On Kernel Derivative Approximation with Random Fourier Features

Zoltán Szabó, Bharath K Sriperumbudur

On Kernel Derivative Approximation with Random Fourier Features

Zoltán Szabó[1] and Bharath K. Sriperumbudur[2]

[1]CMAP, École Polytechnique, Route de Saclay, 91128 Palaiseau, France, zoltan.szabo@polytechnique.edu

[2]Department of Statistics, Pennsylvania State University, 314 Thomas Building, University Park, PA 16802, bks18@psu.edu

Abstract

Random Fourier features (RFF) represent one of the most popular and wide-spread techniques in machine learning to scale up kernel algorithms. Despite the numerous successful applications of RFFs, unfortunately, quite little is understood theoretically on their optimality and limitations of their performance. To the best of our knowledge, the only existing areas where precise statistical-computational trade-offs have been established are approximation of kernel values, kernel ridge regression, and kernel principal component analysis. Our goal is to spark the investigation of optimality of RFF-based approximations in tasks involving not only function values but *derivatives*, which naturally lead to optimization problems with kernel derivatives. Particularly, in this paper, we focus on the approximation quality of RFFs for kernel derivatives and prove that the existing finite-sample guarantees can be improved exponentially in terms of the domain where they hold, using recent tools from unbounded empirical process theory. Our result implies that the same approximation guarantee is achievable for kernel derivatives using RFF as for kernel values.

1 INTRODUCTION

Kernel techniques [3, 30, 17] are among the most influential and widely-applied tools, with significant impact on virtually all areas of machine learning and statistics. Their versatility stems from the function class associated to a kernel called reproducing kernel Hilbert space (RKHS) [2] which shows tremendous success in modelling complex relations.

The key property that makes kernel methods computationally feasible and the optimization over RKHS tractable is the representer theorem [11, 25, 39]. Particularly, given samples $\{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathbb{R}$, consider the regularized empirical risk minimization problem specified by a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the associated RKHS $\mathcal{H}_k \subset \mathbb{R}^{\mathcal{X}}$, a loss function $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{\geq 0}$, and a penalty parameter $\lambda > 0$:

$$\min_{f \in \mathcal{H}_k} J_0(f) := \frac{1}{n} \sum_{i=1}^{n} V(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}_k}^2, \tag{1}$$

where $\mathcal{H}_k$ is the Hilbert space defined by the following two properties:
1. $k(\cdot, x) \in \mathcal{H}_k$ $(\forall x \in \mathcal{X})$,[1] and
2. $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k}$ $(\forall x \in \mathcal{X}, \forall f \in \mathcal{H}_k)$, which is called the reproducing property.

[1] $k(\cdot, x)$ denotes the function $y \in \mathcal{X} \mapsto k(y, x) \in \mathbb{R}$ while keeping $x \in \mathcal{X}$ fixed.

Examples falling under (1) include e.g., kernel ridge regression with the squared loss or soft-classification with the hinge loss:

$$V(f(x_i), y_i) = (f(x_i) - y_i)^2, \qquad V(f(x_i), y_i) = \max(1 - y_i f(x_i), 0).$$

(1) is an optimization problem over a function class ($\mathcal{H}_k$) which could generally be intractable. Thanks to the specific structure of RKHS, however, the representer theorem enables one to parameterize the optimal solution of (1) by finitely many coefficients:

$$f(\cdot) = \sum_{j=1}^{n} c_j k(\cdot, x_j), \quad c_j \in \mathbb{R}. \tag{2}$$

As a result, (1) becomes a finite-dimensional optimization problem determined by the *pairwise similarities* of the samples $[k(x_i, x_j)]$:

$$\min_{\mathbf{c}\in\mathbb{R}^n} \tilde{J}_0(\mathbf{c}) := \frac{1}{n}\sum_{i=1}^{n} V\left(y_i, \sum_{j=1}^{n} c_j k(x_i, x_j)\right) + \lambda \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j), \tag{3}$$

where the second term follows from the reproducing property of kernels.

However, in many learning problems such as nonlinear variable selection [20, 21], (multi-task) gradient learning [38], semi-supervised or Hermite learning with gradient information [41, 26], or density estimation with infinite-dimensional exponential families [28], apparently considering the *derivative information* ($\partial^{\mathbf{p}} f(\mathbf{x}_i) := \frac{\partial^{p_1+\ldots+p_d} f(\mathbf{x}_i)}{\partial_{x_1}^{p_1}\cdots\partial_{x_d}^{p_d}}$, $\mathcal{X} := \mathbb{R}^d$) other than just the function values ($f(\mathbf{x}_i)$) turns out to be *beneficial*. In these tasks containing derivatives, (1) is generalized to the form

$$\min_{f\in\mathcal{H}_k} J(f) := \frac{1}{n}\sum_{i=1}^{n} V\left(y_i, \{\partial^{\mathbf{p}} f(\mathbf{x}_i)\}_{\mathbf{p}\in I_i}\right) + \lambda \|f\|_{\mathcal{H}_k}^2. \tag{4}$$

The solution of this minimization task —similar to (1)—enjoys a finite-dimensional parameterization [41]:

$$f(\cdot) = \sum_{j=1}^{n}\sum_{\mathbf{p}\in I_j} c_{j,\mathbf{p}} \partial^{\mathbf{p},\mathbf{0}} k(\cdot, \mathbf{x}_j), \quad (c_{j,\mathbf{p}} \in \mathbb{R}),$$

where $\partial^{\mathbf{p},\mathbf{q}} k(\mathbf{x}, \mathbf{y}) := \frac{\partial^{\sum_{i=1}^d (p_i+q_i)} k(\mathbf{x},\mathbf{y})}{\partial_{x_1}^{p_1}\cdots\partial_{x_d}^{p_d}\partial_{y_1}^{q_1}\cdots\partial_{y_d}^{q_d}}$. Hence, the optimization in (4) can be reduced to

$$\min_{\mathbf{c}} \tilde{J}(\mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n} V\left(y_i, \left\{\sum_{j=1}^{n}\sum_{\mathbf{p}\in I_j} c_{j,\mathbf{p}} \partial^{\mathbf{p},\mathbf{0}} k(\mathbf{x}_i, \mathbf{x}_j)\right\}_{\mathbf{p}\in I_i}\right) + \lambda \sum_{i=1}^{n}\sum_{\mathbf{p}\in I_i}\sum_{j=1}^{n}\sum_{\mathbf{q}\in I_j} c_{i,\mathbf{p}} c_{j,\mathbf{p}} \partial^{\mathbf{p},\mathbf{q}} k(\mathbf{x}_i, \mathbf{x}_j), \tag{5}$$

where $\mathbf{c} = (c_{i,\mathbf{p}})_{i\in\{1,\ldots,n\},\mathbf{p}\in I_i} \in \mathbb{R}^{\sum_{i=1}^n |I_i|}$, $|I_i|$ denotes the cardinality of I_i, and we used the derivative-reproducing property of kernels

$$\partial^{\mathbf{p}} f(\mathbf{x}) = \left\langle f, \partial^{\mathbf{p},\mathbf{0}} k(\cdot, \mathbf{x})\right\rangle_{\mathcal{H}_k}.$$

Compared to (3) where the kernel values determine the objective, (5) is determined by the kernel derivatives $\partial^{\mathbf{p},\mathbf{q}} k(\mathbf{x}_i, \mathbf{x}_j)$.

While kernel techniques are extremely powerful due to their modelling capabilities, this flexibility comes with a price, often they are computationally expensive. In order to mitigate this computational bottleneck, several approaches have been proposed in the literature such as the Nyström and sub-sampling methods [36, 9, 22], sketching [1, 37], or random Fourier features (RFF) [18, 19] and their approximate memory-reduced variants and structured extensions [12, 7, 4].

The focus of the current submission is probably the conceptually simplest and most influential approximation scheme among these approaches, RFF.[2] The RFF technique implements a rather elementary yet powerful approach: it constructs a random, low-dimensional, explicit Fourier feature map (φ) for a continuous, bounded, shift-invariant kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ relying on the Bochner's theorem:

$$\hat{k}(\mathbf{x}, \mathbf{y}) = \langle \varphi(\mathbf{x}), \varphi(\mathbf{y}) \rangle , \qquad \varphi : \mathbb{R}^d \to \mathbb{R}^m.$$

The advantage of such a feature map becomes apparent after applying the parametrization:

$$\hat{f}(\mathbf{x}) = \langle \mathbf{w}, \varphi(\mathbf{x}) \rangle , \quad \mathbf{w} \in \mathbb{R}^m. \tag{6}$$

This parameterization can be considered as an approximate version of the reproducing property

$$f(\mathbf{x}) = \langle f, k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}_k} ,$$

$f \in \mathcal{H}_k$ is changed to $\mathbf{w} \in \mathbb{R}^m$ and $k(\cdot, \mathbf{x}) \in \mathcal{H}_k$ to $\varphi(\mathbf{x}) \in \mathbb{R}^m$. (6) allows one to leverage fast solvers for kernel machines in the primal [(1) or (4)]. This idea has been applied in a wide range of areas such as causal discovery [15], fast function-to-function regression [16], independence testing [40], convolution neural networks [6], prediction and filtering in dynamical systems [8], or bandit optimization [13].

