating kernels these maps allow scalable inference with simple linear models. Various approximations to different kernels have followed and include polynomial kernels [31, 29], dot-product kernels [21], histogram and γ-homogenous kernels [25, 44], and approximations based upon the so-called *triple-spin* [6, 24, 50, 12], operator-valued kernels [5].

A recent work of note, quasi-Monte Carlo features (QMC) [49], uses deterministic sequences on the hypercube to approximate shift invariant kernels. In [42], the key idea is that periodic kernels can be harmonically decomposed in a deterministic manner using Fourier series decompositions. However, while tractable in the univariate case, it is not immediately extensible to multiple dimensions due to exponential complexity in the number of approximating coefficients.

Connected to our index set based features are quadrature rule based kernel approximations. Often based on grid based solutions, these similarly have exponential dependencies on the input dimension [8, 28, 27] which are countered to some extent via other assumptions such as additivity [28]. Also, works on quadrature are explored only for very specific families of kernels (sub-Gaussian, Gaussian) and explore numerical optimizations therein (e.g. butterfly algorithm, structured matrices). Our work is further distinct in that we explore the use of the *feature set* for parametric Bayesian modelling (regression) in linear *feature space* as opposed to the non-parametric form in *kernel space*. ISFSF are specifically constructed for the space of periodic kernels in the sense of the data lying on some manifold rather than requiring an explicit warping of the input data using standard aperiodic kernels; this has not been investigated previously in the multivariate case. ISFSF can naturally be seen as using an unweighted quadrature scheme, akin to MC and QMC Fourier Features. Thus, our sparse feature construction would directly benefit from quadrature rules applied to periodic spaces [19, 22, 18] however this is not the focus of our study in this paper and deserves future independent investigation.

We compare our body of work with the highy performing Halton and generalised Halton sequence [35]. We also stress that our method is inherently different from methods such as the Spectral Mixture kernel [47] which operate in the full kernel space. Our goal is to represent kernels for inference in a fashion similar to Sparse Spectral Gaussian Processes [23] which make them further amenable to Bayesian inference in streaming domains under hard computational constraints such as in robotics [13]. Specifically we may perform inference in $\mathcal{O}(NM^2)$ time for M features where $M \ll N$.

3 PRELIMINARIES

Notation. Let $\mathbb{R}$ represent the set of real numbers, $\mathbb{Z}$ the set of integers, $\mathbb{Z}_0^+$ the set of non-negative integers, and $\mathbb{N}^+$ the set of positive integers. For any arbitrary set $\mathbb{Y} \neq \emptyset$, let $\mathbb{Y}^D$ be its Cartesian product $\mathbb{Y} \times ... \times \mathbb{Y}$ repeated D-times where $D \geq 1, D \in \mathbb{N}$. Let $\mathbb{T}^D := [a, b]^D$ represent the D-dimensional torus, or the circle with $D = 1$. Throughout this paper, D represents the spatial dimension and $R \in \mathbb{Z}_0^+$ is the maximum *refinement*. The refinement may be interpreted as the multidimensional set of integers that support the the fundamental frequency in the Fourier series expansion for each dimension. In the next section we introduce univariate Fourier series with an illustrative example for deriving an expansion for a univariate periodic kernel.

3.1 UNIVARIATE FOURIER SERIES FOR KERNELS

We demonstrate first how one constructs a stationary periodic kernel with its corresponding Fourier series decomposition. This step is crucial because it is required to obtain Fourier series coefficients corresponding to the kernel being approximated. It is possible to construct a periodic kernel from any stationary kernel by applying the *warping* $\boldsymbol{u}(x) = [\cos(x), \sin(x)]$ to data x and then passing the result into any standard stationary kernel[26]. By performing the warping to a stationary kernel with the general squared distance metric $\|x - x'\|^2$ and replacing x with $\boldsymbol{u}(x)$ we have:

$$\begin{aligned} &\|\boldsymbol{u}(x) - \boldsymbol{u}(x')\|^2 \\ &= (\sin(x) - \sin(x'))^2 + (\cos(x) - \cos(x'))^2 \\ &= 2(1 - \cos(x - x')). \end{aligned} \tag{1}$$

Example. Consider the well known Squared Exponential (SE) kernel [36] $\kappa_{\text{SE}}(\boldsymbol{x} - \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2l^2}\right)$ with lengthscale l. After performing (1) we recover the periodic SE kernel: $\kappa_{\text{perSE}}(x, x') = \exp\left(-\frac{\cos(w\tau)-1}{l^2}\right)$, where $\tau = x - x'$ and ω is the fundamental periodic frequency.

Firstly, κ_{perSE} is both periodic and symmetric over τ. Since it is periodic, the kernel can be represented as a

Fourier Series over the interval $[-L, L]$ where L is the half period $\omega = \frac{\pi}{L}$ is the fundamental frequency. Note the Fourier series representation of some function:

$$f(t) \approx F_k[f(t)] = \sum_{k=-\infty}^{\infty} \mathbf{c}_k e^{ik\omega t}, \tag{2}$$

with coefficients

$$\mathbf{c}_0 = \frac{1}{2L}\int_{-L}^{L} f(t)dt, \tag{3}$$

$$\mathbf{c}_k = \frac{1}{2L}\int_{-L}^{L} f(t)e^{-ik\omega t}dt, \quad \forall k \in \mathbb{N}^+. \tag{4}$$

This reduces to a series of cosines from (2) for even functions, such as stationary periodic kernels. The Fourier series is defined at integer multiples k of the fundamental periodic frequency ω where $k \in \mathbb{N}^+$. To find the k^{th} coefficient $\mathbf{c}_k$, we evaluate the integral:

$$\begin{aligned}\mathbf{c}_k &= \frac{1}{2L}\int_{-L}^{L} e^{l^{-2}(\cos(\omega\tau)-1)}e^{-ik\omega\tau}d\tau \\ &= \frac{e^{-l^{-2}}}{2L}\int_{-L}^{L} e^{l^{-2}(\cos(\omega\tau))}\cos(k\omega\tau)d\tau \\ &= \frac{2\pi I_k(l^{-2})}{e^{l^{-2}}},\end{aligned} \tag{5}$$

using substitution $\omega = \frac{\pi}{L}$, $L = \pi$, where $I_n(z)$ is the Modified Bessel function of the first kind of integer order n and argument z. We obtain the solution using the special function identity $I_n(z) = \frac{1}{\pi}\int_0^{\pi} e^{z\cos(\theta)}\cos(n\theta)d\theta$ [1] which collapses the integral.

