Optimal Rates for Random Fourier Features

Bharath K. Sriperumbudur*
Department of Statistics
Pennsylvania State University
University Park, PA 16802, USA
bks18@psu.edu

Zoltán Szabó*
Gatsby Unit, CSML, UCL
Sainsbury Wellcome Centre, 25 Howland Street
London - W1T 4JG, UK
zoltan.szabo@gatsby.ucl.ac.uk

Abstract

Kernel methods represent one of the most powerful tools in machine learning to tackle problems expressed in terms of function values and derivatives due to their capability to represent and model complex relations. While these methods show good versatility, they are computationally intensive and have poor scalability to large data as they require operations on Gram matrices. In order to mitigate this serious computational limitation, recently randomized constructions have been proposed in the literature, which allow the application of fast linear algorithms. Random Fourier features (RFF) are among the most popular and widely applied constructions: they provide an easily computable, low-dimensional feature representation for shift-invariant kernels. Despite the popularity of RFFs, very little is understood theoretically about their approximation quality. In this paper, we provide a detailed finite-sample theoretical analysis about the approximation quality of RFFs by (i) establishing optimal (in terms of the RFF dimension, and growing set size) performance guarantees in uniform norm, and (ii) presenting guarantees in L^r ($1 \leq r < \infty$) norms. We also propose an RFF approximation to derivatives of a kernel with a theoretical study on its approximation quality.

1 Introduction

Kernel methods [17] have enjoyed tremendous success in solving several fundamental problems of machine learning ranging from classification, regression, feature extraction, dependency estimation, causal discovery, Bayesian inference and hypothesis testing. Such a success owes to their capability to represent and model complex relations by mapping points into high (possibly infinite) dimensional feature spaces. At the heart of all these techniques is the kernel trick, which allows to *implicitly* compute inner products between these high dimensional feature maps, λ via a kernel function k: $k(\mathbf{x}, \mathbf{y}) = \langle \lambda(\mathbf{x}), \lambda(\mathbf{y}) \rangle$. However, this flexibility and richness of kernels has a price: by resorting to implicit computations these methods operate on the Gram matrix of the data, which raises serious computational challenges while dealing with large-scale data. In order to resolve this bottleneck, numerous solutions have been proposed, such as low-rank matrix approximations [25, 6, 1], explicit feature maps designed for additive kernels [23, 11], hashing [19, 9], and random Fourier features (RFF) [13] constructed for shift-invariant kernels, the focus of the current paper.

RFFs implement an extremely simple, yet efficient idea: instead of relying on the implicit feature map λ associated with the kernel, by appealing to Bochner's theorem [24]—any bounded, continuous, shift-invariant kernel is the Fourier transform of a probability measure—-[13] proposed an explicit low-dimensional random Fourier feature map ϕ obtained by empirically approximating the Fourier integral so that $k(\mathbf{x}, \mathbf{y}) \approx \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$. The advantage of this explicit low-dimensional feature representation is that the kernel machine can be efficiently solved in the primal form through fast linear solvers, thereby enabling to handle large-scale data. Through numerical experiments, it has also been demonstrated that kernel algorithms constructed using the approximate kernel do not

---

*Contributed equally.

suffer from significant performance degradation [13]. Another advantage with the RFF approach is that unlike low rank matrix approximation approach [25, 6] which also speeds up kernel machines, it approximates the entire kernel function and not just the kernel matrix. This property is particularly useful while dealing with out-of-sample data and also in online learning applications. The RFF technique has found wide applicability in several areas such as fast function-to-function regression [12], differential privacy preserving [2] and causal discovery [10].

Despite the success of the RFF method, surprisingly, very little is known about its performance guarantees. To the best of our knowledge, the only paper in the machine learning literature providing certain theoretical insight into the accuracy of kernel approximation via RFF is [13, 22]:[1] it shows that $A_m := \sup\{|k(\mathbf{x}, \mathbf{y}) - \langle\phi(\mathbf{x}), \phi(\mathbf{y})\rangle_{\mathbb{R}^{2m}}| : \mathbf{x}, \mathbf{y} \in \mathcal{S}\} = O_p(\sqrt{\log(m)/m})$ for any compact set $\mathcal{S} \subset \mathbb{R}^d$, where m is the number of random Fourier features. However, since the approximation proposed by the RFF method involves empirically approximating the Fourier integral, the RFF estimator can be thought of as an *empirical characteristic function* (ECF). In the probability literature, the systematic study of ECF-s was initiated by [7] and followed up by [5, 4, 27]. While [7] shows the almost sure (a.s.) convergence of A_m to zero, [5, Theorems 1 and 2] and [27, Theorems 6.2 and 6.3] show that the optimal rate is $m^{-1/2}$. In addition, [7] shows that almost sure convergence cannot be attained over the entire space (i.e., $\mathbb{R}^d$) if the characteristic function decays to zero at infinity. Due to this, [5, 27] study the convergence behavior of A_m when the diameter of $\mathcal{S}$ grows with m and show that almost sure convergence of A_m is guaranteed as long as the diameter of $\mathcal{S}$ is $e^{o(m)}$. Unfortunately, all these results (to the best of our knowledge) are asymptotic in nature and the only known finite-sample guarantee by [13, 22] is non-optimal. In this paper (see Section 3), we present a finite-sample probabilistic bound for A_m that holds for any m and provides the optimal rate of $m^{-1/2}$ for any compact set $\mathcal{S}$ along with guaranteeing the almost sure convergence of A_m as long as the diameter of $\mathcal{S}$ is $e^{o(m)}$. Since convergence in uniform norm might sometimes be a too strong requirement and may not be suitable to attain correct rates in the generalization bounds associated with learning algorithms involving RFF,[2] we also study the behavior of $k(\mathbf{x}, \mathbf{y}) - \langle\phi(\mathbf{x}), \phi(\mathbf{y})\rangle_{\mathbb{R}^{2m}}$ in L^r-norm ($1 \leq r < \infty$) and obtain an optimal rate of $m^{-1/2}$. The RFF approach to approximate a translation-invariant kernel can be seen as a special of the problem of approximating a function in the barycenter of a family (say $\mathcal{F}$) of functions, which was considered in [14]. However, the approximation guarantees in [14, Theorem 3.2] do not directly apply to RFF as the assumptions on $\mathcal{F}$ are not satisfied by the cosine function, which is the family of functions that is used to approximate the kernel in the RFF approach. While a careful modification of the proof of [14, Theorem 3.2] could yield $m^{-1/2}$ rate of approximation for any compact set $\mathcal{S}$, this result would still be sub-optimal by providing a linear dependence on $|\mathcal{S}|$ similar to the theorems in [13, 22], in contrast to the optimal logarithmic dependence on $|\mathcal{S}|$ that is guaranteed by our results.

Traditionally, kernel based algorithms involve computing the value of the kernel. Recently, kernel algorithms involving the derivatives of the kernel (i.e., the Gram matrix consists of derivatives of the kernel computed at training samples) have been used to address numerous machine learning tasks, e.g., semi-supervised or Hermite learning with gradient information [28, 18], nonlinear variable selection [15, 16], (multi-task) gradient learning [26] and fitting of distributions in an infinite-dimensional exponential family [20]. Given the importance of these derivative based kernel algorithms, similar to [13], in Section 4, we propose a finite dimensional random feature map approximation to kernel derivatives, which can be used to speed up the above mentioned derivative based kernel algorithms. We present a finite-sample bound that quantifies the quality of approximation in uniform and L^r-norms and show the rate of convergence to be $m^{-1/2}$ in both these cases.

A summary of our *contributions* are as follows. We

1. provide the first detailed finite-sample performance analysis of RFFs for approximating kernels and their derivatives.
2. prove uniform and L^r convergence on fixed compacts sets with optimal rate in terms of the RFF dimension (m);
3. give sufficient conditions for the growth rate of compact sets while preserving a.s. convergence uniformly and in L^r; specializing our result we match the best attainable asymptotic growth rate.

---

[1][22] derived tighter constants compared to [13] and also considered different RFF implementations.

[2]For example, in applications like kernel ridge regression based on RFF, it is more appropriate to consider the approximation guarantee in L^2 norm than in the uniform norm.

Various notations and definitions that are used throughout the paper are provided in Section 2 along with a brief review of RFF approximation proposed by [13]. The missing proofs of the results in Sections 3 and 4 are provided in the supplementary material.

2 Notations & preliminaries

In this section, we introduce notations that are used throughout the paper and then present preliminaries on kernel approximation through random feature maps as introduced by [13].

Definitions & Notation: For a topological space $\mathcal{X}$, $C(\mathcal{X})$ (*resp.* $C_b(\mathcal{X})$) denotes the space of all continuous (*resp.* bounded continuous) functions on $\mathcal{X}$. For $f \in C_b(\mathcal{X})$, $\|f\|_{\mathcal{X}} := \sup_{x \in \mathcal{X}} |f(x)|$ is the supremum norm of f. $M_b(\mathcal{X})$ and $M_+^1(\mathcal{X})$ is the set of all finite Borel and probability measures on $\mathcal{X}$, respectively. For $\mu \in M_b(\mathcal{X})$, $L^r(\mathcal{X}, \mu)$ denotes the Banach space of r-power ($r \geq 1$) μ-integrable functions. For $\mathcal{X} \subseteq \mathbb{R}^d$, we will use $L^r(\mathcal{X})$ for $L^r(\mathcal{X}, \mu)$ if μ is a Lebesgue measure on $\mathcal{X}$. For $f \in L^r(\mathcal{X}, \mu)$, $\|f\|_{L^r(\mathcal{X}, \mu)} := \left(\int_{\mathcal{X}} |f|^r \, \mathrm{d}\mu \right)^{1/r}$ denotes the L^r-norm of f for $1 \leq r < \infty$ and we write it as $\|\cdot\|_{L^r(\mathcal{X})}$ if $\mathcal{X} \subseteq \mathbb{R}^d$ and μ is the Lebesgue measure. For any $f \in L^1(\mathcal{X}, \mathbb{P})$ where $\mathbb{P} \in M_+^1(\mathcal{X})$, we define $\mathbb{P}f := \int_{\mathcal{X}} f(x) \, \mathrm{d}\mathbb{P}(x)$ and $\mathbb{P}_m f := \frac{1}{m} \sum_{i=1}^m f(X_i)$ where $(X_i)_{i=1}^m \overset{i.i.d.}{\sim} \mathbb{P}$, $\mathbb{P}_m := \frac{1}{m} \sum_{i=1}^m \delta_{X_i}$ is the empirical measure and δ_x is a Dirac measure supported on $x \in \mathcal{X}$. $\mathrm{supp}(\mathbb{P})$ denotes the support of $\mathbb{P}$. $\mathbb{P}^m := \otimes_{j=1}^m \mathbb{P}$ denotes the m-fold product measure.

For $\mathbf{v} := (v_1, \ldots, v_d) \in \mathbb{R}^d$, $\|\mathbf{v}\|_2 := \sqrt{\sum_{i=1}^d v_i^2}$. The diameter of $A \subseteq \mathcal{Y}$ where $(\mathcal{Y}, \rho)$ is a metric space is defined as $|A|_\rho := \sup\{\rho(x, y) : x, y \in \mathcal{Y}\}$. If $\mathcal{Y} = \mathbb{R}^d$ with $\rho = \|\cdot\|_2$, we denote the diameter of A as $|A|$; $|A| < \infty$ if A is compact. The volume of $A \subseteq \mathbb{R}^d$ is defined as $\mathrm{vol}(A) = \int_A 1 \, \mathrm{d}\mathbf{x}$. For $A \subseteq \mathbb{R}^d$, we define $A_\Delta := A - A = \{\mathbf{x} - \mathbf{y} : \mathbf{x}, \mathbf{y} \in A\}$. $conv(A)$ is the convex hull of A. For a function g defined on open set $B \subseteq \mathbb{R}^d \times \mathbb{R}^d$, $\partial^{\mathbf{p}, \mathbf{q}} g(\mathbf{x}, \mathbf{y}) := \frac{\partial^{|\mathbf{p}| + |\mathbf{q}|} g(\mathbf{x}, \mathbf{y})}{\partial x_1^{p_1} \cdots \partial x_d^{p_d} \partial y_1^{q_1} \cdots \partial y_d^{q_d}}$, $(\mathbf{x}, \mathbf{y}) \in B$, where $\mathbf{p}, \mathbf{q} \in \mathbb{N}^d$ are multi-indices, $|\mathbf{p}| = \sum_{j=1}^d p_j$ and $\mathbb{N} := \{0, 1, 2, \ldots\}$. Define $\mathbf{v}^{\mathbf{p}} = \prod_{j=1}^d v_j^{p_j}$. For positive sequences $(a_n)_{n \in \mathbb{N}}$, $(b_n)_{n \in \mathbb{N}}$, $a_n = o(b_n)$ if $\lim_{n \to \infty} \frac{a_n}{b_n} = 0$. $X_n = O_p(r_n)$ (*resp.* $O_{a.s.}(r_n)$) denotes that $\frac{X_n}{r_n}$ is bounded in probability (*resp.* almost surely). $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} \, \mathrm{d}x$ is the Gamma function, $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ and $\Gamma(t + 1) = t\Gamma(t)$.