Despite the tremendous practical success of RFF-s, its theoretical understanding is quite limited, with only a few optimal guarantees [29, 23, 27, 14, 32].

- Concerning the approximation quality of kernel values, the uniform finite-sample bounds of [18, 31] show that

$$\left\| k - \widehat{k} \right\|_{L^\infty(\mathcal{S} \times \mathcal{S})} := \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{S}} \left| k(\mathbf{x}, \mathbf{y}) - \widehat{k}(\mathbf{x}, \mathbf{y}) \right| = O_p \left(|\mathcal{S}| \sqrt{m^{-1} \log m} \right),$$

  where $\mathcal{S} \subset \mathbb{R}^d$ is a compact set, $|\mathcal{S}|$ is its diameter, m is the number of RFFs, $O_p(\cdot)$ means convergence in probability. [29] recently proved an exponentially tighter finite-sample bound in terms of $|\mathcal{S}|$ implying

$$\left\| k - \widehat{k} \right\|_{L^\infty(\mathcal{S} \times \mathcal{S})} = O_{a.s.} \left(\sqrt{\log |\mathcal{S}|} / \sqrt{m} \right), \tag{7}$$

  where $O_{a.s.}(\cdot)$ denotes almost sure convergence. This bound is optimal w.r.t. m and $|\mathcal{S}|$, as it is known from the characteristic function literature [5].

- In terms of generalization, [19] showed that $O(1/\sqrt{n})$ generalization error can be attained using $m = O(n)$ RFFs, where n denotes the number of training samples. This bound is somewhat pessimistic, leaving the usefulness of RFFs open. Recently [23] proved that $O\left(1/\sqrt{n}\right)$ generalization performance is attainable in the context of kernel ridge regression, with $m = o(n) = O\left(\sqrt{n} \log n\right)$ RFFs. This result settles RFFs in the simplest least-squares setting with Tikhonov regularization. Recently, the result has been sharpened [14] to $m = O\left(\sqrt{n} \log d_K^\lambda\right)$ with no loss in excess risk, where the effective degrees of freedom d_K^λ can often be significantly smaller than the number of samples.

- [27] has investigated the computational-statistical trade-offs of RFFs in kernel principal component analysis (KPCA). Their result show that depending on the eigenvalue decay behavior of the covariance operator associated to the kernel, $m = O(n^{2/3})$ (polynomial decay) or $m = O(\sqrt{n})$ (exponential decay) RFFs are sufficient to match the statistical performance of KPCA, where n again denotes the number of samples. [32] proved a similar result showing that $m = O\left(\sqrt{n} \log n\right)$ number of RFFs is sufficient for optimal statistical performance provided that the spectrum of the covariance operator follows an exponential decay, and presented a streaming algorithm for KPCA relying on the classical Oja's updates, achieving the same statistical performance.

In contrast to the previous results, the focus of our paper is the investigation of problems involving kernel derivatives [see (4) and (5)]. The idea applied in practice is to formally differentiate (6) giving

$$\widehat{\partial^{\mathbf{p}} f}(\mathbf{x}) := \partial^{\mathbf{p}} \hat{f}(\mathbf{x}) = \langle \mathbf{w}, \partial^{\mathbf{p}} \varphi(\mathbf{x}) \rangle , \tag{8}$$

[2] As a recognition of its influence, the work [18] won the 10-year test-of-time award at NIPS-2017.

which is then used in the primal [(4)], and optimized for $\mathbf{w}$. From the dual point of view [(5)], this means that implicitly the kernel derivatives are approximated via RFFs. The problem we raise in this paper is how accurate these kernel derivative approximations are.

Our **contribution** is to show that the same dependency in terms of m and $|\mathcal{S}|$ can be attained for kernel derivatives as for kernel values depicted in (7). To the best of our knowledge, the tightest available guarantee on kernel derivatives [29] is

$$\left\|\partial^{\mathbf{p},\mathbf{q}}k - \widehat{\partial^{\mathbf{p},\mathbf{q}}k}\right\|_{L^\infty(\mathcal{S}\times\mathcal{S})} = O_{a.s.}\left(|\mathcal{S}|\sqrt{m^{-1}\log m}\right).$$

In this paper, we prove finite sample bounds on the approximation quality of kernel derivatives, which specifically imply that

$$\left\|\partial^{\mathbf{p},\mathbf{q}}k - \widehat{\partial^{\mathbf{p},\mathbf{q}}k}\right\|_{L^\infty(\mathcal{S}\times\mathcal{S})} = O_{a.s.}\left(\sqrt{\log|\mathcal{S}|}/\sqrt{m}\right). \tag{9}$$

The possibility of such an exponentially improved dependence in terms of $|\mathcal{S}|$ is rather surprising, as in case of kernel derivatives the underlying function classes are no longer uniformly bounded. We circumvent this challenge by applying recent tools from unbounded empirical process theory.

Our paper is structured as follows. We formulate our problem in Section 2. The main result on the approximation quality of kernel derivatives is presented in Section 3. Proofs are provided in Section 4.

2 PROBLEM FORMULATION

In this section we formulate our problem after introducing a few notations.

Notations: $\mathbb{N} := \{0,1,2,\ldots\}$, $\mathbb{N}^+ := \mathbb{N}\backslash\{0\}$ and $\mathbb{R}$ denotes the set of natural numbers, positive integers and real numbers respectively. For $n \in \mathbb{N}$, $n!$ denotes its factorial. $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\,\mathrm{d}x$ is the Gamma function ($t > 0$); $\Gamma(n+1) = n!$ ($n \in \mathbb{N}$). Let $n!!$ denote the double factorial of $n \in \mathbb{N}$, that is, the product of all numbers from n to 1 that have the same parity as n; specifically $0!! = 1$. If n is a positive odd integer, then $n!! = \sqrt{\frac{2^{n+1}}{\pi}}\Gamma\left(\frac{n}{2}+1\right)$. For $a \in \mathbb{N}$, $c_a := \cos(\frac{\pi a}{2} + \cdot)$ is the a^{th} derivative of the cos function. For multi-indices $\mathbf{p},\mathbf{q} \in \mathbb{N}^d$ $|\mathbf{p}| = \sum_{j=1}^d p_j$, $\mathbf{v}^{\mathbf{p}} = \prod_{j=1}^d v_j^{p_j}$, and we use $\partial^{\mathbf{p}}h(\mathbf{x}) := \frac{\partial^{|\mathbf{p}|}h(\mathbf{x})}{\partial_{x_1}^{p_1}\cdots\partial_{x_d}^{p_d}}$, $\partial^{\mathbf{p},\mathbf{q}}g(\mathbf{x},\mathbf{y}) := \frac{\partial^{|\mathbf{p}|+|\mathbf{q}|}g(\mathbf{x},\mathbf{y})}{\partial_{x_1}^{p_1}\cdots\partial_{x_d}^{p_d}\partial_{y_1}^{q_1}\cdots\partial_{y_d}^{q_d}}$ to denote partial derivatives. $\langle\mathbf{a},\mathbf{b}\rangle = \sum_{i=1}^d a_i b_i$ is the inner product between $\mathbf{a} \in \mathbb{R}^d$ and $\mathbf{b} \in \mathbb{R}^d$. $\mathbf{a}^T$ is the transpose of $\mathbf{a} \in \mathbb{R}^d$, $\|\mathbf{a}\|_2 = \sqrt{\langle\mathbf{a},\mathbf{a}\rangle}$ is its Euclidean norm, $[\mathbf{a}_1;\ldots;\mathbf{a}_M] \in \mathbb{R}^{\sum_{m=1}^M d_m}$ is the concatenation of the $\mathbf{a}_m \in \mathbb{R}^{d_m}$ vectors. Let $\mathcal{S} \subset \mathbb{R}^d$ be a Borel set. $\mathcal{M}_+^1(\mathcal{S})$ is the set of Borel probability measures on $\mathcal{S}$. $\Lambda^m = \otimes_{i=1}^m \Lambda$ is the m-fold product measure where $\Lambda \in \mathcal{M}_+^1(\mathcal{S})$. $L^r(\mathcal{S})$ is the Banach space of real-valued, r-power Lebesgue integrable functions on $\mathcal{S}$ ($1 \le r < \infty$). $\Lambda f = \int_{\mathcal{S}} f(\boldsymbol{\omega})\mathrm{d}\Lambda(\boldsymbol{\omega})$, where $\Lambda \in \mathcal{M}_+^1(\mathcal{S})$ and $f \in L^1(\mathcal{S})$; specifically for the empirical measure, $\Lambda_m = \frac{1}{m}\sum_{i=1}^m \delta_{\boldsymbol{\omega}_i}$, $\Lambda_m f := \frac{1}{m}\sum_{i=1}^m f(\boldsymbol{\omega}_i)$ where $\boldsymbol{\omega}_{1:m} = (\boldsymbol{\omega}_i)_{i=1}^m \overset{\text{i.i.d.}}{\sim} \Lambda$ and $\delta_{\boldsymbol{\omega}}$ is the Dirac measure supported on $\boldsymbol{\omega} \in \mathcal{S}$. $\mathcal{S}_\Delta = \{\mathbf{s}_1 - \mathbf{s}_2 : \mathbf{s}_1,\mathbf{s}_2 \in \mathcal{S}\}$. For positive sequences $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$, $a_n = O(b_n)$ (resp. $a_n = o(b_n)$) means that $\left(\frac{a_n}{b_n}\right)_{n\in\mathbb{N}}$ is bounded (resp. $\lim_{n\to\infty}\frac{a_n}{b_n} = 0$). Positive sequences $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ are said to be asymptotically equivalent, shortly $a_n \sim b_n$, if $\lim_{n\to\infty}\frac{a_n}{b_n} = 1$. $X_n = O_p(r_n)$ (resp. $O_{a.s.}(r_n)$) denotes that $\frac{X_n}{r_n}$ is bounded in probability (resp. almost surely). The diameter of a compact set $A \subset \mathbb{R}^d$ is defined as $|A| := \sup_{\mathbf{x},\mathbf{y}\in A}\|\mathbf{x}-\mathbf{y}\|_2 < \infty$. The natural logarithm is denoted by ln.