We now have a representation of the kernel as an infinite Fourier series $\kappa(\tau) \approx F_k[\kappa(\tau)]$:

$$\kappa_{perSE}(\tau) = F_k[\kappa(\tau)] = \sum_{k=-\infty}^{\infty} \frac{I_k(l^{-2})}{\exp(l^{-2})}\cos(k\omega\tau). \tag{6}$$

Thus, for the periodic SE kernel, we have Fourier series feature coefficients q_k^2,

$$q_k^2 = \begin{cases} \frac{I_k(l^{-2})}{\exp(l^{-2})} & \text{if } k = 0, \\ \frac{2I_k(l^{-2})}{\exp(l^{-2})} & \text{if } k = 1, 2, ..., K, \end{cases} \tag{7}$$

where K is the truncation factor of the Fourier series, I is the modified Bessel function of the first kind. These coefficients q_k^2 are used on a per-dimension basis for the multivariate feature construction.

3.2 FOURIER SERIES IN MULTIPLE DIMENSIONS

Our goal is to represent multi-dimensional periodic kernels. In the space of the full kernel, such a composition can be represented as a product of D independent periodic kernels on each dimension since is known that product compositions in the space of the kernel have an equivalent cartesian product operation in the feature space [39]. We have various results from harmonic theory on weighted subspaces of the Wiener algebra [3, 20] which allow us to use sparse sampling grids (i.e. index sets) to efficiently represent multivariate periodic kernels. That is to say, if we have functions with Fourier series coefficients that decay sufficiently fast, we can obtain sufficiently accurate approximations with vastly less terms.

Consider a sufficiently smooth multivariate periodic function $f : \mathbb{T}^D \to \mathbb{C}, D \in \mathbb{N}^+$. A function f can formally be represented as its multivariate Fourier series:

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^D} \hat{f}_{\boldsymbol{k}} e^{2\pi j\boldsymbol{k}\cdot\boldsymbol{x}}, \tag{8}$$

with its Fourier series coefficients $\hat{f}_{\boldsymbol{k}}, \boldsymbol{k} \in \mathbb{Z}^D$ defined as $\hat{f}_{\boldsymbol{k}} := \int_{\mathbb{T}^D} f(\boldsymbol{x})e^{2\pi j\boldsymbol{k}\cdot\boldsymbol{x}}d\boldsymbol{x}$. In essence this results in a tensor product of univariate Fourier series. Let $\Pi_{\mathcal{I}}$ denote the space of all multivariate trigonometric polynomials supported on some arbitrary *index set* $\mathcal{I} \subset \mathbb{Z}^D$, which is a finite set of integer vectors. We denote the *cardinality* of the set $\mathcal{I}$ as $|\mathcal{I}|$. Practically, we are interested in a good approximating Fourier partial sum $S_{\mathcal{I}}f \in \Pi_{\mathcal{I}}$ supported on some suitable index set $\mathcal{I}$. One may think of index sets as a multi-dimensional indicator variable. In order to construct such an index set we must have a construction rule or *weight function* w which tells us which index set coordinates to discard.

More formally, we define weighted function spaces of the Wiener algebra: $\mathcal{A}_w(\mathbb{T}^D) := \{f \in L_1(\mathbb{T}^D) : \sum_{\boldsymbol{k}\in\mathbb{Z}^D} \hat{f}_{\boldsymbol{k}} e^{2\pi j\boldsymbol{k}\cdot\boldsymbol{x}}, \sum_{\boldsymbol{k}\in\mathbb{Z}^D} w(\boldsymbol{k})\left|\hat{f}_{\boldsymbol{k}}\right| < \infty\}$ where we assume the function f is in the function space $L_p(\mathbb{T}^D) := \{f : \mathbb{T}^D \to \mathbb{C} \int_{\mathbb{T}^D} \left|f(\boldsymbol{x})\right|^p d\boldsymbol{x} < \infty\}$ with $1 \leq p \leq \infty$, and we define a *weight function* $w : \mathbb{Z}^D \to [1, \infty]$, which characterises the decay of Fourier coefficients $\hat{f}_{\boldsymbol{k}}$ of all functions $f \in \mathcal{A}_w(\mathbb{T}^D)$ such that $\hat{f}_{\boldsymbol{k}}$ decreases faster than the weight function in terms of $\boldsymbol{k}$. That is to say that the decay of the coefficients defines the smoothness of the function f we are approximating with a partial sum. In the next section we will introduce explicit index sets with their corresponding weight function.

3.3 INDEX SETS

While a naive multivariate Fourier series expansion of a univariate Fourier series appears plausible, for explicit tensor products the cardinality grows exponentially fast in dimension D and is therefore computationally intractable in terms of cardinality of the supporting index set. This computational burden is amplified when we consider the

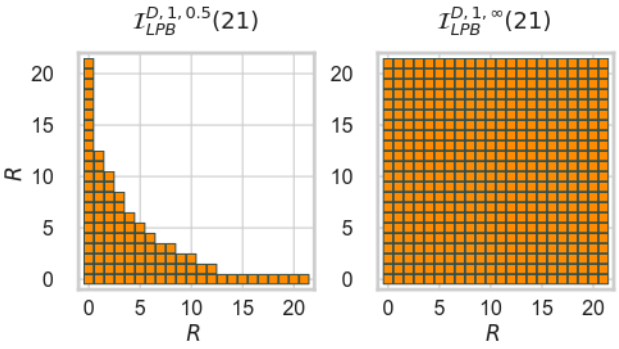

Figure 1: Visualisation of two common instances of the LPB index set in $D = 2$. The left image depicts a sparser index set while the right image depicts a dense tensor index set. Each solid square represents an index set coordinate $\mathcal{I}_i = [r_1, r_2]$ for integer refinement r.

expanded representation required for the separable feature decomposition when the products of harmonic terms of the Fourier series themselves must be expanded into sums of cosines. It would thus be desirable to; i) maintain high function approximation accuracy; and ii) minimise the total number of coefficients.

