Orlicz Random Fourier Features

Linda Chamakh[†‡] LINDA.CHAMAKH@POLYTECHNIQUE.EDU
Emmanuel Gobet[†] EMMANUEL.GOBET@POLYTECHNIQUE.EDU
Zoltán Szabó[†] ZOLTAN.SZABO@POLYTECHNIQUE.EDU
[†] *CMAP, CNRS, École Polytechnique, Institut Polytechnique de Paris*
91128 Palaiseau, France
[‡] *Global Markets Quantitative Research – BNP Paribas*

Abstract

Kernel techniques are among the most widely-applied and influential tools in machine learning with applications at virtually all areas of the field. To combine this expressive power with computational efficiency numerous randomized schemes have been proposed in the literature, among which probably random Fourier features (RFF) are the simplest and most popular. While RFFs were originally designed for the approximation of kernel values, recently they have been adapted to *kernel derivatives*, and hence to the solution of large-scale tasks involving function derivatives. Unfortunately, the understanding of the RFF scheme for the approximation of higher-order kernel derivatives is quite limited due to the challenging polynomial growing nature of the underlying function class in the empirical process. To tackle this difficulty, we establish a finite-sample deviation bound for a general class of polynomial-growth functions under α-exponential Orlicz condition on the distribution of the sample. Instantiating this result for RFFs, our finite-sample uniform guarantee implies a.s. convergence with tight rate for arbitrary kernel with α-exponential Orlicz spectrum and any order of derivative.

Keywords: random Fourier features, kernel derivative, polynomial-growth functions, α-exponential Orlicz norm, unbounded empirical processes

1. Introduction

Kernel machines (Taylor and Cristianini, 2004; Steinwart and Christmann, 2008; Paulsen and Raghupathi, 2016) form one of the most fundamental tools in machine learning and statistics with a wide range of successful applications. The impressive modelling power and flexibility of kernel techniques in capturing complex nonlinear relations originates from the richness of the underlying $\mathcal{H}_k$ function class called reproducing kernel Hilbert space (Aronszajn, 1950, RKHS) associated to a $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel. Kernels extend the classical notion of inner product on $\mathcal{X} = \mathbb{R}^d$ by assuming the existence of a $\phi : \mathcal{X} \to \mathcal{H}$ feature map to a Hilbert space $\mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ for all $x, x' \in \mathcal{X}$. This simple equality (also called the kernel trick) forms the basis of kernel techniques and enables one to compute inner products implicitly without direct access to the feature of the points.

In applications one is often given $\{\mathbf{x}_n\}_{n=1}^N$ samples and is facing with an optimization problem expressed in terms of function values and derivatives[1]

$$\min_{f\in\mathcal{H}_k} l\left(\{\partial^{\mathbf{p}} f(\mathbf{x}_n)\}_{\substack{n\in[N]\\ \mathbf{p}\in D_n}}, \|f\|_{\mathcal{H}_k}^2\right), \tag{1}$$

where $[N]=\{1,\ldots,N\}$, $\partial^{\mathbf{p}} f(\mathbf{x}_n) := \frac{\partial^{p_1+\cdots+p_d} f(\mathbf{x}_n)}{\partial_{x_1}^{p_1}\cdots\partial_{x_d}^{p_d}}$, $D_n \subset \mathbb{N}^d$, $\mathbb{N}:=\{0,1,\ldots\}$ and the RKHS $\mathcal{H}_k$ is characterized by $f(\mathbf{x}) = \langle f, k(\cdot,\mathbf{x})\rangle_{\mathcal{H}_k}$ ($\forall \mathbf{x}\in\mathcal{X}, \forall f\in\mathcal{H}_k$) and $k(\cdot,\mathbf{x})\in\mathcal{H}_k$ ($\forall\mathbf{x}\in\mathcal{X}$).[2] The first property of RKHSs is called the reproducing property, the second one describes basic elements of $\mathcal{H}_k$; combining the two properties makes the canonical feature map and feature space explicit: $k(\mathbf{x},\mathbf{x}') = \langle\phi(\mathbf{x}),\phi(\mathbf{x}')\rangle_{\mathcal{H}_k}$ where $\phi(\mathbf{x}) = k(\cdot,\mathbf{x})\in\mathcal{H}_k$.

For example by taking the quadratic loss, Tikhonov regularization, only function values ($D_n=\{\mathbf{0}\}, \forall n\in[N]$) and $\lambda>0$, (1) reduces to kernel ridge regression

$$\min_{f\in\mathcal{H}_k}\frac{1}{N}\sum_{n\in[N]}[f(\mathbf{x}_n)-y_n]^2 + \lambda\|f\|_{\mathcal{H}_k}^2.$$

Alternatively, one can get back Hermite learning with gradient data (Zhou, 2008; Shi et al., 2010) by additionally including first-order derivatives

$$\min_{f\in\mathcal{H}_k}\frac{1}{N}\sum_{n\in[N]}\left([f(\mathbf{x}_n)-y_n]^2 + \left\|f'(\mathbf{x}_n)-\mathbf{y}_n'\right\|_2^2\right) + \lambda\|f\|_{\mathcal{H}_k}^2, \quad \lambda>0$$

where $f'(\mathbf{x}) = [\partial^{\mathbf{e}_1} f(\mathbf{x});\ldots;\partial^{\mathbf{e}_d} f(\mathbf{x})]\in\mathbb{R}^d$ is the derivative of f, $\mathbf{e}_j\in\mathbb{R}^d$ is the j^{th} canonical basis vector, $\|\cdot\|_2$ is the Euclidean norm and $D_n = \left\{\mathbf{0}, \{\mathbf{e}_j\}_{j=1}^d\right\}$ ($n\in[N]$). Further examples with function derivatives are semi-supervised learning with gradient information (Zhou, 2008), nonlinear variable selection (Rosasco et al., 2010, 2013), learning of piecewise-smooth functions (Lauer et al., 2012), multi-task gradient learning (Ying et al., 2012), structure optimization in parameter-varying ARX (autoregressive with exogenous input) processes (Duijkers et al., 2014), or density estimation with infinite-dimensional exponential families (Sriperumbudur et al., 2017).

An appealing property of RKHSs is that their geometry makes the optimization problem (1) defined over function spaces computationally tractable. Indeed, assuming that l is increasing in its last argument, the $\partial^{\mathbf{p}} f(\mathbf{x}) = \left\langle f, \partial^{\mathbf{p},\mathbf{0}} k(\cdot,\mathbf{x})\right\rangle_{\mathcal{H}_k}$ derivative-reproducing property of kernels and the representer theorem (Zhou, 2008) guarantee that the solution of (1) has a finite-dimensional parameterization $f(\cdot) = \sum_{n\in[N]}\sum_{\mathbf{p}\in D_n} a_{n,\mathbf{p}}\partial^{\mathbf{p},\mathbf{0}} k(\cdot,\mathbf{x}_n)$ ($a_{n,\mathbf{p}}\in\mathbb{R}$) and it is sufficient to solve

$$\min_{\mathbf{a}} l\left(\left\{\sum_{m\in[N]}\sum_{\mathbf{q}\in D_m} a_{m,\mathbf{q}}\partial^{\mathbf{p},\mathbf{q}} k(\mathbf{x}_n,\mathbf{x}_m)\right\}_{\substack{n\in[N]\\ \mathbf{p}\in D_n}}, \sum_{\substack{n,m\in[N]\\ \mathbf{p}\in D_n,\mathbf{q}\in D_m}} a_{n,\mathbf{p}} a_{m,\mathbf{q}}\partial^{\mathbf{p},\mathbf{q}} k(\mathbf{x}_n,\mathbf{x}_m)\right) \tag{2}$$

[1] To have derivatives, in the sequel we assume that $\mathcal{X}=\mathbb{R}^d$.

[2] We use the $k(\cdot,x)$ shorthand to denote the function $y\in\mathcal{X}\mapsto k(y,x)\in\mathbb{R}$ while keeping $x\in\mathcal{X}$ fixed.

determined by the $\partial^{\mathbf{p},\mathbf{q}}k(\mathbf{x},\mathbf{y}) := \frac{\partial^{\sum_{i=1}^{d}(p_i+q_i)}k(\mathbf{x},\mathbf{y})}{\partial_{x_1}^{p_1}\cdots\partial_{x_d}^{p_d}\partial_{y_1}^{q_1}\cdots\partial_{y_d}^{q_d}}$ kernel derivatives; $\mathbf{a} = (a_{n,\mathbf{p}})_{n\in[N],\mathbf{p}\in D_n} \in \mathbb{R}^{\sum_{n\in[N]}|D_n|}$ where $|D_n|$ is the cardinality of the set D_n.

Though kernel methods show impressive modelling power at numerous areas, due to the implicit computation of feature similarities, this flexibility comes with a computational price. Several techniques have been developed in the literature to mitigate this computational challenge such as incomplete Cholesky factorization (Bach and Jordan, 2002), subsampling schemes (Williams and Seeger, 2001; Drineas and Mahoney, 2005; Rudi et al., 2017), sketching (Alaoui and Mahoney, 2015; Yang et al., 2017), random Fourier features (Rahimi and Recht, 2007, 2008, RFF), their quasi-Monte Carlo (Yang et al., 2014), memory-efficient (Le et al., 2013; Dai et al., 2014; Zhang et al., 2019), orthogonal (Yu et al., 2016) or structured (Bojarski et al., 2017) variants.

In this paper we study the RFF technique which is probably the conceptually simplest and most influential approach.[3] By the Bochner theorem (Rudin, 1990) a continuous, bounded, shift-invariant kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ can be written as the Fourier transform of a (finite) measure Λ, called the spectral measure

$$k(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d} \cos\left(\boldsymbol{\omega}^\top(\mathbf{x}-\mathbf{y})\right) \mathrm{d}\Lambda(\boldsymbol{\omega}). \tag{3}$$

The RFF method uses this representation of k to provide an explicit low-dimensional feature map approximation for the kernel values and f

$$k(\mathbf{x},\mathbf{x}') \approx \left\langle \lambda(\mathbf{x}), \lambda(\mathbf{x}')\right\rangle_{\mathbb{R}^{2M}}, \qquad \hat{f}_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \lambda(\mathbf{x})\rangle_{\mathbb{R}^{2M}}, \tag{4}$$

where the integral representation (3) with respect to the measure Λ is replaced by an average over random points; hence the *random Fourier feature* naming. As a result, one can estimate $\mathbf{w}$ by leveraging fast linear primal solvers. The idea has been successfully used in various contexts including differential privacy preserving (Chaudhuri et al., 2011), fast function-to-function regression (Oliva et al., 2015), learning message operators in expectation propagation (Jitkrittum et al., 2015), causal discovery (Lopez-Paz et al., 2015; Strobl et al., 2019), independence testing (Zhang et al., 2017), prediction and filtering in dynamical systems (Downey et al., 2017), convolutional neural networks (Cui et al., 2017), bandit optimization (Li et al., 2018), or estimation of Gaussian mixture models (Keriven et al., 2018).

Similarly to (4), one can consider RFF-based approximation of kernel derivatives when solving optimization tasks involving function derivatives [see (1) and (2)]. This is the strategy followed for example by Strathmann et al. (2015) to fit distributions belonging to the infinite-dimensional exponential family, which boils down to an optimization problem with third-order kernel derivatives (Sriperumbudur et al., 2017, Theorem 5).

The focus of this work is to study the approximation quality of the RFF-based kernel-derivative approximation

$$\left\|\widehat{\partial^{\mathbf{p},\mathbf{q}}k} - \partial^{\mathbf{p},\mathbf{q}}k\right\|_S := \sup_{\mathbf{x},\mathbf{y}\in S}\left|\partial^{\mathbf{p},\mathbf{q}}k(\mathbf{x},\mathbf{y}) - \widehat{\partial^{\mathbf{p},\mathbf{q}}k}(\mathbf{x},\mathbf{y})\right|,$$

[3] Rahimi and Recht (2007) won the 10-year test-of-time award at NIPS-2017 as a recognition of the influence of RFFs.

where $S \subset \mathbb{R}^d$ is a compact set. Despite the large number of successful RFF applications, quite little is understood theoretically on its approximation quality. Below we provide a brief summary with particular focus on optimal guarantees and results related to kernel derivatives.

- **Kernel values ($\mathbf{p} = \mathbf{q} = \mathbf{0}$)**: The uniform finite-sample bounds (Rahimi and Recht, 2007; Sutherland and Schneider, 2015) have recently been improved (Sriperumbudur and Szabó, 2015) exponentially in terms of the diameter of the compact set S_M ($|S_M|$) arriving to[4] $\left\|k - \widehat{k}\right\|_{S_M} = \mathcal{O}_{a.s.}\left(\frac{\sqrt{\log|S_M|}\vee\sqrt{\log M}}{\sqrt{M}}\right)$ from $\left\|k - \widehat{k}\right\|_{S_M} = \mathcal{O}_p\left(\frac{|S_M|\sqrt{\log M}}{\sqrt{M}}\right)$, where $\vee$ denotes the maximum. The result shows that the diameter of the set S_M can grow at a $|S_M| = e^{o(M)}$ rate while still getting a consistent estimate; this rate is optimal as shown in the characteristic function literature (Csörgö and Totik, 1983).

- **Kernel ridge regression**: RFFs have been settled in kernel ridge regression by Rudi and Rosasco (2017) via showing that using $M = o(N) = \mathcal{O}\left(\sqrt{N}\log N\right)$ random Fourier features is sufficient to get $\mathcal{O}\left(1/\sqrt{N}\right)$ generalization error. Under additional γ-capacity ($\gamma \in [0, 1]$) and r-range space conditions ($r \geq \frac{1}{2}$), the same authors showed that even faster, minimax optimal $\mathcal{O}\left(N^{-\frac{2r}{2r+\gamma}}\right)$ rates are achievable with $M = o(N) = \mathcal{O}\left(N^{\frac{1+\gamma(2r-1)}{2r+\gamma}}\log N\right)$ RFFs. The result improves the originally proved (Rahimi and Recht, 2008) guarantee holding under the pessimistic $M = \mathcal{O}(N)$ setting. Recently the analysis has been further sharpened (in terms of the number of required RFFs; Li et al. (2019)) by leveraging the notion of effective degrees of freedom.

- **Classification with 0-1 loss:** In the classification setting with the 0-1 loss and RKHSs, Gilbert et al. (2018) proved that $M = o(N) = \tilde{\mathcal{O}}\left(N^{\frac{2}{2+c}}\right)$ optimized RFF features—optimized in the sense of Bach (2017)—are sufficient to achieve a learning rate of $\tilde{\mathcal{O}}\left(N^{-\frac{c}{2+c}}\right)$ provided that the spectrum of the integral operator associated to the kernel decay polynomially at the rate of $\lambda_i = O\left(i^{-c}\right)$ with $c > 1$.[4] The same authors showed that the learning rate can be improved to $\tilde{\mathcal{O}}\left(N^{-1}\right)$ with $M = \tilde{\mathcal{O}}\left(\ln^d(N)\right)$ RFF-s in case of sub-exponential spectrum, where d denotes the dimension of the inputs in the classification.

- **Kernel PCA**: Sriperumbudur and Sterge (2018) have proved that the statistical performance of kernel principal component analysis (KPCA) can be matched by $M = \mathcal{O}(N^{2/3})$ (polynomial decay) or $M = \mathcal{O}(\sqrt{N})$ (exponential decay) RFFs, depending on the eigenvalue decay of the covariance operator associated to the kernel. Ullah et al. (2018) derived a similar bound for a streaming KPCA algorithm under exponential spectrum decay condition.

- **Kernel derivatives**: Supposing that the support of the spectral measure associated to k is either bounded or it satisfies a Bernstein condition

$$\left\|\widehat{\partial^{\mathbf{p},\mathbf{q}}k} - \partial^{\mathbf{p},\mathbf{q}}k\right\|_{S_M} = \mathcal{O}_{a.s.}\left(\frac{\sqrt{\log|S_M|}\vee\sqrt{\log M}}{\sqrt{M}}\right)$$

[4] The classical $\mathcal{O}(\cdot)$ notation up to logarithmic factors is denoted by $\tilde{\mathcal{O}}(\cdot)$; the extension of $\mathcal{O}(\cdot)$ in almost sure and convergence in probability sense are $\mathcal{O}_p(\cdot)$ and $\mathcal{O}_{a.s.}(\cdot)$.