Random feature maps: Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a bounded, continuous, positive definite, translation-invariant kernel, i.e., there exists a positive definite function $\psi : \mathbb{R}^d \to \mathbb{R}$ such that $k(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x} - \mathbf{y})$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ where $\psi \in C_b(\mathbb{R}^d)$. By Bochner's theorem [24, Theorem 6.6], ψ can be represented as the Fourier transform of a finite non-negative Borel measure Λ on $\mathbb{R}^d$, i.e.,

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

where $(\star)$ follows from the fact that ψ is real-valued and symmetric. Since $\Lambda(\mathbb{R}^d) = \psi(0)$, $k(\mathbf{x}, \mathbf{y}) = \psi(0) \int e^{\sqrt{-1}\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})} \mathrm{d}\mathbb{P}(\boldsymbol{\omega})$ where $\mathbb{P} := \frac{\Lambda}{\psi(0)} \in M_+^1(\mathbb{R}^d)$. Therefore, w.l.o.g., we assume throughout the paper that $\psi(0) = 1$ and so $\Lambda \in M_+^1(\mathbb{R}^d)$. Based on (1), [13] proposed an approximation to k by replacing Λ with its empirical measure, Λ_m constructed from $(\boldsymbol{\omega}_i)_{i=1}^m \overset{i.i.d.}{\sim} \Lambda$ so that resultant approximation can be written as the Euclidean inner product of finite dimensional random feature maps, i.e.,

$$\hat{k}(\mathbf{x}, \mathbf{y}) = \frac{1}{m} \sum_{i=1}^m \cos\left(\boldsymbol{\omega}_i^T(\mathbf{x} - \mathbf{y})\right) \overset{(*)}{=} \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathbb{R}^{2m}}, \qquad (2)$$

where $\phi(\mathbf{x}) = \frac{1}{\sqrt{m}} (\cos(\boldsymbol{\omega}_1^T \mathbf{x}), \ldots, \cos(\boldsymbol{\omega}_m^T \mathbf{x}), \sin(\boldsymbol{\omega}_1^T \mathbf{x}), \ldots, \sin(\boldsymbol{\omega}_m^T \mathbf{x}))$ and $(*)$ holds based on the basic trigonometric identity: $\cos(a - b) = \cos a \cos b + \sin a \sin b$. This elegant approximation to k is particularly useful in speeding up kernel-based algorithms as the finite-dimensional random feature map ϕ can be used to solve these algorithms in the primal thereby offering better computational complexity (than by solving them in the dual) while at the same time not lacking in performance. Apart from these practical advantages, [13, Claim 1] (and similarly, [22, Prop. 1]) provides a theoretical guarantee that $\|\hat{k} - k\|_{\mathcal{S} \times \mathcal{S}} \to 0$ as $m \to \infty$ for any compact set $\mathcal{S} \subset \mathbb{R}^d$. Formally, [13, Claim

1] showed that—note that (3) is slightly different but more precise than the one in the statement of Claim 1 in [13]—for any $\epsilon > 0$,

$$\Lambda^m \left(\left\{ (\boldsymbol{\omega}_i)_{i=1}^m : \|\hat{k} - k\|_{\mathcal{S} \times \mathcal{S}} \geq \epsilon \right\} \right) \leq C_d \left(|\mathcal{S}| \sigma \epsilon^{-1} \right)^{\frac{2d}{d+2}} e^{-\frac{m\epsilon^2}{4(d+2)}}, \tag{3}$$

where $\sigma^2 := \int \|\boldsymbol{\omega}\|^2 \, d\Lambda(\boldsymbol{\omega})$ and $C_d := 2^{\frac{6d+2}{d+2}} \left(\left(\frac{2}{d} \right)^{\frac{d}{d+2}} + \left(\frac{d}{2} \right)^{\frac{2}{d+2}} \right) \leq 2^7 d^{\frac{2}{d+2}}$ when $d \geq 2$. The condition $\sigma^2 < \infty$ implies that ψ (and therefore k) is twice differentiable. From (3) it is clear that the probability has polynomial tails if $\epsilon < |\mathcal{S}|\sigma$ (i.e., small ϵ) and Gaussian tails if $\epsilon \geq |\mathcal{S}|\sigma$ (i.e., large ϵ) and can be equivalently written as

$$\Lambda^m \left(\left\{ (\boldsymbol{\omega}_i)_{i=1}^m : \|\hat{k} - k\|_{\mathcal{S} \times \mathcal{S}} \geq C_d^{\frac{d+2}{2d}} |\mathcal{S}| \sigma \sqrt{m^{-1} \log m} \right\} \right) \leq m^{\frac{\alpha}{4(d+2)}} (\log m)^{-\frac{d}{d+2}}, \tag{4}$$

where $\alpha := 4d - C_d^{\frac{d+2}{d}} |\mathcal{S}|^2 \sigma^2$. For $|\mathcal{S}|$ sufficiently large (i.e., $\alpha < 0$), it follows from (4) that

$$\|\hat{k} - k\|_{\mathcal{S} \times \mathcal{S}} = O_p \left(|\mathcal{S}| \sqrt{m^{-1} \log m} \right). \tag{5}$$

While (5) shows that $\hat{k}$ is a consistent estimator of k in the topology of compact convergence (i.e., $\hat{k}$ convergences to k uniformly over compact sets), the rate of convergence of $\sqrt{(\log m)/m}$ is not optimal. In addition, the order of dependence on $|\mathcal{S}|$ is not optimal. While a faster rate (in fact, an optimal rate) of convergence is desired—better rates in (5) can lead to better convergence rates for the excess error of the kernel machine constructed using $\hat{k}$—, the order of dependence on $|\mathcal{S}|$ is also important as it determines the the number of RFF features (i.e., m) that are needed to achieve a given approximation accuracy. In fact, the order of dependence on $|\mathcal{S}|$ controls the rate at which $|\mathcal{S}|$ can be grown as a function of m when $m \to \infty$ (see Remark 1*(ii)* for a detailed discussion about the significance of growing $|\mathcal{S}|$). In the following section, we present an analogue of (4)—see Theorem 1—that provides optimal rates and has correct dependence on $|\mathcal{S}|$.

3   Main results: approximation of k

As discussed in Sections 1 and 2, while the random feature map approximation of k introduced by [13] has many practical advantages, it does not seem to be theoretically well-understood. The existing theoretical results on the quality of approximation do not provide a complete picture owing to their non-optimality. In this section, we first present our main result (see Theorem 1) that improves upon (4) and provides a rate of $m^{-1/2}$ with logarithm dependence on $|\mathcal{S}|$. We then discuss the consequences of Theorem 1 along with its optimality in Remark 1. Next, in Corollary 2 and Theorem 3, we discuss the L^r-convergence ($1 \leq r < \infty$) of $\hat{k}$ to k over compact subsets of $\mathbb{R}^d$.

Theorem 1. *Suppose $k(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x} - \mathbf{y})$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ where $\psi \in C_b(\mathbb{R}^d)$ is positive definite and $\sigma^2 := \int \|\boldsymbol{\omega}\|^2 \, \mathrm{d}\Lambda(\boldsymbol{\omega}) < \infty$. Then for any $\tau > 0$ and non-empty compact set $\mathcal{S} \subset \mathbb{R}^d$,*

$$\Lambda^m \left(\left\{ (\boldsymbol{\omega}_i)_{i=1}^m : \|\hat{k} - k\|_{\mathcal{S} \times \mathcal{S}} \geq \frac{h(d, |\mathcal{S}|, \sigma) + \sqrt{2\tau}}{\sqrt{m}} \right\} \right) \leq e^{-\tau},$$

where $h(d, |\mathcal{S}|, \sigma) := 32\sqrt{2d \log(2|\mathcal{S}| + 1)} + 32\sqrt{2d \log(\sigma + 1)} + 16\sqrt{2d[\log(2|\mathcal{S}| + 1)]^{-1}}$.

Proof (sketch). Note that $\|\hat{k} - k\|_{\mathcal{S} \times \mathcal{S}} = \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{S}} |\hat{k}(\mathbf{x}, \mathbf{y}) - k(\mathbf{x}, \mathbf{y})| = \sup_{g \in \mathcal{G}} |\Lambda_m g - \Lambda g|$, where $\mathcal{G} := \{ g_{\mathbf{x}, \mathbf{y}}(\boldsymbol{\omega}) = \cos(\boldsymbol{\omega}^T(\mathbf{x} - \mathbf{y})) : \mathbf{x}, \mathbf{y} \in \mathcal{S} \}$, which means the object of interest is the suprema of an empirical process indexed by $\mathcal{G}$. Instead of bounding $\sup_{g \in \mathcal{G}} |\Lambda_m g - \Lambda g|$ by using Hoeffding's inequality on a cover of $\mathcal{G}$ and then applying union bound as carried out in [13, 22], we use the refined technique of applying concentration via McDiarmid's inequality, followed by symmetrization and bound the Rademacher average by Dudley entropy bound. The result is obtained by carefully bounding the $L^2(\Lambda_m)$-covering number of $\mathcal{G}$. The details are provided in Section B.1 of the supplementary material. □

Remark 1. *(i)* Theorem 1 shows that $\hat{k}$ is a consistent estimator of k in the topology of compact convergence as $m \to \infty$ with the rate of a.s. convergence being $\sqrt{m^{-1} \log |\mathcal{S}|}$ (almost sure convergence is guaranteed by the first Borel-Cantelli lemma). In comparison to (4), it is clear that Theorem 1

4

provides improved rates with better constants and logarithmic dependence on $|\mathcal{S}|$ instead of a linear dependence. The logarithmic dependence on $|\mathcal{S}|$ ensures that we need $m = O(\epsilon^{-2} \log |\mathcal{S}|)$ random features instead of $O(\epsilon^{-2} |\mathcal{S}|^2 \log(|\mathcal{S}|/\epsilon))$ random features, i.e., significantly fewer features to achieve the same approximation accuracy of ϵ.

(ii) **Growing diameter:** While Theorem 1 provides almost sure convergence uniformly over compact sets, one might wonder whether it is possible to achieve uniform convergence over $\mathbb{R}^d$. [7, Section 2] showed that such a result is possible if Λ is a discrete measure but not possible for Λ that is absolutely continuous w.r.t. the Lebesgue measure (i.e., if Λ has a density). Since uniform convergence of $\hat{k}$ to k over $\mathbb{R}^d$ is not possible for many interesting k (e.g., Gaussian kernel), it is of interest to study the convergence on $\mathcal{S}$ whose diameter grows with m. Therefore, as mentioned in Section 2, the order of dependence of rates on $|\mathcal{S}|$ is critical. Suppose $|\mathcal{S}_m| \to \infty$ as $m \to \infty$ (we write $|\mathcal{S}_m|$ instead of $|\mathcal{S}|$ to show the explicit dependence on m). Then Theorem 1 shows that $\hat{k}$ is a consistent estimator of k in the topology of compact convergence if $m^{-1} \log |\mathcal{S}_m| \to 0$ as $m \to \infty$ (i.e., $|\mathcal{S}_m| = e^{o(m)}$) in contrast to the result in (4) which requires $|\mathcal{S}_m| = o(\sqrt{m/\log m})$. In other words, Theorem 1 ensures consistency even when $|\mathcal{S}_m|$ grows exponentially in m whereas (4) ensures consistency only if $|\mathcal{S}_m|$ does not grow faster than $\sqrt{m/\log m}$.

(iii) **Optimality:** Note that ψ is the characteristic function of $\Lambda \in M_+^1(\mathbb{R}^d)$ since ψ is the Fourier transform of Λ (by Bochner's theorem). Therefore, the object of interest $\|\hat{k} - k\|_{\mathcal{S} \times \mathcal{S}} = \|\hat{\psi} - \psi\|_{\mathcal{S}_\Delta}$, is the uniform norm of the difference between ψ and the empirical characteristic function $\hat{\psi} = \frac{1}{m} \sum_{i=1}^m \cos(\langle \boldsymbol{\omega}_i, \cdot \rangle)$, when both are restricted to a compact set $\mathcal{S}_\Delta \subset \mathbb{R}^d$. The question of the convergence behavior of $\|\hat{\psi} - \psi\|_{\mathcal{S}_\Delta}$ is not new and has been studied in great detail in the probability and statistics literature (e.g., see [7, 27] for $d = 1$ and [4, 5] for $d > 1$) where the characteristic function is not just a real-valued symmetric function (like ψ) but is Hermitian. [27, Theorems 6.2 and 6.3] show that the optimal rate of convergence of $\|\hat{\psi} - \psi\|_{\mathcal{S}_\Delta}$ is $m^{-1/2}$ when $d = 1$, which matches with our result in Theorem 1. Also Theorems 1 and 2 in [5] show that the logarithmic dependence on $|\mathcal{S}_m|$ is optimal asymptotically. In particular, [5, Theorem 1] matches with the growing diameter result in Remark 1*(ii)*, while [5, Theorem 2] shows that if Λ is absolutely continuous w.r.t. the Lebesgue measure and if $\limsup_{m \to \infty} m^{-1} \log |\mathcal{S}_m| > 0$, then there exists a positive ε such that $\limsup_{m \to \infty} \Lambda^m(\|\hat{\psi} - \psi\|_{\mathcal{S}_{m,\Delta}} \geq \varepsilon) > 0$. This means the rate $|\mathcal{S}_m| = e^{o(m)}$ is not only the best possible in general for almost sure convergence, but if faster sequence $|\mathcal{S}_m|$ is considered then even stochastic convergence cannot be retained for any characteristic function vanishing at infinity along at least one path. While these previous results match with that of Theorem 1 (and its consequences), we would like to highlight the fact that all these previous results are asymptotic in nature whereas Theorem 1 provides a finite-sample probabilistic inequality that holds for any m. We are not aware of any such finite-sample result except for the one in [13, 22]. ∎

Using Theorem 1, one can obtain a probabilistic inequality for the L^r-norm of $\hat{k} - k$ over any compact set $\mathcal{S} \subset \mathbb{R}^d$, as given by the following result.

Corollary 2. *Suppose k satisfies the assumptions in Theorem 1. Then for any $1 \leq r < \infty$, $\tau > 0$ and non-empty compact set $\mathcal{S} \subset \mathbb{R}^d$,*

$$\Lambda^m \left(\left\{ (\boldsymbol{\omega}_i)_{i=1}^m : \|\hat{k} - k\|_{L^r(\mathcal{S})} \geq \left(\frac{\pi^{d/2} |\mathcal{S}|^d}{2^d \Gamma(\frac{d}{2} + 1)} \right)^{2/r} \frac{h(d, |\mathcal{S}|, \sigma) + \sqrt{2\tau}}{\sqrt{m}} \right\} \right) \leq e^{-\tau},$$

where $\|\hat{k} - k\|_{L^r(\mathcal{S})} := \|\hat{k} - k\|_{L^r(\mathcal{S} \times \mathcal{S})} = \left(\int_{\mathcal{S}} \int_{\mathcal{S}} |\hat{k}(\mathbf{x}, \mathbf{y}) - k(\mathbf{x}, \mathbf{y})|^r \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \right)^{\frac{1}{r}}.$

Proof. Note that
$$\|\hat{k} - k\|_{L^r(\mathcal{S})} \leq \|\hat{k} - k\|_{\mathcal{S} \times \mathcal{S}} \mathrm{vol}^{2/r}(\mathcal{S}).$$
The result follows by combining Theorem 1 and the fact that $\mathrm{vol}(\mathcal{S}) \leq \mathrm{vol}(A)$ where $A := \left\{ \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq \frac{|\mathcal{S}|}{2} \right\}$ and $\mathrm{vol}(A) = \frac{\pi^{d/2} |\mathcal{S}|^d}{2^d \Gamma(\frac{d}{2} + 1)}$ (which follows from [8, Corollary 2.55]). □

Corollary 2 shows that $\|\hat{k} - k\|_{L^r(\mathcal{S})} = O_{a.s.}(m^{-1/2} |\mathcal{S}|^{2d/r} \sqrt{\log |\mathcal{S}|})$ and therefore if $|\mathcal{S}_m| \to \infty$ as $m \to \infty$, then consistency of $\hat{k}$ in $L^r(\mathcal{S}_m)$-norm is achieved as long as $m^{-1/2} |\mathcal{S}_m|^{2d/r} \sqrt{\log |\mathcal{S}_m|} \to$

0 as $m \to \infty$. This means, in comparison to the uniform norm in Theorem 1 where $|\mathcal{S}_m|$ can grow exponential in m^δ ($\delta < 1$), $|\mathcal{S}_m|$ cannot grow faster than $m^{\frac{r}{4d}}(\log m)^{-\frac{r}{4d}-\theta}$ ($\theta > 0$) to achieve consistency in L^r-norm.