To this end we introduce a variety of D-dimensional weighted *index sets* $\mathcal{I}$ with their formal definitions. At a high level, one may think of an index set as a generalisation of indicator variables for multi-dimensional tensors which mask only the most important frequencies for the multivariate Fourier series decomposition. The reason index sets are useful is because it is often unnecessary to completely expand all supporting integers due to exponentially decaying coefficients of the function one is approximating. For instance, in Figure 1, we can see two index sets. On the right is the full *tensor* product index set which is dense in the refinement R, while the left index set has significantly fewer components. If the function we are trying to approximate has coefficients that decay sufficiently fast under the coverage of the sparser index set we make a significant saving in the number of terms needed to represent the function. There are various explicit index sets [51, 15, 43, 38], and the first set we introduce is the *l_p-ball* (LPB), $0 < p \leq \infty$ index set $\mathcal{I}^{D,\gamma,p=1}_{\text{LPB}}(R)$. This is a generalised index set and it is constructed with the following weight function,

$$w^{D,\gamma}_{\text{LPB}}(\boldsymbol{k}) = \max\left(1, \|\boldsymbol{k}|l^{D,\gamma}_p\|\right), \quad \text{for } 0 < p \leq \infty \tag{9}$$

with construction parameter $\boldsymbol{\gamma} = (\gamma_d)_{d=1}^{\infty} \in [0,1]^{\mathbb{N}^+}$ controlling the approximation depth for a given dimension d, and where,

$$\|\boldsymbol{k}|l^{D,\gamma}_p\| = \begin{cases} \left(\sum_{d=1}^{D} \left(\gamma_d^{-1} |k_d|\right)^p\right)^{1/p} & \text{for } 0 < p < \infty, \\ \max_{d=1,\ldots,D} \gamma_d^{-1} |k_d| & \text{for } p = \infty. \end{cases} \tag{10}$$

We also have the Energy Norm Hyperbolic Cross (ENHC) index set $\mathcal{I}^{D,\gamma,\zeta}_{\text{ENHC}}(R)$with sparsity parameter $\zeta \in [0, 1)$. It has weight function,

$$w^{D,\gamma,\zeta}_{\text{ENHC}}(\boldsymbol{k}) = \max(1, \|\boldsymbol{k}\|_1)^{\frac{\zeta}{\zeta-1}} \prod_{d=1}^{D} \max(1, \gamma_d^{-1} |k_d|)^{\frac{1}{1-\zeta}}. \tag{11}$$

The ENHC is more suitable for approximating functions of dominating mixed smoothness.

4 INDEX SET FOURIER SERIES FEATURES

The goal of our work is to show how multivariate Fourier series representations of kernels with sparse approximation lattices allow efficient and deterministic feature decompositions of multivariate periodic kernels. Formally, we define the shift invariant multivariate periodic kernel approximation as a Fourier series expansion supported on an arbitrary index set $\mathcal{I}$:

$$\kappa_{\text{per}}(\boldsymbol{x}, \boldsymbol{x}') \approx \sum_{\boldsymbol{k} \in \mathcal{I}} \hat{f}_{\boldsymbol{k}} e^{2\pi j \boldsymbol{k} \cdot (\boldsymbol{x} - \boldsymbol{x}')} = \langle \hat{\Phi}(\boldsymbol{x}), \hat{\Phi}(\boldsymbol{x}') \rangle_{\mathbb{C}^M}, \tag{12}$$

for some explicit feature map $\hat{\Phi}$ and multivariate Fourier series coefficients $\hat{f}_{\boldsymbol{k}}$.

We now continue with our feature construction which we term Index Set Fourier Series Features (ISFSF) and introduce an additionally sparse construction feature count for no loss of accuracy.

4.1 ISFSF FEATURE CONSTRUCTION

This section presents our main contribution for approximating multi-dimensional periodic kernels. We have seen that simply using multivariate Fourier series is not sufficient for tractable decomposition due to an exponential tensor product in the refinement level. Using *index sets* for multivariate Fourier series we present a feature construction using frequency grids, and noting that the resulting feature admits an additionally sparse construction.

We can write the general form of the product expansion for any particular i^{th} index set coordinate $\mathcal{I}_i$ from some index set $\mathcal{I}$ as:

$$\varrho(\mathcal{I}_i) = \prod_{d=1}^{D} q^2_{r_d} \cos(r_d \omega_d (x_d - x'_d)), \tag{13}$$

Algorithm 1: ISFSF feature construction

Input : $\mathcal{I} \in \mathbb{Z}^D$ frequency index set, $|\mathcal{I}| < \infty$
$C = |\mathcal{I}|$ set cardinality (Lemma 1, 2)
$J =$ number of rows in $\boldsymbol{\Xi}$
$\boldsymbol{x} \in \mathbb{R}^D$ raw data to embed into features
$\boldsymbol{\Xi} \in \mathbb{R}^{J \times D}$ cartesian product sign matrix
$\boldsymbol{\omega} \in \mathbb{R}^D$ fundamental frequencies

Initialize: $\hat{\Phi}_{\mathcal{I}} \in \mathbb{R}^{2CJ}$

for $i := 1, ..., C$: **do**
  set $\mathcal{I}_i$ as the i^{th} set coordinate
  $\rho_i = \prod_{d=1}^{D} q_{r_d}$
  for $j := 1, ..., J$: **do**
    set $\boldsymbol{\Xi}_j$ as j^{th} row of $\boldsymbol{\Xi}$
    $\boldsymbol{r}_i = \big[r_1, r_2, ..., r_D\big]$
    prod $= (\boldsymbol{r}_i \odot \boldsymbol{\omega} \odot \boldsymbol{\Xi}_j)\boldsymbol{x}^T$
    append $(\hat{\Phi}_{\mathcal{I}}, \sqrt{\frac{\rho_i}{J}}[\cos(\text{prod}), \sin(\text{prod})])$
  end
end

Output : $\hat{\Phi}_{\mathcal{I}}$

where ω_d is the d^{th} dimension's fundamental frequency, $\mathcal{I}_i$ is the i^{th} index set coordinate, and r_d is the d^{th} dimension integer refinement for a given $\mathcal{I}_i$. For our feature expansion we are interested in the data-dependent trigonometric term made up of a product of data-dependent cosines. It is these that allow us to decompose the series into a sum of cosines. For this we require the product of cosines trigonometric identity $\cos(u)\cos(v) = \frac{1}{2}[\cos(u-v) + \cos(u+v)]$. Applying this identity *recursively* to (13), we obtain the following decomposable form:

$$\varrho(\mathcal{I}_i) = \frac{1}{J}\sum_{j=1}^{J} \rho_i \cos\big((\boldsymbol{r} \odot \boldsymbol{\omega} \odot \boldsymbol{\Xi}_j)\boldsymbol{\Delta}^T\big), \tag{14}$$

