On Kernel Derivative Approximation with Random Fourier Features

Zoltán Szabó
CMAP, École Polytechnique
Route de Saclay, 91128 Palaiseau, France
zoltan.szabo@polytechnique.edu

Bharath K. Sriperumbudur
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. Only recently, precise statistical-computational trade-offs have been established for RFFs in the approximation of kernel values, kernel ridge regression, kernel PCA and SVM classification. 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 attainable for kernel derivatives using RFF as achieved 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, 40]. 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 \( k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \), the associated RKHS \( \mathcal{H}_k \subset \mathbb{R}^\mathcal{X} \), a loss function \( V : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R} \geq 0 \), and a penalty parameter \( \lambda > 0 \):

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

where \( \mathcal{H}_k \) is the Hilbert space defined by the following two properties:

1. \( k(\cdot, x) \in \mathcal{H}_k (\forall x \in \mathcal{X}) \),
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.

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

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

(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:

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

\( k(\cdot, x) \) denotes the function \( y \mapsto k(y, x) \in \mathbb{R} \) while keeping \( x \in \mathcal{X} \) fixed.
As a result, (1) becomes a finite-dimensional optimization problem determined by the pairwise similarities of the samples \( \{k(x_i, x_j)\} \):

\[
\min_{c \in \mathbb{R}^n} J_0(c) := \frac{1}{n} \sum_{i=1}^{n} V \left( y_i, \frac{1}{n} \sum_{j=1}^{n} c_j k(x_i, x_j) \right) + \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j), 
\]

(3)

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

However, in many learning problems such as non-linear variable selection [20, 21], (multi-task) gradient learning [39], semi-supervised or Hermite learning with gradient information [42, 26], or density estimation with infinite-dimensional exponential families [28], apparently considering the derivative information \( \frac{\partial f}{\partial \phi} \) other than just the function values \( \{f(x_j)\} \) turns out to be beneficial. In these tasks containing derivatives, (1) is generalized with loss functions \( V_i : \mathbb{R}^{I_i} \to \mathbb{R}^\geq 0 \) \((i = 1, \ldots, n; I_i \) denotes the cardinality of \( I_i \)) to the form

\[
\min_{f \in \mathcal{H}_k} J(f) := \frac{1}{n} \sum_{i=1}^{n} V_i \left( y_i, \{ \partial^p f(x_i) \}_{p \in I_j} \right) + \lambda \| f \|_{\mathcal{H}_k}^2.
\]

(4)

The solution of this minimization task — similar to (1)—enjoys a finite-dimensional parameterization [42]:

\[
f(x) = \sum_{j=1}^{n} \sum_{p \in I_j} c_{j,p} \partial^0 \partial^0 k(x, x_j), \quad (c_{j,p} \in \mathbb{R}),
\]

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

\[
\min_{c} J(c) = \frac{1}{n} \sum_{i=1}^{n} V_i \left( y_i, \left\{ \sum_{j=1}^{n} \sum_{p \in I_j} c_{j,p} \partial^0 \partial^0 k(x_i, x_j) \right\}_{p \in I_j} \right) + \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \partial^0 \partial^0 k(x_i, x_j),
\]

(5)

where \( c = (c_{j,p})_{j=1}^{n} \) and we used the derivative-reproducing property of kernels

\[
\partial^p f(x) = \langle f, \partial^0 \partial^0 k(\cdot, x) \rangle_{\mathcal{H}_k}.
\]

(6)

Compared to (3) where the kernel values determine the objective, (5) is determined by the kernel derivatives \( \partial^0 \partial^0 k(x_i, x_j) \).

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 [37, 9, 22], sketching [1, 38], 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 on RFF, arguably the simplest and most influential approximation scheme among these approaches. The RFF method constructs a random, low-dimensional, explicit Fourier feature map \( \phi \) for a continuous, bounded, shift-invariant kernel \( k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \) relying on the Bochner’s theorem:

\[
\hat{k}(x, y) = \langle \phi(x), \phi(y) \rangle, \quad \phi : \mathbb{R}^d \to \mathbb{R}^m.
\]

(7)

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

\[
f(x) = \langle w, \phi(x) \rangle, \quad w \in \mathbb{R}^m.
\]

(6)

This parameterization can be considered as an approximate version of the reproducing property \( f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}} \) of \( f \in \mathcal{H}_k \) is changed to \( w \in \mathbb{R}^m \) and \( k(x, \cdot) \in \mathcal{H}_k \) to \( \phi(x) \in \mathbb{R}^m \). (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 [41], convolution neural networks [6], prediction and filtering in dynamical systems [8], or bandit optimization [13].