Instead of using Theorem 1 to obtain a bound on $\|\hat{k} - k\|_{L^r(\mathcal{S})}$ (this bound may be weak as $\|\hat{k} - k\|_{L^r(\mathcal{S})} \leq \|\hat{k} - k\|_{\mathcal{S} \times \mathcal{S}} \mathrm{vol}^{2/r}(\mathcal{S})$ for any $1 \leq r < \infty$), a better bound (for $2 \leq r < \infty$) can be obtained by directly bounding $\|\hat{k} - k\|_{L^r(\mathcal{S})}$, as shown in the following result.

Theorem 3. *Suppose $k(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x} - \mathbf{y})$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ where $\psi \in C_b(\mathbb{R}^d)$ is positive definite. Then for any $1 < r < \infty$, $\tau > 0$ and non-empty compact set $\mathcal{S} \subset \mathbb{R}^d$,*

$$\Lambda^m\left(\left\{(\boldsymbol{\omega}_i)_{i=1}^m : \|\hat{k} - k\|_{L^r(\mathcal{S})} \geq \left(\frac{\pi^{d/2}|\mathcal{S}|^d}{2^d\Gamma(\frac{d}{2}+1)}\right)^{2/r}\left(\frac{C_r'}{m^{1-\max\{\frac{1}{2},\frac{1}{r}\}}} + \frac{\sqrt{2\tau}}{\sqrt{m}}\right)\right\}\right) \leq e^{-\tau},$$

where C_r' is the Khintchine constant given by $C_r' = 1$ for $r \in (1, 2]$ and $C_r' = \sqrt{2}\left[\Gamma\left(\frac{r+1}{2}\right)/\sqrt{\pi}\right]^{\frac{1}{r}}$ for $r \in [2, \infty)$.

Proof (sketch). As in Theorem 1, we show that $\|k - \hat{k}\|_{L^r(\mathcal{S})}$ satisfies the bounded difference property, hence by the McDiarmid's inequality, it concentrates around its expectation $\mathbb{E}\|k - \hat{k}\|_{L^r(\mathcal{S})}$. By symmetrization, we then show that $\mathbb{E}\|k - \hat{k}\|_{L^r(\mathcal{S})}$ is upper bounded in terms of $\mathbb{E}_{\boldsymbol{\varepsilon}}\|\sum_{i=1}^m \varepsilon_i \cos(\langle \boldsymbol{\omega}_i, \cdot - \cdot\rangle)\|_{L^r(\mathcal{S})}$, where $\boldsymbol{\varepsilon} := (\varepsilon_i)_{i=1}^m$ are Rademacher random variables. By exploiting the fact that $L^r(\mathcal{S})$ is a Banach space of type $\min\{r, 2\}$, the result follows. The details are provided in Section B.2 of the supplementary material. $\square$

Remark 2. Theorem 3 shows an improved dependence on $|\mathcal{S}|$ without the extra $\sqrt{\log|\mathcal{S}|}$ factor given in Corollary 2 and therefore provides a better rate for $2 \leq r < \infty$ when the diameter of $\mathcal{S}$ grows, i.e., $\|\hat{k} - k\|_{L^r(\mathcal{S}_m)} \overset{a.s.}{\to} 0$ if $|\mathcal{S}_m| = o(m^{\frac{r}{4d}})$ as $m \to \infty$. However, for $1 < r < 2$, Theorem 3 provides a slower rate than Corollary 2 and therefore it is appropriate to use the bound in Corollary 2. While one might wonder why we only considered the convergence of $\|\hat{k} - k\|_{L^r(\mathcal{S})}$ and not $\|\hat{k} - k\|_{L^r(\mathbb{R}^d)}$, it is important to note that the latter is not well-defined because $\hat{k} \notin L^r(\mathbb{R}^d)$ even if $k \in L^r(\mathbb{R}^d)$. $\blacksquare$

4 Approximation of kernel derivatives

In the previous section we focused on the approximation of the kernel function where we presented uniform and L^r convergence guarantees on compact sets for the random Fourier feature approximation, and discussed how fast the diameter of these sets can grow to preserve uniform and L^r convergence almost surely. In this section, we propose an approximation to derivatives of the kernel and analyze the uniform and L^r convergence behavior of the proposed approximation. As motivated in Section 1, the question of approximating the derivatives of the kernel through finite dimensional random feature map is also important as it enables to speed up several interesting machine learning tasks that involve the derivatives of the kernel [28, 18, 15, 16, 26, 20], see for example the recent infinite dimensional exponential family fitting technique [21], which implements this idea.

To this end, we consider k as in (1) and define $h_a := \cos(\frac{\pi a}{2} + \cdot)$, $a \in \mathbb{N}$ (in other words $h_0 = \cos$, $h_1 = -\sin$, $h_2 = -\cos$, $h_3 = \sin$ and $h_a = h_{a \bmod 4}$). For $\mathbf{p}, \mathbf{q} \in \mathbb{N}^d$, assuming $\int |\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}| \, \mathrm{d}\Lambda(\boldsymbol{\omega}) < \infty$, it follows from the dominated convergence theorem that

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

$$= \int_{\mathbb{R}^d} \boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}\left[h_{|\mathbf{p}|}(\boldsymbol{\omega}^T\mathbf{x})h_{|\mathbf{q}|}(\boldsymbol{\omega}^T\mathbf{y}) + h_{3+|\mathbf{p}|}(\boldsymbol{\omega}^T\mathbf{x})h_{3+|\mathbf{q}|}(\boldsymbol{\omega}^T\mathbf{y})\right] \mathrm{d}\Lambda(\boldsymbol{\omega}),$$

so that $\partial^{\mathbf{p},\mathbf{q}}k(\mathbf{x}, \mathbf{y})$ can be approximated by replacing Λ with Λ_m, resulting in

$$\widehat{\partial^{\mathbf{p},\mathbf{q}}k}(\mathbf{x}, \mathbf{y}) := s^{\mathbf{p},\mathbf{q}}(\mathbf{x}, \mathbf{y}) = \frac{1}{m}\sum_{j=1}^m \boldsymbol{\omega}_j^{\mathbf{p}}(-\boldsymbol{\omega}_j)^{\mathbf{q}} h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}_j^T(\mathbf{x} - \mathbf{y})\right) = \langle \phi^{\mathbf{p}}(\mathbf{x}), \phi^{\mathbf{q}}(\mathbf{y})\rangle_{\mathbb{R}^{2m}}, \quad (6)$$

where $\phi^{\mathbf{P}}(\mathbf{u}) := \frac{1}{\sqrt{m}} \left(\boldsymbol{\omega}_1^{\mathbf{P}} h_{|\mathbf{p}|}(\boldsymbol{\omega}_1^T \mathbf{u}), \cdots, \boldsymbol{\omega}_m^{\mathbf{P}} h_{|\mathbf{p}|}(\boldsymbol{\omega}_m^T \mathbf{u}), \boldsymbol{\omega}_1^{\mathbf{P}} h_{3+|\mathbf{p}|}(\boldsymbol{\omega}_1^T \mathbf{u}), \cdots, \boldsymbol{\omega}_m^{\mathbf{P}} h_{3+|\mathbf{p}|}(\boldsymbol{\omega}_m^T \mathbf{u}) \right)$ and $(\boldsymbol{\omega}_j)_{j=1}^m \overset{i.i.d.}{\sim} \Lambda$. Now the goal is to understand the behavior of $\|s^{\mathbf{p},\mathbf{q}} - \partial^{\mathbf{p},\mathbf{q}} k\|_{\mathcal{S} \times \mathcal{S}}$ and $\|s^{\mathbf{p},\mathbf{q}} - \partial^{\mathbf{p},\mathbf{q}} k\|_{L^r(\mathcal{S})}$ for $r \in [1, \infty)$, i.e., obtain analogues of Theorems 1 and 3.

As in the proof sketch of Theorem 1, while $\|s^{\mathbf{p},\mathbf{q}} - \partial^{\mathbf{p},\mathbf{q}} k\|_{\mathcal{S} \times \mathcal{S}}$ can be analyzed as the suprema of an empirical process indexed by a suitable function class (say $\mathcal{G}$), some technical issues arise because $\mathcal{G}$ is not uniformly bounded. This means McDiarmid or Talagrand's inequality cannot be applied to achieve concentration and bounding Rademacher average by Dudley entropy bound may not be reasonable. While these issues can be tackled by resorting to more technical and refined methods, in this paper, we generalize (see Theorem 4 which is proved in Section B.1 of the supplement) Theorem 1 to derivatives under the restrictive assumption that supp(Λ) is bounded (note that many popular kernels including the Gaussian do not satisfy this assumption). We also present another result (see Theorem 5) by generalizing the proof technique[3] of [13] to *unbounded* functions where the boundedness assumption of supp(Λ) is relaxed but at the expense of a worse rate (compared to Theorem 4).

Theorem 4. *Let* $\mathbf{p}, \mathbf{q} \in \mathbb{N}^d$, $T_{\mathbf{p},\mathbf{q}} := \sup_{\boldsymbol{\omega} \in \mathrm{supp}(\Lambda)} |\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|$, $C_{\mathbf{p},\mathbf{q}} := \mathbb{E}_{\boldsymbol{\omega} \sim \Lambda} \left[|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}| \|\boldsymbol{\omega}\|_2^2 \right]$, *and assume that* $C_{2\mathbf{p},2\mathbf{q}} < \infty$. *Suppose* supp$(\Lambda)$ *is bounded if* $\mathbf{p} \neq \mathbf{0}$ *and* $\mathbf{q} \neq \mathbf{0}$. *Then for any* $\tau > 0$ *and non-empty compact set* $\mathcal{S} \subset \mathbb{R}^d$,

$$\Lambda^m \left(\left\{ (\boldsymbol{\omega}_i)_{i=1}^m : \|\partial^{\mathbf{p},\mathbf{q}} k - s^{\mathbf{p},\mathbf{q}}\|_{\mathcal{S} \times \mathcal{S}} \geq \frac{H(d, \mathbf{p}, \mathbf{q}, |\mathcal{S}|) + T_{\mathbf{p},\mathbf{q}} \sqrt{2\tau}}{\sqrt{m}} \right\} \right) \leq e^{-\tau},$$

where

$$H(d, \mathbf{p}, \mathbf{q}, |\mathcal{S}|) = 32 \sqrt{2d\, T_{2\mathbf{p},2\mathbf{q}}} \left[\sqrt{U(\mathbf{p}, \mathbf{q}, |\mathcal{S}|)} + \frac{1}{2\sqrt{U(\mathbf{p}, \mathbf{q}, |\mathcal{S}|)}} + \sqrt{\log(\sqrt{C_{2\mathbf{p},2\mathbf{q}}} + 1)} \right],$$

$$U(\mathbf{p}, \mathbf{q}, |\mathcal{S}|) = \log \left(2|\mathcal{S}| T_{2\mathbf{p},2\mathbf{q}}^{-1/2} + 1 \right).$$

Remark 3. *(i)* Note that Theorem 4 reduces to Theorem 1 if $\mathbf{p} = \mathbf{q} = 0$, in which case $T_{\mathbf{p},\mathbf{q}} = T_{2\mathbf{p},2\mathbf{q}} = 1$. If $\mathbf{p} \neq \mathbf{0}$ or $\mathbf{q} \neq \mathbf{0}$, then the boundedness of supp(Λ) implies that $T_{\mathbf{p},\mathbf{q}} < \infty$ and $T_{2\mathbf{p},2\mathbf{q}} < \infty$.

(ii) **Growth of** $|\mathcal{S}_m|$**:** By the same reasoning as in Remark 1*(ii)* and Corollary 2, it follows that $\|\partial^{\mathbf{p},\mathbf{q}} k - s^{\mathbf{p},\mathbf{q}}\|_{\mathcal{S}_m \times \mathcal{S}_m} \overset{a.s.}{\to} 0$ if $|\mathcal{S}_m| = e^{o(m)}$ and $\|\partial^{\mathbf{p},\mathbf{q}} k - s^{\mathbf{p},\mathbf{q}}\|_{L^r(\mathcal{S}_m)} \overset{a.s.}{\to} 0$ if $m^{-1/2} |\mathcal{S}_m|^{2d/r} \sqrt{\log |\mathcal{S}_m|} \to 0$ (for $1 \leq r < \infty$) as $m \to \infty$. An exact analogue of Theorem 3 can be obtained (but with different constants) under the assumption that supp(Λ) is bounded and it can be shown that for $r \in [2, \infty)$, $\|\partial^{\mathbf{p},\mathbf{q}} k - s^{\mathbf{p},\mathbf{q}}\|_{L^r(\mathcal{S}_m)} \overset{a.s.}{\to} 0$ if $|\mathcal{S}_m| = o(m^{\frac{r}{4d}})$. ∎

The following result relaxes the boundedness of supp(Λ) by imposing certain moment conditions on Λ but at the expense of a worse rate. The proof relies on applying Bernstein inequality at the elements of a net (which exists by the compactness of $\mathcal{S}$) combined with a union bound, and extending the approximation error from the anchors by a probabilistic Lipschitz argument.