with

$$\rho_i = \prod_{d=1}^{D} q_{r_d}^2, \tag{15}$$

$$\boldsymbol{r} = \big[r_1, r_2, ..., r_D\big], \tag{16}$$

$$\boldsymbol{\omega} = \big[\omega_1, \omega_2, ..., \omega_D\big], \tag{17}$$

$$\boldsymbol{\Xi} = \big[+\mathbf{1}^\frown(+1,-1)^{(D-1)}\big], \tag{18}$$

$$\boldsymbol{\Delta} = \big[x_1 - x_1', x_2 - x_2', ..., x_D - x_D'\big], \tag{19}$$

where $J = 2^{(D-1)}$ is the number of rows in $\boldsymbol{\Xi}$, ρ_i refers to the product of per-dimension Fourier series coefficients corresponding to a given $\mathcal{I}_i$, $(+1,-1)^{(D-1)}$ is the $(D-1)$-times Cartesian combination of the ordered integer set $(+1,-1)$, $U^\frown\Lambda$ denotes a horizontal concatenation between matrix U of length $|\Lambda|$ and every element in some ordered set Λ, and $\odot$ refers to the element-wise (Hadamard) product. To clarify $\boldsymbol{\Xi}$, observe how a two dimensional cosine product $\cos(A)\cos(B)$ expands to $\cos(A+B) + \cos(A-B)$ giving $\boldsymbol{\Xi} = \begin{bmatrix} +1 & +1 \\ +1 & -1 \end{bmatrix}$. In three dimensions one obtains $\cos(A+B+C) + \cos(A+B-C) + \cos(A-B+C) + \cos(A-B-C)$ giving $\boldsymbol{\Xi} = \begin{bmatrix} +1 & +1 & +1 \\ +1 & +1 & -1 \\ +1 & -1 & +1 \\ +1 & -1 & -1 \end{bmatrix}$. Noting equation (12) and using the relation $e^{-j\boldsymbol{\tau}\cdot\boldsymbol{\omega}_k} = \cos(\omega \cdot \boldsymbol{\tau}) - j\sin(\omega \cdot \boldsymbol{\tau})$ and the fact that real kernels have no imaginary part, we can exploit the cosine difference of angles identity $\cos(u-v) = \cos(u)\cos(v) + \sin(u)\sin(v)$ to obtain the complete decomposed feature as:

$$\hat{\Phi}_{\mathcal{I}}^{\text{full}}(\boldsymbol{x}) = \Big[\sqrt{\frac{\rho_i}{J}}\cos\big((\boldsymbol{r}_i \odot \boldsymbol{\omega} \odot \boldsymbol{\Xi}_j)\boldsymbol{x}^T\big), \sqrt{\frac{\rho_i}{J}}\sin\big((\boldsymbol{r}_i \odot \boldsymbol{\omega} \odot \boldsymbol{\Xi}_j)\boldsymbol{x}^T\big)\Big]_{i=1,j=1}^{C,J}, \tag{20}$$

where $\boldsymbol{x} = [x_1, x_2, .., x_D]$.

The feature construction algorithm is depicted explicitly in Algorithm 1 and consists of two loops that iterate over the index set coordinates $\mathcal{I}_i$ and each row of the cartesian combination sign matrix $\boldsymbol{\Xi}$. The construction is embarrassingly parallelizable and is straightforward to implement.

It is useful to determine computational budget of the feature map and for this it is necessary to determine the number of features in the expanded feature map. To do this we can give the cardinality $C_{\mathcal{I}}$ of the resulting *decomposed* feature map $\hat{\Phi}_{\mathcal{I}}^{\text{full}}(\boldsymbol{x})$ over index set $\mathcal{I}$.

Lemma 1. *Cardinality of Index Set Fourier series feature map for some arbitrary index set $\mathcal{I}(R)$. Let $C_{\mathcal{I}} = \big|\mathcal{I}(R)\big|$ be the cardinality of some given index set of refinement R, and let the dimension $D \in \mathbb{N}^+$ be given. Let $C_{\hat{\Phi}}^{full} = \Big|\hat{\Phi}_{\mathcal{I}}^{full}(\boldsymbol{x})\Big|$ denote the cardinality of the decomposed feature. The following holds (see supplementary for proof):*

$$\big|\mathcal{I}(R)\big| \leq \Big|\hat{\Phi}_{\mathcal{I}}^{full}(\boldsymbol{x})\Big| \leq C_{\mathcal{I}} 2^D.$$

4.2 SPARSE CONSTRUCTION

Although suitable, the decomposable form for the multivariate Fourier series features can be improved. An ideal feature representation should not just approximate our kernel *well* but should do it *efficiently*. For ISFSF, this involves the data-dependent term $\cos(\cdot)$, occurrences of which we want to minimise. Scrutinising the product form in (13), the term r_d is an integer $r \in \mathbb{Z}_0^+$ which clearly contains the value 0. This means that for *all* refinement coordinates at the 0^{th} refinement for any dimension we have

$\cos(0\cdot) = 1$, therefore the "feature" is simply 1 times some data-*independent* coefficient. Furthermore, since any *single* $\cos(\cdot)$ term in the product (13) contributes to a multiplier of 2 features before exponentiation due to the trigonometric product identity recursion, we do not unnecessarily want to include features that will simply evaluate to a constant. We now define a *masking* function $\varkappa$ over some function $g(r)$ with $r \in \mathbb{Z}_0^+$:

$$\varkappa\Big(g(r)\Big) = \begin{cases} 1 & \text{for } r = 0, \\ g(r) & \text{otherwise.} \end{cases} \tag{21}$$

This mask acts to identify which redundant harmonic terms to ignore in the feature construction stage. Continuing the decomposition as in the previous section, the sparse feature decomposition is thus:

$$\hat{\Phi}_{\mathcal{I}}^{\text{sparse}}(\boldsymbol{x}) = \Big[\sqrt{\frac{\rho_i}{J}}\varkappa\big(\cos\big((\boldsymbol{r}_i \odot \boldsymbol{\omega} \odot \boldsymbol{\Xi}_j)\boldsymbol{x}^T\big)\big), \sqrt{\frac{\rho_i}{J}}\varkappa\big(\sin\big((\boldsymbol{r}_i \odot \boldsymbol{\omega} \odot \boldsymbol{\Xi}_j)\boldsymbol{x}^T\big)\big)\Big]_{i=1,j=1}^{C,J}, \tag{22}$$

We now give an improved cardinality for the decomposed sparse ISFSF feature map. We emphasise this improved feature map cardinality is for *exactly the same reconstructing* accuracy. To determine the cardinality of the sparse feature map, let:

$$\eta(\mathcal{I}_i) = \sum_{d=1}^{D} \big[\mathcal{I}_i \neq 0\big], \tag{23}$$

define a function that counts for a particular coordinate $\mathcal{I}_i$ the non-zero indexes. Essentially, this counts which $\cos(\cdot)$ terms to keep, which always occur at coordinates with per-dimension refinement not equal to 0.