Despite the tremendous practical success of RFFs, its theoretical understanding is quite limited, with only a few optimal guarantees [29, 23, 27, 14, 33, 31].

- Concerning the approximation quality of kernel values, the uniform finite-sample bounds of [18, 32] show that

\[
\| k - \hat{k} \|_{L^\infty(S_m \times S_m)} := \sup_{x,y \in S_m} | k(x, y) - \hat{k}(x, y) | = O_p \left( \frac{\log m}{m} \right),
\]

where \( S_m \subset \mathbb{R}^d \) is compact, \( |S_m| \) 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 \( |S_m| \) giving

\[
\| k - \hat{k} \|_{L^\infty(S_m \times S_m)} = O_{a.s.} \left( \sqrt{\frac{\log |S_m|}{m}} \right),
\]

(7)

As a recognition of its influence, the work [18] won the 10-year test-of-time award at NIPS-2017.
where \(O_{a,s}(\cdot)\) denotes almost sure convergence. This bound is optimal w.r.t. \(m\) and \(|S_m|\), 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 \(m = m_n = O(n)\) RFFs, where \(n\) denotes the number of training samples. This bound is somewhat pessimistic, leaving the usefulness of RFFs open. Recently [23] proved that \(O(1/\sqrt{n})\) generalization performance is attainable in the context of kernel ridge regression, with \(m_n = o(n) = O(\sqrt{n} \log n)\) RFFs. This result settles RFFs in the least-squares setting with Tikhonov regularization.

- [27] has investigated the computational-statistical trade-offs of RFFs in kernel principal component analysis (KPCA). Their result shows that depending on the spectrum of the covariance operator follows an exponential decay, and optimized the primal [(4)], and optimized \(O\) RFFs. Their result shows that depending on 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.

- Results of similar flavour have recently been showed in SVM classification with the 0-1 loss [31].

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

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

which is then used in the primal [(4)], and optimized for \(w\). 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 \(|S|\) can be achieved for kernel derivatives as attained for kernel values (see (7)). To the best of our knowledge, the tightest available guarantee on kernel derivatives [29] is

\[
\|\partial^p q_k - \partial^p \hat{q}_k\|_{L^\infty(S_m \times S_m)} = O_{a.s.} \left( \frac{\log |S_m|}{m} \right). \tag{9}
\]

The possibility of such an exponentially improved dependence in terms of \(|S_m|\) 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 [35].

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:** \(\mathbb{N} := \{0, 1, 2, \ldots\}, \mathbb{N}^+ := \mathbb{N}\setminus\{0\}\) and \(\mathbb{R}\) denotes the set of natural numbers, positive integers and real numbers respectively. For \(n \in \mathbb{N}\), \(n!\) denotes its factorial. \(\Gamma(t) := \int_0^\infty x^{t-1} e^{-x} dx\) 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 \(n!! = \sqrt{2\pi n^{-\frac{1}{2}} \Gamma \left( \frac{n+1}{2} \right)}\). For \(n \in \mathbb{N}\), \(e_n := \cos(\frac{\pi n}{2}) + i\sin(\frac{\pi n}{2})\) is the \(n\)th derivative of the cosine function. For multi-indices \(p, q \in \mathbb{N}^d\), \(|p| = \sum_{j=1}^d p_j\), \(|p| = \prod_{j=1}^d p_j\), and we use

\[
\partial^p h(x) := \partial^{p_1} h(x) \partial^{p_2} g(x, y) := \frac{\partial^{p_1} h}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} \partial^{p_2} g(x, y)
\]