Assumption on the spectral measure	Conditions on $\mathbf{p}, \mathbf{q}$	Convergence rate for $\left\Vert \partial^{\mathbf{p},\mathbf{q}}k - \widehat{\partial^{\mathbf{p},\mathbf{q}}k} \right\Vert_{S_M}$
2nd moment exists	$\mathbf{p} = \mathbf{q} = \mathbf{0}$	$\mathcal{O}_{a.s.}\left(\frac{\sqrt{\log \lvert S_M \rvert} \vee \sqrt{\log M}}{\sqrt{M}}\right)$
Ref: Sriperumbudur and Szabó (2015, Th. 1)		
bounded support	any $\mathbf{p}, \mathbf{q}$	$\mathcal{O}_{a.s.}\left(\frac{\sqrt{\log \lvert S_M \rvert} \vee \sqrt{\log M}}{\sqrt{M}}\right)$
Ref: Sriperumbudur and Szabó (2015, Th. 4)		
Bernstein condition	small $\mathbf{p}, \mathbf{q}$	$\mathcal{O}_{a.s.}\left(\frac{\sqrt{\log \lvert S_M \rvert} \vee \sqrt{\log M}}{\sqrt{M}}\right)$
Ref: Szabó and Sriperumbudur (2019)		
Orlicz condition	any $\mathbf{p}, \mathbf{q}$	$\mathcal{O}_{a.s.}\left(\frac{\sqrt{\log \lvert S_M \rvert} \vee \sqrt{\log M}}{\sqrt{M}}\right)$
Ref: now		

Table 1: Summary of RFF guarantees on kernel values and derivatives. Last line: it includes any measure Λ with a finite α-exponential moment (for some $\alpha, c > 0$, $\mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left(e^{c\Vert\boldsymbol{\omega}\Vert_2^\alpha}\right) < +\infty$), like the Gaussian and the inverse multiquadratic kernel, see Corollary 4. For further examples see Table 2.

rate is achievable as shown by Sriperumbudur and Szabó (2015) and Szabó and Sriperumbudur (2019), respectively. Unfortunately, the bounded support condition excludes classical kernels such as the Gaussian, while the Bernstein conditions only hold for 'small' (at most 2nd order) derivatives in case of the popular Gaussian kernel (Szabó and Sriperumbudur, 2019). These limitations (summarized in Table 1) of the popular random Fourier features technique motivate our work and the study of widely-applied kernels with unbounded spectral support for the RFF approximation of high-order kernel derivatives. A consequence of our new estimates in Theorem 1 is that the *a.s.* rates previously obtained under stringent conditions (on $\mathbf{p}, \mathbf{q}$ or Λ) are now available for any $\mathbf{p}, \mathbf{q}$ and any spectral measure Λ with α-exponential moments (as defined in (5), $\alpha > 0$). Because Bernstein condition implies exponential moments, our result includes the one given by Szabó and Sriperumbudur (2019).

Particularly, assuming additional smoothness on the bounded shift-invariant kernel, its derivative satisfies a representation similar to (3):

$$\partial^{\mathbf{p},\mathbf{q}}k(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d} \underbrace{\left[\prod_{j=1}^{d} \omega_j^{p_j}(-\omega_j)^{q_j}\right] c_{\left(\sum_{i=1}^{d}\lvert p_i + q_i \rvert\right)}\left(\boldsymbol{\omega}^\top(\mathbf{x}-\mathbf{y})\right)}_{=:f_{\mathbf{x}-\mathbf{y}}(\boldsymbol{\omega})} \mathrm{d}\Lambda(\boldsymbol{\omega}),$$

where c_n is the n^{th} derivative of the $\cos(\cdot)$ function. The primary difficulty is to handle the polynomial growing nature of the

$$\mathcal{F} = \{\boldsymbol{\omega} \mapsto f_{\mathbf{x}-\mathbf{y}}(\boldsymbol{\omega}) : \mathbf{x}, \mathbf{y} \in S\}$$

function class which controls the error $\left\|\widehat{\partial^{\mathbf{p},\mathbf{q}}k}-\partial^{\mathbf{p},\mathbf{q}}k\right\|_S$. We tackle this challenge by imposing the finiteness of the α-exponential Orlicz norm of the spectral measure (Λ) associated to the kernel, in other words

$$\exists\alpha>0,\, c>0 \quad \text{such that} \quad \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left(e^{c\|\boldsymbol{\omega}\|_2^\alpha}\right)<+\infty. \tag{5}$$

Kernels with α-exponential Orlicz spectrum include the popular Gaussian or the inverse multiquadric kernel; for further examples see Table 2 and Remark 5(ii). We establish the consistency and prove finite-sample uniform guarantees of the resulting Orlicz RFF scheme for the approximation of kernel derivatives at any order, as it is briefly illustrated in the last line of Table 1.

To allow this level of generality, we prove a new finite-sample deviation bound for the empirical process related to a general class of functions f with polynomial growth of the sample $\mathbf{X}_m$. The distribution of the latter is assumed to have finite α-exponential Orlicz norm and consequently, the random variables $f(\mathbf{X}_m)$ belong to a γ-exponential Orlicz space with index γ smaller than 1. For deriving such deviation bounds, we have been inspired by the work of Adamczak (2008) which elegantly combines the Klein and Rio (2005) inequality for truncated variables, the Hoffman-Jorgensen inequality to deal with sum of residual of truncated variables, and a Talagrand (1989) inequality in γ-exponential Orlicz norms for sum of centered random variables. However, our work significantly differs from that of Adamczak (2008). First, our aims are different: Adamczak (2008) focuses on getting large deviation bounds while we are looking for all-scale deviation bounds, which leads to a different analysis (in the application of Klein-Rio inequalities for instance). Second, we are concerned by getting upper bounds with quite explicit control. In particular, this requires a careful treatment of Orlicz-type estimates since the function $\Psi_\gamma(x)=e^{x^\gamma}-1$ defining the Orlicz space is not convex for $\gamma<1$ (see Figure 1), as opposed to the usual case; see the results in Section 4. We also derive sharp estimates from the Dudley entropy integral bound (Theorem 9), which enables us to get a tight dependency w.r.t. the diameter of the parameter space. Furthermore, we clarify the use of the Talagrand inequality (Theorem 7); in Adamczak (2008, Theorem 5) it is seemingly invoked for supremum over functions while it is related to sum over centered random variables. With this novel finite-sample deviation bound, the analysis of Orlicz RFFs readily follows, using optimized inequalities.

The paper is structured as follows. Our problem is formulated in Section 2. The main result on the approximation quality of kernel derivatives with random Fourier features is presented in Section 3. Properties of the Orlicz norm are summarized in Section 4. Proofs are provided in Section 5.

2. Problem Formulation

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

Notations: Let the set of natural, real and complex numbers, positive integers, positive reals, non-negative reals and non-positive integers be denoted by $\mathbb{N}=\{0,1,\ldots\}$, $\mathbb{R}$, $\mathbb{C}$, $\mathbb{Z}^+=\{1,2,\ldots\}$, $\mathbb{R}^+=(0,\infty)$, $\mathbb{R}^{\geq 0}=[0,\infty)$ and $\mathbb{Z}^{\leq 0}=\{0,-1,-2,\ldots\}$, respectively. For the maximum of $x,y\in\mathbb{R}$ we use the $x\vee y=\max(x,y)$ shorthand; similarly $x\wedge y=\min(x,y)$. The difference of set A and B is written as $A\backslash B=\{a\in A: a\notin B\}$. The positive value of $x\in\mathbb{R}$ is denoted by $(x)_+=x\vee 0$. The factorial of $n\in\mathbb{N}$ is denoted by $n!$. The Gamma function is $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} \mathrm{d}x$ for $t \in \mathbb{C} \backslash \mathbb{Z}^{\leq 0}$. For $\gamma \in (0,1]$ and $x \in \mathbb{R}^{\geq 0}$, let $I_\gamma(x) = \int_0^x e^{-t^\gamma} \mathrm{d}t$ be the incomplete Gamma function and $\beta_\gamma := \Gamma\left(1 + \frac{1}{\gamma}\right)^{-\gamma}$. The modified Bessel function of the first and second kind are defined as $J_a(z) = \sum_{n \in \mathbb{N}} \frac{1}{n!\Gamma(n+a+1)} \left(\frac{z}{2}\right)^{2n+a}$ and $K_a(z) = \frac{\pi}{2} \frac{J_{-a}(z) - J_a(z)}{\sin(a\pi)}$ for $z \in \mathbb{R}$ and non-integer a; when a is an integer the limit is taken. The notation log stands for the natural logarithm. The polylogarithm function is $\mathrm{Li}_s(z) = \sum_{n \in \mathbb{Z}^+} \frac{z^n}{n^s}$ where $s \in \mathbb{R}$, $z \in \mathbb{R}$ and $|z| < 1$. For $x \in \mathbb{R}$ the hyperbolic sine, cosine and secant function is $\sinh(x) = \frac{e^x - e^{-x}}{2}$, $\cosh(x) = \frac{e^x + e^{-x}}{2}$ and $\mathrm{sech}(x) = \frac{1}{\cosh(x)}$, respectively. The (imaginary) error function is $\mathrm{erfi}(z) = \sum_{n \in \mathbb{N}} \frac{2}{\sqrt{\pi}} \frac{z^{2n+1}}{n!(2n+1)}$ where $z \in \mathbb{R}$. For $n \in \mathbb{N}$ let $a^{(n)}$ be the rising factorial of a defined as $a^{(n)} = \frac{\Gamma(a+n)}{\Gamma(a)}$ where $a \in \mathbb{C} \backslash \mathbb{Z}^{\leq 0}$ and $a + n \in \mathbb{C} \backslash \mathbb{Z}^{\leq 0}$. The ordinary hyperbolic function is ${}_2F_1(a, b; c; z) = \sum_{n \in \mathbb{N}} \frac{a^{(n)} b^{(n)} z^n}{c^{(n)} n!}$ where $a \in \mathbb{C}$, $b \in \mathbb{C}$, $c \in \mathbb{C} \backslash \mathbb{Z}^{\leq 0}$, $z \in \mathbb{C}$ and $|z| < 1$; for $|z| \geq 1$ its analytical continuation is taken. The Kummer's confluent hypergeometric function is ${}_1F_1(a; b; z) = \sum_{n \in \mathbb{N}} \frac{a^{(n)}}{b^{(n)}} \frac{z^n}{n!}$ with $a \in \mathbb{R}^+$, $b \in \mathbb{R}^+$, $z \in \mathbb{R}$. The Fox-Wright generalized hyperbolic function is ${}_1\Psi_1\left((a, A); (b, B); z\right) = \sum_{n \in \mathbb{N}} \frac{\Gamma(a + An)}{\Gamma(b + Bn)} \frac{z^n}{n!}$ where $a \in \mathbb{R}^+$, $b \in \mathbb{R}^+$, $z \in \mathbb{R}$, $A \in \mathbb{R}^+$, $B \in \mathbb{R}^+$ and $1 + B > A$; ${}_1\Psi_1\left((a, 1); (b, 1); z\right) = \frac{\Gamma(a)}{\Gamma(b)} {}_1F_1(a; b; z)$. For an A set $\mathbb{1}_A$ is the indicator function of A: $\mathbb{1}_A(x) = 1$ if $x \in A$, $\mathbb{1}_A(x) = 0$ otherwise. Let $aS + b = \{as + b : s \in S\}$ where $S \subset \mathbb{R}$ and $a, b \in \mathbb{R}$. For an $N \in \mathbb{Z}^+$, $[N] = \{1, \ldots, N\}$. Given an $\mathbf{x} \in \mathbb{R}^d$ vector, its transpose is $\mathbf{x}^\top$; $\|\mathbf{x}\|_p = \left(\sum_{i \in [d]} |x_i|^p\right)^{\frac{1}{p}}$ ($p \in [1, \infty)$) and $\|\mathbf{x}\|_\infty = \max_{i \in [d]} |x_i|$ denotes its p-norm and maximum norm. Let the n^{th} derivative of the $\cos(\cdot)$ function ($n \in \mathbb{N}$) be $c_n = \cos^{(n)}$. For multi-indices $\mathbf{p}, \mathbf{q} \in \mathbb{N}^d$ and $\boldsymbol{\omega} \in \mathbb{R}^d$ let $|\mathbf{p}| = \sum_{j=1}^d p_j$, $\boldsymbol{\omega}^{\mathbf{p}} = \prod_{j=1}^d \omega_j^{p_j}$, $\partial^{\mathbf{p}} f(\mathbf{x}) = \frac{\partial^{|\mathbf{p}|} f(\mathbf{x})}{\partial_{x_1}^{p_1} \cdots \partial_{x_d}^{p_d}}$, $\partial^{\mathbf{p}, \mathbf{q}} g(\mathbf{x}, \mathbf{y}) = \frac{\partial^{|\mathbf{p}| + |\mathbf{q}|} g(\mathbf{x}, \mathbf{y})}{\partial_{x_1}^{p_1} \cdots \partial_{x_d}^{p_d} \partial_{y_1}^{q_1} \cdots \partial_{y_d}^{q_d}}$. The diameter of a compact set $T \subset \mathbb{R}^d$ is denoted by $|T| = \sup_{\mathbf{x}, \mathbf{y} \in T} \|\mathbf{x} - \mathbf{y}\|_2 < \infty$. Let $S \subset \mathbb{R}^d$ be a Borel set. $S_\Delta = S - S = \{a - b : a \in S, b \in S\}$. The set of Borel probability measures on S is written as $\mathcal{M}_1^+(S)$. Let the Banach space of real-valued, r-power μ-integrable functions on S ($1 \leq r < \infty$) be $L^r(S, \mu)$, with $\|f\|_{L^r(S,\mu)} = \left[\int_S |f(x)|^r \mathrm{d}\mu(x)\right]^{\frac{1}{r}}$. We use the shorthand $\mu f = \int_S f(x) \mathrm{d}\mu(x)$ where $\mu \in \mathcal{M}_1^+(S)$ and $f \in L^1(S, \mu)$. The product measure of $\mu_1, \ldots, \mu_M \in \mathcal{M}_1^+(S)$ is $\otimes_{m=1}^M \mu_m$; specifically when all the components coincide we use the shorthand $\mu^M = \otimes_{m \in [M]} \mu$. The empirical measure is $\mathbb{P}_M = \frac{1}{M} \sum_{m=1}^M \delta_{X_m}$ with δ_X being the Dirac measure concentrated on X and $X_1, \ldots, X_M \sim \otimes_{m=1}^M \mu_m$. Let $(r_n)_{n \in \mathbb{N}}$ be a positive sequence. The boundedness of $\frac{X_n}{r_n}$ almost surely is denoted by $X_n = \mathcal{O}_{a.s.}(r_n)$. The expectation is $\mathbb{E}$. Let $n \in \mathbb{R}^+$. We say that an $f : \mathbb{R}^d \to \mathbb{R}$ function is of polynomial growth of order n (shortly $f \in \mathcal{F}_{\mathcal{P}(n)}$) if $\sup_{\mathbf{x} \in \mathbb{R}^d} \frac{|f(\mathbf{x})|}{1 + \|\mathbf{x}\|_2^n} < \infty$; $\mathcal{F}_{\mathcal{P}} = \cup_{n \in \mathbb{R}^+} \mathcal{F}_{\mathcal{P}(n)}$. Let us assume that $\Psi : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ is a continuous, strictly increasing mapping, $\Psi(0) = 0$ and $\lim_{x \to \infty} \Psi(x) = \infty$. The set of $\mathbb{R}^d$-valued random variables having finite Ψ-Orlicz norm is defined as $L_\Psi = \left\{\mathbf{X} : \|\mathbf{X}\|_\Psi := \inf\left\{c > 0 : \mathbb{E}\Psi\left(\frac{\|\mathbf{X}\|_2}{c}\right) \leq 1\right\} < +\infty\right\}$. Throughout the paper we will be particularly interested in (see Fig. 1)

$$\Psi_\alpha : x \in \mathbb{R}^{\geq 0} \mapsto e^{x^\alpha} - 1 \in \mathbb{R}^{\geq 0} \quad (\alpha > 0),$$

in other words in random variables having finite α-exponential Orlicz norm. $\mathbf{X} \in L_{\Psi_\alpha}$ is equivalent to the existence of an $s > 0$ constant such that $\mathbb{E}\left[e^{s\|\mathbf{X}\|_2^\alpha}\right] < \infty$. Random variables $\mathbf{X} \in L_{\Psi_2}$ and $\mathbf{X} \in L_{\Psi_1}$ are called sub-Gaussian and sub-exponential, respectively. For $f \in \mathcal{F}_{\mathcal{P}}$ and random variable $\mathbf{X}$ having α-exponential moment ($\mathbf{X} \in L_{\Psi_\alpha}$) $\mathbb{E}f(\mathbf{X}) < \infty$. Normal random variables with mean m and variance σ^2 are denoted by $\mathcal{N}\left(m, \sigma^2\right)$. Let (Z, m) be a semi-metric space and $\epsilon \in \mathbb{R}^+$. $S \subseteq Z$ is said to be an ϵ-net of Z if for any $z \in Z$ there exists $s \in S$ such that $m(s, z) \le \epsilon$. The ϵ-covering number of Z is defined as the size of the smallest ϵ-net, i.e., $N(\epsilon, m, Z) = \inf\left\{\ell \ge 1 : \exists s_1, \ldots, s_l \in Z \text{ such that } Z \subseteq \cup_{j=1}^{\ell} B_m(s_j, \epsilon)\right\}$, where $B_m(s, \epsilon) = \{z \in Z : m(z, s) \le \epsilon\}$ is the closed ball with center $s \in Z$ and radius ϵ.

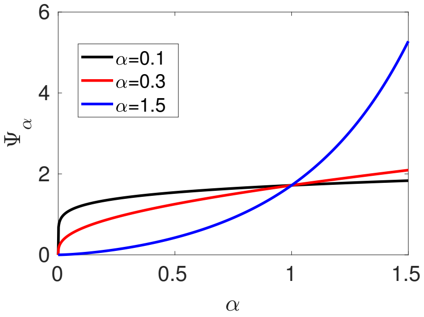

Figure 1: Ψ_α for different α values.