We continue with the formulation of our task. Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a continuous, bounded, shift-invariant kernel. By the Bochner theorem [24], it is the Fourier transform of a finite, non-negative Borel measure Λ:

$$k(\mathbf{x},\mathbf{y}) = \tilde{k}(\mathbf{x}-\mathbf{y}) = \int_{\mathbb{R}^d} e^{\sqrt{-1}\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})}\mathrm{d}\Lambda(\boldsymbol{\omega}) \overset{(a)}{=} \int_{\mathbb{R}^d}\cos\left(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})\right)\mathrm{d}\Lambda(\boldsymbol{\omega}) \tag{10}$$

$$\overset{(b)}{=} \int_{\mathbb{R}^d}\left[\cos\left(\boldsymbol{\omega}^T\mathbf{x}\right)\cos\left(\boldsymbol{\omega}^T\mathbf{y}\right) + \sin\left(\boldsymbol{\omega}^T\mathbf{x}\right)\sin\left(\boldsymbol{\omega}^T\mathbf{y}\right)\right]\mathrm{d}\Lambda(\boldsymbol{\omega}) = \int_{\mathbb{R}^d}\langle\phi_{\boldsymbol{\omega}}(\mathbf{x}),\phi_{\boldsymbol{\omega}}(\mathbf{y})\rangle_{\mathbb{R}^2}\mathrm{d}\Lambda(\boldsymbol{\omega}), \tag{11}$$

where

$$\phi_{\boldsymbol{\omega}}(\mathbf{x}) = [\cos\left(\boldsymbol{\omega}^T\mathbf{x}\right); \sin\left(\boldsymbol{\omega}^T\mathbf{x}\right)].$$

(a) follows from the real-valued property of k, and (b) is a consequence of the trigonometric identity $\cos(\alpha - \beta) = \cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta)$. Without loss of generality, it can be assumed that $\Lambda \in \mathcal{M}_+^1\left(\mathbb{R}^d\right)$ since $\tilde{k}(\mathbf{0}) = \Lambda\left(\mathbb{R}^d\right)$ and the normalization $\frac{k(\mathbf{x},\mathbf{y})}{\tilde{k}(\mathbf{0})}$ yields

$$\frac{k(\mathbf{x},\mathbf{y})}{\tilde{k}(\mathbf{0})} = \int_{\mathbb{R}^d} \cos\left(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})\right) \mathrm{d}\underbrace{\frac{\Lambda(\boldsymbol{\omega})}{\Lambda\left(\mathbb{R}^d\right)}}_{=:\mathbb{P}(\boldsymbol{\omega})\in\mathcal{M}_+^1(\mathbb{R}^d)} .$$

Let $\mathbf{p}, \mathbf{q} \in \mathbb{N}^d$. By differentiating[3] (11) one gets

$$\partial^{\mathbf{p},\mathbf{q}} k(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d} \langle \partial^{\mathbf{p}}\phi_{\boldsymbol{\omega}}(\mathbf{x}), \partial^{\mathbf{q}}\phi_{\boldsymbol{\omega}}(\mathbf{y})\rangle_{\mathbb{R}^2} \mathrm{d}\Lambda(\boldsymbol{\omega}). \tag{12}$$

The resulting expectation can be approximated by the Monte-Carlo technique using $\boldsymbol{\omega}_{1:m} = (\boldsymbol{\omega}_j)_{j=1}^m \overset{\text{i.i.d.}}{\sim} \Lambda$ as

$$\begin{aligned}\widehat{\partial^{\mathbf{p},\mathbf{q}} k}(\mathbf{x},\mathbf{y}) &= \int_{\mathbb{R}^d} \langle \partial^{\mathbf{p}}\phi_{\boldsymbol{\omega}}(\mathbf{x}), \partial^{\mathbf{q}}\phi_{\boldsymbol{\omega}}(\mathbf{y})\rangle_{\mathbb{R}^2} \mathrm{d}\Lambda_m(\boldsymbol{\omega}) = \frac{1}{m}\sum_{j=1}^m \left\langle \partial^{\mathbf{p}}\phi_{\boldsymbol{\omega}_j}(\mathbf{x}), \partial^{\mathbf{q}}\phi_{\boldsymbol{\omega}_j}(\mathbf{y})\right\rangle_{\mathbb{R}^2} \\ &= \langle \varphi_{\mathbf{p}}(\mathbf{x}), \varphi_{\mathbf{q}}(\mathbf{y})\rangle_{\mathbb{R}^{2m}} , \end{aligned} \tag{13}$$

where $\Lambda_m = \frac{1}{m}\sum_{j=1}^m \delta_{\boldsymbol{\omega}_j}$,

$$\varphi_{\mathbf{p}}(\mathbf{x}) = \frac{1}{\sqrt{m}}\left(\partial^{\mathbf{p}}\phi_{\boldsymbol{\omega}_j}(\mathbf{x})\right)_{j=1}^m = \underbrace{\partial^{\mathbf{p}}\varphi_{\mathbf{0}}(\mathbf{x})}_{\in\mathbb{R}^{2m}} = \frac{1}{\sqrt{m}}\left(\boldsymbol{\omega}_j^{\mathbf{p}}\left[c_{|\mathbf{p}|}\left(\boldsymbol{\omega}_j^T\mathbf{x}\right); c_{3+|\mathbf{p}|}\left(\boldsymbol{\omega}_j^T\mathbf{x}\right)\right]\right)_{j=1}^m . \tag{14}$$

Specifically, if $\mathbf{p} = \mathbf{q} = \mathbf{0}$ then (13) boils down to the celebrated RFF technique [18]:

$$\widehat{k}(\mathbf{x},\mathbf{y}) = \langle \varphi_{\mathbf{0}}(\mathbf{x}), \varphi_{\mathbf{0}}(\mathbf{y})\rangle_{\mathbb{R}^{2m}} , \qquad \varphi_{\mathbf{0}}(\mathbf{x}) = \frac{1}{\sqrt{m}}\left(\cos\left(\boldsymbol{\omega}_j^T\mathbf{x}\right); \sin\left(\boldsymbol{\omega}_j^T\mathbf{x}\right)\right)_{j=1}^m .$$

Our **goal** is to prove that similarly to $\mathbf{p} = \mathbf{q} = \mathbf{0}$ [(7)], fast approximation of kernel derivatives [(9)] is attainable. Alternatively, we establish that the derivative (see $\varphi_{\mathbf{p}}$ and (13)-(14)) of the RFF feature map ($\varphi_{\mathbf{0}}$) is as efficient for kernel derivative approximation as $\varphi_{\mathbf{0}}$ for kernel value approximation.