denote partial derivatives. \([a_1, \ldots, a_M] = \sum_{i=1}^d a_i b_i\) is the inner product between \(a \in \mathbb{R}^d\) and \(b \in \mathbb{R}^d\). \(a^T\) is the transpose of \(a \in \mathbb{R}^d, \|a\|_2 = \sqrt{\langle a, a \rangle}\) is its Euclidean norm, \([a_1, \ldots, a_M] \in \mathbb{R}^{\sum_{i=1}^d d_i}\) is the concatenation of vectors \(a_i \in \mathbb{R}^{d_i}\). Let \(S \subset \mathbb{R}^d\) be a Borel set. \(M_1^+(S)\) is the set of Borel probability measures on \(S\). \(\Lambda^m = \otimes_{i=1}^m \Lambda\) is the \(m\)-fold product measure where \(\Lambda = M_1^+(S)\). \(L^1(S)\) is the Banach space of real-valued, \(r\)-power Lebesgue integrable functions on \(S\) (\(1 \leq r < \infty\), \(A f = \int_S f(\omega) d\Lambda(\omega), \lambda = \int_S f(\omega) d\Lambda(\omega)\)) and \(f \in L^1(S)\); specifically for the empirical measure, \(A_m = \frac{1}{m} \sum_{i=1}^m \delta_\omega, A_m f := \frac{1}{m} \sum_{i=1}^m f(\omega)\) is the \(1\)-dimensional empirical measure supported on \(\omega \in S\). \(\mathcal{A}_\delta = \{\omega_1 - \omega_2 : \delta_1, \delta_2 \in \mathcal{S}\}\). For positive sequences \(a_n \in \mathbb{N}, (b_n)_{n \in \mathbb{N}}, a_n = o(b_n)\) (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 \(a_n \in \mathbb{N}, (b_n)_{n \in \mathbb{N}}\) are said to be asymptotically equivalent, shortly \(a_n \sim b_n\), if \(\lim_{n \to \infty} \frac{a_n}{b_n} = 1\).
On Kernel Derivative Approximation with Random Fourier Features

denotes that $\frac{X_n}{n}$ is bounded in probability (resp. almost surely). The diameter of a compact set $A \subset \mathbb{R}^d$ is defined as $|A| := \sup_{x, y \in A} \|x - y\|_2 < \infty$. 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 $\Lambda$:

$$k(x, y) = \hat{k}(x - y) = \int_{\mathbb{R}^d} e^{\sqrt{-1} \omega^T (x - y)} d\Lambda(\omega)$$

where $\phi_\omega(x) = \cos(\omega^T x); \sin(\omega^T x)$. Theorem

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 $p, q \in \mathbb{N}^d$, assume $C_{p, q} = \sqrt{\int_{\mathbb{R}^d} \|\omega+p\|^2 \|\omega+q\|^2 d\Lambda(\omega)} / \sigma_{p, q} < \infty$ and for some constant $K \geq 1$, the following Bernstein condition holds:

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

where $\sigma_{p, q} = \sqrt{\int_{\mathbb{R}^d} \|\omega+p\|^2 \|\omega+q\|^2 d\Lambda(\omega)}$. Let $L_m = \frac{\sqrt{5}}{2\sqrt{m}}$, $C_1 = 14\sqrt{6 \ln(2)} + 1$, $C_2 = 36k[\ln(2) + 1]$ and $C_3 = 7\sqrt{6} \left(1 + \frac{\pi}{\ln \frac{7}{2}}(2)\right)$. Then for any $t > 0$ and compact set $S \subset \mathbb{R}^d$,

$$\Lambda^m \left( \left\{ \omega_{1:m} : \|\hat{\omega}_{1:m} - \hat{\hat{\omega}}_{1:m} \|_{L^\infty(S \times S)} \geq \sigma_{p, q} \right\} \right) \leq 2e^{-t}. (16)$$

Remarks.