Lemma 2. *Cardinality of sparse Index Set feature map for arbitrary index set $\mathcal{I}(R)$. Let $C_{\mathcal{I}} = \big|\mathcal{I}(R)\big|$ be the cardinality of some given index set of refinement R, and let the dimension $D \in \mathbb{N}^+$ be given. Let $C_{\hat{\Phi}}^{sparse} = \Big|\hat{\Phi}_{\mathcal{I}}^{sparse}(\boldsymbol{x})\Big|$ then denote the cardinality of the decomposed index set Fourier series feature. The following holds (see supplementary for proof):*

$$\big|\mathcal{I}(R)\big| \leq \Big|\hat{\Phi}_{\mathcal{I}}^{sparse}(\boldsymbol{x})\Big| = \sum_{i=1}^{|\mathcal{I}|} 2^{\eta(\mathcal{I}_i)} \leq \Big|\hat{\Phi}_{\mathcal{I}}^{full}(\boldsymbol{x})\Big| \leq C_{\mathcal{I}} 2^D.$$

4.3 MULTIVARIATE TRUNCATION ERROR

To better understand the effect of Fourier series approximations on kernels we can analyse the truncation error as a function of kernel hyperparameters. We know that the univariate truncation error [41] for $k \in \mathbb{N}^+$ is $\big|\cos(\omega k \tau)\big| \leq 1$, and $\sum_{k=0}^{\infty} q_k^2 = 1$, since the sum of the coefficients converges to 1. We extend this to the multivariate case by considering the tensor index set product expansion. We have $\prod_{d=1}^{D} \big|\cos(\omega_d r_d \tau_d)\big| \leq 1$. Since $\max(\kappa(\boldsymbol{\tau})) = 1$ we obtain the multivariate truncation error:

$$\epsilon(R, l) = 1 - \prod_{d=1}^{D} \Big[\sum_{r=0}^{R-1} q_{r_d}^2\Big], \tag{24}$$

where R is the refinement, l is the kernel lengthscale. q_{r_d} refers to the approximated kernel's Fourier coefficient at refinement index r, with subscript d referring to evaluation on the d^{th} dimension. A visualization of

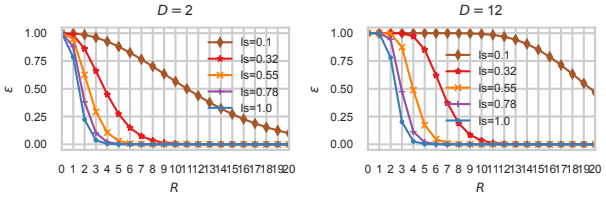

Figure 2: Visualisation of the multivariate truncation error, for the periodic SE kernel, for refinements $R = [0, 20]$, dimensions $D = \{2, 12\}$ and isotropic lengthscales $ls = \{0.1, 1.0\}$.

the truncation error for different dimensions, for the periodic Squared Exponential kernel, is presented in Figure 2. An intuitive explanation is, that with the same refinement factor (and therefore the same number of features), larger lengthscales are easier to approximate than smaller lengthscales. This can be understood from the frequency domain perspective in which a kernel with the smaller lengthscale has a larger spectrum of frequencies. Our kernel Gram approximation as well as downstream error corroborate these results - data that requires larger lengthscales converge in predictive error faster with fewer features.

5 EXPERIMENTS

We evaluate three aspects of our proposed periodic kernel approximation with the ENHC index set. First, we compare our approximation with the full kernel in terms of Gram matrix reconstruction. Second, we compare our features to state of the art approximations on large scale real world textured image data. Next, we perform a comparison of predictive qualities against an analytic periodic function in higher dimensions. Finally, we demonstrate the kernel on predicting a periodic trajectory of various walking robots used commonly in Reinforcement Learning and Control tasks.

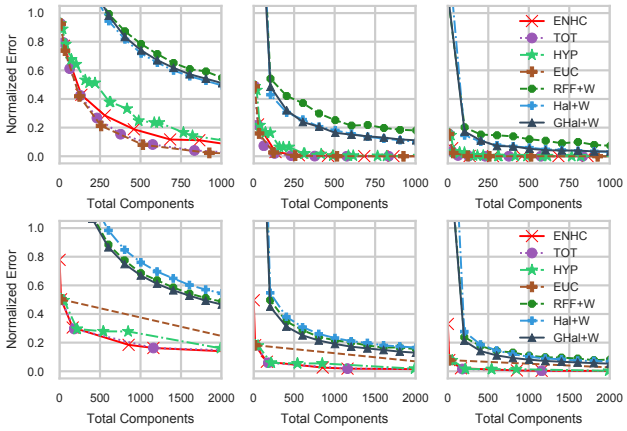

Figure 3: Reconstruction error on simulated data with the normalised Frobenius error between Fourier Feature methods (RFF, Hal, GHal) with periodic warpings, and Index Set Fourier Series with various index sets. **Row 1**: $D = 3$, $ls = \{0.5, 1.0, 1.5\}$, **Row 2**: $D = 9, ls = \{1.5, 2.0, 2.5\}$. ENHC with weighting $\gamma = \frac{2}{3}$ for $D = 9$.

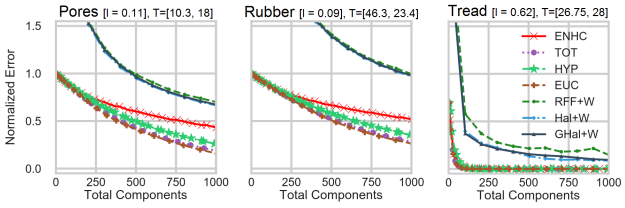

Figure 4: Reconstruction error on real texture datasets using learned hyperparameters. Note how smaller lengthscales in Pores and Rubber require more features than larger lengthscales in Tread.