3 MAIN RESULT

In this section we present our main result on the uniform approximation quality of kernel derivatives using RFFs. Its proof is available in Section 4.

Theorem (Uniform guarantee on kernel derivative approximation)**.** *Suppose that $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a continuous, bounded and shift-invariant kernel. For $\mathbf{p}, \mathbf{q} \in \mathbb{N}^d$, assume $C_{\mathbf{p},\mathbf{q}} = \sqrt{\int_{\mathbb{R}^d} |\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|^2 \|\boldsymbol{\omega}\|_2^2 \mathrm{d}\Lambda(\boldsymbol{\omega})}/\sigma_{\mathbf{p},\mathbf{q}} < \infty$ and for some constant $K \geq 1$, the following Bernstein condition holds:*

$$\int_{\mathbb{R}^d} \frac{|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|^n}{(\sigma_{\mathbf{p},\mathbf{q}})^n} \mathrm{d}\Lambda(\boldsymbol{\omega}) \leq \frac{n!}{2} K^{n-2}, \quad n = 2, 3, \ldots, \tag{15}$$

[3]By the dominated convergence theorem, the differentiation is valid if $\int_{\mathbb{R}^d} |\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}| \mathrm{d}\Lambda(\boldsymbol{\omega}) < \infty$.

where $\sigma_{\mathbf{p},\mathbf{q}} = \sqrt{\int_{\mathbb{R}^d} |\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|^2 \, \mathrm{d}\Lambda(\boldsymbol{\omega})}$. *Let* $L_m = \frac{\sqrt{6}K}{2\sqrt{m}}$, $C_1 = 14\sqrt{6 \ln 2} + 1$, $C_2 = 36K[\ln 2 + 1]$ *and* $C_3 = 7\sqrt{6}\left(1 + \frac{\sqrt{\pi}}{\ln^{\frac{3}{2}} 2}\right)$. *Then for any* $t > 0$ *and compact set* $\mathcal{S} \subset \mathbb{R}^d$

$$\Lambda^m\left(\left\{\boldsymbol{\omega}_{1:m} : \left\|\partial^{\mathbf{p},\mathbf{q}} k - \widehat{\partial^{\mathbf{p},\mathbf{q}} k}\right\|_{L^\infty(\mathcal{S}\times\mathcal{S})} \geq \sigma_{\mathbf{p},\mathbf{q}}\left(\frac{C_3\sqrt{d \ln\left(16|\mathcal{S}|C_{\mathbf{p},\mathbf{q}} + 4\right)}}{\sqrt{m}} + \frac{C_1}{\sqrt{m}} + \frac{C_2}{m} + + \frac{24\sqrt{6}}{\sqrt{m}}\left[\sqrt{t} + \frac{L_m t}{2}\right]\right)\right\}\right)$$

with probablity at most $2e^{-t}$.

Remarks.

- **Growth of** $|\mathcal{S}_m|$: The theorem proves the same dependence [(9)] on m and $|\mathcal{S}|$ as is known for kernel values ($\mathbf{p} = \mathbf{q} = \mathbf{0}$). The result implies that

  $$\left\|\partial^{\mathbf{p},\mathbf{q}} k - \widehat{\partial^{\mathbf{p},\mathbf{q}} k}\right\|_{L^\infty(\mathcal{S}_m\times\mathcal{S}_m)} \xrightarrow{a.s.} 0$$

  if $|\mathcal{S}_m| = e^{o(m)}$.

- **Requirements for** $\mathbf{p} = \mathbf{q} = \mathbf{0}$**:** In this case $\sigma_{\mathbf{0},\mathbf{0}} = 1$,
  - $\int_{\mathbb{R}^d} \frac{|\boldsymbol{\omega}^{\mathbf{0}+\mathbf{0}}|^n}{(\sigma_{\mathbf{0},\mathbf{0}})^n} \mathrm{d}\Lambda(\boldsymbol{\omega}) = 1$, thus (15) holds ($K = 1$).
  - The only requirement is the finiteness of $C_{\mathbf{0},\mathbf{0}} = \int_{\mathbb{R}^d} \|\boldsymbol{\omega}\|_2^2 \, \mathrm{d}\Lambda(\boldsymbol{\omega})$, which is identical to that imposed in [29, Theorem 1] for kernel values.

- $L^\infty(\mathcal{S}\times\mathcal{S})$**-based** $L^r(\mathcal{S}\times\mathcal{S})$ **guarantee**: From the theorem above one can also get (see Section 4) the following $L^r(\mathcal{S}\times\mathcal{S})$ guarantee, where $r \in [1, \infty)$.

  Under the same conditions and notations as in the theorem, for any $t > 0$

  $$\Lambda^m\left(\left\{\boldsymbol{\omega}_{1:m} : \left\|\partial^{\mathbf{p},\mathbf{q}} k - \widehat{\partial^{\mathbf{p},\mathbf{q}} k}\right\|_{L^r(\mathcal{S}\times\mathcal{S})} \geq \sigma_{\mathbf{p},\mathbf{q}}\left[\frac{\pi^{d/2}|\mathcal{S}|^d}{2^d\Gamma\left(\frac{d}{2}+1\right)}\right]^{\frac{2}{r}}\left(\frac{C_3\sqrt{2d \ln\left(16|\mathcal{S}|C_{\mathbf{p},\mathbf{q}} + 4\right)}}{\sqrt{m}} + \right.\right.$$
  $$\left.\left. + \frac{C_1}{\sqrt{m}} + \frac{C_2}{m} + \frac{24\sqrt{6}}{\sqrt{m}}\left[\sqrt{t} + \frac{L_m t}{2}\right]\right)\right\}\right) \leq 2e^{-t}.$$

  This shows that

  $$\left\|\partial^{\mathbf{p},\mathbf{q}} k - \widehat{\partial^{\mathbf{p},\mathbf{q}} k}\right\|_{L^r(\mathcal{S}\times\mathcal{S})} = O_{a.s.}\left(m^{-\frac{1}{2}}|\mathcal{S}|^{\frac{2d}{r}}\sqrt{\log|\mathcal{S}|}\right).$$

  Consequently, if $|\mathcal{S}_m| \to \infty$ as $m \to \infty$ then $\widehat{\partial^{\mathbf{p},\mathbf{q}} k}$ is a consistent estimator of $\partial^{\mathbf{p},\mathbf{q}} k$ in $L^r(\mathcal{S}_m \times \mathcal{S}_m)$-norm provided that $m^{-\frac{1}{2}}|\mathcal{S}_m|^{\frac{2d}{r}}\sqrt{\log|\mathcal{S}_m|} \xrightarrow{m\to\infty} 0$.

- **Bernstein condition with** $[\mathbf{p};\mathbf{q}] \neq \mathbf{0}$**:** Next we illustrate how the Bernstein condition [(15)] translates to the efficient estimation of 'not too large'-order kernel derivatives in case of the Gaussian kernel. For simplicity let us consider the Gaussian kernel in one dimension ($d = 1$); in this case $\Lambda = N\left(0, \sigma^2\right)$ is a normal distribution with mean zero and variance σ^2. Let $r = p + q \in \mathbb{N}^+$ and denote the l.h.s. of (15) as

  $$A_{r,n}(\Lambda) = \frac{\int_{\mathbb{R}} |\omega|^{rn} \mathrm{d}\Lambda(\omega)}{\left[\sqrt{\int_{\mathbb{R}} |\omega|^{2r} \, \mathrm{d}\Lambda(\omega)}\right]^n}.$$

  By the analytical formula for the absolute moments of normal random variables

  $$A_{r,n}(\Lambda) = \frac{\sigma^{nr}(nr-1)!! \begin{cases} 1 & \text{if } nr \text{ is even} \\ \sqrt{\frac{2}{\pi}} & \text{if } nr \text{ is odd} \end{cases}}{\left[\sigma^{2r}(2r-1)!!\right]^{\frac{n}{2}}} = \frac{(nr-1)!! \begin{cases} 1 & \text{if } nr \text{ is even} \\ \sqrt{\frac{2}{\pi}} & \text{if } nr \text{ is odd} \end{cases}}{\left[(2r-1)!!\right]^{\frac{n}{2}}}. \tag{16}$$

Since $A_{r,n}(\Lambda)$ does not depend on σ, one can assume that $\sigma = \sqrt{\int_{\mathbb{R}} |\omega|^2 \mathrm{d}\Lambda(\omega)} = 1$ and $\Lambda = N(0,1)$. Exploiting the analytical expression obtained for $A_{r,n}(\Lambda)$ one can show (Section 4) that for
– $r = 1$: Since $A_{1,n}(\Lambda) \leq \frac{n!}{2}$, (15) holds with $K = 1$.
– $r = 2$: $K = 2$ is a suitable choice in (15).
– $r = 3$ and $r = 4$: Asymptotic argument shows that (15) can not hold.
It is an interesting open question whether one can relax (15) while maintaining similar rates, and what are the trade-offs.