- Growth of $|S_m|$: The theorem proves the same dependence on $m$ and $|S_m|$ as is known [see (9)] for kernel values $(p = q = 0)$. The result implies that

$$\|\hat{\hat{\omega}}_{1:m} - \hat{\hat{\omega}}_{1:m} \|_{L^\infty(S \times S)} \to 0 \quad \text{a.s.}$$

if $|S_m| = \epsilon(m)$. Specifically, if $p = q = 0$ then (13) reduces to the celebrated RFF technique [18]:

$$\tilde{k}(x, y) = \langle \varphi_0(x), \varphi_0(y) \rangle_{\mathbb{R}^m},$$

$$\varphi_0(x) = \frac{1}{\sqrt{m}} \left( (\omega_j^T x) \sin(\omega_j^T x) \right)_{j=1}^m.$$
• **Requirements for** $p = q = 0$: In this case $\sigma_{0,0} = 1$,
  
  \[ \int_{\mathbb{R}^d} |\omega|^{s+0} d\Lambda(\omega) = 1, \text{ thus } (15) \text{ holds } (K = 1). \]
  
  The only requirement is the finiteness of $C_{0,0} = \int_{\mathbb{R}^d} |\omega|^2 d\Lambda(\omega)$, which is identical to that imposed in [29, Theorem 1] for kernel values.

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

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

\[
\Lambda^n \left( \left\{ \omega_{1:m} : \left\| \partial^p \partial^q k - \partial^p \hat{k} \right\|_{L^r(\mathbb{R}^d)} \leq \sigma_{p,q} \right\} \right) \geq \frac{C_1 \sqrt{2d \ln (16|C_{p,q} + 4|)} + C_2 + C_3 \sqrt{2d \ln (16|C_{p,q} + 4|)}}{\sqrt{m}} + 2\pi^t.
\]

This shows that $\| \partial^p \partial^q k - \partial^p \hat{k} \|_{L^r(\mathbb{R}^d \times \mathbb{R}^d)} = O_{a.s.}(m^{-\frac{1}{2}}).$ Consequently, if $|S_m| \to \infty$ as $m \to \infty$ then $\partial^p \partial^q k$ is a consistent estimator of $\partial^p \partial^q k$ in $L^2(\mathbb{R}^d \times \mathbb{R}^d)$-norm provided $m^{-\frac{1}{2}} \left[ \frac{2d}{\sqrt{\log |S_m|}} \right]^{\frac{m}{m-\infty}} = 0$.

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

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

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

\[
\sigma^{nr} (nr - 1)!! \left\{ \begin{array}{ll} 1 & \text{if } nr \text{ is even} \\ \sqrt{\frac{2}{\pi}} & \text{if } nr \text{ is odd} \end{array} \right. \frac{1}{[\sigma^{2r} (2r-1)!!]^{\frac{1}{2}}} = \frac{(nr - 1)!! \left\{ \begin{array}{ll} 1 & \text{if } nr \text{ is even} \\ \sqrt{\frac{2}{\pi}} & \text{if } nr \text{ is odd} \end{array} \right.}{[(2r - 1)!!]^{\frac{1}{2}}}. (16)
\]

Since $A_{r,n}(\Lambda)$ does not depend on $\sigma$, one can assume that $\sigma = \sqrt{\int_{\mathbb{R}^d} |\omega|^2 d\Lambda(\omega)} = 1$ and $\Lambda = N(0, 1)$.

Exploiting the analytical expression obtained for $A_{r,n}(\Lambda)$ one can show (Section 4) that for
  
  \[ -r = 1: (15) \text{ holds with } K = 1 \text{ since } A_{1,n}(\Lambda) \leq \frac{1}{2}. \]
  \[ -r = 2: K = 2 \text{ is a suitable choice in } (15). \]
  \[ -r = 3 \text{ and } r = 4: \text{ Asymptotic argument shows that } (15) \text{ can not hold.} \]

It is an interesting open question whether one can relax (15) while maintaining similar rates, and what are the trade-offs.

• **Higher-order derivatives**: In the Gaussian example we saw that (15) holds for $r \leq 2$, but it is not satisfied for $r > 2$. For kernels with spectral densities proportional to $e^{-\omega^2}$ ($\omega \in \mathbb{R}$), it turns out that (15) is fulfilled with $r \leq 2\ell$-order derivatives; for completeness the proof is available in Section A (supplement). In other words, kernels with faster decaying spectral densities can guarantee the efficient RFF-based estimation of kernel derivatives, without deterioration in the $L^r$ and $m$-dependence.