5.1 QUALITY OF KERNEL APPROXIMATION

We first analyse the proposed feature in terms of the reconstruction error between a true Gram matrix $\mathbf{K}$, using the analytic periodic kernel, and its approximated Gram matrix $\tilde{\mathbf{K}}_{i,j} = \hat{\kappa}(x_i, y_j)$. For all comparisons the metric we use is the normalized Frobenius error $\frac{\|\tilde{\mathbf{K}}-\mathbf{K}\|_F}{\|\mathbf{K}\|_F}$ using $N = 4000$ uniformly drawn samples from $[-2, 2]$. The primary comparison in Figure 3 compares the effects of various index set constructions, RFFs, QMC (Halton, Generalised Halton), and the following index sets: Energy Norm Hyperbolic Cross (ENHC), Total Order (TOT), Hyperbolic (HYP), and Euclidean (EUC). The supplementary contains an extended comparison of index set parameters, dimensionality, and nuances of the reconstruction. The first observation we can make from Figure 3 is that for lower dimensions $D \approx 3$ the best performing features are those with the Euclidean degree or Total order index sets. Of the index sets the Hyperbolic and Energy Norm Hyperbolic Cross perform the worst in particular for smaller lengthscales. Overall the index sets all perform significantly better than the warped Fourier Feature methods, amongst which, the original MC based RFF performs the worst and the standard Halton sequence appears to perform marginally better than the generalised Halton.

As the number of dimensions increases the Total and Euclidean index sets become intractable due to their heavy dependency on cross-dimensional products of data-dependent harmonic terms. Considering FF methods, the approximation accuracy of the standard Halton sequence falls behind even RFFs while the generalised Halton remains consistently ahead. As we have seen, as the dimensionality increases, the Total and Euclidean index sets have no parameterisation that allows them to scale properly; indeed their flexibility is due entirely being a specific instance of the LPBall index set which can give sparser *Hyperbolic* index sets. On the other hand, the ENHC can be parameterised by sparsity ζ and weighting γ giving additional flexibility.

An interesting observation in the Frobenius norms of the real texture datasets is the errors and the connection between the truncation error defined in 24. We can see how the smaller lengthscales (Pores and Rubber) result in more difficult inference in terms of number of features required for better predictions, than larger lengthscales (Tread) - this can be seen in the generalisation experiments in the next section where we evaluate predictive error on an increasing range of features.

We conjecture that the significantly improved Gram matrix approximation performance of ISFSF is not just from our deterministic construction, but also due to a large suppression of *negative covariances* which we have observed in reconstruction plots. This is a known issue for RFFs with Gaussian Processes which can negatively affect predictive uncertainties. An extended discourse on this behaviour is provided in the supplementary.

5.2 GENERALISATION ERROR

We evaluate predictive performance with Root Mean Square Error (RMSE) and Mean Negative Log Loss (MNLL). The MNLL accounts for the model's predictive mean and uncertainty. It is defined as MNLL $= \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}\log(2\pi\sigma_i^*)) + \frac{(\mu_i^* - f_i^*)^2}{2\sigma_i^*}$ where σ_i^*, μ_i^*, and f_i^* are respectively the predictive standard deviation, predictive mean, and true value at the i^{th} test point. This section demonstrates the generalisation performance of a single multidimensional periodic kernel on image texture data. We use images from [48] with the same 12675 train and 4225 test set pixel locations $\boldsymbol{x} \in \mathbb{R}^2$. A Bayesian Linear

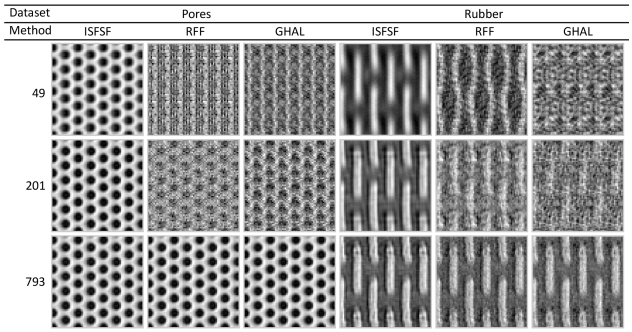

Figure 5: Predicted missing area for the *Pores* and *Rubber* datasets. Left to right, each column represents predictions made using ISFSF, RFF, and GHAL. Top to bottom, each row represents an increasing number of features used at $49, 201, 793$ respectively.

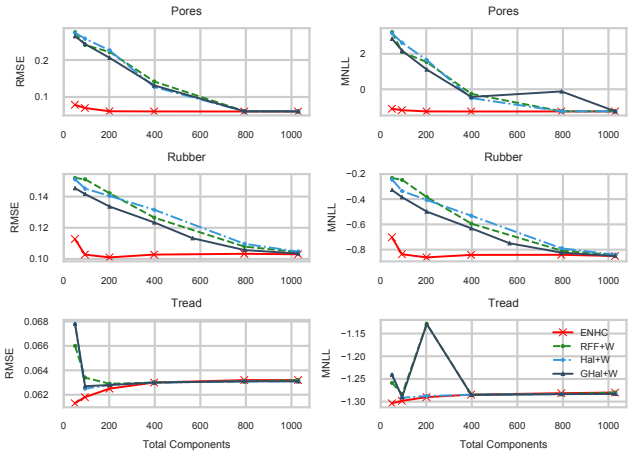

Figure 6: Comparison of predictive RMSE and MNLL with our method alongside Fourier Feature methods, for increasing number of components.

Regression [2] model is used as the regression model. We fix the hyperparameters across different kernel representations in order for the comparison to be consistent for the same underlying kernel. Overall, for the same kernel, both qualitatively and quantitatively the results show clear advantages of ISFSF over the alternative feature methods. The visual effect is demonstrated in in Figure 7 and empirically in Figure 5.

Textures. In both the *pores* and *rubber* datasets we can see the RMSE and MNLL for the ISFSF based features (using the ENHC) perform the best in all cases. The RMSE performs exceedingly well even with only 49 features almost equaling the performance of RFF and QMC methods which require 794 features. In both RMSE and MNLL, the ISFSF with 201 features outperforms

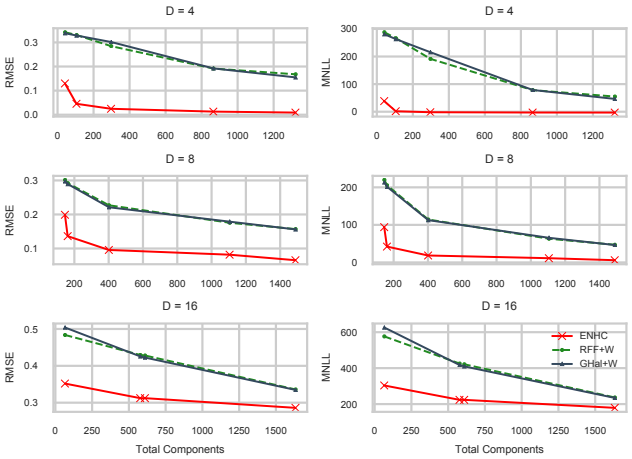

Figure 7: Higher dimensional comparison of predictive RMSE and MNLL with our method using the ENHC alongside Fourier Feature methods, for increasing number of components. We train on 8000 points from [-5,5] and test on 4000 points.