- **Difficulty:** The fundamental difficulty one has to tackle to arrive at the stated theorem is as follows.

  By differentiating (10) one gets

$$\partial^{\mathbf{p},\mathbf{q}} k(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d} \boldsymbol{\omega}^{\mathbf{p}} (-\boldsymbol{\omega})^{\mathbf{q}} c_{|\mathbf{p}+\mathbf{q}|} \left(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})\right) \mathrm{d}\Lambda(\boldsymbol{\omega}).$$

  By defining

$$g_{\mathbf{z}}(\boldsymbol{\omega}) = \boldsymbol{\omega}^{\mathbf{p}} (-\boldsymbol{\omega})^{\mathbf{q}} c_{|\mathbf{p}+\mathbf{q}|} \left(\boldsymbol{\omega}^T \mathbf{z}\right), \tag{17}$$

  the error we would like to control can be rewritten as the supremum of the empirical process

$$\sup_{\mathbf{x},\mathbf{y}\in\mathcal{S}} \left|\partial^{\mathbf{p},\mathbf{q}} k(\mathbf{x},\mathbf{y}) - \widehat{\partial^{\mathbf{p},\mathbf{q}} k}(\mathbf{x},\mathbf{y})\right| = \sup_{\mathbf{z}\in\mathcal{S}_\Delta} |(\Lambda - \Lambda_m) g_{\mathbf{z}}|,$$

  where $\mathcal{G} := \{g_{\mathbf{z}} : \mathbf{z} \in \mathcal{S}_\Delta\}$. For $\mathbf{p} = \mathbf{q} = \mathbf{0}$ (i.e., the classical RFF-based kernel approximation)

$$g_{\mathbf{z}}(\boldsymbol{\omega}) = \cos\left(\boldsymbol{\omega}^T \mathbf{z}\right) \quad (\mathbf{z} \in \mathcal{S}_\Delta)$$

  which is a *uniformly bounded* family of functions:

$$\sup_{\mathbf{z}\in\mathcal{S}_\Delta} \|g_{\mathbf{z}}\|_{L^\infty(\mathbb{R}^d)} \leq 1.$$

  This uniform boundedness is the classical assumption of empirical process theory, and allowed one [29] to get the optimal rates. For $\mathbf{p}, \mathbf{q} \in \mathbb{N}^d \backslash \{\mathbf{0}\}$, however, the functions $g_{\mathbf{z}}$ are unbounded and so $\mathcal{G}$ is no longer uniformly bounded in $L^\infty\left(\mathbb{R}^d\right)$. Therefore, one has to control unbounded empirical processes for which only few tools are available.

  The key idea of our paper is to apply a recent technique which bounds the supremum as a weighted sum of bracketing entropies of $\mathcal{G}$ at multiple scales. By estimating these bracketing entropies and optimizing the scale the result will follow. This is what we detail in the next section.

4 PROOFS

We provide the proofs of the results (main theorem and its consequence, remark on the Bernstein condition) presented in Section 3. We start by introducing a few additional notations specific to this section.

Notations: The volume of $A \subseteq \mathbb{R}^d$ is defined as $\mathrm{vol}(A) = \int_A 1 \, \mathrm{d}\mathbf{x}$. $\gamma(a,b) = \int_0^b e^{-t} t^{a-1} \mathrm{d}t$ is the incomplete Gamma function ($a > 0$, $b \geq 0$) that satisfies $\gamma(a+1,b) = a\gamma(a,b) - b^a e^{-b}$ and $\gamma\left(\frac{1}{2}, b\right) = \sqrt{\pi}\mathrm{erf}\left(\sqrt{b}\right)$, where $\mathrm{erf}(b) = \frac{2}{\sqrt{\pi}} \int_0^b e^{-t^2} \mathrm{d}t$ is the error function ($b \geq 0$). Let $(\mathcal{F}, \rho)$ be a metric space. The r-covering number of $\mathcal{F}$ is defined as the size of the smallest r-net, i.e., $N(r, \mathcal{F}, \rho) = \inf\left\{\ell \geq 1 : \exists\, (f_j)_{j=1}^{\ell} \text{ s.t. } \mathcal{F} \subseteq \cup_{j=1}^{\ell} B_\rho(f_j, r)\right\}$, where $B_\rho(s,r) = \{f \in \mathcal{F} : \rho(f,s) \leq r\}$ is the closed ball with center $s \in \mathcal{F}$ and radius r. For a set of real-valued functions $\mathcal{F}$ and $r > 0$, the cardinality of the minimal r-bracketing of $\mathcal{F}$ is defined as $N_{[]}(r, \mathcal{F}, \rho) = \inf\{n \geq 1 : \exists\, \{(f_{j,L}, f_{j,U})\}_{j=1}^n, f_{j,L}, f_{j,U} \in \mathcal{F}\ (\forall j) \text{ such that } \rho\left(f_{j_L}, f_{j,U}\right) \leq r \text{ and } \forall f \in \mathcal{F}\ \exists j\ f_{j,L} \leq f \leq f_{j,U}\}$.

The **proof of the main theorem** is structured as follows.

1. First, we rescale and reformulate the approximation error as the suprema of unbounded empirical processes, for which bounds in terms of bracketing entropies at multiple scales can be obtained.
2. Then, we bound the bracketing entropies via Lipschitz continuity.
3. Finally, the scale is optimized.

Step 1. It follows from (17) that,

$$\|g_{\mathbf{z}}\| := \|g_{\mathbf{z}}\|_{L^2(\mathbb{R}^d,\Lambda)} = \sqrt{\Lambda g_{\mathbf{z}}^2} \leq \underbrace{\sqrt{\int_{\mathbb{R}^d} |\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|^2 \,\mathrm{d}\Lambda(\boldsymbol{\omega})}}_{=:\sigma_{\mathbf{p},\mathbf{q}}}.$$

Define $f_{\mathbf{z}}(\boldsymbol{\omega}) := \frac{g_{\mathbf{z}}(\boldsymbol{\omega})}{\sigma_{\mathbf{p},\mathbf{q}}}$ so that

$$\|f_{\mathbf{z}}\| \leq 1 \quad \forall \mathbf{z} \in \mathcal{S}_\Delta \;\Rightarrow\; \sup_{f\in\mathcal{F}} \|f\| \leq 1, \tag{18}$$

where $\mathcal{F} := \{f_{\mathbf{z}} : \mathbf{z} \in \mathcal{S}_\Delta\}$. The target quantity can be rewritten in supremum of empirical process form as

$$\sup_{\mathbf{x},\mathbf{y}\in\mathcal{S}} \left|\partial^{\mathbf{p},\mathbf{q}} k(\mathbf{x},\mathbf{y}) - \widehat{\partial^{\mathbf{p},\mathbf{q}} k}(\mathbf{x},\mathbf{y})\right| = \sup_{\mathbf{z}\in\mathcal{S}_\Delta} |\Lambda g_{\mathbf{z}} - \Lambda_m g_{\mathbf{z}}| = \sigma_{\mathbf{p},\mathbf{q}} \sup_{f\in\mathcal{F}} |(\Lambda - \Lambda_m) f| =: \sigma_{\mathbf{p},\mathbf{q}} \|\Lambda - \Lambda_m\|_{\mathcal{F}}.$$

By the Bernstein condition [(15)] the following uniform bound holds:

$$\sup_{f_{\mathbf{z}}:\mathbf{z}\in\mathcal{S}_\Delta} \Lambda|f_{\mathbf{z}}|^n \leq \int_{\mathbb{R}^d} \frac{|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|^n}{(\sigma_{\mathbf{p},\mathbf{q}})^n} \mathrm{d}\Lambda(\boldsymbol{\omega}) \leq \frac{n!}{2} K^{n-2} \quad (n = 2,3,\ldots). \tag{19}$$

The uniform $L^2(\Lambda)$ boundedness of $\mathcal{F}$ [(18)] with its Bernstein property [(19)] imply by [34, Theorem 8] that for all $t > 0$ and for all scale $S \in \mathbb{N}$

$$\Lambda^m \left(\left\{ \boldsymbol{\omega}_{1:m} : \sup_{f\in\mathcal{F}} \left|\sqrt{m}(\Lambda - \Lambda_m) f\right| \geq \min_S E_S + \frac{36K}{\sqrt{m}} + 24\sqrt{6}\left[\sqrt{t} + \frac{L_m t}{2}\right] \right\} \right) \leq 2e^{-t}, \tag{20}$$

where

$$E_S := 2^{-S}\sqrt{m} + 14\sum_{s=0}^{S} 2^{-s}\sqrt{6H_s} + \frac{36KH_0}{\sqrt{m}}, \qquad L_m := \frac{\sqrt{6}K}{2\sqrt{m}}, \quad H_s := \ln(N_s+1),$$
$$N_s := N_{[\,]}(2^{-s}, \mathcal{F}, \|\cdot\|), \qquad H_0 = \ln(N_0+1),$$

and N_0 is the cardinality of the minimal generalized bracketing set of $\mathcal{F}$. Formally, $N_0 = N_0(K) := \inf\{n \geq 1 : \exists f_{j,L}, f_{j,U} \in \mathcal{F}\,(j = 1,\ldots,n), \Lambda\,|f_{j,L} - f_{j,U}|^n \leq \frac{n!}{2}(2K)^{n-2}\,(n = 2,3,\ldots)$, and for $\forall f \in \mathcal{F}$, $\exists j \in \{1,\ldots,n\}$ such that $f_{j,L} \leq f \leq f_{j,U}\}$.