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

By differentiating (10) one gets

\[
\partial^p \partial^q k(x, y) = \int_{\mathbb{R}^d} \omega^p (-\omega^q) c_{[p+q]} (\omega^T(x - y)) d\Lambda(\omega).
\]

By defining

\[
g_{\phi}(\omega) = \omega^{p-q} c_{[p+q]} (\omega^T z), \quad (17)
\]

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

\[
\sup_{x, y \in \mathcal{S}} |\partial^p \partial^q k(x, y) - \partial^p \hat{k}(x, y)| = \sup_{x, y \in \mathcal{S}} |\Lambda - \Lambda_m| g_{\phi},
\]

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

\[
g_{\phi}(\omega) = \cos (\omega^T z), \quad (z \in \mathcal{S}_x)
\]

which is a uniformly bounded family of functions:

\[
\sup_{x \in \mathcal{S}_x} |g_{\phi}|_{L^\infty(\mathbb{R}^d)} \leq 1.
\]

This uniform boundedness is the classical assumption of empirical process theory, which was exploited by [29] to get the optimal rates. For $p, q \in \mathbb{N}^d \setminus \{0\}$, however, the functions $g_{\phi}$ are unbounded and so $\mathcal{G}$ is no longer uniformly bounded in $L^\infty(\mathbb{R}^d)$. 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 for Gaussian kernel) 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 $\text{vol}(A) = \int_A 1 \, dx$. Let $\gamma(a,b) = \int_0^b e^{-t} t^{a-1} \, dt$ be the incomplete Gamma function $(a > 0, b \geq 0)$ that satisfies $\gamma(a+1,b) = a\gamma(a,b) - b^a e^{-b}$ and $\gamma \left( \frac{1}{2}, b \right) = \sqrt{\pi} \text{erf} \left( \sqrt{b} \right)$, where $\text{erf}(b) = \frac{2}{\sqrt{\pi}} \int_0^b e^{-t^2} \, dt$ is the error function $(b \geq 0)$. Let $(F, \rho)$ be a metric space. The $r$-covering number of $F$ is defined as the size of the smallest $r$-net, i.e., $N(r, F, \rho) = \inf \{ t \geq 1 : \exists \{ f_j \}_{j=1}^t \text{ s.t. } F \subseteq \bigcup_{j=1}^t B_r(f_j, r) \}$, where $B_r(s, r) = \{ f \in F : \rho(f, s) \leq r \}$ is the closed ball with center $s \in F$ and radius $r$. For a set of real-valued functions $F$ and $r > 0$, the cardinality of the minimal $r$-bracketing of $F$ is defined as $N_1(r, F, \rho) = \inf \{ n \geq 1 : \exists \{ (f_{j,L}, f_{j,U}) \}_{j=1}^n \text{ s.t. } f_{j,L}, f_{j,U} \in F (\forall j) \}$ such that $\rho(f_{j,L}, f_{j,U}) \leq r$ and for all $f \in F$, $\exists j : f_j \leq f \leq f_j$.

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,

\[
\| g_z \| : = \| g_z \|_{L^2(\mathbb{R}^d, \Lambda)} = \sqrt{\Lambda g_z^2} \leq \sqrt{\int_{\mathbb{R}^d} |\omega P + q|^2 \, d\Lambda(\omega)},
\]

Define $f_z(\omega) : = \frac{g_z(\omega)}{\sigma_{p, q}}$ so that

\[
\| f_z \| \leq 1 \quad \forall z \in \mathcal{S}_\Lambda \quad \Rightarrow \sup_{f \in F} \| f \| \leq 1,
\]

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

\[
\sup_{x,y \in \mathcal{S}} |\partial^p q_k(x, y) - \partial^p q_k(x, y)| = \sup_{z \in \mathcal{S}_\Lambda} |\Lambda g_z - \Lambda_m g_z| = \sigma_{p, q} \sup_{f \in \mathcal{F}} \| (\Lambda - \Lambda_m) f \| =: \sigma_{p, q} \| \Lambda - \Lambda_m \|_{\mathcal{F}}.
\]

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

\[
\sup_{f \in F, z \in \mathcal{S}_\Lambda} \| f_z \| \leq \left\{ \begin{array}{c}
\int_{\mathbb{R}^d} |\omega P + q|^n \, d\Lambda(\omega)
\leq \frac{n!}{2} K^{n-2} \quad (n = 2, 3, \ldots).
\end{array} \right.
\]