Theorem 5. *Let* $\mathbf{p}, \mathbf{q} \in \mathbb{N}^d$, ψ *be continuously differentiable,* $\mathbf{z} \mapsto \nabla_{\mathbf{z}} \left[\partial^{\mathbf{p},\mathbf{q}} k(\mathbf{z}) \right]$ *be continuous,* $\mathcal{S} \subset \mathbb{R}^d$ *be any non-empty compact set,* $D_{\mathbf{p},\mathbf{q},\mathcal{S}} := \sup_{\mathbf{z} \in conv(\mathcal{S}_\Delta)} \|\nabla_{\mathbf{z}} \left[\partial^{\mathbf{p},\mathbf{q}} k(\mathbf{z}) \right]\|_2$ *and* $E_{\mathbf{p},\mathbf{q}} := \mathbb{E}_{\boldsymbol{\omega} \sim \Lambda} \left[|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}| \|\boldsymbol{\omega}\|_2 \right]$. *Assume that* $E_{\mathbf{p},\mathbf{q}} < \infty$. *Suppose* $\exists L > 0, \sigma > 0$ *such that*

$$\mathbb{E}_{\boldsymbol{\omega} \sim \Lambda} \left[|f(\mathbf{z}; \boldsymbol{\omega})|^M \right] \leq \frac{M! \sigma^2 L^{M-2}}{2} \quad (\forall M \geq 2, \forall \mathbf{z} \in \mathcal{S}_\Delta), \tag{7}$$

---

[3]We also correct some technical issues in the proof of [13, Claim 1], where (i) a shift-invariant argument was applied to the *non-shift invariant* kernel estimator $\hat{k}(\mathbf{x}, \mathbf{y}) = \frac{1}{m} \sum_{j=1}^m 2\cos(\boldsymbol{\omega}_j^T \mathbf{x} + b_j) \cos(\boldsymbol{\omega}_j^T \mathbf{y} + b_j) = \frac{1}{m} \sum_{j=1}^m \left[\cos(\boldsymbol{\omega}_j^T (\mathbf{x} - \mathbf{y})) + \cos(\boldsymbol{\omega}_j^T (\mathbf{x} + \mathbf{y}) + 2b_j) \right]$, (ii) the *convexity* of $\mathcal{S}$ was not imposed leading to possibly undefined Lipschitz constant (L) and (iii) the randomness of $\boldsymbol{\Delta}^* = \arg\max_{\boldsymbol{\Delta} \in \mathcal{S}_\Delta} \|\nabla [k(\boldsymbol{\Delta}) - \hat{k}(\boldsymbol{\Delta})]\|_2$ was not taken into account, thus the upper bound on the expectation of the squared Lipschitz constant ($\mathbb{E}[L^2]$) does not hold.

where $f(\mathbf{z}; \boldsymbol{\omega}) = \partial^{\mathbf{p},\mathbf{q}} k(\mathbf{z}) - \boldsymbol{\omega}^{\mathbf{p}}(-\boldsymbol{\omega})^{\mathbf{q}} h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T \mathbf{z}\right)$. *Define* $F_d := d^{-\frac{d}{d+1}} + d^{\frac{1}{d+1}}$.[4] *Then*

$$\Lambda^m\left(\{(\boldsymbol{\omega}_i)_{i=1}^m : \|\partial^{\mathbf{p},\mathbf{q}} k - s^{\mathbf{p},\mathbf{q}}\|_{\mathcal{S}\times\mathcal{S}} \geq \epsilon\}\right) \leq$$

$$\leq 2^{d-1} e^{-\frac{m\epsilon^2}{8\sigma^2\left(1+\frac{\epsilon L}{2\sigma^2}\right)}} + F_d 2^{\frac{4d-1}{d+1}}\left[\frac{|\mathcal{S}|(D_{\mathbf{p},\mathbf{q},\mathcal{S}} + E_{\mathbf{p},\mathbf{q}})}{\epsilon}\right]^{\frac{d}{d+1}} e^{-\frac{m\epsilon^2}{8(d+1)\sigma^2\left(1+\frac{\epsilon L}{2\sigma^2}\right)}}. \quad (8)$$

Remark 4. (i) The compactness of $\mathcal{S}$ implies that of $\mathcal{S}_\Delta$. Hence, by the continuity of $\mathbf{z} \mapsto \nabla_{\mathbf{z}}[\partial^{\mathbf{p},\mathbf{q}} k(\mathbf{z})]$, one gets $D_{\mathbf{p},\mathbf{q},\mathcal{S}} < \infty$. (7) holds if $|f(\mathbf{z};\boldsymbol{\omega})| \leq \frac{L}{2}$ and $\mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[|f(\mathbf{z};\boldsymbol{\omega})|^2\right] \leq \sigma^2$ ($\forall \mathbf{z} \in \mathcal{S}_\Delta$). If $\operatorname{supp}(\Lambda)$ is bounded, then the boundedness of f is guaranteed (see Section B.4 in the supplement).

(ii) In the special case when $\mathbf{p} = \mathbf{q} = \mathbf{0}$, our requirement boils down to the continuously differentiability of ψ, $E_{\mathbf{0},\mathbf{0}} = \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\|\boldsymbol{\omega}\|_2 < \infty$, and (7).

(iii) Note that (8) is similar to (3) and therefore based on the discussion in Section 2, one has $\|\partial^{\mathbf{p},\mathbf{q}} k - s^{\mathbf{p},\mathbf{q}}\|_{\mathcal{S}\times\mathcal{S}} = O_{a.s.}(|\mathcal{S}|\sqrt{m^{-1}\log m})$. But the advantage with Theorem 5 over [13, Claim 1] and [22, Prop. 1] is that it can handle unbounded functions. In comparison to Theorem 4, we obtain worse rates and it will be of interest to improve the rates of Theorem 5 while handling unbounded functions. ∎

5 Discussion

In this paper, we presented the first detailed theoretical analysis about the approximation quality of random Fourier features (RFF) that was proposed by [13] in the context of improving the computational complexity of kernel machines. While [13, 22] provided a probabilistic bound on the uniform approximation (over compact subsets of $\mathbb{R}^d$) of a kernel by random features, the result is not optimal. We improved this result by providing a finite-sample bound with optimal rate of convergence and also analyzed the quality of approximation in L^r-norm ($1 \leq r < \infty$). We also proposed an RFF approximation for derivatives of a kernel and provided theoretical guarantees on the quality of approximation in uniform and L^r-norms over compact subsets of $\mathbb{R}^d$.

While all the results in this paper (and also in the literature) dealt with the approximation quality of RFF over only compact subsets of $\mathbb{R}^d$, it is of interest to understand its behavior over entire $\mathbb{R}^d$. However, as discussed in Remark 1*(ii)* and in the paragraph following Theorem 3, RFF cannot approximate the kernel uniformly or in L^r-norm over $\mathbb{R}^d$. By truncating the Taylor series expansion of the exponential function, [3] proposed a non-random finite dimensional representation to approximate the Gaussian kernel which also enjoys the computational advantages of RFF. However, this representation also does not approximate the Gaussian kernel uniformly over $\mathbb{R}^d$. Therefore, the question remains whether it is possible to approximate a kernel uniformly or in L^r-norm over $\mathbb{R}^d$ but still retaining the computational advantages associated with RFF.

References

[1] A. E. Alaoui and M. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *NIPS*, 2015.

[2] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

[3] A. Cotter, J. Keshet, and N. Srebro. Explicit approximations of the Gaussian kernel. Technical report, 2011. http://arxiv.org/pdf/1109.4603.pdf.

[4] S. Csörgő. Multivariate empirical characteristic functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 55:203–229, 1981.

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

---

[4] F_d is monotonically decreasing in d, $F_1 = 2$.

[6] P. Drineas and M. 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.

[7] A. Feuerverger and R. A. Mureika. The empirical characteristic function and its applications. *Annals of Statistics*, 5(1):88–98, 1977.

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

[9] B. Kulis and K. Grauman. Kernelized locality-sensitive hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:1092–1104, 2012.

[10] D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin. Towards a learning theory of cause-effect inference. *JMLR W&CP – ICML*, pages 1452–1461, 2015.

[11] S. Maji, A. C. Berg, and J. Malik. Efficient classification for additive kernel SVMs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:66–77, 2013.

[12] J. Oliva, W. Neiswanger, B. Póczos, E. Xing, and J. Schneider. Fast function to function regression. *JMLR W&CP – AISTATS*, pages 717–725, 2015.

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

[14] A. Rahimi and B. Recht. Uniform approximation of functions with random bases. In *Allerton*, pages 555–561, 2008.

[15] L. Rosasco, M. Santoro, S. Mosci, A. Verri, and S. Villa. A regularization approach to nonlinear variable selection. *JMLR W&CP – AISTATS*, 9:653–660, 2010.

[16] L. Rosasco, S. Villa, S. Mosci, M. Santoro, and A. Verri. Nonparametric sparsity and regularization. *Journal of Machine Learning Research*, 14:1665–1714, 2013.

[17] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[18] L. Shi, X. Guo, and D.-X. Zhou. Hermite learning with gradient data. *Journal of Computational and Applied Mathematics*, 233:3046–3059, 2010.

[19] Q. Shi, J. Petterson, G. Dror, J. Langford, A. Smola, A. Strehl, and V. Vishwanathan. Hash kernels. *AISTATS*, 5:496–503, 2009.

[20] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, A. Hyvärinen, and R. Kumar. Density estimation in infinite dimensional exponential families. Technical report, 2014. http://arxiv.org/pdf/1312.3516.pdf.

[21] H. Strathmann, D. Sejdinovic, S. Livingstone, Z. Szabó, and A. Gretton. Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families. In *NIPS*, 2015.

[22] D. J. Sutherland and J. Schneider. On the error of random Fourier features. In *UAI*, pages 862–871, 2015.

[23] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:480–492, 2012.

[24] H. Wendland. *Scattered Data Approximation*. Cambridge University Press, 2005.

[25] C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *NIPS*, pages 682–688, 2001.

[26] Y. Ying, Q. Wu, and C. Campbell. Learning the coordinate gradients. *Advances in Computational Mathematics*, 37:355–378, 2012.

[27] J. E. Yukich. Some limit theorems for the empirical process indexed by functions. *Probability Theory and Related Fields*, 74:71–90, 1987.

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

Supplement

A Definitions & notation

Let (Z, ρ) be a metric space, $(\Omega, \mathcal{A})$ a measurable space and $L_0(\Omega, \mathcal{A})$ denotes the set of $(\Omega, \mathcal{A}) \mapsto \mathbb{R}$ measurable functions. A family of maps $\mathcal{G} = \{g_z\}_{z \in Z} \subseteq L_0(\Omega, \mathcal{A})$ is called a separable Carathéodory family w.r.t. Z if (Z, ρ) is separable and $z \mapsto g_z(\omega)$ is continuous for all $\omega \in \Omega$. Let $\mathcal{G} \subseteq L_0(\Omega, \mathcal{A})$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_m)$ be a Rademacher sequence, i.e., ε_j-s are i.i.d. and $\mathbb{P}(\varepsilon_j = 1) = \mathbb{P}(\varepsilon_j = -1) = \frac{1}{2}$, and $(\omega_j)_{j=1}^m \in \Omega^m$. The Rademacher average of $\mathcal{G}$ is defined as $\mathcal{R}(\mathcal{G}, \boldsymbol{\omega}_{1:m}) := \mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{g \in \mathcal{G}} \left| \frac{1}{m} \sum_{j=1}^m \varepsilon_j g(\omega_j) \right|$; we use the shorthand $\boldsymbol{\omega}_{1:m} = (\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_m)$. $S \subseteq Z$ is said to be an r-net of Z if for any $z \in Z$ there is an $s \in S$ such that $\rho(s, z) \leq r$. The r-covering number of Z is defined as the size of the smallest r-net, i.e., $\mathcal{N}(Z, \rho, r) = \inf \left\{ \ell \geq 1 : \exists\, s_1, \dots, s_\ell \text{ such that } Z \subseteq \cup_{j=1}^\ell B_\rho(s_j, r) \right\}$, where $B_\rho(s, r) = \{z \in Z : \rho(z, s) \leq r\}$ is the closed ball with center $s \in Z$ and radius r. $\log \mathcal{N}(Z, \rho, r)$ is called the metric entropy. A $(Z, \|\cdot\|)$ Banach space is said to be of type $q \in (1, 2]$ if there exists a constant $C \in \mathbb{R}$ such that the $\mathbb{E}_{\boldsymbol{\varepsilon}} \left\| \sum_{j=1}^m \varepsilon_j f_j \right\| \leq C \left(\sum_{j=1}^m \|f_j\|^q \right)^{\frac{1}{q}}$ holds for every finite set of vectors $\{f_j\}_{j=1}^m \subseteq Z$. For example, $L^r(\Omega, \mathcal{A}, \mu)$ spaces are of type $q = \min(2, r)$ [6, page 73], where the C constant only depends on r ($C = C_r$). For a $(Z, \|\cdot\|)$ normed space, Z^* denotes the space of continuous linear functionals on Z.

B Proofs

We provide proofs of the results presented in Sections 3 and 4. Lemmas used in the proofs are enlisted in Section C.

B.1 Proof of Theorems 1 and 4

Below we prove Theorem 4, thereby Theorem 1 ($\mathbf{p} = \mathbf{q} = \mathbf{0}$). The idea of the proof is as follows: *(i)* We note that

$$\|\partial^{\mathbf{p},\mathbf{q}} k - s^{\mathbf{p},\mathbf{q}}\|_{S \times S} = \sup_{\mathbf{x}, \mathbf{y} \in S} |\partial^{\mathbf{p},\mathbf{q}} k(\mathbf{x}, \mathbf{y}) - s^{\mathbf{p},\mathbf{q}}(\mathbf{x}, \mathbf{y})| = \sup_{g \in \mathcal{G}} |\Lambda g - \Lambda_m g| =: \|\Lambda - \Lambda_m\|_{\mathcal{G}}, \tag{B.1}$$

where $\mathcal{G} := \{g_{\mathbf{z}} : \mathbf{z} \in S_\Delta\}$ and $g_{\mathbf{z}} : \operatorname{supp}(\Lambda) \to \mathbb{R}$, $\boldsymbol{\omega} \mapsto \boldsymbol{\omega}^{\mathbf{p}} (-\boldsymbol{\omega})^{\mathbf{q}} h_{|\mathbf{p}+\mathbf{q}|} \left(\boldsymbol{\omega}^T \mathbf{z} \right)$, which means the object of interest is the suprema of an empirical process indexed by $\mathcal{G}$. *(ii)* We show that $\|\Lambda - \Lambda_m\|_{\mathcal{G}}$ is measurable w.r.t. Λ^m by verifying that $\mathcal{G}$ is a separable Carathéodory family (see the discussion following Definition 7.4 in [9]). *(iii)* (B.1) can be shown to satisfy the bounded difference property in C.1 and therefore by McDiarmid's inequality (Lemma C.1), $\|\Lambda - \Lambda_m\|_{\mathcal{G}}$ concentrates around its expectation. *(iv)* By applying the symmetrization lemma [9, Proposition 7.10] for the uniformly bounded function family $\mathcal{G}$, we obtain an upper bound in terms of the expected Rademacher average of $\mathcal{G}$. *(v)* The Rademacher average is bounded by the metric entropy of $\mathcal{G}$ (making use of the Dudley's entropy integral [2, Equation 4.4]), for which we can get an estimate by showing that $\mathcal{G}$ is a smoothly parametrized function class using the compactness of S_Δ.