FF based methods using 794. In the pores dataset the generalised Halton marginally outperforms all methods in the case of 794 features, and the Halton at 94 features. For the *tread* dataset the resulting performance is interesting because for all features, the RMSE performances are alike across all methods with the ISFSF slightly outperforming for lower features. The asymptotic performance of both ISFSF and Fourier feature methods are similar and for rubber and tread become marginally worse as we increase the feature count. This is expected because the datasets contain small amounts of non-stationary information which the stationary periodic RBF is unable to completely capture. As the feature count increases, the modelling fidelity of the approximated kernel increases resulting in the slightly larger error.

High Dimensional Tensor Function. To analyse the method's efficacy in higher dimensions we perform experiments on the D dimensional tensor-product function $G^D(\boldsymbol{x}) := \prod_d^D g(x_d)$ from [20] where the one-dimensional function g is defined as

$$g(x) := 8\sqrt{6\pi/(6369\pi - 4096)}\big(4 + \text{sgn}(x \bmod 1) - \frac{1}{2}(\sin(2\pi x)^3 + \sin(2\pi x)^4)\big) \tag{25}$$

In this experiment we also observe improved performance of the ISFSF over standard Fourier Feature methods. Since we know the function is stationary, we observe a steady convergence in predictive accuracy unlike the non-stationary texture datasets. As we increase the dimensionality the predictive error degrades slightly; this is

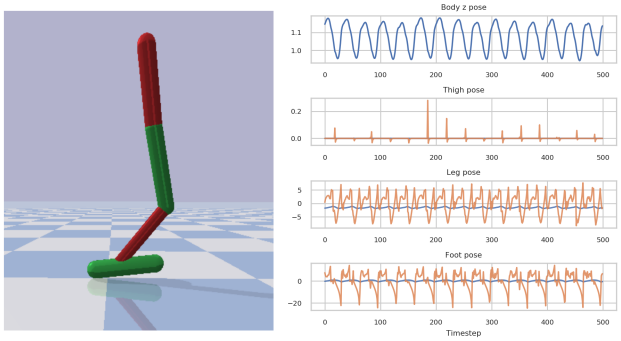

Figure 8: "Hopper" multi-jointed robot with visualisation of joint and robot trajectories.

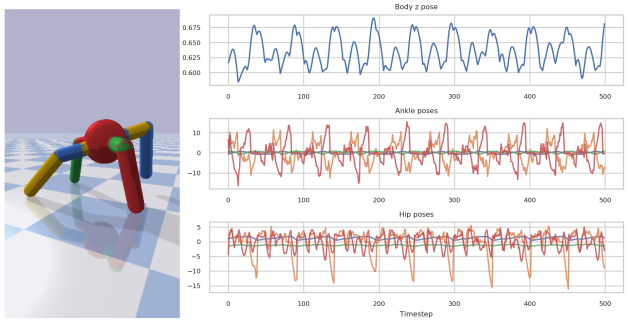

Figure 9: "Ant" multi-jointed robot with visualisation of joint and robot trajectories.

expected when we consider multivariate truncation error which is affected by dimension.

Periodic trajectory tracking of jointed robots. Robotics is an area in which various hardware platforms of interest are often constructed with jointed actuators. We demonstrate an application of multivariate periodic kernels for periodic motion tracking in two simulated jointed robots commonly used in Reinforcement Learning (RL) and Control tasks [10]. We consider the problem of regressing on the vertical trajectory of the robot as a function of the input joints. 500 timesteps of trajectory were collected alongside a 6 ("Hopper", Figure 8) and 16 ("Ant", Figure 9) dimensional position and orientation vector of the joints from a pre-learned policy for the "Hopper" and "Ant" environments [7]. 300 steps were used for training. Simulation was performed in the open source simulator PyBullet. The results are visualized in Figure 10. For both "Hopper" and "Ant" robots, we can see a significant improvement of the ISFSF features over the standard periodic kernel formulation. This is most significant in the RMSE, and holds in the MNLL. The early convergence of predictions suggest that ISFSFs

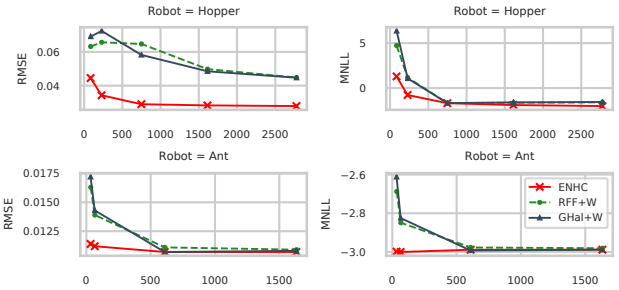

Figure 10: Periodic trajectory prediction error for increasing number of features on various robots.

are both better able to represent the signal as well as the uncertainty of the prediction, as evidenced in the MNLL. An explanation for the faster convergence of the "Ant" task for both feature methods could be observed in the lengthscale of the kernels used in both - "Ant" uses an isotropic lengthscale of 13 and "Hopper" an isotropic lengthscale of 0.5. These results suggest that ISFSF may be used to improve policy learning in RL, as well as play a role in improved system identification and prediction of periodic systems.

6 CONCLUSION

Feature approximations have been a large component of scaling kernel methods such as GPs. An important issue with kernel approximation methods is their efficiency in their approximation and has been a focus of more deterministic construction methods. Having effective periodic kernel approximations enables numerous applications in various domains in a much more efficient manner. Crucially, we introduce effective sparse approximation to multivariate periodic kernels using multivariate Fourier series with sparse index set based sampling grids for efficient feature space periodic kernel decompositions. We demonstrate experimentally on a range of datasets that our features result in predictive models of greater accuracy with vastly less components. Future directions include how to construct kernel-dependent index sets as well as direct application to learning policies for jointed robotic systems.

Acknowledgements

We are grateful to Richard Scalzo for early meandering discussions and Lionel Ott for valuable guidance. We would also like to thank the anonymous reviewers for their comments and helpful suggestions.

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, book section "Modified Bessel Functions I and K." §9.6. 1972.

[2] C. Bishop. Pattern recognition and machine learning (information science and statistics). *Springer, New York*, 2007.