The uniform $L^2(\Lambda)$ boundedness of $\mathcal{F}$ [(18)] with its Bernstein property [(19)] imply by [35, Theorem 8] that for all $t > 0$ and for all scale $S \in \mathbb{N}$

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

where

\[
E_S := 2^{-S} \sqrt{m} + 14 \sum_{s=0}^S 2^{-s} \sqrt{6H_s} + \frac{36KH_0}{\sqrt{m}},
\]

and $H_0 = \ln(N_0 + 1)$, and $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}| \leq \frac{\eta}{2} (2K)^{-n-2} (n = 2, 3, \ldots) \}

for all $f \in \mathcal{F}$, $\exists j \in \{1, \ldots, n\}$ 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

\[
\Lambda \left( \left\{ \sup_{f \in \mathcal{F}} \| f \|^n \right\} \right) = \Lambda \left( \sup_{f \in \mathcal{F}} |f|^n \right) \leq \int_{\mathbb{R}^d} |\omega P + q|^n \, d\Lambda(\omega) \leq \frac{n!}{2} K^{n-2}, \quad n = 2, 3, \ldots.
\]

Hence $F$ also satisfies the weaker Bernstein condition: $\Lambda \left( \left\{ \sup_{f \in \mathcal{F}} |f|^n \right\} \right) \leq \frac{n!}{2} (2K)^{-n-2} (n = 2, 3, \ldots)$. Consequently, one can choose $N_0 = 1$ [35, 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 $(f_{z_1}, f_{z_2} \in \mathcal{F})$

\[
|f_{z_1}(\omega) - f_{z_2}(\omega)|
\]

\[
= |\omega P(\omega) q_c p q | (\omega^T z_1) - \omega P(\omega) q_c p q | (\omega^T z_2)|
\]

\[
= |\omega P + q | c_p q | (\omega^T z_1) - c_p q | (\omega^T z_2)|
\]

\[
\leq \frac{\sigma_{p, q}}{\alpha} |\omega P + q | |\omega^T (z_1 - z_2)| \leq \frac{\sigma_{p, q}}{\alpha} \| \omega \|_2 \| z_1 - z_2 \|_2,
\]

where $G(\omega)$. 
where we used the Lipschitz property of \(u \rightarrow c_{p+q}(u)\) (with Lipschitz constant 1) in (a) and the Cauchy-Bunyakowski-Schwarz inequality in (b). Thus, by [36, Theorem 2.7.11, page 164] for any \(\delta > 0\),

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

(21)

where

\[
\|G\| = \sqrt{\int_{\mathbb{R}^d} G^2(\omega)d\Lambda(\omega)}
= \sqrt{\int_{\mathbb{R}^d} \frac{|\omega_p+q|^2}{\sigma_{p,q}^2} \|\omega\|^2 d\Lambda(\omega)} =: C_{p,q}.
\]

From Lemma 2.5 in [34] it follows that

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

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

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

where \(K_\mathcal{S} = 8|\mathcal{S}|C_{p,q} + 1\). Thus for any \(s \geq 1\),

\[
H_s = \ln(N_s + 1) \leq d \ln \left( \frac{2^sK_\mathcal{S} + 1}{2^{s+1}C_{p,q}} \right) \leq 2^s(K_{\mathcal{S}} + 1)
\leq d \left[ s \ln(2) + \ln \left( K_{\mathcal{S}} + 1 \right) \right]
\leq s d \left[ \ln(2) + \ln \left( K_{\mathcal{S}} + 1 \right) \right] = d \ln(2K_{\mathcal{S}}+2) =: K_s.
\]

Hence,

\[
E_S \leq 14 \sum_{s=1}^S 2^{-s} \sqrt{6sK_{\mathcal{S}}} + 2^{-S} \sqrt{m}
\leq 14 \sqrt{6K_{\mathcal{S}}} \sum_{s=1}^S 2^{-s} \sqrt{s}
+ 14 \sqrt{6 \ln(2)} + \frac{36K_{\mathcal{S}} \ln(2)}{\sqrt{m}}. \quad \text{(22)}
\]