We proceed by formally defining our task. Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a continuous, bounded and shift-invariant kernel. Then, by the Bochner theorem (Rudin, 1990) one can assume w.l.o.g. the existence of a $\Lambda \in \mathcal{M}_1^+\left(\mathbb{R}^d\right)$ spectral measure such that

$$\begin{aligned} k(\mathbf{x}, \mathbf{y}) &= \int_{\mathbb{R}^d} \cos\left(\boldsymbol{\omega}^\top(\mathbf{x} - \mathbf{y})\right) \mathrm{d}\Lambda(\boldsymbol{\omega}) \\ &= \int_{\mathbb{R}^d} \cos\left(\boldsymbol{\omega}^\top \mathbf{x}\right) \cos\left(\boldsymbol{\omega}^\top \mathbf{y}\right) + \sin\left(\boldsymbol{\omega}^\top \mathbf{x}\right) \sin\left(\boldsymbol{\omega}^\top \mathbf{y}\right) \mathrm{d}\Lambda(\boldsymbol{\omega}). \end{aligned}$$

Let $\mathbf{p}, \mathbf{q} \in \mathbb{N}^d$ and assume that $\int_{\mathbb{R}^d} |\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}| \, \mathrm{d}\Lambda(\boldsymbol{\omega}) < \infty$. In this case $\partial^{\mathbf{p},\mathbf{q}} k(\mathbf{x}, \mathbf{y})$ exists, and by the dominated convergence theorem one arrives at

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

The integral can be estimated by Monte-Carlo technique replacing Λ with $\Lambda_M = \frac{1}{M}\sum_{m=1}^M \delta_{\boldsymbol{\omega}_m}$, $(\boldsymbol{\omega}_m)_{m=1}^M \overset{\text{i.i.d.}}{\sim} \Lambda$:

$$\begin{aligned} \widehat{\partial^{\mathbf{p},\mathbf{q}} k}(\mathbf{x}, \mathbf{y}) &= \frac{1}{M} \sum_{m=1}^M \partial^{\mathbf{p}} \cos\left(\boldsymbol{\omega}_m^\top \mathbf{x}\right) \partial^{\mathbf{q}} \cos\left(\boldsymbol{\omega}_m^\top \mathbf{y}\right) + \partial^{\mathbf{p}} \sin\left(\boldsymbol{\omega}_m^\top \mathbf{x}\right) \partial^{\mathbf{q}} \sin\left(\boldsymbol{\omega}_m^\top \mathbf{y}\right) \\ &= \langle \lambda_{\mathbf{p}}(\mathbf{x}), \lambda_{\mathbf{q}}(\mathbf{y}) \rangle_{\mathbb{R}^{2M}}, \end{aligned} \tag{6}$$

where $\lambda_{\mathbf{p}}(\mathbf{x}) = \frac{1}{\sqrt{M}}\left[\left(\partial^{\mathbf{p}} \cos\left(\boldsymbol{\omega}_m^\top \mathbf{x}\right)\right)_{m\in[M]}; \left(\partial^{\mathbf{p}} \sin\left(\boldsymbol{\omega}_m^\top \mathbf{x}\right)\right)_{m\in[M]}\right] \in \mathbb{R}^{2M}$; this is the RFF feature approximation $\lambda_{\mathbf{p}}$ in (4). For $\mathbf{p} = \mathbf{q} = \mathbf{0}$, the construction reduces to the traditional RFF technique (Rahimi and Recht, 2007).

This form implies that our target quantity can be written as

$$\left\|\widehat{\partial^{\mathbf{p},\mathbf{q}}k} - \partial^{\mathbf{p},\mathbf{q}}k\right\|_S = \sup_{\mathbf{z}\in S_\Delta} |(\Lambda_M - \Lambda)(f_{\mathbf{z}})|, \qquad f_{\mathbf{z}}(\boldsymbol{\omega}) = \boldsymbol{\omega}^{\mathbf{p}}(-\boldsymbol{\omega})^{\mathbf{q}} c_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^\top \mathbf{z}\right), \tag{7}$$

$$= \sup_{f\in\mathcal{F}} |(\Lambda_M - \Lambda)(f)|, \qquad \mathcal{F} = \{f_{\mathbf{z}} : \mathbf{z} \in S_\Delta\}, \tag{8}$$

thus the problem boils down to the study of supremum of empirical processes with $\mathcal{F} \subset \mathcal{F}_{\mathcal{P}(n)}$ where $n = \sum_{j\in[d]}(p_j + q_j) + 1$. In the next section we detail our main result about the fluctuation of such processes.

3. Main Result

In this section we present our main result on the supremum of empirical processes of polynomial growth, and specialize it to the approximation quality of RFFs for kernel derivatives. The proofs are given in Section 5.

We investigate the concentration of the $\sup_{f\in\mathcal{F}} |\frac{1}{M}\sum_{m=1}^M f(\mathbf{X}_m)|$ quantity under the following assumptions:

1. **Compact parameterization**: $\mathcal{F} = \{f_{\mathbf{t}} : \mathbf{t} \in T\}$ where $f_{\mathbf{t}} : \mathbb{R}^d \to \mathbb{R}$ is parameterized by a compact set $T \subset \mathbb{R}^{d'}$.
2. **Lipschitz condition**: There exists $n \in \mathbb{R}^+$ and function $L : \mathbb{R}^d \to \mathbb{R}^{\geq 0}$, $L \in \mathcal{F}_{\mathcal{P}(n)}$ such that
   (a) $|f_{\mathbf{t}_0}(\mathbf{x})| \leq L(\mathbf{x})$ for some $\mathbf{t}_0 \in T$,
   (b) $|f_{\mathbf{t}_1}(\mathbf{x}) - f_{\mathbf{t}_2}(\mathbf{x})| \leq L(\mathbf{x})\rho\left(\|\mathbf{t}_1 - \mathbf{t}_2\|_2\right)$ for all $\mathbf{x} \in \mathbb{R}^d, \mathbf{t}_1 \in T, \mathbf{t}_2 \in T$,
   (c) with $\rho : [0, |T|] \to \mathbb{R}^{\geq 0}$ continuous strictly increasing mapping with $\rho(0) = 0$ such that $I_\rho(|T|) := \rho\left(|T|\right)\int_0^1 \sqrt{\log\left(1 + \frac{2|T|}{\rho^{-1}(u\rho(|T|))}\right)}\mathrm{d}u < \infty$.
3. **Independence**, **finite α-exponential Orlicz norm**:
   (a) $(\mathbf{X}_m)_{m\in[M]}$ are independent $\mathbb{R}^d$-valued random variables; shortly, $(\mathbf{X}_m)_{m\in[M]} \sim \otimes_{m\in[M]}\mu_m$ with $\mu_m \in \mathcal{M}_1^+\left(\mathbb{R}^d\right)$.
   (b) $\exists\alpha \in \mathbb{R}^+$ such that $\|\mathbf{X}_m\|_{\Psi_\alpha} < \infty$ for all $m \in [M]$.
4. **Centering**: $\mathbb{E}\left[f(\mathbf{X}_m)\right] = 0$ for all $f \in \mathcal{F}$ and $m \in [M]$.

Under these conditions, our main result is as follows.

Theorem 1 (Concentration of processes with polynomial growth) *Assume that $\mathcal{F}$ and $(\mathbf{X}_m)_{m\in[M]}$ satisfy Assumptions 1-4 and $\gamma := \frac{\alpha}{n} \leq 1$. Let $\mathbb{P} = \otimes_{m\in[M]}\mu_m$, and $\|L\|_{L^2(\mathbf{X}_{1:M})} := \sqrt{\frac{1}{M}\sum_{m\in[M]} L^2(\mathbf{X}_m)}$. Let $\Psi_\gamma^{(l)}$ be the convexification[5] of Ψ_γ, $A_\gamma := \frac{\left(\Psi_\gamma^{(l)}\right)^{-1}(1)}{\Psi_\gamma^{-1}(1)}$, $B_\gamma := \left(\Psi_\gamma^{(l)}\right)^{-1}(1)$, C_γ and C_D be the constants defined in (44) and (45), and $K_\gamma := 2^{\left(\frac{1}{\gamma}-1\right)}\left(C_\gamma\left[16B_\gamma + 2^{\left(\frac{1}{\gamma}-1\right)}\left(1 + A_\gamma\right)\right] + 8A_\gamma\right)$. Then for any $\varepsilon > 0$ satisfying*

$$\varepsilon \geq 6B, \qquad B := 2C_D\sqrt{d'}\frac{\mathbb{E}\left[\|L\|_{L^2(\mathbf{X}_{1:M})}\right]}{\sqrt{M}} I_\rho(|T|), \tag{9}$$

[5] Ψ_γ is not convex for $\gamma < 1$. We convexify Ψ_γ and use the Section 4(v) based integral control property holding for *convex* Ψ-s; for details on $\Psi_\gamma^{(l)}$ see Section 5.3.

we have

$$\mathbb{P}\left(\sup_{\mathbf{t}\in T}\frac{1}{M}\sum_{m\in[M]} f_{\mathbf{t}}(\mathbf{X}_m)\geq\varepsilon\right)\leq 2e^{-\left(\frac{M\varepsilon}{3K_\gamma\left\|\max_{m\in[M]}\sup_{\mathbf{t}\in T}|f_{\mathbf{t}}(\mathbf{X}_m)|\right\|_{\Psi_\gamma}}\right)^{\gamma}}+e^{-\frac{M\,\varepsilon^2}{72\sigma^2+84c\,\varepsilon}}, \quad (10)$$

where

$$\sigma^2 := \sup_{\mathbf{t}\in T}\frac{1}{M}\sum_{m\in[M]}\mathbb{E}\left[f_{\mathbf{t}}^2(\mathbf{X}_m)\right],$$

$$c := \max_{m\in[M]}\sup_{\mathbf{t}\in T}\|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma}\left[\frac{1}{\beta_\gamma}\log\left(\frac{6\Gamma\left(1+\frac{1}{\gamma}\right)\max_{m\in[M]}\sup_{\mathbf{t}\in T}\|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma}}{\gamma\varepsilon}\right)\right]^{\frac{1}{\gamma}}$$
$$\vee\, 8\mathbb{E}\left[\max_{m\in[M]}\sup_{\mathbf{t}\in T}|f_{\mathbf{t}}(\mathbf{X}_m)|\right]\in[0,+\infty).$$

Remark 2

(i) ***Two-sided bound****: For $\mathbb{P}\left(\inf_{\mathbf{t}\in T}\frac{1}{M}\sum_{m\in[M]} f_{\mathbf{t}}(\mathbf{X}_m)\leq-\varepsilon\right)$ the same one-sided deviation bound can be obtained by replacing $f_{\mathbf{t}}$ with $-f_{\mathbf{t}}$. As a result one can estimate $\mathbb{P}\left(\sup_{\mathbf{t}\in T}\left|\frac{1}{M}\sum_{m\in[M]} f_{\mathbf{t}}(\mathbf{X}_m)\right|\geq\varepsilon\right)$ by twice the bound above.*

(ii) ***Assumption*** (3)*: Assumption* (3a) *with Assumption* (4) *is weaker than being i.i.d.: for example $\mathbb{E}\left[f_{\mathbf{t}}(X_m)\right]=0$ holds for $X_m=\mathcal{N}\left(0,\sigma_m^2\right)$ and $f_{\mathbf{t}}(x)=c_{\mathbf{t}}x^3$, but X_m-s can differ in their variance.*

(iii) ***Assumption*** $\alpha/n\leq 1$*: This condition holds without loss of generality. Indeed, in case of $\alpha/n>1$, one can get a modified (α',n') pair satisfying $\alpha'/n'\leq 1$ by either increasing n to the value $n'=\alpha$ using that $\mathcal{F}_{\mathcal{P}(n)}\subset\mathcal{F}_{\mathcal{P}(n')}$, or by decreasing α to the value $\alpha'=n$ using that $\|\mathbf{X}_m\|_{\Psi_\alpha}<\infty$ implies $\|\mathbf{X}_m\|_{\Psi_{\alpha'}}<\infty$ for any $\alpha'\in(0,\alpha)$.*

(iv) ***Proof-related remarks:***

1. ***Compactness of*** T*: This compactness with the Lipschitz property enables one to control the covering number of $\mathcal{F}$.*
2. ***Truncated functions****: The Lipschitz property of $\mathcal{F}$ implies that of the truncated functions: for $\forall\mathbf{x}\in\mathbb{R}^d$, $\mathbf{s}$ and $\mathbf{t}\in T$*

$$|\mathcal{T}_c f_{\mathbf{t}}(\mathbf{x})-\mathcal{T}_c f_{\mathbf{s}}(\mathbf{x})|\leq|f_{\mathbf{t}}(\mathbf{x})-f_{\mathbf{s}}(\mathbf{x})|\leq L(\mathbf{x})\rho\left(\|\mathbf{t}-\mathbf{s}\|_2\right), \quad (11)$$

where $\mathcal{T}_c f(\mathbf{x}):=f(\mathbf{x})\mathbb{1}_{|f(\mathbf{x})|\leq c}+c\mathbb{1}_{f(\mathbf{x})>c}-c\mathbb{1}_{f(\mathbf{x})<-c}$ is f soft-thresholded at level c.

3. *$\mathcal{F}\subset\mathcal{F}_{\mathcal{P}(n)}$: This property is inherited (Section 5.5) from $L\in\mathcal{F}_{\mathcal{P}(n)}$ by the Lipschitz conditions* (2a)-(2b)*.*
4. ***Finiteness of the terms in Theorem 1****: $\left\|\max_{m\in[M]}\sup_{\mathbf{t}\in T}|f_{\mathbf{t}}(\mathbf{X}_m)|\right\|_{\Psi_{\frac{\alpha}{n}}}$ and $\mathbb{E}\left[\max_{m\in[M]}\sup_{\mathbf{t}\in T}|f_{\mathbf{t}}(\mathbf{X}_m)|\right]$ are finite (see Section 5.5) in Theorem 1 by the Lipschitz assumption* (2a)-(2b)*, $\|\mathbf{X}_m\|_{\Psi_\alpha}<\infty$ (Assumption* (3b)*) and $L\in\mathcal{F}_{\mathcal{P}(n)}$ (Assumption* (2)*).*

(v) ***RFF specialization**: Assuming that the α-exponential Orlicz condition holds for the Λ spectral measure associated to k ($\exists \alpha \in \mathbb{R}^+$ such that $\|\boldsymbol{\omega}\|_{\Psi_\alpha} < \infty$, $\boldsymbol{\omega} \sim \Lambda$),[6] one can see (Section 5.1) that RFFs are covered by choosing*

$$d' = d, \quad f_{\mathbf{t}}(\mathbf{x}) \leftarrow f_{\mathbf{z}}(\boldsymbol{\omega}) - \Lambda f_{\mathbf{z}}, \quad \mathbf{t} \leftarrow \mathbf{z}, \quad T \leftarrow S_\Delta, \quad \mathbf{X}_m \leftarrow \boldsymbol{\omega}_m,$$
$$\rho(u) = u^\beta, \quad \beta = \frac{1}{1 + (\log|S_\Delta|)_+} \in (0,1], \quad n \leftarrow |\mathbf{p}+\mathbf{q}| + \beta.$$

While any value of $\beta \in (0,1]$ would met the assumptions, allowing β to depend on the diameter of S_Δ enables us to get optimal convergence rates w.r.t. the diameter (see Corollary 4).
The terms driving the guarantee for RFF can be bounded (Section 5.7) as follows: there is a constant $C_{\mathtt{RFF}} \in \mathbb{R}^+$, depending only on Λ, $|\mathbf{p}+\mathbf{q}|$, but not on $|S_\Delta|$ and M, such that

$$\begin{aligned}
B &\le \frac{C_{\mathtt{RFF}}\sqrt{1 + (\log|S_\Delta|)_+}}{\sqrt{M}}, \\
\sigma^2 &\le C_{\mathtt{RFF}}, \\
\max_{m \in [M]} \sup_{\mathbf{z} \in S_\Delta} \|g_{\mathbf{z}}(\boldsymbol{\omega}_m)\|_{\Psi_\gamma} &\le C_{\mathtt{RFF}}, \\
\left\| \max_{m \in [M]} \sup_{\mathbf{z} \in S_\Delta} |g_{\mathbf{z}}(\boldsymbol{\omega}_m)| \right\|_{\Psi_\gamma} &\le C_{\mathtt{RFF}} \left[\log(1+M)\right]^{n/\alpha}, \\
\mathbb{E}\left[\max_{m \in [M]} \sup_{\mathbf{z} \in S_\Delta} |g_{\mathbf{z}}(\boldsymbol{\omega}_m)| \right] &\le C_{\mathtt{RFF}} \left[\log(1+M)\right]^{n/\alpha}.
\end{aligned} \tag{12}$$

Using these bounds, our finite-sample uniform guarantee on Orlicz RFFs is as follows.