- **$\mathcal{G}$ is a separable Carathéodory family:** $\mathcal{G}$ is a separable Carathéodory family w.r.t. S_Δ since
  1. $g_{\mathbf{z}} : \operatorname{supp}(\Lambda) \to \mathbb{R}$, $\boldsymbol{\omega} \mapsto \boldsymbol{\omega}^{\mathbf{p}} (-\boldsymbol{\omega})^{\mathbf{q}} h_{|\mathbf{p}+\mathbf{q}|} \left(\boldsymbol{\omega}^T \mathbf{z} \right)$ is measurable for all $\mathbf{z} \in S_\Delta$.
  2. $S_\Delta \subseteq \mathbb{R}^d$ is separable since $\mathbb{R}^d$ is separable.
  3. $\mathbf{z} \mapsto \boldsymbol{\omega}^{\mathbf{p}} (-\boldsymbol{\omega})^{\mathbf{q}} h_{|\mathbf{p}+\mathbf{q}|} \left(\boldsymbol{\omega}^T \mathbf{z} \right)$ is continuous for all $\boldsymbol{\omega} \in \operatorname{supp}(\Lambda)$.

- **Concentration of $\|\Lambda - \Lambda_m\|_{\mathcal{G}}$ by its bounded difference property**: By defining $f(\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_m) := \|\Lambda - \Lambda_m\|_{\mathcal{G}}$, we have that for $\forall i \in \{1, \dots, m\}$,

$$|f(\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_{i-1}, \boldsymbol{\omega}_i, \boldsymbol{\omega}_{i+1}, \dots, \boldsymbol{\omega}_m) - f(\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_{i-1}, \boldsymbol{\omega}_i', \boldsymbol{\omega}_{i+1}, \dots, \boldsymbol{\omega}_m)| =$$

$$= \left| \sup_{g \in \mathcal{G}} \left| \Lambda g - \frac{1}{m} \sum_{j=1}^m g(\boldsymbol{\omega}_j) \right| - \sup_{g \in \mathcal{G}} \left| \Lambda g - \frac{1}{m} \sum_{j=1}^m g(\boldsymbol{\omega}_j) + \frac{1}{m} \left[g(\boldsymbol{\omega}_i) - g(\boldsymbol{\omega}_i') \right] \right| \right| \leq \frac{1}{m} \sup_{g \in \mathcal{G}} |g(\boldsymbol{\omega}_i) - g(\boldsymbol{\omega}_i')|$$

$$\leq \frac{1}{m} \sup_{g \in \mathcal{G}} (|g(\boldsymbol{\omega}_i)| + |g(\boldsymbol{\omega}_i')|) \leq \frac{1}{m} \left[\sup_{g \in \mathcal{G}} |g(\boldsymbol{\omega}_i)| + \sup_{g \in \mathcal{G}} |g(\boldsymbol{\omega}_i')| \right] \leq \frac{1}{m} \left[|\boldsymbol{\omega}_i^{\mathbf{p}+\mathbf{q}}| + |(\boldsymbol{\omega}_i')^{\mathbf{p}+\mathbf{q}}| \right] \leq \frac{2T_{\mathbf{p},\mathbf{q}}}{m}.$$

Applying McDiarmid's inequality (Lemma C.1) to f, for any $\tau > 0$, with probability at least $1 - e^{-\tau}$ over the choice of $(\boldsymbol{\omega}_i)_{i=1}^m \overset{i.i.d.}{\sim} \Lambda$,

$$\|\Lambda - \Lambda_m\|_{\mathcal{G}} \leq \mathbb{E}_{\boldsymbol{\omega}_{1:m}} \|\Lambda - \Lambda_m\|_{\mathcal{G}} + T_{\mathbf{p},\mathbf{q}} \sqrt{\frac{2\tau}{m}}. \tag{B.2}$$

- **Bounding $\mathbb{E}_{\boldsymbol{\omega}_{1:m}} \|\Lambda - \Lambda_m\|_{\mathcal{G}}$:** By the symmetrization lemma [9, Proposition 7.10] applied for the uniformly bounded function family $\mathcal{G}$ ($\sup_{g \in \mathcal{G}} \|g\|_\infty \leq T_{\mathbf{p},\mathbf{q}} < \infty$), we have

$$\mathbb{E}_{\boldsymbol{\omega}_{1:m}} \|\Lambda - \Lambda_m\|_{\mathcal{G}} \leq 2 \, \mathbb{E}_{\boldsymbol{\omega}_{1:m}} \mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right). \tag{B.3}$$

- **Bounding $\mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right)$:** Using Dudley's entropy integral [2, Equation 4.4], we have

$$\mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right) \leq \frac{8\sqrt{2}}{\sqrt{m}} \int_0^{|\mathcal{G}|_{L^2(\Lambda_m)}} \sqrt{\log \mathcal{N}(\mathcal{G}, L^2(\Lambda_m), r)} \, \mathrm{d}r. \tag{B.4}$$

The upper limit of the integral can be bounded as

$$|\mathcal{G}|_{L^2(\Lambda_m)} = \sup_{g_1, g_2 \in \mathcal{G}} \|g_1 - g_2\|_{L^2(\Lambda_m)} \leq \sup_{g_1, g_2 \in \mathcal{G}} \left(\|g_1\| + \|g_2\|_{L^2(\Lambda_m)}\right) \leq 2 \sup_{g \in \mathcal{G}} \|g\|_{L^2(\Lambda_m)} \overset{(*)}{\leq} 2\sqrt{T_{2\mathbf{p},2\mathbf{q}}}, \tag{B.5}$$

where $(*)$ follows from

$$\sup_{g \in \mathcal{G}} \|g\|_{L^2(\Lambda_m)} = \sup_{\mathbf{z} \in \mathcal{S}_\Delta} \sqrt{\frac{1}{m} \sum_{j=1}^m g_{\mathbf{z}}^2(\boldsymbol{\omega}_j)} = \sup_{\mathbf{z} \in \mathcal{S}_\Delta} \sqrt{\frac{1}{m} \sum_{j=1}^m \left[\boldsymbol{\omega}_j^{\mathbf{P}}(-\boldsymbol{\omega}_j)^{\mathbf{q}} h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}_j^T \mathbf{z}\right)\right]^2} \leq \sqrt{\frac{1}{m} \sum_{j=1}^m \boldsymbol{\omega}_j^{2(\mathbf{p}+\mathbf{q})}} \leq \sqrt{T_{2\mathbf{p},2\mathbf{q}}}.$$

- **Bounding $\mathcal{N}(\mathcal{G}, L^2(\Lambda_m), r)$ by the compactness of $\mathcal{S}_\Delta$:** For any $g_{\mathbf{z}_1}, g_{\mathbf{z}_2} \in \mathcal{G}$,

$$\|g_{\mathbf{z}_1} - g_{\mathbf{z}_2}\|_{L^2(\Lambda_m)} = \left\|\boldsymbol{\omega} \mapsto \boldsymbol{\omega}^{\mathbf{P}}(-\boldsymbol{\omega})^{\mathbf{q}} \left(h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T \mathbf{z}_1\right) - h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T \mathbf{z}_2\right)\right)\right\|_{L^2(\Lambda_m)}.$$

By the mean value theorem, there exists $c \in (0, 1)$ such that

$$\left|h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T \mathbf{z}_1\right) - h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T \mathbf{z}_2\right)\right| \leq \left\|\nabla_{\mathbf{z}} h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T(c\mathbf{z}_1 + (1-c)\mathbf{z}_2))\right)\right\|_2 \|\mathbf{z}_1 - \mathbf{z}_2\|_2,$$

where

$$\left\|\nabla_{\mathbf{z}} h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T(c\mathbf{z}_1 + (1-c)\mathbf{z}_2))\right)\right\|_2 \leq \|\boldsymbol{\omega}\|_2.$$

Therefore,

$$\|g_{\mathbf{z}_1} - g_{\mathbf{z}_2}\|_{L^2(\Lambda_m)} \leq \sqrt{\frac{1}{m} \sum_{j=1}^m \left(\left|\boldsymbol{\omega}_j^{\mathbf{p}+\mathbf{q}}\right| \|\boldsymbol{\omega}_j\|_2 \|\mathbf{z}_1 - \mathbf{z}_2\|_2\right)^2} = \|\mathbf{z}_1 - \mathbf{z}_2\|_2 \sqrt{\frac{1}{m} \sum_{j=1}^m \left|\boldsymbol{\omega}_j^{2(\mathbf{p}+\mathbf{q})}\right| \|\boldsymbol{\omega}_j\|_2^2}. \tag{B.6}$$

(B.6) shows that the existence of an ϵ-net on $(\mathcal{S}_\Delta, \|\cdot\|_2)$ implies an $r = \epsilon\sqrt{\frac{1}{m} \sum_{j=1}^m \left|\boldsymbol{\omega}_j^{2(\mathbf{p}+\mathbf{q})}\right| \|\boldsymbol{\omega}_j\|_2^2}$-net on $(\mathcal{G}, L^2(\Lambda_m))$. In other words,

$$\mathcal{N}\left(\mathcal{G}, L^2(\Lambda_m), r\right) \leq \mathcal{N}\left(\mathcal{S}_\Delta, \|\cdot\|_2, r\left(\frac{1}{m} \sum_{j=1}^m \left|\boldsymbol{\omega}_j^{2(\mathbf{p}+\mathbf{q})}\right| \|\boldsymbol{\omega}_j\|_2^2\right)^{-\frac{1}{2}}\right).$$

Define

$$A_{\mathbf{p},\mathbf{q}} := \sqrt{\frac{1}{m} \sum_{j=1}^m \left|\boldsymbol{\omega}_j^{2(\mathbf{p}+\mathbf{q})}\right| \|\boldsymbol{\omega}_j\|_2^2}.$$

By using the fact that $\mathcal{S}_\Delta \subseteq B_{\|\cdot\|_2}\left(\mathbf{t}, \frac{|\mathcal{S}_\Delta|}{2}\right)$ for some $\mathbf{t} \in \mathbb{R}^d$ and $\mathcal{N}(B_{\|\cdot\|_2}(\mathbf{s}, R), \|\cdot\|_2, \epsilon) \leq \left(\frac{4R}{\epsilon} + 1\right)^d$ for any $\mathbf{s} \in \mathbb{R}^d$ [10, Lemma 2.5, page 20], we obtain

$$\mathcal{N}\left(\mathcal{G}, L^2(\Lambda_m), r\right) \leq \left(\frac{4|\mathcal{S}|A_{\mathbf{p},\mathbf{q}}}{r} + 1\right)^d, \tag{B.7}$$

by noting that $|\mathcal{S}_\Delta| \leq 2|\mathcal{S}|$. Using (B.5) and (B.7) in (B.4), we have

$$\mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right) \leq \frac{8\sqrt{2d}}{\sqrt{m}} \int_0^{2\sqrt{T_{2\mathbf{p},2\mathbf{q}}}} \sqrt{\log\left(\frac{4|\mathcal{S}|A_{\mathbf{p},\mathbf{q}}}{r} + 1\right)} \, \mathrm{d}r \leq \frac{8\sqrt{2d}}{\sqrt{m}} \int_0^{2\sqrt{T_{2\mathbf{p},2\mathbf{q}}}} \sqrt{\log\left(\frac{4|\mathcal{S}|A_{\mathbf{p},\mathbf{q}} + 2\sqrt{T_{2\mathbf{p},2\mathbf{q}}}}{r}\right)} \, \mathrm{d}r, \tag{B.8}$$

where in the last inequality we used the fact that $r \leq 2\sqrt{T_{2\mathbf{p},2\mathbf{q}}}$. By bounding $2|\mathcal{S}|A_{\mathbf{p},\mathbf{q}} + \sqrt{T_{2\mathbf{p},2\mathbf{q}}} \leq (2|\mathcal{S}| + \sqrt{T_{2\mathbf{p},2\mathbf{q}}})(A_{\mathbf{p},\mathbf{q}} + 1)$, (B.8) reduces to

$$
\begin{aligned}
\mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right) &\leq \frac{8\sqrt{2d}}{\sqrt{m}} \left(\int_0^{2\sqrt{T_{2\mathbf{p},2\mathbf{q}}}} \sqrt{\log \frac{2\left(2|\mathcal{S}| + \sqrt{T_{2\mathbf{p},2\mathbf{q}}}\right)}{r}} \, \mathrm{d}r + 2\sqrt{T_{2\mathbf{p},2\mathbf{q}} \log(A_{\mathbf{p},\mathbf{q}}+1)} \right) \\
&= \frac{16\sqrt{2d}}{\sqrt{m}} \sqrt{T_{2\mathbf{p},2\mathbf{q}}} \left(\int_0^1 \sqrt{\log \frac{B_{\mathbf{p},\mathbf{q}}+1}{r}} \, \mathrm{d}r + \sqrt{\log(A_{\mathbf{p},\mathbf{q}}+1)} \right),
\end{aligned} \tag{B.9}
$$

where the last equality is obtained by changing the variable of integration and defining $B_{\mathbf{p},\mathbf{q}} := \frac{2|\mathcal{S}|}{\sqrt{T_{2\mathbf{p},2\mathbf{q}}}}$. By applying Lemma C.2 to bound the integral in (B.9), we obtain

$$
\mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right) \leq \frac{16\sqrt{2d}}{\sqrt{m}} \sqrt{T_{2\mathbf{p},2\mathbf{q}}} \left(\sqrt{\log(B_{\mathbf{p},\mathbf{q}}+1)} + \frac{1}{2\sqrt{\log(B_{\mathbf{p},\mathbf{q}}+1)}} + \sqrt{\log(A_{\mathbf{p},\mathbf{q}}+1)} \right). \tag{B.10}
$$