Step 2. We continue the proof by bounding the entropies H_0 and H_s ($s \geq 1$) in (20). Using (15) for the envelope function $F := \sup_{f\in\mathcal{F}} |f|$, we get

$$\Lambda\left(F^n\right) = \Lambda\left(\left[\sup_{f\in\mathcal{F}} |f|\right]^n\right) = \Lambda\left(\sup_{f\in\mathcal{F}} |f|^n\right) \leq \int_{\mathbb{R}^d} \frac{|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|^n}{(\sigma_{\mathbf{p},\mathbf{q}})^n} \mathrm{d}\Lambda(\boldsymbol{\omega}) \leq \frac{n!}{2} K^{n-2}, \; n = 2,3,\ldots$$

Hence F also satisfies the weaker $\Lambda\left(F^n\right) \leq \frac{n!}{2}(2K)^{n-2}$ $(n = 2,3,\ldots)$ Bernstein condition. Consequently, one can choose $N_0 = 1$ [34, remark after Definition 8], and $H_0 = \ln(N_0+1) = \ln 2$.

Next we bound H_s ($s \geq 1$). The $\mathcal{F}$ function class is Lipschitz continuous in the parameters ($f_{\mathbf{z}_1}, f_{\mathbf{z}_2} \in \mathcal{F}$):

$$\begin{aligned}
|f_{\mathbf{z}_1}(\boldsymbol{\omega}) - f_{\mathbf{z}_2}(\boldsymbol{\omega})| &= \frac{\left|\boldsymbol{\omega}^{\mathbf{p}}(-\boldsymbol{\omega})^{\mathbf{q}} c_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T\mathbf{z}_1\right) - \boldsymbol{\omega}^{\mathbf{p}}(-\boldsymbol{\omega})^{\mathbf{q}} c_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T\mathbf{z}_2\right)\right|}{\sigma_{\mathbf{p},\mathbf{q}}} \\
&= \frac{|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|\left|c_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T\mathbf{z}_1\right) - c_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T\mathbf{z}_2\right)\right|}{\sigma_{\mathbf{p},\mathbf{q}}} \overset{(a)}{\leq} \frac{|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|}{\sigma_{\mathbf{p},\mathbf{q}}}\left|\boldsymbol{\omega}^T(\mathbf{z}_1 - \mathbf{z}_2)\right| \\
&\overset{(b)}{\leq} \underbrace{\frac{|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|}{\sigma_{\mathbf{p},\mathbf{q}}}\|\boldsymbol{\omega}\|_2}_{=:G(\boldsymbol{\omega})}\|\mathbf{z}_1 - \mathbf{z}_2\|_2,
\end{aligned}$$

where we used the Lipschitz property of $u \mapsto c_{|\mathbf{p}+\mathbf{q}|}(u)$ (with Lipschitz constant 1) in (a) and the Cauchy-Bunyakovskii-Schwarz inequality in (b). Thus, by [35, Theorem 2.7.11, page 164] for any $\delta > 0$,

$$N_{[\,]}(\delta, \mathcal{F}, \|\cdot\|) \leq N\left(\frac{\delta}{2\|G\|}, \mathcal{S}_\Delta, \|\cdot\|_2\right), \tag{21}$$

where

$$\|G\| = \sqrt{\int_{\mathbb{R}^d} G^2(\boldsymbol{\omega})\mathrm{d}\Lambda(\boldsymbol{\omega})} = \sqrt{\int_{\mathbb{R}^d} \frac{|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|^2}{\sigma^2_{\mathbf{p},\mathbf{q}}}\|\boldsymbol{\omega}\|_2^2\,\mathrm{d}\Lambda(\boldsymbol{\omega})} =: C_{\mathbf{p},\mathbf{q}}.$$

From Lemma 2.5 in [33] it follows that

$$N(r, M, \|\cdot\|_2) \leq \left(\frac{2|M|}{r} + 1\right)^d, \quad \forall r > 0$$

for any compact $M \subset \mathbb{R}^d$. Choosing $M = \mathcal{S}_\Delta$, $\delta = 2^{-s}$ and using that $|\mathcal{S}_\Delta| \leq 2|\mathcal{S}|$, one can bound the l.h.s. in (21) as

$$N_s = N_{[\,]}\left(2^{-s}, \mathcal{F}, \|\cdot\|\right) \leq N\left(\frac{1}{2^{s+1}C_{\mathbf{p},\mathbf{q}}}, \mathcal{S}_\Delta, \|\cdot\|_2\right) \leq \Big(\underbrace{2^{s+3}|\mathcal{S}|C_{\mathbf{p},\mathbf{q}} + 1}_{\leq 2^s\tilde{K}_{|S|}}\Big)^d,$$

where $\tilde{K}_{|S|} = 8|\mathcal{S}|C_{\mathbf{p},\mathbf{q}} + 1$. Thus for any $s \geq 1$,

$$H_s = \ln(N_s + 1) \leq d\ln\big(\underbrace{2^s\tilde{K}_{|S|} + 1}_{\leq 2^s(\tilde{K}_{|S|}+1)}\big) \leq d\left[s\ln 2 + \ln\left(\tilde{K}_{|\mathcal{S}|} + 1\right)\right] \leq \underbrace{s\,d\left[\ln 2 + \ln\left(\tilde{K}_{|\mathcal{S}|} + 1\right)\right]}_{d\ln\left(2\tilde{K}_{|\mathcal{S}|}+2\right)=:K_{|\mathcal{S}|}}.$$

Hence,

$$E_S \leq 2^{-S}\sqrt{m} + 14\sqrt{6\ln 2} + \underbrace{14\sum_{s=1}^{S} 2^{-s}\sqrt{6sK_{|\mathcal{S}|}}}_{14\sqrt{6K_{|\mathcal{S}|}}\sum_{s=1}^{S}2^{-s}\sqrt{s}} + \frac{36K\ln 2}{\sqrt{m}}. \tag{22}$$

Step 3. By (22), to control E_S as a function of the scale S, we study the behaviour of the $h(t) = 2^{-t}\sqrt{t}$ function. It is easy to verify that h is monotonically decreasing on $\left[\frac{1}{2\ln 2}, \infty\right)$ as its derivative

$$h'(t) = \frac{\frac{1}{2}t^{-\frac{1}{2}}2^t - \sqrt{t}2^t\ln 2}{2^{2t}} \leq 0$$

on $\left[\frac{1}{2\ln 2},\infty\right)$. Using this monotonicity, one gets $h(s)\le\int_{s-1}^{s}h(x)\mathrm{d}x$ for any s such that $\frac{1}{2\ln 2}\le s-1\Leftrightarrow\frac{1}{2\ln 2}+1\le s$, specifically for all $2\le s$ since $\frac{1}{2\ln 2}<1$. Hence, applying change of variables ($2^{-x}=e^{-t}$, i.e. $x=\frac{t}{\ln 2}$) we arrive at

$$\begin{aligned}\sum_{s=1}^{S}2^{-s}\sqrt{s}\quad&=\underbrace{h(1)}_{\frac{1}{2}}+\sum_{s=2}^{S}\underbrace{h(s)}_{\le\int_{s-1}^{s}h(x)\mathrm{d}x}\le\tfrac{1}{2}+\int_{1}^{S}h(x)\mathrm{d}x=\tfrac{1}{2}+\tfrac{1}{\ln^{\frac{3}{2}}(2)}\int_{\ln 2}^{S\ln 2}e^{-t}\sqrt{t}\mathrm{d}t\\&\le\tfrac{1}{2}+\tfrac{1}{\ln^{\frac{3}{2}}(2)}\int_{0}^{S\ln 2}e^{-t}\sqrt{t}\mathrm{d}t=\tfrac{1}{2}+\tfrac{1}{\ln^{\frac{3}{2}}(2)}\left[\tfrac{\sqrt{\pi}}{2}\mathrm{erf}(\sqrt{S\ln 2})-2^{-S}\sqrt{S\ln 2}\right].\end{aligned}$$