Corollary 3 (Orlicz RFFs for kernel derivative approximation) *Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a continuous, bounded, shift-invariant kernel with spectral measure Λ. Suppose that Λ satisfies the α-exponential Orlicz assumption ($\exists \alpha \in \mathbb{R}^+$ such that $\|\boldsymbol{\omega}\|_{\Psi_\alpha} < \infty$, $\boldsymbol{\omega} \sim \Lambda$) and let $S \subset \mathbb{R}^d$ be a compact set. Let $\beta = \frac{1}{1+(\log|S_\Delta|)_+} \in (0,1]$, let $\mathbf{p}, \mathbf{q} \in \mathbb{N}^d$, $n := |\mathbf{p}+\mathbf{q}| + \beta$, and assume that $\gamma := \frac{\alpha}{n} \le 1$. Let $\widehat{\partial^{\mathbf{p},\mathbf{q}} k}$ be the RFF estimate of $\partial^{\mathbf{p},\mathbf{q}} k$ using $(\boldsymbol{\omega}_m)_{m \in [M]} \overset{i.i.d.}{\sim} \Lambda$ samples as given in Eq. (6). Then, there exists a constant $\tilde{C} \in \mathbb{R}^+$ (depending only on Λ, $|\mathbf{p}+\mathbf{q}|$, but not on S and M) such that for any $\epsilon \ge \frac{\tilde{C}\sqrt{1+(\log|S_\Delta|)_+}}{\sqrt{M}}$,*

$$\Lambda^M \left(\left\| \widehat{\partial^{\mathbf{p},\mathbf{q}} k} - \partial^{\mathbf{p},\mathbf{q}} k \right\|_S \ge \epsilon \right) \le 2 e^{-\frac{(M\varepsilon)^\gamma}{\tilde{C}\log(1+M)}} + e^{-\frac{M\,\varepsilon^2}{\tilde{C}\left(1+\varepsilon\left[\log(\tilde{C}/\varepsilon) \vee \log(1+M)\right]^{1/\gamma}\right)}}. \tag{13}$$

Corollary 4 (Almost sure convergence for kernel derivative approximation) *Let $\mathbf{p}, \mathbf{q} \in \mathbb{N}^d$ and $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a continuous, bounded, shift-invariant kernel with spectral measure Λ which satisfies the α-exponential Orlicz assumption for some $\alpha > 0$. Then, for any*

[6] This requirement implies that $\int_{\mathbb{R}^d} |\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}| \, \mathrm{d}\Lambda(\boldsymbol{\omega}) < \infty$ and thus the existence of $\partial^{\mathbf{p},\mathbf{q}} k$ for any $\mathbf{p}, \mathbf{q} \in \mathbb{N}^d$.

sequence of compact sets $(S_M)_{M=2}^{\infty}$ *such that* $(\log|S_M|)_+ = o(M)$*, we have*

$$\left\| \widehat{\partial^{\mathbf{p},\mathbf{q}} k} - \partial^{\mathbf{p},\mathbf{q}} k \right\|_{S_M} = \mathcal{O}_{a.s.}\left(\frac{\sqrt{(\log|S_M|)_+ \vee \log M}}{\sqrt{M}} \right). \tag{14}$$

Remark 5

(i) ***Spectral measure*** **(Λ)** ***examples****: Our result assumes the* α*-exponential Orlicz property of the spectral measure* Λ *associated to* k*. In Table 2 we provide various examples for* Λ *(with the relevant case of unbounded support) satisfying this requirement; their relations is summarized in Fig. 2. While for the RFF approximation it is not necessary, in many of these examples the corresponding kernel value can also be computed, see Table 3.*

(ii) α***-exponential Orlicz assumption for tensor product kernels****: Using the* α*-exponential Orlicz spectral measures of Table 2 on* $\mathbb{R}$*, one can immediately construct Orlicz spectral measures on* $\mathbb{R}^d$*. Indeed, assume that (i)* k *is a product kernel, i.e.* $k(\mathbf{x},\mathbf{y}) = \prod_{i\in[d]} k_i(x_i, y_i)$*,* $\Lambda = \otimes_{i\in[d]}\Lambda_i$*, and (ii)* Λ_i*, the spectral measure associated to* k_i*, satisfies the* α_i*-exponential Orlicz assumption* $(\alpha_i \in \mathbb{R}^+)$*. Then* $\boldsymbol{\omega} \sim \Lambda$ *is* α*-exponential Orlicz with* $\alpha = \min_{i\in[d]} \alpha_i$*; see Section 5.9.*

(iii) α***-exponential Orlicz vs. Bernstein assumption****: Our result complements Szabó and Sriperumbudur (2019)'s work, where the authors showed that for* $d = 1$ *and spectral densities* $f_\lambda(\omega) \propto e^{-\omega^{2\ell}}$ *the Bernstein condition (and hence fast rates) holds for* $|\mathbf{p}+\mathbf{q}| \leq 2\ell = \alpha$*. Indeed, we proved under the more general* α*-exponential Orlicz assumption the same a.s. convergence rates for any arbitrary order (see Corollary 4) kernel derivatives.*

Spectrum	Spectral density: $f_\Lambda(\omega)$	Parameters	α
Gaussian	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\omega^2}{2\sigma^2}}$	$\sigma > 0$	2
Laplace	$\frac{\sigma}{2} e^{-\sigma\lvert\omega\rvert}$	$\sigma > 0$	1
generalized Gaussian	$\frac{\alpha}{2\beta\Gamma\left(\frac{1}{\alpha}\right)} e^{-\frac{\lvert\omega\rvert}{\beta}^{\alpha}}$	$\alpha > 0$, $\beta > 0$	α
variance Gamma	$\frac{\sigma^{2b}\lvert\omega\rvert^{b-\frac{1}{2}} K_{b-\frac{1}{2}}(\sigma\lvert\omega\rvert)}{\sqrt{\pi}\Gamma(b)(2\sigma)^{b-\frac{1}{2}}}$	$\sigma > 0$, $b > \frac{1}{2}$	1
Weibull (S)	$\frac{s}{2\lambda}\left(\frac{\lvert\omega\rvert}{\lambda}\right)^{s-1} e^{-\left(\frac{\lvert\omega\rvert}{\lambda}\right)^s}$	$s > 0$, $\lambda > 0$	s
exponentiated exponential (S)	$\frac{\alpha}{2\lambda}\left(1 - e^{-\frac{\lvert\omega\rvert}{\lambda}}\right)^{\alpha-1} e^{-\frac{\lvert\omega\rvert}{\lambda}}$	$\lambda > 0$, $\alpha > 0$	1
exponentiated Weibull (S)	$\frac{\alpha s}{2\lambda}\left(\frac{\lvert\omega\rvert}{\lambda}\right)^{s-1}\left[1 - e^{-\left(\frac{\lvert\omega\rvert}{\lambda}\right)^s}\right]^{\alpha-1} \times$ $\times e^{-\left(\frac{\lvert\omega\rvert}{\lambda}\right)^s}$	$s > 0$, $\lambda > 0$, $\alpha > 0$	s

Spectrum	Spectral density: $f_\Lambda(\omega)$	Parameters	α
Nakagami (S)	$\frac{m^m}{\Gamma(m)\Omega^m}\lvert\omega\rvert^{2m-1}e^{-\frac{m\omega^2}{\Omega}}$	$m \geq \frac{1}{2}$, $\Omega > 0$	2
chi-squared (S)	$\frac{1}{2^{\frac{s}{2}+1}\Gamma\left(\frac{s}{2}\right)}\lvert\omega\rvert^{\frac{s}{2}-1}e^{-\frac{\lvert\omega\rvert}{2}}$	$s \in \mathbb{Z}^+$	1
Erlang (S)	$\frac{\lambda^s\lvert\omega\rvert^{s-1}e^{-\lambda\lvert\omega\rvert}}{2(s-1)!}$	$s \in \mathbb{Z}^+$, $\lambda > 0$	1
Gamma (S)	$\frac{1}{2\Gamma(s)\theta^s}\lvert\omega\rvert^{s-1}e^{-\frac{\lvert\omega\rvert}{\theta}}$	$s > 0$, $\theta > 0$	1
generalized Gamma (S)	$\frac{p/a^D}{2\Gamma\left(\frac{D}{p}\right)}\lvert\omega\rvert^{D-1}e^{-\left(\frac{\lvert\omega\rvert}{a}\right)^p}$	$a > 0$, $D > 0$, $p > 0$	p
Rayleigh (S)	$\frac{\lvert\omega\rvert}{2\sigma^2}e^{-\frac{\omega^2}{2\sigma^2}}$	$\sigma > 0$	2
Maxwell-Boltzmann (S)	$\frac{1}{\sqrt{2\pi}}\frac{\omega^2 e^{-\frac{\omega^2}{2a^2}}}{a^3}$	$a > 0$	2
chi (S)	$\frac{1}{2^{\frac{s}{2}}\Gamma\left(\frac{s}{2}\right)}\lvert\omega\rvert^{s-1}e^{-\frac{\omega^2}{2}}$	$s > 0$	2
exponential-logarithmic (S)	$-\frac{1}{2\log(p)}\frac{\beta(1-p)e^{-\beta\lvert\omega\rvert}}{1-(1-p)e^{-\beta\lvert\omega\rvert}}$	$p \in (0,1)$, $\beta > 0$	1
Weibull-logarithmic (S)	$-\frac{1}{2\log(p)}\frac{\alpha\beta(1-p)\lvert\omega\rvert^{\alpha-1}e^{-\beta\lvert\omega\rvert^\alpha}}{1-(1-p)e^{-\beta\lvert\omega\rvert^\alpha}}$	$p \in (0,1)$, $\beta > 0$, $\alpha > 0$	α
Gamma/Gompertz (S)	$\frac{bse^{b\lvert\omega\rvert}\beta^s}{2\left(\beta-1+e^{b\lvert\omega\rvert}\right)^{s+1}}$	$b > 0$, $\beta > 0$, $s > 0$	bs
hyperbolic secant	$\frac{1}{2}\mathrm{sech}\left(\frac{\pi}{2}\omega\right)$		1
logistic	$\frac{e^{-\frac{\omega}{s}}}{s\left[1+e^{-\frac{\omega}{s}}\right]^2}$	$s > 0$	1
normal-inverse Gaussian	$\frac{\alpha\delta K_1\left(\alpha\sqrt{\delta^2+\omega^2}\right)}{\pi\sqrt{\delta^2+\omega^2}}e^{\delta\alpha}$	$\alpha > 0$, $\delta \in \mathbb{R}$	1
hyperbolic	$\frac{1}{2\delta K_1(\delta\alpha)}e^{-\alpha\sqrt{\delta^2+\omega^2}}$	$\alpha > 0$, $\delta \in \mathbb{R}$	1
generalized hyperbolic	$\frac{(\alpha/\delta)^\lambda}{\sqrt{2\pi}K_\lambda(\delta\gamma)}\frac{K_{\lambda-\frac{1}{2}}\left(\alpha\sqrt{\delta^2+\omega^2}\right)}{\left(\frac{\sqrt{\delta^2+\omega^2}}{\alpha}\right)^{\frac{1}{2}-\lambda}}$	$\alpha > 0$, $\lambda \in \mathbb{R}$, $\delta \in \mathbb{R}$	1

Table 2: Kernel spectrum examples in one dimension ($d = 1$) obeying the α-exponential Orlicz assumption. '(S)' stands for symmetrized. The symmetrization guarantees that the kernel associated to Λ is real-valued. Last column: Orlicz exponent. For the variance Gamma distribution the Orlicz exponent follows from the known $K_u(z) \sim \sqrt{\pi/(2z)}e^{-z}$ asymptotics (Barndorff-Nielsen et al., 2001, page 297). Notice that the 'normal-inverse Gaussian $\xrightarrow{\delta=\sigma^2\alpha,\ \alpha\to\infty}$ Gaussian' limit (see Fig. 2) changed the Orlicz exponent from 1 to 2.

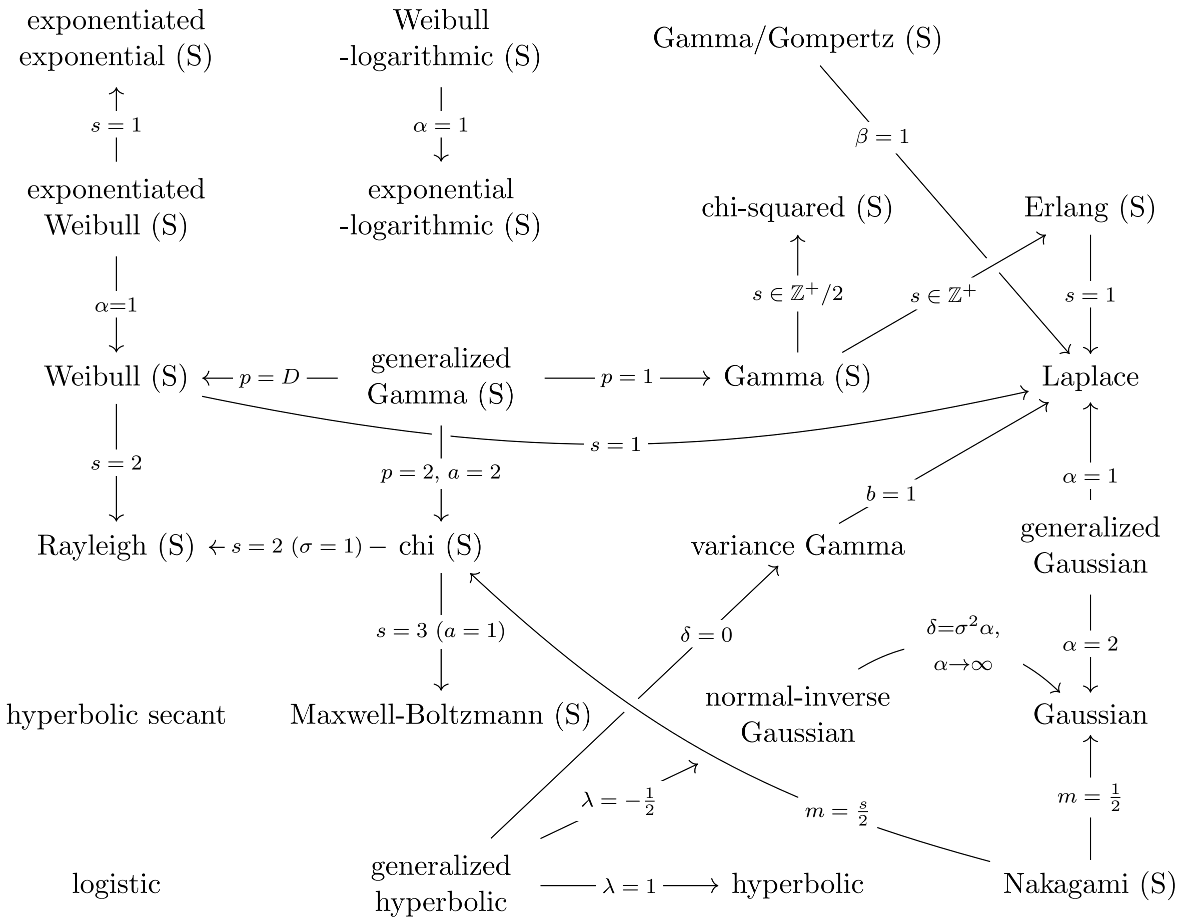

Figure 2: Relation of the spectral density examples of Table 2. '(S)' stands for symmetrized.

Kernel name	Kernel value: $k(x,y)$	Spectrum
Gaussian	$e^{-\frac{\sigma^2(x-y)^2}{2}}$	Gaussian
inverse quadric	$\frac{\sigma^2}{\sigma^2+(x-y)^2}$	Laplace
–	$\frac{\sqrt{\pi}}{\Gamma(1/\alpha)}\ {}_1\Psi_1\left(\left(\frac{1}{\alpha},\frac{2}{\alpha}\right);\left(\frac{1}{2};1\right),\frac{-[\beta(x-y)]^2}{4}\right)$	generalized Gaussian[a]
inverse multiquadric	$\left[\frac{\sigma^2}{\sigma^2+(x-y)^2}\right]^b$	variance Gamma
–	$\sum_{n\in 2\mathbb{N}}\frac{(-1)^{\frac{n}{2}}(x-y)^n\lambda^n}{n!}\Gamma\left(1+\frac{n}{s}\right)$	Weibull (S)[b]
–	$\frac{[1-2i(x-y)]^{-\frac{s}{2}}+[1+2i(x-y)]^{-\frac{s}{2}}}{2}$	chi-squared (S)[b]
–	$\frac{\left[1-\frac{i(x-y)}{\lambda}\right]^{-s}+\left[1+\frac{i(x-y)}{\lambda}\right]^{-s}}{2}$	Erlang (S)[b]
–	$\frac{[1-\theta i(x-y)]^{-s}+[1+\theta i(x-y)]^{-s}}{2}$	Gamma (S)[b]
–	$1-\sigma(x-y)e^{-\frac{\sigma^2(x-y)^2}{2}}\sqrt{\frac{\pi}{2}}\mathrm{erfi}\left(\frac{\sigma(x-y)}{\sqrt{2}}\right)$	Rayleigh (S)[b]
–	${}_1F_1\left(\frac{s}{2};\frac{1}{2};\frac{-(x-y)^2}{2}\right)$	chi (S)[b]
–	$\sum_{n\in 2\mathbb{N}}\frac{(-1)^{\frac{n}{2}}(x-y)^n\Gamma\left(\frac{n}{\alpha}+1\right)}{-\log(p)n!\beta^{\frac{\alpha}{n}}}\mathrm{Li}_{\frac{n}{\alpha}+1}(1-p)$	Weibull-logarithmic (S)[b,c]
–	$\frac{1}{2}\left[c_\Lambda(x-y)+c_\Lambda(y-x)\right]$, with $c_\Lambda(t)=$	Gamma/Gompertz (S)[b]
	$=\beta^s\frac{sb}{sb-ti}\ {}_2F_1\left(s+1;-\frac{ti}{b}+s;-\frac{ti}{b}+s+1;1-\beta\right)$	
–	$\mathrm{sech}(x-y)$	hyperbolic secant
–	$\frac{\pi s(x-y)}{\sinh(\pi s(x-y))}$	logistic
–	$e^{\delta\left[\alpha-\sqrt{\alpha^2+(x-y)^2}\right]}$	normal-inverse Gaussian
–	$\frac{\alpha K_1\left(\delta\sqrt{\alpha^2+(x-y)^2}\right)}{\sqrt{\alpha^2+(x-y)^2}K_1(\delta\alpha)}$	hyperbolic
–	$\left[\frac{\alpha}{\sqrt{\alpha^2+(x-y)^2}}\right]^\lambda\frac{K_\lambda\left(\delta\sqrt{\alpha^2+(x-y)^2}\right)}{K_\lambda(\delta\alpha)}$	generalized hyperbolic

a. The analytical computation of the characteristic function (and hence the kernel value) was carried out for $\alpha > 1$ (Pogány and Nadarajah, 2010).

b. In case of symmetrization (S): $k(x,y)=\frac{1}{2}\left[c_\Lambda(x-y)+c_\Lambda(y-x)\right]$ where $c_\Lambda(t)=\mathbb{E}_{\omega\sim\Lambda}[e^{it\omega}]$ is the characteristic function of the spectral measure (on $\mathbb{R}^{\geq 0}$) before symmetrization; $i=\sqrt{-1}$.

c. The characteristic function was obtained by Ciumara and Preda (2009).