- **Bounding the expectation of the Rademacher average**: From (B.10), we have

$$
\mathbb{E}_{\boldsymbol{\omega}_{1:m}} \mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right) \leq \frac{16\sqrt{2d}}{\sqrt{m}} \sqrt{T_{2\mathbf{p},2\mathbf{q}}} \left[\sqrt{\log(B_{\mathbf{p},\mathbf{q}}+1)} + \frac{1}{2\sqrt{\log(B_{\mathbf{p},\mathbf{q}}+1)}} + \sqrt{\log\left(\sqrt{C_{2\mathbf{p},2\mathbf{q}}}+1\right)} \right], \tag{B.11}
$$

which is obtained by repeated applications of Jensen's inequality to bound $\mathbb{E}_{\boldsymbol{\omega}_{1:m}} \sqrt{\log(A_{\mathbf{p},\mathbf{q}}+1)} \leq \sqrt{\mathbb{E}_{\boldsymbol{\omega}_{1:m}} \log(A_{\mathbf{p},\mathbf{q}}+1)} \leq \sqrt{\log(\mathbb{E}_{\boldsymbol{\omega}_{1:m}} A_{\mathbf{p},\mathbf{q}}+1)}$ where $\mathbb{E}_{\boldsymbol{\omega}_{1:m}} A_{\mathbf{p},\mathbf{q}} \leq \sqrt{\frac{1}{m} \sum_{j=1}^m \mathbb{E}_{\boldsymbol{\omega}_j} \left[\left| \boldsymbol{\omega}_j^{2(\mathbf{p}+\mathbf{q})} \right| \|\boldsymbol{\omega}_j\|_2^2 \right]} \leq \sqrt{C_{2\mathbf{p},2\mathbf{q}}}$.

- **Final bound:** Combining (B.2), (B.3) and (B.11) yields the result. $\qquad\square$

B.2 Proof of Theorem 3

Below we prove Theorem 3: (i) We show that $f(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_m) := \|k - \hat{k}\|_{L^r(\mathcal{S})}$ satisfies the bounded difference property, hence by the McDiarmid's inequality (Lemma C.1) it concentrates around its expectation $\mathbb{E}\|k - \hat{k}\|_{L^r(\mathcal{S})}$. (ii) By $L^r(\mathcal{S}) = \left[L^{\tilde{r}}(\mathcal{S})\right]^*$ ($\frac{1}{r} + \frac{1}{\tilde{r}} = 1$), the separability of $L^{\tilde{r}}(\mathcal{S})$ and the symmetrization lemma [11, Lemma 2.3.1] the value of $\mathbb{E}\|k - \hat{k}\|_{L^r(\mathcal{S})}$ is upper bounded in terms of $\mathbb{E}_{\boldsymbol{\varepsilon}} \|\sum_{i=1}^m \varepsilon_i \cos(\langle \boldsymbol{\omega}_i, \cdot - \cdot \rangle)\|_{L^r(\mathcal{S})}$. (iii) Exploiting that $L^r(\mathcal{S})$ is of type $\min(r, 2)$ with a constant independent of $\mathcal{S}$, we get the result.

- **Concentration of $\|k - \hat{k}\|_{L^r(\mathcal{S})}$ by its bounded difference property**: Define $\hat{k}_i(\mathbf{x}, \mathbf{y}) = \frac{1}{m} \sum_{j \neq i} \cos(\boldsymbol{\omega}_j^T(\mathbf{x} - \mathbf{y})) + \frac{1}{m} \cos(\tilde{\boldsymbol{\omega}}_i^T(\mathbf{x} - \mathbf{y}))$ where $\tilde{\boldsymbol{\omega}}_i$ is an i.i.d. copy of $\boldsymbol{\omega}_i$. Then $\|k - \hat{k}\|_{L^r(\mathcal{S})}$ satisfies the bounded difference property in (C.1):

$$
\sup_{(\boldsymbol{\omega}_i)_{i=1}^m, \tilde{\boldsymbol{\omega}}_i} \left| \|k - \hat{k}\|_{L^r(\mathcal{S})} - \|k - \hat{k}_i\|_{L^r(\mathcal{S})} \right| \leq \sup_{(\boldsymbol{\omega}_i)_{i=1}^m, \tilde{\boldsymbol{\omega}}_i} \|\hat{k}_i - \hat{k}\|_{L^r(\mathcal{S})} \leq \frac{2}{m} \sup_{\boldsymbol{\omega}_i} \|\cos(\langle \boldsymbol{\omega}_i, \cdot - \cdot \rangle)\|_{L^r(\mathcal{S})} \leq \frac{2}{m} \mathrm{vol}^{2/r}(\mathcal{S})
$$

and therefore by McDiarmid's inequality (Lemma C.1), for any $\tau > 0$, with probability at least $1 - e^{-\tau}$ over the choice of $(\boldsymbol{\omega}_i)_{i=1}^m \sim \Lambda$, we have

$$
\|k - \hat{k}\|_{L^r(\mathcal{S})} \leq \mathbb{E}_{\boldsymbol{\omega}_{1:m}} \|k - \hat{k}\|_{L^r(\mathcal{S})} + \mathrm{vol}^{2/r}(\mathcal{S}) \sqrt{\frac{2\tau}{m}}. \tag{B.12}
$$

- **Symmetrization, reduction to $\mathbb{E}_{\boldsymbol{\varepsilon}} \|\sum_{i=1}^m \varepsilon_i \cos(\langle \boldsymbol{\omega}_i, \cdot - \cdot \rangle)\|_{L^r(\mathcal{S})}$**: Let $\tilde{r}$ be the dual exponent of r, in other words $\frac{1}{r} + \frac{1}{\tilde{r}} = 1$. Then, by $L^r(\mathcal{S}) = \left[L^{\tilde{r}}(\mathcal{S})\right]^*$ and the separability of $L^{\tilde{r}}(\mathcal{S})$, there exists (see Lemma C.4) a countable $\mathcal{G} \subseteq L^{\tilde{r}}(\mathcal{S})$ ($\forall g \in \mathcal{G}$, $\|g\|_{L^{\tilde{r}}(\mathcal{S})} = 1$) such that

$$
\|k - \hat{k}\|_{L^r(\mathcal{S})} = \sup_{g \in \mathcal{G}} \left| \int_{\mathcal{S} \times \mathcal{S}} g(\mathbf{x}, \mathbf{y}) \left[k(\mathbf{x}, \mathbf{y}) - \hat{k}(\mathbf{x}, \mathbf{y}) \right] \mathrm{d}\mathbf{x} \mathrm{d}\mathbf{y} \right|. \tag{B.13}
$$

One can rewrite the argument of this supremum by Eqs. (1)-(2) as

$$
\begin{aligned}
\int_{\mathcal{S} \times \mathcal{S}} g(\mathbf{x}, \mathbf{y}) \left[k(\mathbf{x}, \mathbf{y}) - \hat{k}(\mathbf{x}, \mathbf{y}) \right] \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} &= \int_{\mathcal{S} \times \mathcal{S}} g(\mathbf{x}, \mathbf{y}) \left[\int_{\mathbb{R}^d} \cos(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})) \mathrm{d}(\Lambda - \Lambda_m)(\boldsymbol{\omega}) \right] \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \\
&= \int_{\mathbb{R}^d} \left[\int_{\mathcal{S} \times \mathcal{S}} g(\mathbf{x}, \mathbf{y}) \cos(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \right] \mathrm{d}(\Lambda - \Lambda_m)(\boldsymbol{\omega}),
\end{aligned}
$$

and thus

$$\|k - \hat{k}\|_{L^r(\mathcal{S})} = \sup_{\tilde{g} \in \tilde{\mathcal{G}}} |(\Lambda - \Lambda_m)\tilde{g}|, \tag{B.14}$$

where $\tilde{\mathcal{G}} := \{\tilde{g}_g : g \in \mathcal{G}\}$, $\tilde{g}_g(\boldsymbol{\omega}) = \int_{\mathcal{S} \times \mathcal{S}} g(\mathbf{x}, \mathbf{y}) \cos(\boldsymbol{\omega}^T(\mathbf{x} - \mathbf{y})) d\mathbf{x} d\mathbf{y}$ and $\tilde{g}_g$ is continuous. Hence, using (B.14) with the symmetrization lemma [11, Lemma 2.3.1] and (B.13), we have

$$\mathbb{E}_{\boldsymbol{\omega}_{1:m}}\|k - \hat{k}\|_{L^r(\mathcal{S})} \leq 2\mathbb{E}_{\boldsymbol{\omega}_{1:m}}\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{\tilde{g} \in \tilde{\mathcal{G}}} \left| \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i \tilde{g}(\boldsymbol{\omega}_i) \right| = \frac{2}{m} \mathbb{E}_{\boldsymbol{\omega}_{1:m}}\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{g \in \mathcal{G}} \left| \sum_{i=1}^{m} \varepsilon_i \int_{\mathcal{S} \times \mathcal{S}} g(\mathbf{x}, \mathbf{y}) \cos\left(\boldsymbol{\omega}_i^T(\mathbf{x} - \mathbf{y})\right) d\mathbf{x} d\mathbf{y} \right|$$

$$= \frac{2}{m} \mathbb{E}_{\boldsymbol{\omega}_{1:m}}\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{g \in \mathcal{G}} \left| \int_{\mathcal{S} \times \mathcal{S}} g(\mathbf{x}, \mathbf{y}) \left[\sum_{i=1}^{m} \varepsilon_i \cos\left(\boldsymbol{\omega}_i^T(\mathbf{x} - \mathbf{y})\right) \right] d\mathbf{x} d\mathbf{y} \right| = \frac{2}{m} \mathbb{E}_{\boldsymbol{\omega}_{1:m}}\mathbb{E}_{\boldsymbol{\varepsilon}} \left\| \sum_{i=1}^{m} \varepsilon_i \cos(\langle \boldsymbol{\omega}_i, \cdot - \cdot \rangle) \right\|_{L^r(\mathcal{S})} \tag{B,15}$$

where $(\varepsilon_i)_{i=1}^{m}$ is a Rademacher sequence and $\mathbb{E}_{\boldsymbol{\varepsilon}}$ is the conditional expectation w.r.t. $(\varepsilon_i)_{i=1}^{m}$ with $(\boldsymbol{\omega}_i)_{i=1}^{m}$ being the conditioning random variables. Notice that the measurability of $\tilde{g}_g$-s with the countable cardinality of $\tilde{\mathcal{G}}$ enabled us to write expectations instead of outer expectations in [11, Lemma 2.3.1, page 108-110], and hence in Eq. (B.15).

- **Bounding** $\mathbb{E}_{\varepsilon} \left\| \sum_{i=1}^{m} \varepsilon_i \cos(\langle \boldsymbol{\omega}_i, \cdot - \cdot \rangle) \right\|_{L^r(\mathcal{S})}$ **by the type of** $L^r(\mathcal{S})$:

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \left\| \sum_{i=1}^{m} \varepsilon_i \cos(\langle \boldsymbol{\omega}_i, \cdot - \cdot \rangle) \right\|_{L^r(\mathcal{S})} \overset{(*)}{\leq} C_r' \left(\sum_{i=1}^{m} \| \cos(\langle \boldsymbol{\omega}_i, \cdot - \cdot \rangle) \|_{L^r(\mathcal{S})}^{\min\{r,2\}} \right)^{\frac{1}{\min\{r,2\}}} \leq C_r' \text{vol}^{2/r}(\mathcal{S}) m^{\max\{\frac{1}{2}, \frac{1}{r}\}}, \tag{B.16}$$

since $L^r(\mathcal{S})$ is of type $\min(2, r)$ [6, page 73] and there exists a universal constant C_r' independent of $\mathcal{S}$ (the so-called Khintchine constant) [5, page 247] such that $(*)$ holds; in addition we used

$$\sum_{i=1}^{m} \| \cos(\langle \boldsymbol{\omega}_i, \cdot - \cdot \rangle) \|_{L^r(\mathcal{S})}^{\min\{2,r\}} = \sum_{i=1}^{m} \left(\int_{\mathcal{S} \times \mathcal{S}} \left| \cos(\boldsymbol{\omega}_i^T(\mathbf{x} - \mathbf{y})) \right|^r d\mathbf{x} d\mathbf{y} \right)^{\frac{\min\{2,r\}}{r}} \leq m \left[\text{vol}^2(\mathcal{S}) \right]^{\frac{\min\{2,r\}}{r}},$$

and $\frac{1}{\min\{2,r\}} = \max\left\{ \frac{1}{2}, \frac{1}{r} \right\}$.

Combining (B.12)–(B.16) and using the bound on vol($\mathcal{S}$) given in the proof of Corollary 2 yields the result. $\square$

B.3 Proof of Theorem 5

Below we give the detailed proof of Theorem 5. At high-level the proof goes as follows: *(i)* By the compactness of $\mathcal{S}_\Delta$ (implied by that of $\mathcal{S}$) one can take an r-net covering $\mathcal{S}_\Delta$ (for any $r > 0$). *(ii)* Small approximation error can be guaranteed at the centers of the r-net by Bernstein's inequality combined with a union bound. *(iii)* Propagation of the error from the centers to arbitrary points is achieved by Lipschitzness. *(iv)* The Lipschitz constant is, however, a random quantity and we show with high probability that it is 'not too large'. *(v)* Union bounding the two events (small errors at the centers and small Lipschitz constant) leads to a uniform bound for arbitrary r, which holds with high probability. *(vi)* Optimizing over r gives the stated result.