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

\[
h'(t) = \frac{\frac{1}{2} t^{-\frac{1}{2}} - 2 t^{-\frac{3}{2}} \ln(2)}{2^t} \leq 0
\]
on \(\left[ \frac{1}{2\ln(2)}, \infty \right)\). Using this monotonicity, one gets \(h(s) \leq \int_{s-1}^s h(x)dx\) for any \(s\) such that \(\frac{1}{2\ln(2)} \leq s-1 \leftrightarrow \frac{1}{2\ln(2)} + 1 \leq s\), specifically for all \(2 \leq s\) since \(\frac{1}{2\ln(2)} < 1\). Hence, applying change of variables \((2^{-x} = e^{-t}\), i.e. \(x = \frac{1}{2\ln(2)} - e^{-t})\) we arrive at

\[
\sum_{s=1}^S 2^{-s} \sqrt{s} = h(1) + \sum_{s=2}^S h(s) \leq \frac{1}{2} + \int_1^S h(x)dx
= \frac{1}{2} + \frac{1}{\ln^2(2)} \int_{t=0}^{S \ln(2)} e^{-t} \sqrt{t} dt
\leq \frac{1}{2} + \frac{1}{\ln^2(2)} \int_0^{S \ln(2)} e^{-t} \sqrt{t} dt
= \frac{1}{2} + \frac{1}{\ln^2(2)} \left[ \frac{\sqrt{\pi}}{2} \text{erf}(\sqrt{S \ln(2)}) - 2^{-S} \sqrt{S \ln(2)} \right].
\]

Plugging this estimate in (22) results in

\[
E_S \leq \frac{\sqrt{m}}{2^S} + 14 \sqrt{6 \ln(2)} + \frac{36K_{\mathcal{S}} \ln(2)}{\sqrt{m}} + 14 \sqrt{6 K_{\mathcal{S}}} \times \left( \frac{1}{2} + \frac{1}{\ln^2(2)} \left[ \frac{\sqrt{\pi}}{2} \text{erf}(\sqrt{S \ln(2)}) - \frac{\sqrt{S \ln(2)}}{2^S} \right] \right)
\leq \frac{\sqrt{m}}{2^S} + 14 \sqrt{6 K_{\mathcal{S}}} \times \left( \frac{1}{2} + \frac{1}{\ln^2(2)} \left[ \frac{\sqrt{\pi}}{2} \right] \right)
+ C_1 + \frac{C_2}{\sqrt{m}}
\leq \frac{\sqrt{m}}{2^S} + 14 \sqrt{6} \sqrt{d \ln(16|\mathcal{S}|C_{p,q} + 4)} \times \left( \frac{1}{2} + \frac{1}{\ln^2(2)} \left( \frac{\sqrt{\pi}}{2} \right) \right) + C_1 + \frac{C_2}{\sqrt{m}} =: (*)
\]

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

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

Combining this result with (20), we obtain

\[
\Lambda_m \left( \left\{ \omega_1 : \omega_1 : \|\Lambda - \Lambda_m\|_p \geq \frac{C_0}{\sqrt{m}} \sqrt{d \ln(16|\mathcal{S}|C_{p,q} + 4)} + C_1 + \frac{C_2}{\sqrt{m}} \right\} \right) \leq 2e^{-t}.
\]

By redefining \(C_1\) and \(C_2\) as \(C_1 = 14 \sqrt{6 \ln(2)} + 1\), \(C_2 = 36K_{\mathcal{S}} \ln(2) + 1\) and taking into account the \(\sigma_{p,q}\) normalization, the claimed result follows. \(\square\)
The proof of the consequence is as follows. Let \( r \in [1, \infty) \) be fixed. Then
\[
\| \partial^p q_k - \partial^p \hat{q}_k \|_{L^r(\mathbb{S} \times \mathbb{S})} = \left( \int_{\mathbb{S} \times \mathbb{S}} \left\| \partial^p q_k(x, y) - \partial^p \hat{q}_k(x, y) \right\|^r \, dx \, dy \right)^{\frac{1}{r}}
\]
\[
\leq \left( \int_{\mathbb{S} \times \mathbb{S}} \left\| \partial^p q_k - \partial^p \hat{q}_k \right\|_{L^\infty(\mathbb{S} \times \mathbb{S})} \, dx \, dy \right)^{\frac{1}{r}}
\]
\[
= \left\| \partial^p q_k - \partial^p \hat{q}_k \right\|_{L^\infty(\mathbb{S} \times \mathbb{S})} \frac{\pi d/2 |\mathbb{S}|^d}{2^{d+1} (d+1)} .
\]
Using the fact (which follows from [10, Corollary 2.55]) that
\[
\text{vol}(\mathbb{S}) \leq \frac{\pi d/2 |\mathbb{S}|^d}{2^{d+1} (d+1)},
\]
we obtain
\[
\| \partial^p q_k - \partial^p \hat{q}_k \|_{L^r(\mathbb{S} \times \mathbb{S})} \leq \frac{\pi d/2 |\mathbb{S}|^d}{2^{d+1} (d+1)} .
\]
Hence the main theorem implies the claimed \( L^r(\mathbb{S} \times \mathbb{S}) \) bound.