Table 3: Kernel examples for the spectral densities given in Table 2.

4. Properties of the Orlicz Norm

In this section, for self-containedness we summarize the properties of $\|\cdot\|_\Psi$ which hold independently of the convexity/non-convexity of Ψ (unless explicitly required).

Let $X, X' \in \mathbb{R}^d$ be random variables, and assume that $\Psi : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ (and similarly Φ below) is continuous, strictly increasing, $\Psi(0) = 0$ and $\lim_{x\to\infty} \Psi(x) = \infty$.

(i) Normalization: If $X \in L_\Psi$ then $\mathbb{E}\left[\Psi\left(\frac{\|X\|_2}{\|X\|_\Psi}\right)\right] \leq 1$.

(ii) Constant: For a $\lambda \in \mathbb{R}$ constant $\|\lambda\|_\Psi = |\lambda|/\Psi^{-1}(1)$.

(iii) Monotonicity in Ψ: $\Psi \leq \Phi$ implies $\|X\|_\Psi \leq \|X\|_\Phi$.

(iv) Monotonicity in the argument: If $d = 1$ and $X \leq X'$ a.s., then $\|X\|_\Psi \leq \|X'\|_\Psi$.

(v) Finite $\|\cdot\|_\Psi$ implies integrability: If Ψ is convex and $X \in L_\Psi$, then $\mathbb{E}\left[\|X\|_2\right] \leq \|X\|_\Psi \ \Psi^{-1}(1)$.

(vi) Generalized triangle inequality: Let $X, X' \in L_{\Psi_\alpha}$ and $\alpha \in \mathbb{R}^+$. Then $X + X' \in L_{\Psi_\alpha}$ and

$$\left\|X + X'\right\|_{\Psi_\alpha} \leq 2^{\left(\frac{1}{\alpha}-1\right)_+}\left(\|X\|_{\Psi_\alpha} + \left\|X'\right\|_{\Psi_\alpha}\right).$$

(vii) Deviation inequality from $\|\cdot\|_\Psi$: If $X \in L_\Psi$ then $\mathbb{P}\left(\|X\|_2 \geq c\right) \leq \frac{2}{\Psi(c/\|X\|_\Psi)+1}$ for any $c \geq 0$.

(viii) Maximal inequality for $\|\cdot\|_{\Psi_\alpha}$ and $\alpha \in \mathbb{R}^+$: for any sequence $(X_m)_{m=1}^M$ of random variables in L_{Ψ_α}, we have

$$\left\|\max_{m\in[M]} \|X_m\|_2\right\|_{\Psi_\alpha} \leq \max_{m\in[M]} \|X_m\|_{\Psi_\alpha}\left[\frac{\log(1+M)}{\log(3/2)}\right]^{1/\alpha}.$$

The proofs of these properties are available in Section 5.10.

5. Proofs

We provide the proofs of our results and remarks presented in Sections 3 and 4. External statements used in the proofs are summarized in Section 5.11.

5.1 Proof of Remark 2(v)

In view of (7)-(8), we need to check Assumptions 1-4 with the parameterized function class

$$g_{\mathbf{z}}(\boldsymbol{\omega}) := f_{\mathbf{z}}(\boldsymbol{\omega}) - \Lambda f_{\mathbf{z}} = \boldsymbol{\omega}^{\mathbf{p}}(-\boldsymbol{\omega})^{\mathbf{q}} c_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^\top \mathbf{z}\right) - \Lambda f_{\mathbf{z}}, \quad (\mathbf{z} \in S_\Delta).$$

Thanks to the α-exponential Orlicz condition on Λ and the i.i.d. property of $(\boldsymbol{\omega}_m)_{m=1}^M$ in (6), Assumption 3 is trivially fulfilled. Assumption 4 holds by the definition of $g_{\mathbf{z}}(\cdot)$ and because the distribution of $\boldsymbol{\omega}_m$ is Λ. Assumption 1 is satisfied since S_Δ is a compact set of $\mathbb{R}^d$. Therefore, it remains to prove Assumption 2, with the existence of $n \in \mathbb{R}^+$ and $L \in \mathcal{F}_{\mathcal{P}(n)}$. First, notice that

$$|f_{\mathbf{z}}(\boldsymbol{\omega})| \leq \prod_{i\in[d]} |\omega_i|^{p_i+q_i} \leq \|\boldsymbol{\omega}\|_2^{|\mathbf{p}+\mathbf{q}|}. \tag{15}$$

- **Order**: (15) implies that

$$|g_{\mathbf{z}}(\boldsymbol{\omega})| \leq |f_{\mathbf{z}}(\boldsymbol{\omega})| + \Lambda |f_{\mathbf{z}}| \leq \|\boldsymbol{\omega}\|_2^{|\mathbf{p}+\mathbf{q}|} + \Lambda \left[\| \cdot \|_2^{|\mathbf{p}+\mathbf{q}|}\right] =: L_1(\boldsymbol{\omega}). \tag{16}$$

- **Lipschitz condition**: Let $[\mathbf{z}_1, \mathbf{z}_2] = \{a\mathbf{z}_1 + (1-a)\mathbf{z}_2 : a \in [0,1]\}$ denote the segment connecting $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^d$. By using the mean value theorem

$$|g_{\mathbf{z}_1}(\boldsymbol{\omega}) - g_{\mathbf{z}_2}(\boldsymbol{\omega})| \leq \max_{\mathbf{z} \in [\mathbf{z}_1, \mathbf{z}_2]} \left\| \frac{\partial g_{\mathbf{z}}(\boldsymbol{\omega})}{\partial \mathbf{z}} \right\|_2 \|\mathbf{z}_1 - \mathbf{z}_2\|_2 \,, \tag{17}$$

$\frac{\partial g_{\mathbf{z}}(\boldsymbol{\omega})}{\partial \mathbf{z}} = \frac{\partial f_{\mathbf{z}}(\boldsymbol{\omega})}{\partial \mathbf{z}} - \Lambda \frac{\partial f_{\mathbf{z}}(\boldsymbol{\omega})}{\partial \mathbf{z}}$ with $\frac{\partial f_{\mathbf{z}}(\boldsymbol{\omega})}{\partial \mathbf{z}} = \boldsymbol{\omega}^{\mathbf{p}}(-\boldsymbol{\omega})^{\mathbf{q}} c_{|\mathbf{p}+\mathbf{q}|+1}\left(\boldsymbol{\omega}^\top \mathbf{z}\right)\boldsymbol{\omega}$, and by using similar computations as before, one gets

$$\left\| \frac{\partial g_{\mathbf{z}}(\boldsymbol{\omega})}{\partial \mathbf{z}} \right\|_2 \leq \|\boldsymbol{\omega}\|_2^{|\mathbf{p}+\mathbf{q}|+1} + \Lambda \left[\| \cdot \|_2^{|\mathbf{p}+\mathbf{q}|+1}\right] =: L_2(\boldsymbol{\omega}). \tag{18}$$

As a result, to fulfill Assumption 2, we can take $L(\boldsymbol{\omega}) = \max(L_1(\boldsymbol{\omega}), L_2(\boldsymbol{\omega}))$ and $\rho(u) = u$. For such L, we have $n = |\mathbf{p}+\mathbf{q}| + 1$.

Refined L and ρ: We now derive refined L and ρ, by interpolating different bounds. From (16), we can obtain the crude estimate $|g_{\mathbf{z}_1}(\boldsymbol{\omega}) - g_{\mathbf{z}_2}(\boldsymbol{\omega})| \leq 2L_1(\boldsymbol{\omega})$, which combined with (17)-(18) gives

$$|g_{\mathbf{z}_1}(\boldsymbol{\omega}) - g_{\mathbf{z}_2}(\boldsymbol{\omega})| \leq (2L_1(\boldsymbol{\omega}))^{1-\beta} \Big[\|\mathbf{z}_1 - \mathbf{z}_2\|_2 \, L_2(\boldsymbol{\omega}) \Big]^\beta \tag{19}$$

for any $\beta \in (0,1]$. Here we have used that if $0 \leq x \leq \min(x_1, x_2)$ then $x \leq x_1^{1-\beta} x_2^\beta$. It follows that one can take

$$\rho(u) = u^\beta, \quad n = |\mathbf{p}+\mathbf{q}| + \beta, \quad L(\boldsymbol{\omega}) = \max\left(L_1(\boldsymbol{\omega}), (2L_1(\boldsymbol{\omega}))^{1-\beta} L_2^\beta(\boldsymbol{\omega})\right) \in \mathcal{F}_{\mathcal{P}(n)}. \tag{20}$$

For $\beta = 1$, we retrieve the former choice of L and ρ. Furthermore, we have

$$I_\rho(|T|) = |T|^\beta \int_0^1 \sqrt{\log\left(1 + \frac{2|T|}{(u|T|^\beta)^{1/\beta}}\right)} \mathrm{d}u = |T|^\beta \int_0^1 \sqrt{\log\left(1 + \frac{2}{u^{1/\beta}}\right)} \mathrm{d}u < +\infty. \tag{21}$$

Notice that the advantage of having the additional degree-of-freedom β is two-fold, and it is striking when $\beta \to 0$ (compared to $\beta = 1$). Firstly, it gives a smaller n, which has a (slight) positive impact on the control of statistical fluctuations; secondly, the dependence of $I_\rho(|T|)$ in the diameter $|T|$ is smaller through the growth exponent.

To conclude, we have proved that Orlicz RFFs fulfill the assumptions of Theorem 1. Later in Section 5.7, we will establish that $I_\rho(|T|)$ satisfies a (tight) bound w.r.t. $\sqrt{1 + (\log|T|)_+}$.

5.2 Proof that Polynomial Growth Preserves the Exponential Orlicz Property

We show that $\|f(\mathbf{X})\|_{\Psi_\gamma} < \infty$ for $\|\mathbf{X}\|_{\Psi_\alpha} < \infty$, $f \in \mathcal{F}_{\mathcal{P}(n)}$, $n \in \mathbb{R}^+$, $\gamma = \frac{\alpha}{n}$. Indeed, by the definition of $f \in \mathcal{F}_{\mathcal{P}(n)}$, there exists $C \in \mathbb{R}^+$ such that $|f(\mathbf{x})| \leq C(1 + \|\mathbf{x}\|_2^n)$ for all $\mathbf{x} \in \mathbb{R}^d$. Hence for any $\gamma > 0$

$$|f(\mathbf{x})|^\gamma \leq C^\gamma \left(1 + \|\mathbf{x}\|_2^n\right)^\gamma \overset{(*)}{\leq} 2^{(\gamma-1)_+} C^\gamma \left(1 + \|\mathbf{x}\|_2^{n\gamma}\right), \tag{22}$$

where in $(*)$ we used that

$$(a+b)^\gamma \leq 2^{(\gamma-1)_+}\left(a^\gamma+b^\gamma\right), \quad a,b\geq 0, \gamma>0. \tag{23}$$

Since $\mathbf{X} \in L_{\Psi_\alpha}$ there is some $s \in \mathbb{R}^+$ for which $\mathbb{E}\left[e^{s\|\mathbf{X}\|_2^\alpha}\right] < \infty$. Combining this property with (22) and recalling that $n\gamma = \alpha$ yields

$$e^{s'|f(\mathbf{x})|^\gamma} \leq e^{s' 2^{(\gamma-1)_+}C^\gamma\left(1+\|\mathbf{x}\|_2^\alpha\right)} \qquad \Rightarrow \quad \mathbb{E}\left[e^{s'|f(\mathbf{X})|^\gamma}\right] \leq e^{s' 2^{(\gamma-1)_+}C^\gamma}\mathbb{E}\left[e^{s\|\mathbf{X}\|_2^\alpha}\right] < \infty$$

with $s' = \frac{s}{2^{(\gamma-1)_+}C^\gamma}$; this shows that $f(\mathbf{X}) \in L_{\Psi_\gamma}$.

5.3 $\Psi_\gamma^{(l)}$, the Convexification of Ψ_γ

In the proof of Theorem 1 an integral control with *convex* Ψ (see Section 4(v)) is beneficial/applied. However, Ψ_γ is not convex for $\gamma \in (0,1)$. To handle this issue, we convexify $\Psi_\gamma(x) = e^{x^\gamma}-1$ in case of $\gamma \in (0,1)$ for 'small' values of the argument.[7]

- By computing the derivatives of Ψ_γ we get that it is convex iff $x \geq x_\gamma := \left(\frac{1-\gamma}{\gamma}\right)^{\frac{1}{\gamma}}$. Indeed,

$$\begin{aligned}
&\Psi_\gamma'(x) = \gamma x^{\gamma-1}e^{x^\gamma},\\
&\Psi_\gamma''(x) = \gamma e^{x^\gamma}\left[(\gamma-1)x^{\gamma-2}+x^{\gamma-1}\gamma x^{\gamma-1}\right] \Rightarrow\\
&\Psi_\gamma''(x) = 0 \Leftrightarrow x = x_\gamma, \quad \Psi_\gamma''(x) > 0 \Leftrightarrow x > x_\gamma, \quad \Psi_\gamma''(x) < 0 \Leftrightarrow x < x_\gamma.
\end{aligned}$$

- We also have to make sure that $\Psi_\gamma^{(l)}$, constructed as the line connecting $(0,0)$ with $(x, \Psi_\gamma(x))$ glued to $\Psi_\gamma|_{[x,\infty)}$, gives a convex function, for a suitable choice of x. A geometric argument shows that it is enough to choose $x \geq x_\gamma (>0)$ such that

$$\frac{\Psi_\gamma(x)}{x} \leq \Psi_\gamma'(x) \ \Leftrightarrow\ e^{x^\gamma}-1 \leq \gamma x^\gamma e^{x^\gamma}.$$

Since the r.h.s. is higher order than the l.h.s., the requirement can be satisfied for large enough x; we can choose

$$\tilde{x}_\gamma := \inf\left\{x \geq x_\gamma : e^{x^\gamma}-1 \leq \gamma x^\gamma e^{x^\gamma}\right\},$$

and define

$$\Psi_\gamma^{(l)}(x) := \begin{cases} \frac{\Psi_\gamma(\tilde{x}_\gamma)}{\tilde{x}_\gamma}x & \text{if } x \in [0,\tilde{x}_\gamma),\\ \Psi_\gamma(x) & \text{if } x \in [\tilde{x}_\gamma,\infty). \end{cases}$$

Notice that by construction $\Psi_\gamma^{(l)} \leq \Psi_\gamma$.

[7] For $\gamma = 1$, $\Psi_\gamma^{(l)} = \Psi_\gamma$.

5.4 Proof of Theorem 1

By introducing the $\mathcal{R}_c f(\mathbf{x}) := f(\mathbf{x}) - \mathcal{T}_c f(\mathbf{x})$ notation of residuals obtained at level $c \in \mathbb{R}^+$ (the value of c will be specified later), we bound the target quantity by using the subadditivity of supremum

$$
\begin{aligned}
\sup_{\mathbf{t}\in T} \frac{1}{M} \sum_{m\in[M]} & \underbrace{f_{\mathbf{t}}(\mathbf{X}_m)}_{\mathcal{T}_c f_{\mathbf{t}}(\mathbf{X}_m) + \mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)} \\
&= \sup_{\mathbf{t}\in T} \frac{1}{M} \sum_{m\in[M]} \left(\mathcal{T}_c f_{\mathbf{t}}(\mathbf{X}_m) - \mathbb{E}\left[\mathcal{T}_c f_{\mathbf{t}}(\mathbf{X}_m)\right] + \mathbb{E}\left[\mathcal{T}_c f_{\mathbf{t}}(\mathbf{X}_m)\right] + \mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)\right) \\
&\le \underbrace{\sup_{\mathbf{t}\in T} \frac{1}{M} \sum_{m\in[M]} \Big(\mathcal{T}_c f_{\mathbf{t}}(\mathbf{X}_m) - \mathbb{E}\left[\mathcal{T}_c f_{\mathbf{t}}(\mathbf{X}_m)\right]\Big)}_{\overline{Z}^{\mathcal{T}_c}} + \underbrace{\sup_{\mathbf{t}\in T} \mathbb{E}\left[\frac{1}{M} \sum_{m\in[M]} \mathcal{T}_c f_{\mathbf{t}}(\mathbf{X}_m)\right]}_{\mathcal{E}^{\mathcal{T}_c}} \\
&\quad \underbrace{+ \sup_{\mathbf{t}\in T} \frac{1}{M} \sum_{m\in[M]} \mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)}_{Z^{\mathcal{R}_c}} .
\end{aligned}
$$

This means that using c for which $\mathcal{E}^{\mathcal{T}_c} \le \frac{\varepsilon}{3}$,

$$
\mathbb{P}\left(\sup_{\mathbf{t}\in T} \frac{1}{M} \sum_{m\in[M]} f_{\mathbf{t}}(\mathbf{X}_m) \ge \varepsilon\right) \le \mathbb{P}\left(Z^{\mathcal{R}_c} \ge \varepsilon/3\right) + \mathbb{P}\left(\overline{Z}^{\mathcal{T}_c} \ge \varepsilon/3\right). \tag{24}
$$

The structure of our proof is as follows.