Formally, the proof is as follows. Let us define

$$B_{\mathbf{p},\mathbf{q},\mathcal{S}} := \mathbb{E}_{\boldsymbol{\omega} \sim \Lambda} \left[\sup_{\mathbf{z} \in conv(\mathcal{S}_\Delta)} \|\nabla_{\mathbf{z}} f(\mathbf{z}; \boldsymbol{\omega})\|_2 \right],$$

where $f(\mathbf{z}; \boldsymbol{\omega}) = \partial^{\mathbf{p},\mathbf{q}} k(\mathbf{z}) - \boldsymbol{\omega}^{\mathbf{p}}(-\boldsymbol{\omega})^{\mathbf{q}} h_{|\mathbf{p}+\mathbf{q}|}(\boldsymbol{\omega}^T \mathbf{z})$. Let us notice that since $conv(\mathcal{S}_\Delta)$ is compact (by the compactness of $\mathcal{S}_\Delta$, implied by that of $\mathcal{S}$) and $\mathbf{z} \mapsto \|\nabla_{\mathbf{z}} f(\mathbf{z}; \boldsymbol{\omega})\|_2$ is continuous, the supremum inside the expectation in $B_{\mathbf{p},\mathbf{q},\mathcal{S}}$ is finite for any $\boldsymbol{\omega}$.

- **Covering of** $\mathcal{S}_\Delta$: By the compactness of $\mathcal{S}_\Delta$ there exist an r-net with at most

$$N = \left(\frac{2|\mathcal{S}_\Delta|}{r} + 1 \right)^d \leq \left(\frac{4|\mathcal{S}|}{r} + 1 \right)^d \tag{B.17}$$

balls covering $\mathcal{S}_\Delta$ [10, Lemma 2.5, page 20], where we used that $|\mathcal{S}_\Delta| \leq 2|\mathcal{S}|$. Let us denote the centers of this r-net by $\mathbf{c}_1, \ldots, \mathbf{c}_N$.

13

- **Bounding** $\bar{f}(\mathbf{b};\boldsymbol{\omega}_{1:m}) - \bar{f}(\mathbf{a};\boldsymbol{\omega}_{1:m})$, **where** $\mathbf{a}, \mathbf{b} \in \mathcal{S}_\Delta$; $\boldsymbol{\omega}_{1:m} = (\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_m)$ **is fixed:** Let

$$\bar{f}(\mathbf{z};\boldsymbol{\omega}_{1:m}) = \frac{1}{m}\sum_{j=1}^m f(\mathbf{z};\boldsymbol{\omega}_j) = \frac{1}{m}\sum_{j=1}^m \left[\partial^{\mathbf{p},\mathbf{q}}k(\mathbf{z}) - \boldsymbol{\omega}_j^{\mathbf{p}}(-\boldsymbol{\omega}_j)^{\mathbf{q}}h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}_j^T\mathbf{z}\right)\right].$$

$\mathbf{z} \mapsto \bar{f}(\mathbf{z};\boldsymbol{\omega}_{1:m})$ is continuously differentiable since ψ is so. Thus by the mean value theorem $\exists\, t \in (0,1)$ such that

$$\bar{f}(\mathbf{b};\boldsymbol{\omega}_{1:m}) - \bar{f}(\mathbf{a};\boldsymbol{\omega}_{1:m}) = \left\langle \nabla_{\mathbf{z}}\bar{f}(t\mathbf{a}+(1-t)\mathbf{b};\boldsymbol{\omega}_{1:m}), \mathbf{b}-\mathbf{a}\right\rangle.$$

Hence by the Cauchy-Bunyakovsky-Schwarz inequality, we get

$$|\bar{f}(\mathbf{b};\boldsymbol{\omega}_{1:m}) - \bar{f}(\mathbf{a};\boldsymbol{\omega}_{1:m})| \le \left\|\nabla_{\mathbf{z}}\bar{f}(t\mathbf{a}+(1-t)\mathbf{b};\boldsymbol{\omega}_{1:m})\right\|_2 \|\mathbf{b}-\mathbf{a}\|_2 \le \sup_{\mathbf{z}\in conv(\mathcal{S}_\Delta)} \left\|\nabla_{\mathbf{z}}\bar{f}(\mathbf{z};\boldsymbol{\omega}_{1:m})\right\|_2 \|\mathbf{b}-\mathbf{a}\|_2$$

$$=: L(\boldsymbol{\omega}_{1:m})\|\mathbf{b}-\mathbf{a}\|_2, \tag{B.18}$$

where we used the compactness of $conv(\mathcal{S}_\Delta)$ (implied by that of $\mathcal{S}_\Delta$) and the continuity of the $\mathbf{z}\mapsto \left\|\nabla_{\mathbf{z}}\bar{f}(\mathbf{z};\boldsymbol{\omega}_{1:m})\right\|_2$ mapping to guarantee that $L(\boldsymbol{\omega}_{1:m})$ exists, and it is finite for any $\boldsymbol{\omega}_{1:m}$.

- **Bound on** $\mathbb{E}_{\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_m}[L(\boldsymbol{\omega}_{1:m})]$: Using the definition of $\bar{f}(\mathbf{z};\boldsymbol{\omega}_{1:m})$, the linearity of differentiation, and the triangle inequality, we get

$$\left\|\nabla_{\mathbf{z}}\bar{f}(\mathbf{z};\boldsymbol{\omega}_{1:m})\right\|_2 = \left\|\nabla_{\mathbf{z}}\left[\frac{1}{m}\sum_{j=1}^m f(\mathbf{z};\boldsymbol{\omega}_j)\right]\right\|_2 = \left\|\frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{z}}f(\mathbf{z};\boldsymbol{\omega}_j)\right\|_2 \le \frac{1}{m}\sum_{j=1}^m \|\nabla_{\mathbf{z}}f(\mathbf{z};\boldsymbol{\omega}_j)\|_2.$$

Therefore,

$$\sup_{\mathbf{z}\in conv(\mathcal{S}_\Delta)} \left\|\nabla_{\mathbf{z}}\bar{f}(\mathbf{z};\boldsymbol{\omega}_{1:m})\right\|_2 \le \frac{1}{m}\sum_{j=1}^m \sup_{\mathbf{z}\in conv(\mathcal{S}_\Delta)} \|\nabla_{\mathbf{z}}f(\mathbf{z};\boldsymbol{\omega}_j)\|_2$$

and

$$\mathbb{E}_{\boldsymbol{\omega}_{1:m}}[L(\boldsymbol{\omega}_{1:m})] = \mathbb{E}_{\boldsymbol{\omega}_{1:m}}\left[\sup_{\mathbf{z}\in conv(\mathcal{S}_\Delta)}\|\nabla_{\mathbf{z}}f(\mathbf{z};\boldsymbol{\omega}_{1:m})\|_2\right] \le \frac{1}{m}\sum_{j=1}^m \mathbb{E}_{\boldsymbol{\omega}_{1:m}}\left[\sup_{\mathbf{z}\in conv(\mathcal{S}_\Delta)}\|\nabla_{\mathbf{z}}f(\mathbf{z};\boldsymbol{\omega}_j)\|_2\right]$$

$$= \frac{1}{m}\sum_{j=1}^m B_{\mathbf{p},\mathbf{q},\mathcal{S}} = B_{\mathbf{p},\mathbf{q},\mathcal{S}}. \tag{B.19}$$

- **Bound on** $B_{\mathbf{p},\mathbf{q},\mathcal{S}}$: Note that

$$\sup_{\mathbf{z}\in conv(\mathcal{S}_\Delta)}\|\nabla_{\mathbf{z}}f(\mathbf{z};\boldsymbol{\omega})\|_2 = \sup_{\mathbf{z}\in conv(\mathcal{S}_\Delta)}\left\|\nabla_{\mathbf{z}}\left[\partial^{\mathbf{p},\mathbf{q}}k(\mathbf{z}) - \boldsymbol{\omega}^{\mathbf{p}}(-\boldsymbol{\omega})^{\mathbf{q}}h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T\mathbf{z}\right)\right]\right\|_2$$

$$\le \sup_{\mathbf{z}\in conv(\mathcal{S}_\Delta)}\left(\left\|\nabla_{\mathbf{z}}\left[\partial^{\mathbf{p},\mathbf{q}}k(\mathbf{z})\right]\right\|_2 + \left\|\nabla_{\mathbf{z}}\left[\boldsymbol{\omega}^{\mathbf{p}}(-\boldsymbol{\omega})^{\mathbf{q}}h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T\mathbf{z}\right)\right]\right\|_2\right)$$

$$\le \sup_{\mathbf{z}\in conv(\mathcal{S}_\Delta)}\left\|\nabla_{\mathbf{z}}\left[\partial^{\mathbf{p},\mathbf{q}}k(\mathbf{z})\right]\right\|_2 + \sup_{\mathbf{z}\in conv(\mathcal{S}_\Delta)}\left\|\nabla_{\mathbf{z}}\left[\boldsymbol{\omega}^{\mathbf{p}}(-\boldsymbol{\omega})^{\mathbf{q}}h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T\mathbf{z}\right)\right]\right\|_2$$

$$= D_{\mathbf{p},\mathbf{q},\mathcal{S}} + \sup_{\mathbf{z}\in conv(\mathcal{S}_\Delta)}\left\|\nabla_{\mathbf{z}}\left[\boldsymbol{\omega}^{\mathbf{p}}(-\boldsymbol{\omega})^{\mathbf{q}}h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T\mathbf{z}\right)\right]\right\|_2. \tag{B.20}$$

By the homogenity of norms ($\|a\mathbf{v}\| = |a|\|\mathbf{v}\|$), the chain rule, and $|h_a(v)| \le 1$ ($\forall a, \forall v$)

$$\left\|\nabla_{\mathbf{z}}\left[\boldsymbol{\omega}^{\mathbf{p}}(-\boldsymbol{\omega})^{\mathbf{q}}h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T\mathbf{z}\right)\right]\right\|_2 = |\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|\left\|h_{|\mathbf{p}+\mathbf{q}|+1}\left(\boldsymbol{\omega}^T\mathbf{z}\right)\boldsymbol{\omega}\right\|_2 \le |\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|\|\boldsymbol{\omega}\|_2. \tag{B.21}$$

Combining Eq. (B.20) and (B.21) results in the bound

$$B_{\mathbf{p},\mathbf{q},\mathcal{S}} = \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[\sup_{\mathbf{z}\in conv(\mathcal{S}_\Delta)}\|\nabla_{\mathbf{z}}f(\mathbf{z};\boldsymbol{\omega})\|_2\right] \le D_{\mathbf{p},\mathbf{q},\mathcal{S}} + \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|\|\boldsymbol{\omega}\|_2\right] = D_{\mathbf{p},\mathbf{q},\mathcal{S}} + E_{\mathbf{p},\mathbf{q}}. \tag{B.22}$$

- **Error propagation from the net centers:** We will use the following note to propagate the error from the net centers ($\mathbf{c}_j$, $j = 1,\ldots,N$) to an arbitrary $\mathbf{z} \in \mathcal{S}_\Delta$ point. Note: If $|\bar{f}(\mathbf{c}_j;\boldsymbol{\omega}_{1:m})| < \frac{\epsilon}{2}$ ($\forall j$) and $L(\boldsymbol{\omega}_{1:m}) < \frac{\epsilon}{2r}$, then

$$|\bar{f}(\mathbf{z};\boldsymbol{\omega}_{1:m})| < \epsilon \quad (\forall \mathbf{z} \in \mathcal{S}_\Delta). \tag{B.23}$$

Indeed

$$\left||\bar{f}(\mathbf{z};\boldsymbol{\omega}_{1:m})| - \underbrace{|\bar{f}(\mathbf{c}_j;\boldsymbol{\omega}_{1:m})|}_{<\frac{\epsilon}{2}}\right| \le |\bar{f}(\mathbf{z};\boldsymbol{\omega}_{1:m}) - \bar{f}(\mathbf{c}_j;\boldsymbol{\omega}_{1:m})| \le \underbrace{L(\boldsymbol{\omega}_{1:m})}_{<\frac{\epsilon}{2r}}\underbrace{\|\mathbf{z}-\mathbf{c}_j\|_2}_{\le r} < \frac{\epsilon}{2},$$

where we used (B.18) and our assumptions in the note, thereby yielding (B.23).

- **Guaranteeing the conditions of** (B.23) **with high probability:**
  - Notice that $\mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}[f(\mathbf{z};\boldsymbol{\omega})] = \mathbf{0}$ ($\forall \mathbf{z}$). Also since (7) holds, applying Bernstein's inequality for the individual $\mathbf{c}_j$ points (Lemma C.3; $\xi_n := f(\mathbf{c}_j;\boldsymbol{\omega}_n)$, $n = 1,\ldots,m$; $S := \sqrt{m}\sigma$) gives that for any $\eta > 0$

$$\Lambda^m\left(|\bar{f}(\mathbf{c}_j;\boldsymbol{\omega}_{1:m})| \geq \frac{\eta\sigma}{\sqrt{m}}\right) \leq e^{-\frac{1}{2}\frac{\eta^2}{1+\frac{\eta L}{\sqrt{m}\sigma}}}. \tag{B.24}$$

  Setting $\epsilon = \frac{2\eta\sigma}{\sqrt{m}}$, (B.24) is written as

$$\Lambda^m\left(|\bar{f}(\mathbf{c}_j;\boldsymbol{\omega}_{1:m})| < \frac{\epsilon}{2}\right) \geq 1 - e^{-\frac{1}{2}\frac{\left(\frac{\sqrt{m}\epsilon}{2\sigma}\right)^2}{1+\frac{\frac{\sqrt{m}\epsilon}{2\sigma}L}{\sqrt{m}\sigma}}} = 1 - e^{-\frac{m\epsilon^2}{8\sigma^2\left(1+\frac{\epsilon L}{2\sigma^2}\right)}}.$$

  By union bounding ($j = 1,\ldots,N$), we get

$$\Lambda^m\left(\cap_{j=1}^N\left\{|\bar{f}(\mathbf{c}_j;\boldsymbol{\omega}_{1:m})| < \frac{\epsilon}{2}\right\}\right) \geq 1 - Ne^{-\frac{m\epsilon^2}{8\sigma^2\left(1+\frac{\epsilon L}{2\sigma^2}\right)}}. \tag{B.25}$$