Plugging this estimate to (22) results in

$$\begin{aligned}E_S&\le 2^{-S}\sqrt{m}+14\sqrt{6\ln 2}+\frac{36K\ln 2}{\sqrt{m}}+14\sqrt{6K_{|\mathcal{S}|}}\left(\frac{1}{2}+\frac{1}{\ln^{\frac{3}{2}}(2)}\left[\frac{\sqrt{\pi}}{2}\mathrm{erf}\left(\sqrt{S\ln 2}\right)-2^{-S}\sqrt{S\ln 2}\right]\right)\\&\le 2^{-S}\sqrt{m}+14\sqrt{6}\sqrt{K_{|\mathcal{S}|}}\times\left(\frac{1}{2}+\frac{1}{\ln^{\frac{3}{2}}(2)}\frac{\sqrt{\pi}}{2}\right)+C_1+\frac{C_2}{\sqrt{m}}\\&\le 2^{-S}\sqrt{m}+14\sqrt{6}\sqrt{d\ln\left(16|\mathcal{S}|C_{\mathbf{p},\mathbf{q}}+4\right)}\left(\frac{1}{2}+\frac{1}{\ln^{\frac{3}{2}}(2)}\frac{\sqrt{\pi}}{2}\right)+C_1+\frac{C_2}{\sqrt{m}}=:(*),\end{aligned}$$

where we used the fact that $\mathrm{erf}(b)\le 1$ for any $b\ge 0$, $2^{-S}\sqrt{S}\ge 0$, $C_1=14\sqrt{6\ln 2}$, $C_2=36K\ln 2$ and $K_{|\mathcal{S}|}=d\ln\left(2\tilde{K}_{|\mathcal{S}|}+2\right)=d\ln\left(16|\mathcal{S}|C_{\mathbf{p},\mathbf{q}}+4\right)$. Let us choose the scale S such that $2^{-S}\sqrt{m}\le 1$, i.e. $\frac{\ln m}{2\ln 2}\le S$. In this case, by defining $C_3=7\sqrt{6}\left(1+\frac{\sqrt{\pi}}{\ln^{\frac{3}{2}}(2)}\right)$, we have

$$(*)=1+C_3\sqrt{d\ln\left(16|\mathcal{S}|C_{\mathbf{p},\mathbf{q}}+4\right)}+C_1+\frac{C_2}{\sqrt{m}}.$$

Combining this result with (20), we obtain

$$\Lambda^m\left(\left\{\boldsymbol{\omega}_{1:m}:\|\Lambda-\Lambda_m\|_{\mathcal{F}}\ge\frac{C_3\sqrt{d\ln\left(16|\mathcal{S}|C_{\mathbf{p},\mathbf{q}}+4\right)}}{\sqrt{m}}+\frac{C_1+1}{\sqrt{m}}+\frac{C_2+36K}{m}+\frac{24\sqrt{6}}{\sqrt{m}}\left[\sqrt{t}+\frac{L_m t}{2}\right]\right\}\right)\le 2e^{-t}.$$

By redefining C_1 and C_2 as $C_1=14\sqrt{6\ln 2}+1$, $C_2=36K[\ln 2+1]$ and taking into account the $\sigma_{\mathbf{p},\mathbf{q}}$ normalization, the claimed result follows. □

The **proof of the consequence** is as follows. Let $r\in[1,\infty)$ be fixed. Then

$$\begin{aligned}\left\|\partial^{\mathbf{p},\mathbf{q}}k-\widehat{\partial^{\mathbf{p},\mathbf{q}}k}\right\|_{L^r(\mathcal{S}\times\mathcal{S})}&=\left(\int_{\mathcal{S}}\int_{\mathcal{S}}\left|\partial^{\mathbf{p},\mathbf{q}}k(\mathbf{x},\mathbf{y})-\widehat{\partial^{\mathbf{p},\mathbf{q}}k}(\mathbf{x},\mathbf{y})\right|^r\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\right)^{\frac{1}{r}}\\&\le\left(\int_{\mathcal{S}}\int_{\mathcal{S}}\left\|\partial^{\mathbf{p},\mathbf{q}}k-\widehat{\partial^{\mathbf{p},\mathbf{q}}k}\right\|^r_{L^\infty(\mathcal{S}\times\mathcal{S})}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\right)^{\frac{1}{r}}=\left[\left\|\partial^{\mathbf{p},\mathbf{q}}k-\widehat{\partial^{\mathbf{p},\mathbf{q}}k}\right\|^r_{L^\infty(\mathcal{S}\times\mathcal{S})}\mathrm{vol}^2(\mathcal{S})\right]^{\frac{1}{r}}\\&=\left\|\partial^{\mathbf{p},\mathbf{q}}k-\widehat{\partial^{\mathbf{p},\mathbf{q}}k}\right\|_{L^\infty(\mathcal{S}\times\mathcal{S})}\mathrm{vol}^{\frac{2}{r}}(\mathcal{S}).\end{aligned}$$

Using the fact (which follows from [10, Corollary 2.55]) that $\mathrm{vol}(\mathcal{S})\le\frac{\pi^{d/2}|\mathcal{S}|^d}{2^d\Gamma\left(\frac{d}{2}+1\right)}$ one arrives at

$$\left\|\partial^{\mathbf{p},\mathbf{q}}k-\widehat{\partial^{\mathbf{p},\mathbf{q}}k}\right\|_{L^r(\mathcal{S}\times\mathcal{S})}\le\left\|\partial^{\mathbf{p},\mathbf{q}}k-\widehat{\partial^{\mathbf{p},\mathbf{q}}k}\right\|_{L^\infty(\mathcal{S}\times\mathcal{S})}\left[\frac{\pi^{d/2}|\mathcal{S}|^d}{2^d\Gamma\left(\frac{d}{2}+1\right)}\right]^{\frac{2}{r}}.$$

Hence the main theorem implies the claimed $L^r(\mathcal{S}\times\mathcal{S})$ bound. □

The **Bernstein condition**-related results can be obtained as follows. Recall that the goal is to check (15) and we apply the expression for $A_{r,n}(\Lambda)$ given in (16).

- For $r=1$:

$$A_{1,n}(\Lambda)=\int_{\mathbb{R}}|\omega|^n \mathrm{d}\Lambda(\omega)=(n-1)!!\begin{cases}1 & \text{if } n \text{ is even}\\ \sqrt{\frac{2}{\pi}} & \text{if } n \text{ is odd}\end{cases}\leq (n-1)!!\leq (n-1)!\leq \frac{n!}{2},$$

where the last inequality is equivalent to $2\leq n$. Hence, (15) is satisfied with $K=1$.

- For $r=2$: In this case nr is even and $A_{2,n}(\Lambda)=\frac{(2n-1)!!}{3^{\frac{n}{2}}}$ by (16). For (15), it is enough ($K^{n-2}\leq K^n$) that for some $K\geq 1$ and for $n=2,3,\ldots$

$$A_{2,n}(\Lambda)\leq \frac{n!}{2}K^n \Leftrightarrow \underbrace{(2n-1)!!}_{\frac{2^n}{\sqrt{\pi}}\Gamma\left(n+\frac{1}{2}\right)}\leq \underbrace{n!}_{\Gamma(n+1)}\frac{1}{2}\left(\sqrt{3}K\right)^n \overset{(a)}{\Leftarrow} \frac{2^n}{\sqrt{\pi}}\leq \frac{1}{2}\left(\sqrt{3}K\right)^n \Leftrightarrow \frac{2}{\sqrt{\pi}}\leq \left(\frac{\sqrt{3}K}{2}\right)^n. \quad (23)$$

In (a) we used that $\Gamma\left(n+\frac{1}{2}\right)\leq \Gamma(n+1)$ for $n\geq 2$. (23) holds e.g. with $K=2$ since $1<\frac{2}{\sqrt{\pi}}<\sqrt{3}$.