1. Unbounded part ($Z^{\mathcal{R}_c}$): Based on the Talagrand and the Hoffman-Jorgensen inequalities, for large enough c (referred to as $c_{\mathtt{HJ}}$) we will derive an exponential control over $\mathbb{P}\left(Z^{\mathcal{R}_c} \ge \varepsilon/3\right)$ expressed with $\left\|\max_{m\in[M]} \sup_{\mathbf{t}\in T} |f_{\mathbf{t}}(\mathbf{X}_m)|\right\|_{\Psi_\gamma}$ which is finite by Section 5.5.
2. Bounded part ($\overline{Z}^{\mathcal{T}_c}$): We handle this term using the Klein-Rio inequality and the Dudley entropy integral bound. In addition, this part will give rise to the constraint (9) on ε.
3. Truncation ($\mathcal{E}^{\mathcal{T}_c}$): As $\mathbb{E}[f_{\mathbf{t}}(\mathbf{X}_m)] = 0$, $\mathcal{T}_c f_{\mathbf{t}} \approx f_{\mathbf{t}}$ and $\mathbb{E}[\mathcal{T}_c f_{\mathbf{t}}(\mathbf{X}_m)] \approx 0$ for large c (called $c_{\min}$). The $\mathcal{E}^{\mathcal{T}_c} \le \frac{\varepsilon}{3}$ requirement can be controlled via the integral form of the expectation of non-negative random variables and the incomplete Gamma function.

The bounding of the $Z^{\mathcal{R}_c}$, $\overline{Z}^{\mathcal{T}_c}$ and $\mathcal{E}^{\mathcal{T}_c}$ quantities is detailed in the following sections. Plugging the (25) and (27) results of the computations into (24) gives the final bound (10). The ε constraint comes from (32), provided that $c \ge c_{\min} \vee c_{\mathtt{HJ}}$. The constants $c_{\min}$ and $c_{\mathtt{HJ}}$ are defined in (34) and (37), respectively.

5.4.1 Bounding $Z^{\mathcal{R}_c}$

$\mathbb{P}\left(Z^{\mathcal{R}_c} \ge \varepsilon/3\right)$ is bounded as

$$
\mathbb{P}\left(Z^{\mathcal{R}_c} \ge \varepsilon/3\right) \le \mathbb{P}\left(\sup_{\mathbf{t}\in T} \sum_{m\in[M]} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)| \ge M\varepsilon/3\right) \overset{(a)}{\le} \mathbb{P}\left(\sum_{m\in[M]} \sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)| \ge M\varepsilon/3\right)
$$

$$\overset{(b)}{\leq} 2e^{-\left(\frac{M\varepsilon/3}{\left\|\sum_{m\in[M]}\sup_{\mathbf{t}\in T}|\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right\|_{\Psi_\gamma}}\right)^\gamma} \overset{(c)}{\leq} 2e^{-\left(\frac{M\varepsilon}{3K_\gamma\left\|\max_{m\in[M]}\sup_{\mathbf{t}\in T}|f_{\mathbf{t}}(\mathbf{X}_m)|\right\|_{\Psi_\gamma}}\right)^\gamma}, \tag{25}$$

where in (a) we used the the sub-additivity of the supremum, in (b) the deviation inequality Section (4)(vii) was applied, (c) holds by Section 5.6 for $c \geq c_{\mathtt{HJ}}$ (the value of $c_{\mathtt{HJ}}$ is defined in Section 5.6).

5.4.2 Bounding $\overline{Z}^{\mathcal{T}_c}$

Below we will invoke the Klein-Rio inequality and control the expectation $\mathbb{E}\left[\overline{Z}^{\mathcal{T}_c}\right]$.

- **Klein-Rio inequality**: Let $g_{m,\mathbf{t}} : \mathbf{x} \in \mathbb{R}^d \mapsto \mathcal{T}_c f_{\mathbf{t}}(\mathbf{x}) - \mathbb{E}\left[\mathcal{T}_c f_{\mathbf{t}}(\mathbf{X}_m)\right]$ and let us define the function classes

  $$\mathcal{T}_c\mathcal{F}^{[M]} := \{\mathbf{g}_{\mathbf{t}} := (g_{1,\mathbf{t}}, \dots, g_{M,\mathbf{t}}) : \mathbf{t} \in T\}, \qquad \mathcal{T}_c\mathcal{F} := \{\mathcal{T}_c f_{\mathbf{t}} : \mathbf{t} \in T\}.$$

  - $g_{m,\mathbf{t}} \in [-2c, 2c]$ are measurable and bounded functions.
  - Centering: by construction $\mathbb{E}[g_{m,\mathbf{t}}(\mathbf{X}_m)] = 0$ ($\forall m \in [M]$).
  - Countability: Since $\mathbf{t} \mapsto f_{\mathbf{t}}$ is continuous, the $\sup_{\mathbf{t}\in T}$ can be restricted to rational numbers ($T \cap \mathbb{Q}^d$), one can take $T \leftarrow T \cap \mathbb{Q}^d$, and assume that $\mathcal{T}_c\mathcal{F}^{[M]}$ is countable.

  If

  $$\mathbb{E}\left[\overline{Z}^{\mathcal{T}_c}\right] \leq \varepsilon/6, \tag{26}$$

  then the Klein-Rio inequality (Theorem 8 where the $\sup_{\mathbf{t}\in T}$ and $\sup_{\mathbf{f}\in\mathcal{T}_c\mathcal{F}^{[M]}}$ coincide) implies that

  $$\begin{aligned}\mathbb{P}\left(\overline{Z}^{\mathcal{T}_c} \geq \varepsilon/3\right) &\overset{(26)}{\leq} \mathbb{P}\left(\overline{Z}^{\mathcal{T}_c} - \mathbb{E}\left[\overline{Z}^{\mathcal{T}_c}\right] \geq \varepsilon/6\right) \leq e^{-\frac{M\,(\varepsilon/6)^2}{2\left(\bar{\sigma}^2 + 4c\mathbb{E}\left[\overline{Z}^{\mathcal{T}_c}\right]\right) + 6c\,\varepsilon/6}} \\ &\overset{(26)}{\leq} e^{-\frac{M\,(\varepsilon/6)^2}{2\bar{\sigma}^2 + 14c\varepsilon/6}} = e^{-\frac{M\,\varepsilon^2}{72\bar{\sigma}^2 + 84c\,\varepsilon}} \leq e^{-\frac{M\,\varepsilon^2}{72\sigma^2 + 84c\,\varepsilon}},\end{aligned} \tag{27}$$

  where the weak variance $\bar{\sigma}^2$ is defined and bounded by

  $$\begin{aligned}\bar{\sigma}^2 := \sup_{\mathbf{t}\in T}\frac{1}{M}\sum_{m\in[M]}\mathbb{E}\left[\left(\mathcal{T}_c f_{\mathbf{t}}(\mathbf{X}_m) - \mathbb{E}\left[\mathcal{T}_c f_{\mathbf{t}}(\mathbf{X}_m)\right]\right)^2\right] &\leq \sup_{\mathbf{t}\in T}\frac{1}{M}\sum_{m\in[M]}\mathbb{E}\left[\left(\mathcal{T}_c f_{\mathbf{t}}(\mathbf{X}_m)\right)^2\right] \\ &\leq \sup_{\mathbf{t}\in T}\frac{1}{M}\sum_{m\in[M]}\mathbb{E}\left[f_{\mathbf{t}}^2(\mathbf{X}_m)\right] =: \sigma^2.\end{aligned}$$

- **Bounding** $\mathbb{E}\left[\overline{Z}^{\mathcal{T}_c}\right]$: We control $\mathbb{E}\left[\overline{Z}^{\mathcal{T}_c}\right]$ in (26) by the Dudley entropy integral bound. In this bound the covering number of $\mathcal{T}_c\mathcal{F}$ is estimated by that of the compact set $T \subset \mathbb{R}^{d'}$ with propagation relying on Assumption (2b).
  - **Dudley entropy integral bound**: Slight modification (without absolute value) of (van der Vaart and Wellner, 1996, Lemma 2.3.1) gives

    $$\mathbb{E}\left[\overline{Z}^{\mathcal{T}_c}\right] \leq 2\mathbb{E}\left[\mathcal{R}(\mathbf{X}_{1:M}, \mathcal{T}_c\mathcal{F})\right], \tag{28}$$

where $\mathcal{R}(\mathbf{x}_{1:M}, \mathcal{T}_c\mathcal{F}) := \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{\mathbf{t}\in T}\frac{1}{M}\sum_{m\in[M]}\varepsilon_m \mathcal{T}_c f_{\mathbf{t}}(\mathbf{x}_m)\right]$ is the Rademacher average of $\mathcal{T}_c\mathcal{F}$, $\mathbf{X}_{1:M} := (\mathbf{X}_m)_{m\in[M]}$, $\mathbf{x}_{1:M} := (\mathbf{x}_m)_{m\in[M]}$, $\boldsymbol{\varepsilon} :=(\varepsilon_m)_{m\in[M]}$ contains independent Rademacher variables (i.e. $\mathbb{P}\left(\varepsilon_m = \pm 1\right) = \frac{1}{2}$), and $\boldsymbol{\varepsilon}$ is independent of $\mathbf{X}_{1:M}$.

Let $Z_{\mathbf{t}}(\mathbf{x}_{1:M}) := \frac{1}{M}\sum_{m\in[M]}\varepsilon_m \mathcal{T}_c f_{\mathbf{t}}(\mathbf{x}_m)$, so $\mathcal{R}(\mathbf{x}_{1:M}, \mathcal{T}_c\mathcal{F}) = \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{\mathbf{t}\in T} Z_{\mathbf{t}}(\mathbf{x}_{1:M})\right]$, and define the pseudo-metric on T as $d(\mathbf{t},\mathbf{s}) := \left(\frac{1}{M}\sum_{m\in[M]}\left[\mathcal{T}_c f_{\mathbf{t}}(\mathbf{x}_m) - \mathcal{T}_c f_{\mathbf{s}}(\mathbf{x}_m)\right]^2\right)^{1/2}$.

The $\{Z_{\mathbf{t}} : \mathbf{t}\in T\}$ process is

∗ separable since it is continuous, and $T\subset\mathbb{R}^{d'}$ is separable,

∗ centered thanks to the Rademacher variables,

∗ sub-Gaussian with respect to $M^{-1/2}d$: indeed, for any $\lambda\in\mathbb{R}^+$ and $\mathbf{t},\mathbf{s}\in T$

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[e^{\lambda(Z_{\mathbf{t}}-Z_{\mathbf{s}})}\right] \overset{(a)}{=} \prod_{m=1}^{M}\mathbb{E}_{\varepsilon_m}\left[e^{\varepsilon_m\frac{\lambda}{M}\left[\mathcal{T}_c f_{\mathbf{t}}(\mathbf{x}_m)-\mathcal{T}_c f_{\mathbf{s}}(\mathbf{x}_m)\right]}\right] \overset{(b)}{\leq} \prod_{m=1}^{M} e^{\frac{\lambda^2}{2M^2}\left[\mathcal{T}_c f_{\mathbf{t}}(\mathbf{x}_m)-\mathcal{T}_c f_{\mathbf{s}}(\mathbf{x}_m)\right]^2}$$
$$= e^{\frac{\lambda^2\left(M^{-1/2}d(\mathbf{t},\mathbf{s})\right)^2}{2}}.$$

In (a) we used the independence of ε_m-s, (b) follows from $\mathbb{E}_{\varepsilon_m}\left[e^{a\varepsilon_m}\right] = \cosh(a) \overset{(*)}{\leq} e^{\frac{a^2}{2}}$ ($\forall a\in\mathbb{R}$), where $(*)$ can be obtained from the power series expansion of the $\cosh(\cdot)$ and the exponential function.

Hence Theorem 9 can be applied:

$$\mathcal{R}(\mathbf{x}_{1:M}, \mathcal{T}_c\mathcal{F}) \leq C_D\int_0^\infty \sqrt{\log(N(\varepsilon, M^{-1/2}d, T))}\mathrm{d}\varepsilon$$
$$= \frac{C_D}{\sqrt{M}}\int_0^\infty \sqrt{\log(N(\varepsilon, d, T))}\mathrm{d}\varepsilon, \tag{29}$$

where we used the $N(\varepsilon, M^{-1/2}d, T) = N(M^{1/2}\varepsilon, d, T)$ identity, and applied an $\tilde{\varepsilon} = M^{1/2}\varepsilon$ substitution. We note that the above infinite integral can be truncated at $|T|_d := \sup_{\mathbf{t},\mathbf{s}\in T} d(\mathbf{t},\mathbf{s})$, the d-diameter of T, since $N(\varepsilon, d, T) = 1$ for $\varepsilon \geq |T|_d$.

– **Covering number**: By (11) one can relate $d(\mathbf{t},\mathbf{s})$ and $\|\mathbf{t}-\mathbf{s}\|_2$ as

$$d(\mathbf{t},\mathbf{s}) \leq \left[\frac{1}{M}\sum_{m\in[M]} L^2(\mathbf{x}_m)\right]^{1/2}\rho\left(\|\mathbf{t}-\mathbf{s}\|_2\right) := \|L\|_{L^2(\mathbf{x}_{1:M})}\,\rho\left(\|\mathbf{t}-\mathbf{s}\|_2\right),$$

which implies

$$N(\varepsilon, d, T) \leq N\left(\rho^{-1}\left(\frac{\varepsilon}{\|L\|_{L^2(\mathbf{x}_{1:M})}}\right), \|\cdot\|_2, T\right), \tag{30}$$
$$|T|_d \leq \|L\|_{L^2(\mathbf{x}_{1:M})}\sup_{\mathbf{t},\mathbf{s}\in T}\rho\left(\|\mathbf{t}-\mathbf{s}\|_2\right) \leq \|L\|_{L^2(\mathbf{x}_{1:M})}\,\rho(|T|). \tag{31}$$

In the last inequality the increasing property of ρ was exploited. Combining (30)-(31) with the well-known bound (van de Geer, 2000, Lemma 2.5) on the covering number[8]

[8] In our definition of the covering number, in its bound on compact sets in $\mathbb{R}^d$ (van de Geer, 2000, Lemma 2.5) and in the final Dudley entropy bound (Bartlett, 2013, Lecture 11, 14) the elements of the ϵ-net are assumed to belong to the set covered.

of a compact set $T \subset \mathbb{R}^{d'}$

$$N\left(\varepsilon', \|\cdot\|_2, T\right) \leq \left(\frac{2|T|}{\varepsilon'} + 1\right)^{d'}, \forall \varepsilon' > 0,$$

(29) can be estimated further as

$$\begin{aligned}\mathcal{R}(\mathbf{x}_{1:M}, \mathcal{T}_c\mathcal{F}) &\leq \frac{C_D\sqrt{d'}}{\sqrt{M}} \int_0^{\|L\|_{L^2(\mathbf{x}_{1:M})}\rho(|T|)} \sqrt{\log\left(\frac{2|T|}{\rho^{-1}\left(\frac{\varepsilon}{\|L\|_{L^2(\mathbf{x}_{1:M})}}\right)} + 1\right)} \mathrm{d}\varepsilon \\ &= C_D\sqrt{d'}\frac{\|L\|_{L^2(\mathbf{x}_{1:M})}\rho\left(|T|\right)}{\sqrt{M}} \int_0^1 \sqrt{\log\left(1 + \frac{2|T|}{\rho^{-1}\left(u\rho\left(|T|\right)\right)}\right)} \mathrm{d}u,\end{aligned}$$

where we introduced the new variable $u = \frac{\varepsilon}{\|L\|_{L^2(\mathbf{x}_{1:M})}\rho(|T|)}$. Substituting this bound into (28) we arrive at

$$\mathbb{E}\left[\overline{Z}^{\mathcal{T}_c}\right] \leq 2C_D\sqrt{d'}\frac{\mathbb{E}\left[\|L\|_{L^2(\mathbf{X}_{1:M})}\right]}{\sqrt{M}} I_\rho(|T|) =: B. \tag{32}$$

To guarantee $\mathbb{E}\left[\overline{Z}^{\mathcal{T}_c}\right] \leq \varepsilon/6$, we solve $B \leq \frac{\varepsilon}{6}$; this gives the (9) bound on ε.