  - **Condition** $L(\boldsymbol{\omega}_{1:m}) < \frac{\epsilon}{2r}$: Applying Markov's inequality to $L(\boldsymbol{\omega}_{1:m})$ (note that $L(\boldsymbol{\omega}_{1:m})$ is non-negative), for any $t > 0$, we obtain

$$\Lambda^m\left(L(\boldsymbol{\omega}_{1:m}) \geq t\right) \leq \frac{\mathbb{E}_{\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_m}[L(\boldsymbol{\omega}_{1:m})]}{t} \leq \frac{D_{\mathbf{p},\mathbf{q},\mathcal{S}} + E_{\mathbf{p},\mathbf{q}}}{t},$$

  by invoking (B.19) and (B.22). Choosing $t = \frac{\epsilon}{2r}$, we have

$$\Lambda^m\left(L(\boldsymbol{\omega}_{1:m}) < \frac{\epsilon}{2r}\right) \geq 1 - \frac{2r}{\epsilon}(D_{\mathbf{p},\mathbf{q},\mathcal{S}} + E_{\mathbf{p},\mathbf{q}}). \tag{B.26}$$

- **Final bound for any** $r > 0$: By (B.25) and (B.26), and substituting the explicit form of N in (B.17), we get

$$\Lambda^m\left(\sup_{\mathbf{z}\in\mathcal{S}_\Delta}|\bar{f}(\mathbf{z};\boldsymbol{\omega}_{1:m})| < \epsilon\right) \geq \Lambda^m\left(\left\{L(\boldsymbol{\omega}_{1:m}) < \frac{\epsilon}{2r}\right\}\bigcap\cap_{j=1}^N\left\{|\bar{f}(\mathbf{c}_j;\boldsymbol{\omega}_{1:m})| < \frac{\epsilon}{2}\right\}\right)$$

$$\geq 1 - \left(\frac{4|\mathcal{S}|}{r} + 1\right)^d e^{-\frac{m\epsilon^2}{8\sigma^2\left(1+\frac{\epsilon L}{2\sigma^2}\right)}} - \frac{2r}{\epsilon}(D_{\mathbf{p},\mathbf{q},\mathcal{S}} + E_{\mathbf{p},\mathbf{q}}) \overset{(\dagger)}{\geq} 1 - c_* - \kappa_1 r^{-d} - \kappa_2 r, \tag{B.27}$$

  where we invoked the

$$\left(\frac{4|\mathcal{S}|}{r} + 1\right)^d = \left[2\left(\frac{\frac{4|\mathcal{S}|}{r}}{2} + \frac{1}{2}\right)\right]^d = 2^d\left(\frac{\frac{4|\mathcal{S}|}{r}}{2} + \frac{1}{2}\right)^d \overset{(\dagger)}{\leq} 2^d\frac{1}{2}\left[\left(\frac{4|\mathcal{S}|}{r}\right)^d + 1^d\right] = 2^{d-1}\left[\left(\frac{4|\mathcal{S}|}{r}\right)^d + 1\right]$$

  Jensen's inequality in ($\dagger$), $c_* := 2^{d-1}e^{-\frac{m\epsilon^2}{8\sigma^2\left(1+\frac{\epsilon L}{2\sigma^2}\right)}}$, $\kappa_1 := 4^d|\mathcal{S}|^d c_*$ and $\kappa_2 := \frac{2}{\epsilon}(D_{\mathbf{p},\mathbf{q},\mathcal{S}} + E_{\mathbf{p},\mathbf{q}})$.

- **Matching the two terms to choose** r: Maximizing w.r.t. r in (B.27)

$$f(r) = \kappa_1 r^{-d} + \kappa_2 r \Rightarrow f'(r) = \kappa_1(-d)r^{-d-1} + \kappa_2 = 0 \Rightarrow \frac{d\kappa_1}{\kappa_2} = r^{d+1}$$

we note that $r = \left(\frac{d\kappa_1}{\kappa_2}\right)^{\frac{1}{d+1}}$ maximizes it. Using this in (B.27), we have

$$\Lambda^m\left(\sup_{\mathbf{z}\in\mathcal{S}_\Delta}|\bar{f}(\mathbf{z};\boldsymbol{\omega}_{1:m})| \geq \epsilon\right) \leq c_* + \kappa_1\left(\frac{d\kappa_1}{\kappa_2}\right)^{-\frac{d}{d+1}} + \kappa_2\left(\frac{d\kappa_1}{\kappa_2}\right)^{\frac{1}{d+1}} = c_* + F_d\kappa_1^{\frac{1}{d+1}}\kappa_2^{\frac{d}{d+1}}$$

$$= 2^{d-1}e^{-\frac{m\epsilon^2}{8\sigma^2\left(1+\frac{\epsilon L}{2\sigma^2}\right)}} + F_d\left[2^{3d-1}|\mathcal{S}|^d e^{-\frac{m\epsilon^2}{8\sigma^2\left(1+\frac{\epsilon L}{2\sigma^2}\right)}}\right]^{\frac{1}{d+1}}\left[\frac{2}{\epsilon}(D_{\mathbf{p},\mathbf{q},\mathcal{S}} + E_{\mathbf{p},\mathbf{q}})\right]^{\frac{d}{d+1}}$$

$$= 2^{d-1}e^{-\frac{m\epsilon^2}{8\sigma^2\left(1+\frac{\epsilon L}{2\sigma^2}\right)}} + F_d 2^{\frac{4d-1}{d+1}}\left[\frac{|\mathcal{S}|(D_{\mathbf{p},\mathbf{q},\mathcal{S}} + E_{\mathbf{p},\mathbf{q}})}{\epsilon}\right]^{\frac{d}{d+1}}e^{-\frac{m\epsilon^2}{8(d+1)\sigma^2\left(1+\frac{\epsilon L}{2\sigma^2}\right)}},$$

where $F_d := d^{-\frac{d}{d+1}} + d^{\frac{1}{d+1}}$.

B.4 Proof of bounded supp(Λ) $\Rightarrow$ (7)

We prove that the boundedness of supp(Λ) implies that of f [see (B.28)], specifically (7).

Proof: Indeed, let

$$f(\mathbf{z};\boldsymbol{\omega}) = \partial^{\mathbf{p},\mathbf{q}}k(\mathbf{z}) - \boldsymbol{\omega}^{\mathbf{p}}(-\boldsymbol{\omega})^{\mathbf{q}}h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T\mathbf{z}\right) = \left[\int_{\mathbb{R}^d}\boldsymbol{\omega}^{\mathbf{p}}(-\boldsymbol{\omega})^{\mathbf{q}}h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T\mathbf{z}\right)\mathrm{d}\Lambda(\boldsymbol{\omega})\right] - \boldsymbol{\omega}^{\mathbf{p}}(-\boldsymbol{\omega})^{\mathbf{q}}h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T\mathbf{z}\right). \quad \text{(B.28)}$$

Applying the triangle inequality and $|h_a(v)| \le 1$ $(\forall a, \forall v)$ we have

$$|f(\mathbf{z};\boldsymbol{\omega})| \le \left|\int_{\mathbb{R}^d}\mathbf{s}^{\mathbf{p}}(-\mathbf{s})^{\mathbf{q}}h_{|\mathbf{p}+\mathbf{q}|}\left(\mathbf{s}^T\mathbf{z}\right)\mathrm{d}\Lambda(\mathbf{s})\right| + \left|\boldsymbol{\omega}^{\mathbf{p}}(-\boldsymbol{\omega})^{\mathbf{q}}h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T\mathbf{z}\right)\right| \le \int_{\mathbb{R}^d}\left|\mathbf{s}^{\mathbf{p}}(-\mathbf{s})^{\mathbf{q}}h_{|\mathbf{p}+\mathbf{q}|}\left(\mathbf{s}^T\mathbf{z}\right)\right|\mathrm{d}\Lambda(\mathbf{s}) + \left|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}\right|$$

$$\le \int_{\mathbb{R}^d}\left|\mathbf{s}^{\mathbf{p}+\mathbf{q}}\right|\mathrm{d}\Lambda(\mathbf{s}) + \left|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}\right| = \int_{\mathrm{supp}(\Lambda)}\left|\mathbf{s}^{\mathbf{p}+\mathbf{q}}\right|\mathrm{d}\Lambda(\mathbf{s}) + \left|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}\right| \le 2\sup_{\mathbf{s}\in\mathrm{supp}(\Lambda)}\left|\mathbf{s}^{\mathbf{p}+\mathbf{q}}\right|.$$

$K := \sup_{\mathbf{s}\in\mathrm{supp}(\Lambda)}\left|\mathbf{s}^{\mathbf{p}+\mathbf{q}}\right|$ is finite since supp(Λ) is bounded, thus $|f(\mathbf{z};\boldsymbol{\omega})|$ is bounded.

C Supplementary results

In this section, we present some technical results that are used in the proofs.

Lemma C.1 (McDiarmid Inequality [7]). *Let $(X_i)_{i=1}^m$ be $\mathcal{X}$-valued independent random variables. Suppose $f : \mathcal{X}^m \to \mathbb{R}$ satisfies the bounded difference property,*

$$\sup_{u_1,\ldots,u_m,u_r'\in\mathcal{X}}|f(u_1,\ldots,u_m) - f(u_1,\ldots,u_{r-1},u_r',u_{r+1},\ldots,u_m)| \le c_r \;\; (\forall r = 1,\ldots,m). \quad \text{(C.1)}$$

Then for any $\epsilon > 0$,

$$\mathbb{P}\left(f(X_1,\ldots,X_m) - \mathbb{E}\left[f(X_1,\ldots,X_m)\right] \ge \epsilon\right) \le e^{-\frac{2\epsilon^2}{\sum_{r=1}^m c_r^2}}.$$

Note: specifically, if $c = c_r$ $(\forall r)$ then applying a $\tau = \frac{2\epsilon^2}{\sum_{r=1}^m c_r^2} = \frac{2\epsilon^2}{mc^2} \Leftrightarrow \epsilon = c\sqrt{\frac{\tau m}{2}}$ reparameterization one gets $\mathbb{P}\left(f(X_1,\ldots,X_m) < \mathbb{E}\left[f(X_1,\ldots,X_m)\right] + c\sqrt{\frac{\tau m}{2}}\right) \ge 1 - e^{-\tau}$.

Lemma C.2. *For $a > 1$, $\int_0^1 \sqrt{\log\frac{a}{\epsilon}}\,\mathrm{d}\epsilon \le \sqrt{\log a} + \frac{1}{2\sqrt{\log a}}$.*

Proof. By change of variables, we have $\int_0^1 \sqrt{\log\frac{a}{\epsilon}}\,\mathrm{d}\epsilon = a\int_{\log a}^\infty \sqrt{t}e^{-t}\,\mathrm{d}t$. Applying partial integration, we have

$$\int_{\log a}^\infty \sqrt{t}e^{-t}\,\mathrm{d}t = [\sqrt{t}e^{-t}]_\infty^{\log a} + \int_{\log a}^\infty \frac{1}{2\sqrt{t}}e^{-t}\,\mathrm{d}t \le \frac{\sqrt{\log a}}{a} + \frac{1}{2\sqrt{\log a}}\int_{\log a}^\infty e^{-t}\,\mathrm{d}t = \frac{\sqrt{\log a}}{a} + \frac{1}{2a\sqrt{\log a}},$$

thereby yields the result. $\square$

Lemma C.3 (Bernstein inequality [12]). *Let $\xi \in \mathbb{R}$ be a random variable, $\mathbb{E}_{\xi\sim\mathbb{P}}[\xi] = 0$, and assume that $\exists L > 0, S > 0$ satisfying*

$$\sum_{j=1}^m \mathbb{E}_{\xi_j\sim\mathbb{P}}\left[|\xi_j|^M\right] \le \frac{M!S^2L^{M-2}}{2} \quad (\forall M \ge 2),$$

where $(\xi_j)_{j=1}^m \overset{i.i.d.}{\sim} \mathbb{P}$. Then for any $0 < m \in \mathbb{N}$, $\eta > 0$,

$$\mathbb{P}^m\left(\left|\sum_{j=1}^m \xi_j\right| \ge \eta S\right) \le e^{-\frac{1}{2}\frac{\eta^2}{1+\frac{\eta L}{S}}}.$$

Lemma C.4 (L^r norm as countable supremum). *Assume that $1 < \tilde{r} < \infty$. If $(X,\mathcal{A},\mu)$, $\mu(X) < \infty$, $\frac{1}{r} + \frac{1}{\tilde{r}} = 1$, then $\left[L^{\tilde{r}}(X,\mathcal{A},\mu)\right]^* = \{F_f : f \in L^r(X,\mathcal{A},\mu)\}$, where $F_f(u) = \int_X uf\mathrm{d}\mu$, and $\|f\|_{L^r} = \|F_f\|\ (= \sup_{\|g\|_{L^{\tilde{r}}}=1}|F_f(g)|)$; see [8, Theorem 4.1]. Specifically, if $X = \mathcal{S} \subseteq \mathbb{R}^d$ compact and it is endowed with the Borel σ-algebra, then by the separability of $\mathcal{S}$, $L^{\tilde{r}}(\mathcal{S})$ is also separable [4, Prop. 3.4.5] since the Borel σ-algebra is countably generated [1, page 17 (vol. 2)], thus there exists a countable $\mathcal{G} \subseteq L^{\tilde{r}}(\mathcal{S})$, [3, Lemma 6.7] such that $\|g\|_{L^{\tilde{r}}(\mathcal{S})} = 1$ $(\forall g \in \mathcal{G})$ and $\|F_f\| = \sup_{g\in\mathcal{G}}|F_f(g)|$.*

Note: the σ-algebra of Lebesgue *measurable sets is typically* not *countably generated [1, page 106 (vol. I)].*

References

[1] V. I. Bogachev. *Measure Theory*. Springer, 2007.

[2] O. Bousquet. New approaches to statistical learning theory. *Annals of the Institute of Statistical Mathematics*, 55:371–389, 2003.

[3] N. L. Carothers. *A Short Course on Banach Space Theory*. Cambridge University Press, 2004.

[4] D. L. Cohn. *Measure Theory: Second Edition*. Birkhäuser Basel, 2013.

[5] M. Ledoux and M. Talagrand. *Probability in Banach Spaces – Isoperimetry and Processes*. Springer-Verlag, 2002.

[6] J. Lindenstrauss and L. Tzafriri. *Classical Banach Spaces II – Function Spaces*. Springer-Verlag, 1979.

[7] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[8] E. M. Stein and R. Shakarchi. *Functional Analysis: Introduction to Further Topics in Analysis*. Princeton University Press, 2011.

[9] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

[10] S. A. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2009.

[11] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, 1996.

[12] V. Yurinsky. *Sums and Gaussian Vectors*. Springer, 1995.