[3] F. F. Bonsall and J. Duncan. *Complete normed algebras*. Springer Science & Business Media, 2012.

[4] W. Boomsma and J. Frellsen. Spherical convolutions and their application in molecular modelling. In *Advances in Neural Information Processing Systems*, 2017.

[5] R. Brault, M. Heinonen, and F. Buc. Random fourier features for operator-valued kernels. In *Asian Conference on Machine Learning*, 2016.

[6] K. Choromanski, F. Fagan, C. Gouy-Pailler, A. Morvan, T. Sarlos, and J. Atif. Triplespin-a generic compact paradigm for fast machine learning computations. *arXiv preprint arXiv:1605.09046*, 2016.

[7] E. Coumans and Y. Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. *GitHub repository*, 2016.

[8] T. Dao, C. M. De Sa, and C. Ré. Gaussian quadrature for kernel features. In *Advances in neural information processing systems*, 2017.

[9] J. C. Dunlap, J. J. Loros, and P. J. DeCoursey. *Chronobiology: biological timekeeping*. Sinauer Associates, 2004.

[10] T. Erez, Y. Tassa, and E. Todorov. Infinite horizon model predictive control for nonlinear periodic tasks. *Manuscript under review*, 4, 2011.

[11] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento. Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry*, 2015.

[12] 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*, 2016.

[13] A. Gijsberts and G. Metta. Real-time model learning using incremental sparse spectrum gaussian process regression. *Neural Networks*, 2013.

[14] A. Gittens and M. W. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. *Journal of Machine Learning Research*, 2013.

[15] K. Hallatschek. Fouriertransformation auf dünnen Gittern mit hierarchischen basen. *Numerische Mathematik*, 1992.

[16] G. W. Henry and J. N. Winn. The rotation period of the planet-hosting star hd 189733. *The Astronomical Journal*, 2008.

[17] J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In *Uncertainty in Artificial Intelligence*. Citeseer, 2013.

[18] C. Kacwin, J. Oettershagen, M. Ullrich, and T. Ullrich. Numerical performance of optimized frolov lattices in tensor product reproducing kernel sobolev spaces. *arXiv preprint arXiv:1802.08666*, 2018.

[19] L. Kämmerer. Multiple rank-1 lattices as sampling schemes for multivariate trigonometric polynomials. *Journal of Fourier Analysis and Applications*, 2018.

[20] L. Kämmerer, D. Potts, and T. Volkmer. Approximation of multivariate periodic functions by trigonometric polynomials based on rank-1 lattice sampling. *Journal of Complexity*, 2015.

[21] P. Kar and H. Karnick. Random feature maps for dot product kernels. In *International Conference on Artificial Intelligence and Statistics*, 2012.

[22] P. Kritzer, F. Y. Kuo, D. Nuyens, and M. Ullrich. Lattice rules with random n achieve nearly the optimal o (n- α- 1/ 2) error independently of the dimension. *Journal of Approximation Theory*, 2018.

[23] M. Lázaro-Gredilla, J. Quiñonero-Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal. Sparse spectrum gaussian process regression. *Journal of Machine Learning Research*, 2010.

[24] Q. Le, T. Sarlós, and A. Smola. Fastfood-approximating kernel expansions in loglinear time. In *Proceedings of the International Conference on Machine Learning*, 2013.

[25] F. Li, C. Ionescu, and C. Sminchisescu. Random fourier approximations for skewed multiplicative histogram kernels. In *Joint Pattern Recognition Symposium*. Springer, 2010.

[26] D. J. MacKay. Introduction to Gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 1998.

[27] M. Munkhoeva, Y. Kapushev, E. Burnaev, and I. Oseledets. Quadrature-based features for kernel approximation. In *Advances in Neural Information Processing Systems*, 2018.

[28] M. Mutny and A. Krause. Efficient high dimensional bayesian optimization with additivity and quadrature fourier features. In *Advances in Neural Information Processing Systems*, 2018.

[29] J. Pennington, X. Y. Felix, and S. Kumar. Spherical random features for polynomial kernels. In *Advances in Neural Information Processing Systems*, 2015.

[30] L. Peternel, T. Noda, T. Petrič, A. Ude, J. Morimoto, and J. Babič. Adaptive control of exoskeleton robots for periodic assistive behaviours based on emg feedback minimisation. *PloS one*, 2016.

[31] N. Pham and R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2013.

[32] C. L. Phillips and G. A. Voth. Discovering crystals using shape matching and machine learning. *Soft Matter*, 2013.

[33] A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 2007.

[34] A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems*, 2008.

[35] D. Rainville, C. Gagné, O. Teytaud, D. Laurendeau, et al. Evolutionary optimization of low-discrepancy sequences. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 2012.

[36] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[37] K. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. Müller, and E. Gross. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Physical Review B*, 2014.

[38] P. Seshadri, A. Narayan, and S. Mahadevan. Effectively subsampled quadratures for least squares polynomial approximations. *SIAM/ASA Journal on Uncertainty Quantification*, 2017.

[39] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[40] E. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. In *Advance in Neural Information Processing Systems*, 2006.

[41] A. Solin and S. Särkkä. Explicit link between periodic covariance functions and state space models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, 2014.

[42] A. Tompkins and F. Ramos. Fourier feature approximations for periodic kernels in time-series modelling. In *AAAI Conference on Artificial Intelligence*, 2018.

[43] L. Trefethen. Multivariate polynomial approximation in the hypercube. *Proceedings of the American Mathematical Society*, 2017.

[44] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE transactions on pattern analysis and machine intelligence*, 2012.

[45] L. A. Wasserman. *All of nonparametric statistics: with 52 illustrations*. Springer, 2006.

[46] C. K. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, 2001.

[47] A. Wilson and R. Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, 2013.

[48] A. G. Wilson, E. Gilboa, A. Nehorai, and J. P. Cunningham. Fast kernel learning for multidimensional pattern extrapolation. In *Advances in Neural Information Processing Systems*, 2014.

[49] J. Yang, V. Sindhwani, H. Avron, and M. Mahoney. Quasi-monte carlo feature maps for shift-invariant kernels. In *International Conference on Machine Learning*, 2014.

[50] Z. Yang, A. Wilson, A. Smola, and L. Song. A la carte–learning fast kernels. In *Artificial Intelligence and Statistics*, 2015.

[51] S. Zaremba. La méthode des "bons treillis" pour le calcul des intégrales multiples. In *Applications of number theory to numerical analysis*. Elsevier, 1972.