- For $r=3$: Let us restrict n to even numbers ($n=2\ell$, $\ell\in\mathbb{N}^+$) in (15). In this case by by (16) $A_{3,n}(\Lambda)=\frac{(3n-1)!!}{(5!!)^{\frac{n}{2}}}$, and (15) can be written as

$$\underbrace{\frac{(6\ell-1)!!}{15^\ell}}_{\frac{2^{3\ell}}{\sqrt{\pi}}\Gamma\left(3\ell+\frac{1}{2}\right)\frac{1}{15^\ell}}\overset{?}{\leq}\frac{(2\ell)!}{2}K^{2\ell-2},\quad \forall \ell\in\mathbb{N}^+.$$

Using the bound, $\Gamma\left(3\ell+\frac{1}{2}\right)\geq\Gamma\left(3\ell\right)=(3\ell-1)!$, we have

$$\left(\frac{8}{15}\right)^\ell\frac{1}{\sqrt{\pi}}(3\ell-1)!\overset{?}{\leq}\frac{(2\ell)!}{2}K^{2\ell-2},\quad \forall \ell\in\mathbb{N}^+$$

should also hold. By the Stirling's formula $u!\sim\sqrt{2\pi u}\left(\frac{u}{e}\right)^u$, hence

$$\left(\frac{8}{15}\right)^\ell\frac{1}{\sqrt{\pi}}\sqrt{2\pi(3\ell-1)}\left(\frac{3\ell-1}{e}\right)^{3\ell-1}\overset{?}{\leq}\frac{\sqrt{2\pi(2\ell)}}{2}\left(\frac{2\ell}{e}\right)^{2\ell}$$

as $\ell\to\infty$. Taking $\ln(\cdot)$ however

$$\ln(\text{l.h.s.})=\ell\ln\left(\frac{8}{15}\right)+\ln\left(1/\sqrt{\pi}\right)+\ln\left(\sqrt{2\pi(3\ell-1)}\right)+(3\ell-1)\left[\ln(3\ell-1)-1\right]\sim(3\ell-1)\ln(3\ell-1),$$

$$\ln(\text{r.h.s.})=\ln\left(\sqrt{2\pi(2\ell)}\right)-\ln(2)+2\ell[\ln(2\ell)-1]\sim 2\ell\ln(2\ell).$$

Since ln(l.h.s.) is asymptotically larger than ln(r.h.s.), (15) can not hold.

- For $r=4$: nr is even, $A_{4,n}(\Lambda)=\frac{(4n-1)!!}{[7!!]^{\frac{n}{2}}}$ by (16), and (15) is equivalent to

$$\underbrace{(4n-1)!!}_{\frac{2^{2n}}{\sqrt{\pi}}\Gamma\left(2n+\frac{1}{2}\right)}\leq\frac{n!}{2}K^{n-2}\left(\sqrt{7\times5\times3}\right)^n.$$

By using the $\Gamma(z+1)=z\Gamma(z)$ recursion:

$$\Gamma\left(2n+\frac{1}{2}\right)=\underbrace{\underbrace{\left(2n-\frac{1}{2}\right)}_{1.}\underbrace{\left(2n-\frac{3}{2}\right)}_{2.}\cdots\underbrace{\left(n+\frac{3}{2}\right)}_{n-1.}}_{\geq(n-1)^{n-1}}\Gamma\left(n+\frac{3}{2}\right)\overset{?}{\leq}\underbrace{n!}_{\Gamma(n+1)}\quad\underbrace{K^{n-2}}_{K^{-1}K^{n-1}}.$$

Since $\Gamma\left(n+\frac{3}{2}\right)>\Gamma(n+1)$ for all $n\in\mathbb{N}^+$ and $f(n)=n^n$ grows faster than $g(n)=K^n$ for any fixed K, (15) can not be satisfied for all $n\geq 2$.

Acknowledgements

This work was started and partially carried out while ZSz was visiting BKS at the Department of Statistics, Pennsylvania State University; ZSz thanks for their generous support. BKS is supported by NSF-DMS-1713011.

References

[1] Ahmed El Alaoui and Michael Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems (NIPS)*, pages 775–783, 2015.

[2] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[3] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

[4] Mariusz Bojarski, Anna Choromanska, Krzysztof Choromanski, Francois Fagan, Cedric Gouy-Pailler, Anne Morvan, Nouri Sakr, Tamás Sarlós, and Jamal Atif. Structured adaptive and random spinners for fast machine learning computations. *International Conference on Artificial Intelligence and Statistics (AISTATS; PMLR)*, 54:1020–1029, 2017.

[5] Sándor Csörgö and Vilmos Totik. On how long interval is the empirical characteristic function uniformly consistent? *Acta Scientiarum Mathematicarum*, 45:141–149, 1983.

[6] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. Kernel pooling for convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2930, 2017.

[7] Bo Dai, Bo Xie, Niao He, Yingyu Liang, Anant Raj, Maria-Florina F. Balcan, and Le Song. Scalable kernel methods via doubly stochastic gradients. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3041–3049, 2014.

[8] Carlton Downey, Ahmed Hefny, Byron Boots, Boyue Li, and Geoff Gordon. Predictive state recurrent neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6053–6064, 2017.

[9] Petros Drineas and Michael W. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.

[10] Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley-Interscience, 1999.

[11] George Kimeldorf and Grace Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33:82–95, 1971.

[12] Quoc Le, Tamás Sarlós, and Alexander Smola. Fastfood - computing Hilbert space expansions in loglinear time. *International Conference on Machine Learning (ICML; PMLR)*, 28:244–252, 2013.

[13] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.

[14] Zhu Li, Jean-Franǫis Ton, Dino Oglic, and Dino Sejdinovic. A unified analysis of random Fourier features. Technical report, 2018. (`https://arxiv.org/abs/1806.09178`).

[15] David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Ilya Tolstikhin. Towards a learning theory of cause-effect inference. *International Conference on Machine Learning (ICML; PMLR)*, pages 1452–1461, 2015.

[16] Junier Oliva, Willie Neiswanger, Barnabás Póczos, Eric Xing, and Jeff Schneider. Fast function to function regression. *International Conference on Artificial Intelligence and Statistics (AISTATS; PMLR)*, pages 717–725, 2015.

[17] Vern I. Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, 2016.

[18] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1177–1184, 2007.

[19] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1313–1320, 2008.

[20] Lorenzo Rosasco, Matteo Santoro, Sofia Mosci, Alessandro Verri, and Silvia Villa. A regularization approach to nonlinear variable selection. *International Conference on Artificial Intelligence and Statistics (AISTATS; PMLR)*, 9:653–660, 2010.

[21] Lorenzo Rosasco, Silvia Villa, Sofia Mosci, Matteo Santoro, and Alessandro Verri. Nonparametric sparsity and regularization. *Journal of Machine Learning Research*, 14:1665–1714, 2013.

[22] Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. FALKON: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3891–3901, 2017.

[23] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3218–3228, 2017.

[24] Walter Rudin. *Fourier Analysis on Groups*. Wiley-Interscience, 1990.

[25] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *Conference on Learning Theory (COLT)*, pages 416–426, 2001.

[26] Lei Shi, Xin Guo, and Ding-Xuan Zhou. Hermite learning with gradient data. *Journal of Computational and Applied Mathematics*, 233:3046–3059, 2010.

[27] Bharath Sriperumbudur and Nicholas Sterge. Approximate kernel PCA using random features: Computational vs. statistical trade-off. Technical report, Pennsylvania State University, 2018. (`https://arxiv.org/abs/1706.06296`).

[28] Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(1-59), 2017.

[29] Bharath K. Sriperumbudur and Zoltán Szabó. Optimal rates for random Fourier features. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1144–1152, 2015.

[30] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

[31] Dougal J. Sutherland and Jeff Schneider. On the error of random Fourier features. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 862–871, 2015.

[32] Enayat Ullah, Poorya Mianjy, Teodor V. Marinov, and Raman Arora. Streaming kernel PCA with $\tilde{O}(n)$ random features. Technical report, 2018. (`https://arxiv.org/abs/1808.00934`).

[33] Sara van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2009.

[34] Sara van de Geer and Johannes Lederer. The Bernstein-Orlicz norm and deviation inequalities. *Probability Theory and Related Fields*, 157:225–250, 2013.

[35] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

[36] Christopher K. I. Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, pages 682–688, 2001.

[37] Yun Yang, Mert Pilanci, and Martin J. Wainwright. Randomized sketches for kernels: Fast and optimal non-parametric regression. *The Annals of Statistics*, 45:991–1023, 2017.

[38] Yiming Ying, Qiang Wu, and Colin Campbell. Learning the coordinate gradients. *Advances in Computational Mathematics*, 37:355–378, 2012.

[39] Yao-Liang Yu, Hao Cheng, Dale Schuurmans, and Csaba Szepesvári. Characterizing the representer theorem. *International Conference on Machine Learning (ICML; PMLR)*, 28:570–578, 2013.

[40] Qinyi Zhang, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, pages 1–18, 2017.

[41] Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220:456–463, 2008.