5.4.3 Bounding $\mathcal{E}^{\mathcal{T}_c}$

- **Bounding $\mathcal{E}^{\mathcal{T}_c}$ by the incomplete Gamma function (I_γ):**

$$\begin{aligned}\mathbb{E}\left[\mathcal{T}_c f_{\mathbf{t}}(\mathbf{X}_m)\right] &\overset{(a)}{=} -\mathbb{E}\left[\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)\right] \overset{(b)}{\leq} \mathbb{E}\left[\left(-f_{\mathbf{t}}(\mathbf{X}_m) - c\right)\mathbb{1}_{f_{\mathbf{t}}(\mathbf{X}_m)\leq -c}\right] \\ &\overset{(c)}{=} \int_c^\infty \mathbb{P}\left(-f_{\mathbf{t}}(\mathbf{X}_m) \geq y\right)\mathrm{d}y.\end{aligned} \tag{33}$$

In (a) we used that $\mathcal{T}_c f_{\mathbf{t}}(\mathbf{x}) = f_{\mathbf{t}}(\mathbf{x}) - \mathcal{R}_c f_{\mathbf{t}}(\mathbf{x})$ and $\mathbb{E}[f_{\mathbf{t}}(\mathbf{X}_m)] = 0$, (b) follows from

$$\mathcal{R}_c f_{\mathbf{t}}(\mathbf{x}) = [f_{\mathbf{t}}(\mathbf{x}) + c]\mathbb{1}_{f_{\mathbf{t}}(\mathbf{x})\leq -c} + [f_{\mathbf{t}}(\mathbf{x}) - c]\mathbb{1}_{f_{\mathbf{t}}(\mathbf{x})\geq c} \geq [f_{\mathbf{t}}(\mathbf{x}) + c]\mathbb{1}_{f_{\mathbf{t}}(\mathbf{x})\leq -c}.$$

(c) holds by using that for a $Z \geq 0$ random variable, $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}\left(Z \geq z\right)\mathrm{d}z$; we choose $Z = \max\left(0, -f_{\mathbf{t}}(\mathbf{X}_m) - c\right)$. Therefore

$$\begin{aligned}&\mathcal{E}^{\mathcal{T}_c} = \sup_{\mathbf{t}\in T}\mathbb{E}\left[\frac{1}{M}\sum_{m\in[M]}\mathcal{T}_c f_{\mathbf{t}}(\mathbf{X}_m)\right] \overset{(a)}{\leq} \max_{m\in[M]}\sup_{\mathbf{t}\in T}\int_c^\infty \mathbb{P}\left(-f_{\mathbf{t}}(\mathbf{X}_m) \geq y\right)\mathrm{d}y \\ &\overset{(b)}{\leq} 2\max_{m\in[M]}\sup_{\mathbf{t}\in T}\int_c^\infty e^{-\left(\frac{y}{\|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma}}\right)^\gamma}\mathrm{d}y \\ &\overset{(c)}{=} 2\max_{m\in[M]}\sup_{\mathbf{t}\in T}\left(\|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma}\int_{\frac{c}{\|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma}}}^\infty e^{-u^\gamma}\mathrm{d}u\right)\end{aligned}$$

$$\begin{aligned}
&\overset{(d)}{=} 2 \max_{m\in[M]} \sup_{\mathbf{t}\in T} \left(\|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma} \left[\int_0^\infty e^{-u^\gamma} \mathrm{d}u - \int_0^{\frac{c}{\|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma}}} e^{-u^\gamma} \mathrm{d}u \right] \right) \\
&\overset{(e)}{=} 2 \max_{m\in[M]} \sup_{\mathbf{t}\in T} \left(\|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma} \left[\Gamma\left(1+\frac{1}{\gamma}\right) - I_\gamma\left(\frac{c}{\|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma}}\right) \right] \right) \\
&\overset{(f)}{\leq} 2 \max_{m\in[M]} \sup_{\mathbf{t}\in T} \|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma} \left[\Gamma\left(1+\frac{1}{\gamma}\right) - I_\gamma\left(\frac{c}{\max_{m'\in[M]} \sup_{\mathbf{t}'\in T} \|f_{\mathbf{t}'}(\mathbf{X}_{m'})\|_{\Psi_\gamma}}\right) \right] \\
&\overset{(g)}{\leq} 2\Gamma\left(1+\frac{1}{\gamma}\right) \left(1 - \left[1 - e^{-\beta_\gamma \left(\frac{c}{\max_{m\in[M]} \sup_{\mathbf{t}\in T} \|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma}} \right)^\gamma} \right]^{\frac{1}{\gamma}} \right) \max_{m\in[M]} \sup_{\mathbf{t}\in T} \|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma} \\
&=: \tilde{B},
\end{aligned}$$

where (a) holds by taking maximum over $m \in [M]$ and using (33), (b) follows from the $\mathbb{P}\left(-f_{\mathbf{t}}(\mathbf{X}_m) \geq y\right) \leq 2e^{-\left(\frac{y}{\|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma}}\right)^\gamma}$ deviation inequality implied by Section 4(vii). (c) was obtained from a $u = \frac{y}{\|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma}}$ substitution, in (d) we decomposed the integral to have the incomplete Gamma function appear. (e) is a consequence of the definition of I_γ and the limit

$$I_\gamma(x) = \int_0^x e^{-t^\gamma} \mathrm{d}t = \frac{1}{\gamma} \int_0^{x^{\frac{1}{\gamma}}} u^{\frac{1}{\gamma}-1} e^{-u} \mathrm{d}u \xrightarrow{x\to\infty} \frac{1}{\gamma}\Gamma\left(\frac{1}{\gamma}\right) = \Gamma\left(1+\frac{1}{\gamma}\right),$$

where we applied an $u = t^\gamma$ substitution and the $\Gamma(z+1) = z\Gamma(z)$ recursion. (f) comes from the monotonicity of I_γ ($I_\gamma(x) \leq I_\gamma(y)$ if $x \leq y$). (g) follows from applying the lower bound on I_γ from Theorem 10 with $x = \frac{c}{\max_{m\in[M]} \sup_{\mathbf{t}\in T} \|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma}}$.

- **Additional truncation level bound on** c: Guaranteeing $\tilde{B} \leq \frac{\varepsilon}{3}$ (and thus $\mathcal{E}^{\mathcal{T}_c} \leq \frac{\varepsilon}{3}$) is equivalent to choosing c large enough such that

$$1 - \left[1 - e^{-\beta_\gamma \left(\frac{c}{\max_{m\in[M]} \sup_{\mathbf{t}\in T} \|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma}} \right)^\gamma} \right]^{\frac{1}{\gamma}} \leq \frac{\varepsilon}{6\Gamma\left(1+\frac{1}{\gamma}\right) \max_{m\in[M]} \sup_{\mathbf{t}\in T} \|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma}}.$$

Because $\gamma \leq 1$, the function $x \mapsto 1 - (1-x)^{\frac{1}{\gamma}}$ is concave on $[0,1]$, and thus it is below its tangent line computed at $(0, h(0))$, i.e. $1 - (1-x)^{\frac{1}{\gamma}} \leq \frac{1}{\gamma}x$. Therefore choosing $x = e^{-\beta_\gamma \left(\frac{c}{\max_{m\in[M]} \sup_{\mathbf{t}\in T} \|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma}} \right)^\gamma}$ it is enough to use c such that

$$\frac{1}{\gamma} e^{-\beta_\gamma \left(\frac{c}{\max_{m\in[M]} \sup_{\mathbf{t}\in T} \|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma}} \right)^\gamma} \leq \frac{\epsilon}{6\Gamma\left(1+\frac{1}{\gamma}\right) \max_{m\in[M]} \sup_{\mathbf{t}\in T} \|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma}}.$$

Solving this inequality for c means that

$$c \geq c_{\min} := \max_{m\in[M]} \sup_{\mathbf{t}\in T} \|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma} \left[\frac{1}{\beta_\gamma} \log\left(\frac{6\Gamma\left(1+\frac{1}{\gamma}\right) \max_{m\in[M]} \sup_{\mathbf{t}\in T} \|f_{\mathbf{t}}(\mathbf{X}_m)\|_{\Psi_\gamma}}{\gamma\varepsilon} \right) \right]^{\frac{1}{\gamma}}. \tag{34}$$

5.5 Proof of $\mathcal{F} \subset \mathcal{F}_{\mathcal{P}(n)}$, $\mathbb{E}\left[\max_{m\in[M]} \sup_{\mathbf{t}\in T}|f_{\mathbf{t}}(\mathbf{X}_m)|\right] < \infty$, and $\left\|\max_{m\in[M]} \sup_{\mathbf{t}\in T} |f_{\mathbf{t}}(\mathbf{X}_m)|\right\|_{\Psi_{\frac{\alpha}{n}}} < \infty$

By Assumption (2a)-(2b), the triangle inequality and the monotonicity of ρ, one gets

$$|f_{\mathbf{t}}(\mathbf{x})| \leq |f_{\mathbf{t}}(\mathbf{x}) - f_{\mathbf{t}_0}(\mathbf{x})| + |f_{\mathbf{t}_0}(\mathbf{x})| \leq L(\mathbf{x})\left[\rho\left(\|\mathbf{t}-\mathbf{t}_0\|_2\right)+1\right] \leq L(\mathbf{x})[\rho(|T|)+1], \tag{35}$$

for any $\mathbf{t} \in T, \mathbf{x} \in \mathbb{R}^d$. The individual statements now can be proved as follows.

- $\mathcal{F} \subset \mathcal{F}_{\mathcal{P}(n)}$: By $L \in \mathcal{F}_{\mathcal{P}(n)}$ and (35), $f_{\mathbf{t}} \in \mathcal{F}_{\mathcal{P}(n)}$ for all $\mathbf{t} \in T$, in other words $\mathcal{F} \subset \mathcal{F}_{\mathcal{P}(n)}$.
- **Finiteness of** $\left\|\max_{m\in[M]} \sup_{\mathbf{t}\in T} |f_{\mathbf{t}}(\mathbf{X}_m)|\right\|_{\Psi_{\frac{\alpha}{n}}}$: Using (35), we get

$$\max_{m\in[M]} \sup_{\mathbf{t}\in T} |f_{\mathbf{t}}(\mathbf{X}_m)| \leq [\rho(|T|)+1] \sum_{m\in[M]} L(\mathbf{X}_m). \tag{36}$$

  Thanks to Section 5.2, each $L(\mathbf{X}_m)$ belongs to $L_{\Psi_{\frac{\alpha}{n}}}$. Combining this with the generalized triangular inequality Section 4(vi) gives the claim.
- **Finiteness of** $\mathbb{E}\left[\max_{m\in[M]} \sup_{\mathbf{t}\in T} |f_{\mathbf{t}}(\mathbf{X}_m)|\right]$: Each $L(\mathbf{X}_m)$ is integrable (because L has a polynomial growth and the distribution of X_m satisfies the α-Orlicz exponential assumption). Thus, the statement follows from (36).

5.6 Control if $c \geq c_{HJ}$

We show that under the assumptions of Theorem 1 with

$$c \geq c_{\mathtt{HJ}} := 8\mathbb{E}\left[\max_{m\in[M]} \sup_{\mathbf{t}\in T}|f_{\mathbf{t}}(\mathbf{X}_m)|\right] \tag{37}$$

one has

$$\left\|\sum_{m\in[M]} \sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right\|_{\Psi_\gamma} \leq K_\gamma \left\|\max_{m\in[M]} \sup_{\mathbf{t}\in T} |f_{\mathbf{t}}(\mathbf{X}_m)|\right\|_{\Psi_\gamma}. \tag{38}$$

Notice that $c_{\mathtt{HJ}}$ is finite by Section 5.5. We bound the l.h.s. of (38):

$$\begin{aligned}
&\left\|\sum_{m\in[M]} \sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right\|_{\Psi_\gamma} = \\
&= \left\|\sum_{m\in[M]} \left(\sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)| - \mathbb{E}\left[\sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right] + \mathbb{E}\left[\sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right]\right)\right\|_{\Psi_\gamma} \\
&\overset{(a)}{\leq} 2^{\frac{1}{\gamma}-1} \left(\left\|\sum_{m\in[M]} \left(\sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)| - \mathbb{E}\left[\sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right]\right)\right\|_{\Psi_\gamma} + \right. \\
&\qquad \left. + \left\|\mathbb{E}\left[\sum_{m\in[M]} \sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right]\right\|_{\Psi_\gamma}\right)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{\leq} 2^{\frac{1}{\gamma}-1}\left(C_\gamma\left(\mathbb{E}\left[\left|\sum_{m\in[M]}\left(\sup_{\mathbf{t}\in T}|\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)| - \mathbb{E}\left[\sup_{\mathbf{t}\in T}|\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right]\right)\right|\right]+\right.\right.\\
&\qquad\left.+\left\|\max_{m\in[M]}\left|\sup_{\mathbf{t}\in T}|\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)| - \mathbb{E}\left[\sup_{\mathbf{t}\in T}|\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right]\right|\right\|_{\Psi_\gamma}\right)+\\
&\qquad\left.+\frac{1}{\Psi_\gamma^{-1}(1)}\mathbb{E}\left[\sum_{m\in[M]}\sup_{\mathbf{t}\in T}|\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right]\right)\\
&=: 2^{\frac{1}{\gamma}-1}\left[C_\gamma(E_1+E_2)+\frac{1}{\Psi_\gamma^{-1}(1)}E_3\right]. \qquad (39)
\end{aligned}
$$

In (a) we applied the generalized triangle inequality Section 4(vi) and $\left(\frac{1}{\gamma}-1\right)_+ = \frac{1}{\gamma}-1$ as $\gamma = \frac{\alpha}{n} \in (0,1]$. In (b) the Talagrand inequality (44) was invoked with the $Y_m := \sup_{\mathbf{t}\in T}|\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)| - \mathbb{E}\left[\sup_{\mathbf{t}\in T}|\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right]$ centered variables and $B := \mathbb{R}$, followed by taking the γ-Orlicz norm of the constant $\lambda := \mathbb{E}\left[\sum_{m\in[M]}\sup_{\mathbf{t}\in T}|\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right]$ according to Section 4(ii).

We continue the derivation with bounding the E_1, E_2 and E_3 terms in (39).

- **Bounding** E_1:

$$
\begin{aligned}
E_1 &= \mathbb{E}\left[\left|\sum_{m\in[M]}\left(\sup_{\mathbf{t}\in T}|\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)| - \mathbb{E}\left[\sup_{\mathbf{t}\in T}|\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right]\right)\right|\right]\\
&\overset{(a)}{\leq} 2\mathbb{E}\left[\sum_{m\in[M]}\sup_{\mathbf{t}\in T}|\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right] \overset{(b)}{\leq} 16\mathbb{E}\left[\max_{m\in[M]}\sup_{\mathbf{t}\in T}|\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right]\\
&\overset{(c)}{\leq} 16\mathbb{E}\left[\max_{m\in[M]}\sup_{\mathbf{t}\in T}|f_{\mathbf{t}}(\mathbf{X}_m)|\right] \overset{(d)}{\leq} 16\left\|\max_{m\in[M]}\sup_{\mathbf{t}\in T}|f_{\mathbf{t}}(\mathbf{X}_m)|\right\|_{\Psi_\gamma^{(l)}}\left(\Psi_\gamma^{(l)}\right)^{-1}(1)\\
&\overset{(e)}{\leq} 16\left\|\max_{m\in[M]}\sup_{\mathbf{t}\in T}|f_{\mathbf{t}}(\mathbf{X}_m)|\right\|_{\Psi_\gamma}\left(\Psi_\gamma^{(l)}\right)^{-1}(1),
\end{aligned}
$$

where in (a) we used the triangle inequality, in (b) we applied the Hoffman-Jorgensen inequality (Theorem 6; $t_0 = 0$, $p = 1$, $B = \mathbb{R}$, $Y_m = \sup_{\mathbf{t}\in T}|\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|$) with

$$
\begin{aligned}
\mathbb{P}\left(\sum_{m\in[M]}\sup_{\mathbf{t}\in T}|\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)| > 0\right) &\overset{(f)}{=} \mathbb{P}\left(\max_{j\in[M]}\sum_{m\in[j]}\sup_{\mathbf{t}\in T}|\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)| > 0\right)\\
&= \mathbb{P}\left(\max_{m\in[M]}\sup_{\mathbf{t}\in T}|f_{\mathbf{t}}(\mathbf{X}_m)| > c\right)\\
&\overset{(g)}{\leq} \mathbb{P}\left(\max_{m\in[M]}\sup_{\mathbf{t}\in T}|f_{\mathbf{t}}(\mathbf{X}_m)| \geq c_{\texttt{HJ}}\right) \overset{(h)}{\leq} \frac{1}{8} = \frac{1}{2\times 4^p} \text{ with } p = 1.
\end{aligned}
$$

In (f) the non-negativity of Y_m was exploited; in (g) $c \geq c_{\texttt{HJ}}$ was used. We applied the Markov inequality and the definition of $c_{\texttt{HJ}}$ in (h). (c) holds by $|\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)| \leq |f_{\mathbf{t}}(\mathbf{X}_m)|$.

In (d) and (e) we applied Section 4(v) with the convex $\Psi_\gamma^{(l)}$ and the monotonicity property Section 4(iii) with $\Psi_\gamma^{(l)} \le \Psi_\gamma$, respectively.