The result on the Bernstein condition for the Gaussian kernel 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 \( r = 1 \):
  \[
  A_{1,n}(\Lambda) = \int_{\mathbb{R}} |\omega|^n d\Lambda(\omega) = (n-1)!! \left\{ \begin{array}{ll}
  1 & \text{if } n \text{ is even} \\
  \sqrt{\pi} & \text{if } n \text{ is odd}
\end{array} \right.
  \leq (n-1)!! \leq (n-1)! \leq \frac{n!}{2} .
\]
  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 \( A_{2,n}(\Lambda) = \frac{(2n-1)!!}{2^n} \) by (16). (For (15), it is enough \( (K^{n-2} \leq K^n) \) that for some \( K \geq 1 \) and for \( n = 2, 3, \ldots \)
  \[
  A_{2,n}(\Lambda) \leq \frac{n!}{2} K^n \leq \frac{n!}{2} \left( \sqrt{3} K \right)^n .
\]

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). By (16), \( A_{3,n}(\Lambda) = \frac{(3n-1)!!}{2^n} \), and (15) can be written as
  \[
  \frac{(6\ell - 1)!!}{2n} \leq \frac{2\ell}{15} K^{2\ell-2}, \quad \forall \ell \in \mathbb{N}^+ .
\]

Using the bound, \( \Gamma(3\ell + \frac{1}{2}) \geq \Gamma(3\ell) = (3\ell - 1)! \), we have that
  \[
  \left( \frac{8}{15} \right)^{\frac{\ell}{2}} \frac{1}{\sqrt{\pi}} (3\ell - 1) \leq \frac{(2\ell)!}{2^{2\ell} K^{2\ell-2}}, \quad \forall \ell \in \mathbb{N}^+ .
\]

Hence, \( 1 \) should also hold. By the Stirling’s formula \( u! \sim \sqrt{2\pi u} \left( \frac{u}{e} \right)^u \), we have
  \[
  \left( \frac{8}{15} \right)^{\frac{\ell}{2}} \frac{1}{\sqrt{\pi}} (3\ell - 1) \leq \frac{(2\ell)!}{2^{2\ell} K^{2\ell-2}} \leq \frac{2\ell}{15} K^{2\ell-2}.
\]

As \( \ell \to \infty \). Taking \( \ln() \) yields
  \[
  \ln(\text{r.h.s.}) \leq \frac{1}{\ell} \ln \left( \frac{8}{15} \right) + \ln \left( \frac{2\ell}{e} \right) \sim \ln(2) + 2\ell \left( \ln(2\ell) - 1 \right) \sim 2\ell \ln(2) .
\]

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

- For \( r = 4 \): \( nr \) is even, \( A_{4,n}(\Lambda) = \frac{(4n-1)!!}{2^n} \) by (16), and (15) is equivalent to
  \[
  \frac{(4n-1)!!}{2^n} \leq \frac{n!}{2^n} K^{n-2} \left( \sqrt{7 \times 5 \times 3} \right)^n .
\]

By using the \( \Gamma(z+1) = z\Gamma(z) \) recursion, we obtain
  \[
  \frac{\Gamma \left( 2n + \frac{1}{2} \right)}{\Gamma \left( n+1 \right)} \leq \frac{(2n-1)!!}{2^n} K^{n-2} \left( \sqrt{7 \times 5 \times 3} \right)^n .
\]

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