- **Bounding** E_2:

$$\begin{aligned}
E_2 &= \left\| \max_{m\in[M]} \left| \sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)| - \mathbb{E}\left[\sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right] \right| \right\|_{\Psi_\gamma} \\
&\overset{(a)}{\le} \left\| \max_{m\in[M]} \left(\sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)| + \mathbb{E}\left[\sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right] \right) \right\|_{\Psi_\gamma} \\
&\overset{(b)}{\le} \left\| \max_{m\in[M]} \sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)| + \max_{m\in[M]} \mathbb{E}\left[\sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right] \right\|_{\Psi_\gamma} \\
&\overset{(c)}{\le} 2^{\frac{1}{\gamma}-1} \left(\left\| \max_{m\in[M]} \sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)| \right\|_{\Psi_\gamma} + \left\| \max_{m\in[M]} \mathbb{E}\left[\sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right] \right\|_{\Psi_\gamma} \right) \\
&\overset{(d)}{\le} 2^{\frac{1}{\gamma}-1} \left(\left\| \max_{m\in[M]} \sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)| \right\|_{\Psi_\gamma} + \frac{1}{\Psi_\gamma^{-1}(1)} \mathbb{E}\left[\max_{m\in[M]} \sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right] \right) \\
&\overset{(e)}{\le} 2^{\frac{1}{\gamma}-1} \left(\left\| \max_{m\in[M]} \sup_{\mathbf{t}\in T} |f_{\mathbf{t}}(\mathbf{X}_m)| \right\|_{\Psi_\gamma} + \frac{\left(\Psi_\gamma^{(l)}\right)^{-1}(1)}{\Psi_\gamma^{-1}(1)} \left\| \max_{m\in[M]} \sup_{\mathbf{t}\in T} |f_{\mathbf{t}}(\mathbf{X}_m)| \right\|_{\Psi_\gamma^{(l)}} \right) \\
&\overset{(f)}{\le} 2^{\frac{1}{\gamma}-1} \left(1 + \frac{\left(\Psi_\gamma^{(l)}\right)^{-1}(1)}{\Psi_\gamma^{-1}(1)} \right) \left\| \max_{m\in[M]} \sup_{\mathbf{t}\in T} |f_{\mathbf{t}}(\mathbf{X}_m)| \right\|_{\Psi_\gamma}.
\end{aligned}$$

In (a) we used the triangle inequality with the monotonicity Section 4(iv), (b) holds by the sub-additivity of the maximum and again the monotonicity Section 4(iv), in (c) we applied the generalized triangle inequality Section 4(vi) and that $\left(\frac{1}{\gamma}-1\right)_+ = \frac{1}{\gamma}-1$ as $\gamma \in (0,1]$, (d) holds by Section 4(ii) with the constant $\lambda = \mathbb{E}\left[\max_{m\in[M]} \sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right]$, (e) is by the monotonicity Section 4(iv) as $|\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)| \le |f_{\mathbf{t}}(\mathbf{X}_m)|$, and by Section 4(v), (f) follows from $\Psi_\gamma^{(l)} \le \Psi_\gamma$ combined with the monotonicity Section 4(iii).

- **Bounding** E_3: By (b)-(e) of the E_1 derivation we have that

$$E_3 = \mathbb{E}\left[\sum_{m\in[M]} \sup_{\mathbf{t}\in T} |\mathcal{R}_c f_{\mathbf{t}}(\mathbf{X}_m)|\right] \le 8 \left\| \max_{m\in[M]} \sup_{\mathbf{t}\in T} |f_{\mathbf{t}}(\mathbf{X}_m)| \right\|_{\Psi_\gamma} \left(\Psi_\gamma^{(l)}\right)^{-1}(1).$$

By adding the obtained E_1, E_2 and E_3 bounds, we get (38) with K_γ defined in Theorem 1.

5.7 Bounding the Driving Terms of Theorem 1 for RFF

We bound the constants of Theorem 1 in the RFF case described in Remark 2(v).

- **The term** B: It is defined in (9). Recalling the expression (21) for $I_\rho(|T|)$ and using the Cauchy-Schwarz inequality for bounding $\mathbb{E}\left[\|L\|_{L^2(\boldsymbol{\omega}_{1:M})}\right]$ by $\sqrt{\mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[L^2(\boldsymbol{\omega})\right]}$ gives

$$B \le 2C_D\sqrt{d}\frac{\sqrt{\mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[L^2(\boldsymbol{\omega})\right]}}{\sqrt{M}}|S_\Delta|^\beta \int_0^1 \sqrt{\log\left(1+2u^{-1/\beta}\right)}\mathrm{d}u.$$

We now aim at showing a tight bound for $|S_\Delta|^\beta \int_0^1 \sqrt{\log\left(1+2u^{-1/\beta}\right)}\mathrm{d}u$ w.r.t. $|S_\Delta|$ with an appropriate choice of $\beta = \beta(|S_\Delta|)$. Indeed, let $\beta = \frac{1}{1+(\log|S_\Delta|)_+} \in (0,1]$. We start by proving the bound

$$I_\beta := \int_0^1 \sqrt{\log\left(1+2u^{-1/\beta}\right)}\mathrm{d}u \le \frac{4}{\sqrt{\beta}}, \qquad \forall \beta \in (0,1]. \tag{40}$$

By the change of variable $t = \beta\log\left(1+2u^{-\frac{1}{\beta}}\right)$ (i.e. $u = \left(\frac{e^{t/\beta}-1}{2}\right)^{-\beta}$), we get

$$I_\beta = \frac{2^\beta}{\sqrt{\beta}} \int_{\beta\log(3)}^{\infty} \frac{\sqrt{t}\, e^{t/\beta}}{(e^{t/\beta}-1)^{\beta+1}}\mathrm{d}t = \frac{2^\beta}{\sqrt{\beta}} \int_{\beta\log(3)}^{\infty} \frac{\sqrt{t}}{e^t(1-e^{-t/\beta})^{\beta+1}}\mathrm{d}t.$$

Using the fact that $1-e^{-t/\beta} \ge \frac{2}{3}$ on $[\beta\log(3),+\infty)$, we arrive at

$$I_\beta \le \frac{3^{\beta+1}}{2\sqrt{\beta}} \int_{\beta\log(3)}^{\infty} \sqrt{t}e^{-t}\mathrm{d}t \overset{(*)}{\le} \frac{9}{2\sqrt{\beta}}\Gamma\left(\frac{3}{2}\right) \le \frac{4}{\sqrt{\beta}},$$

where the inequality $(*)$ is obtained by taking $\beta = 1$ in $3^{\beta+1}$ and $\beta = 0$ in the integral; hence (40) is proved. Now, using (40) with $\beta = \frac{1}{1+(\log|S_\Delta|)_+}$ and its $|S_\Delta|^\beta = e^{\frac{\log|S_\Delta|}{1+(\log|S_\Delta|)_+}}$ implication, we get

$$|S_\Delta|^\beta \int_0^1 \sqrt{\log\left(1+2u^{-1/\beta}\right)}\mathrm{d}u \le 4e^{\frac{\log|S_\Delta|}{1+(\log|S_\Delta|)_+}}\sqrt{1+(\log|S_\Delta|)_+} \le 4e\sqrt{1+(\log|S_\Delta|)_+},$$

and therefore

$$B \le \frac{8eC_D\sqrt{d}\sqrt{\mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[L^2(\boldsymbol{\omega})\right]}\sqrt{1+(\log|S_\Delta|)_+}}{\sqrt{M}}.$$

- **The term** σ^2: It is defined in Theorem 1. Since the variance is bounded by the second moment, $\mathbb{E}\left[g_{\mathbf{z}}^2(\boldsymbol{\omega}_m)\right] \le \mathbb{E}\left[f_{\mathbf{z}}^2(\boldsymbol{\omega}_m)\right]$. Furthermore, since $(\boldsymbol{\omega}_m)_{m=1}^M$ are i.i.d., the previous expectation can bounded by $\mathbb{E}\left[\|\boldsymbol{\omega}\|_2^{2|\mathbf{p}+\mathbf{q}|}\right]$ using (15). As a result, we get

$$\sigma^2 \le \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[\|\boldsymbol{\omega}\|_2^{2|\mathbf{p}+\mathbf{q}|}\right].$$

- **The term** $\max_{m\in[M]}\sup_{\mathbf{z}\in S_\Delta}\|g_{\mathbf{z}}(\boldsymbol{\omega}_m)\|_{\Psi_\gamma}$ **with** $\gamma = \alpha/n \le 1$ **and** $n = |\mathbf{p}+\mathbf{q}| + \beta$: It appears in the definition of c (in Theorem 1). In view of the bound (16) which is uniform in $\mathbf{z}$ and using property (iv) of Section 4, we get

$$\max_{m\in[M]}\sup_{\mathbf{z}\in S_\Delta}\|g_{\mathbf{z}}(\boldsymbol{\omega}_m)\|_{\Psi_\gamma} \le \left\|\|\boldsymbol{\omega}\|_2^{|\mathbf{p}+\mathbf{q}|} + \Lambda\left[\|\cdot\|_2^{|\mathbf{p}+\mathbf{q}|}\right]\right\|_{\Psi_{\alpha/n}}.$$

- **The term** $\left\|\max_{m\in[M]}\sup_{\mathbf{z}\in S_\Delta}|g_{\mathbf{z}}(\boldsymbol{\omega}_m)|\right\|_{\Psi_\gamma}$: It shows up in the exponential bound (10). We invoke the maximal inequality for the γ-Orlicz norm (item (viii) of Section 4) with the previous estimate to obtain

$$\left\|\max_{m\in[M]}\sup_{\mathbf{z}\in S_\Delta}|g_{\mathbf{z}}(\boldsymbol{\omega}_m)|\right\|_{\Psi_\gamma} \le \left[\frac{\log(1+M)}{\log(3/2)}\right]^{n/\alpha}\left\|\|\boldsymbol{\omega}\|_2^{|\mathbf{p}+\mathbf{q}|} + \Lambda\left[\|\cdot\|_2^{|\mathbf{p}+\mathbf{q}|}\right]\right\|_{\Psi_{\alpha/n}}.$$

- **The term** $\mathbb{E}\left[\max_{m\in[M]}\sup_{\mathbf{z}\in S_\Delta}|g_{\mathbf{z}}(\boldsymbol{\omega}_m)|\right]$: It appears in the definition of c. Using properties (iii) and (v) of Section 4 and the convexification of Ψ_γ, we directly get

$$\mathbb{E}\left[\max_{m\in[M]}\sup_{\mathbf{z}\in S_\Delta}|g_{\mathbf{z}}(\boldsymbol{\omega}_m)|\right] \le \left(\Psi^{(l)}_{\alpha/n}\right)^{-1}(1)\left[\frac{\log(1+M)}{\log(3/2)}\right]^{n/\alpha}\left\|\|\boldsymbol{\omega}\|_2^{|\mathbf{p}+\mathbf{q}|}+\Lambda\left[\|\cdot\|_2^{|\mathbf{p}+\mathbf{q}|}\right]\right\|_{\Psi_{\alpha/n}}.$$

Collecting the different bounds we obtain (12) by setting

$$\begin{aligned} C_{\mathtt{RFF}}(n) := \max\Bigg(& 8eC_D\sqrt{d}\sqrt{\mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[L^2(\boldsymbol{\omega})\right]}, \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[\|\boldsymbol{\omega}\|_2^{2|\mathbf{p}+\mathbf{q}|}\right], \\ & 1\vee\left(\left(\Psi^{(l)}_{\alpha/n}\right)^{-1}(1)\left[\log(3/2)\right]^{-n/\alpha}\right)\left\|\|\boldsymbol{\omega}\|_2^{|\mathbf{p}+\mathbf{q}|}+\Lambda\left[\|\cdot\|_2^{|\mathbf{p}+\mathbf{q}|}\right]\right\|_{\Psi_{\alpha/n}}\Bigg). \end{aligned} \tag{41}$$

Long (but standard) computations show that $C_{\mathtt{RFF}}(n)$ is uniformly bounded for $n\in[|\mathbf{p}+\mathbf{q}|,|\mathbf{p}+\mathbf{q}|+1]$, and thus we can set $C_{\mathtt{RFF}} := \sup_{n\in[|\mathbf{p}+\mathbf{q}|,|\mathbf{p}+\mathbf{q}|+1]} C_{\mathtt{RFF}}(n)$.

5.8 Proofs of Corollaries 3 and 4

Corollary 3 is a direct consequence of Theorem 1 combined with Remark 2(v), in particular because $C_{\mathtt{RFF}}$ does not depend on S_Δ and M, and K_γ can be bounded uniformly in $n\in[|\mathbf{p}+\mathbf{q}|,|\mathbf{p}+\mathbf{q}|+1]$. The Talagrand constant C_γ is uniformly bounded w.r.t. γ provided that γ is bounded away from 0, see the proof of (Talagrand, 1989, Theorem 3).

Now let us prove Corollary 4; set $\varepsilon_M = \frac{\left(\sqrt{6\tilde{C}}\vee\tilde{C}\right)\sqrt{\left(1+[\log|(S_M)_\Delta|]_+\right)\vee\log(1+M)}}{\sqrt{M}}$. Observe that

(i) ε_M satisfies the lower bound requirement on ε in Corollary 3;

(ii) by assumption $\varepsilon_M\to 0$ as $M\to 0$ by using that $|S_\Delta|\le 2|S|$;

(iii) therefore

$$\begin{aligned} 1+\varepsilon_M\left[\log\left(\frac{\tilde{C}}{\varepsilon_M}\right)\vee\log(1+M)\right]^{1/\gamma} &\le 1+\varepsilon_M\left(\left[\log\left(\frac{\tilde{C}}{\varepsilon_M}\right)\right]^{1/\gamma}+\left[\log(1+M)\right]^{1/\gamma}\right) \\ &\le 2+\varepsilon_M\left[\log(1+M)\right]^{1/\gamma} \end{aligned}$$

for M large enough;

(iv) $\varepsilon_M \ge \frac{\left(\sqrt{6\tilde{C}}\vee\tilde{C}\right)\sqrt{\log(1+M)}}{\sqrt{M}}$.

As a consequence of (iii) and (iv), setting $\delta_M := \frac{1+(\log|(S_M)_\Delta|)_+}{\log(1+M)}\vee 1$, we get (for M large enough)

$$\frac{M\,\varepsilon_M^2}{\tilde{C}\left(1+\varepsilon_M\left[\log\left(\frac{\tilde{C}}{\varepsilon_M}\right)\vee\log(1+M)\right]^{1/\gamma}\right)} \ge \frac{6\tilde{C}\log(1+M)\delta_M}{\tilde{C}\left[2+\frac{\left(\sqrt{6\tilde{C}}\vee\tilde{C}\right)\sqrt{\delta_M}\left[\log(1+M)\right]^{1/2+1/\gamma}}{\sqrt{M}}\right]}$$

$$= 6\log(1+M)\ \frac{\delta_M}{2+z_M\sqrt{\delta_M}},$$

where $z_M = \frac{\left(\sqrt{6\tilde{C}}\vee\tilde{C}\right)[\log(1+M)]^{1/2+1/\gamma}}{\sqrt{M}} \xrightarrow{M\to\infty} 0$. Since the function $\delta\in\mathbb{R}^+\mapsto\frac{\delta}{2+z_M\sqrt{\delta}}$ is increasing and $\delta_M\geq 1$, we get (for M large enough)

$$\frac{M\,\varepsilon_M^2}{\tilde{C}\left(1+\varepsilon_M\left[\log\left(\frac{\tilde{C}}{\varepsilon_M}\right)\vee\log(1+M)\right]^{1/\gamma}\right)}\geq 6\log(1+M)\frac{1}{3}.$$

On the other hand, using (iv), we easily get $\frac{(M\varepsilon_M)^\gamma}{\tilde{C}\log(1+M)}\geq\frac{\left[\left(\sqrt{6\tilde{C}}\vee\tilde{C}\right)\sqrt{\log(1+M)}\sqrt{M}\right]^\gamma}{\tilde{C}\log(1+M)}\geq 2\log(1+M)$ for M large enough.
To sum up, in view of (13), we have proved (still for large enough M)

$$\Lambda^M\left(\left\|\widehat{\partial^{\mathbf{p},\mathbf{q}}k}-\partial^{\mathbf{p},\mathbf{q}}k\right\|_{S_M}\geq\epsilon_M\right)\leq\frac{2}{(1+M)^2}+\frac{1}{(1+M)^2}$$

and by the Borell-Cantelli lemma, we conclude to the a.s. convergence (14).

5.9 Proof of Remark 5(ii)

$\omega_i\in L_{\Psi_{\alpha_i}}$ means that $\mathbb{E}_{\omega_i\sim\Lambda_i}\left[e^{s_i|\omega_i|^{\alpha_i}}\right]<\infty$ for some $s_i\in\mathbb{R}^+$ ($\forall i\in[d]$). Let $\alpha=\min_{i\in[d]}\alpha_i$ and $\|\boldsymbol{\omega}\|_\alpha:=\left(\sum_{i\in[d]}|\omega_i|^\alpha\right)^{\frac{1}{\alpha}}$. Then $\|\boldsymbol{\omega}\|_2\leq\sqrt{d}\sup_{i\in[d]}|\omega_i|\leq\sqrt{d}\,\|\boldsymbol{\omega}\|_\alpha$ ($\forall\boldsymbol{\omega}\in\mathbb{R}^d$). Notice that $|\omega_i|^\alpha\leq|\omega_i|^{\alpha_i}$ if $|\omega_i|\geq 1$ and $|\omega_i|^\alpha\leq 1$ otherwise, i.e. we have $|\omega_i|^\alpha\leq|\omega_i|^{\alpha_i}+1$ for any $\omega_i\in\mathbb{R}$. This means that taking $s=\min_{i\in[d]}s_i$ and $\tilde{s}:=\frac{s}{d^{\alpha/2}}>0$ gives

$$\|\boldsymbol{\omega}\|_2^\alpha\leq d^{\alpha/2}\,\|\boldsymbol{\omega}\|_\alpha^\alpha=d^{\alpha/2}\sum_{i\in[d]}|\omega_i|^\alpha\leq d^{\alpha/2}\sum_{i\in[d]}(|\omega_i|^{\alpha_i}+1),$$

$$\mathbb{E}_{\boldsymbol{\omega}}\left[e^{\tilde{s}\|\boldsymbol{\omega}\|_2^\alpha}\right]\leq\mathbb{E}_{\boldsymbol{\omega}}\left[e^{s\sum_{i\in[d]}(|\omega_i|^{\alpha_i}+1)}\right]$$

$$\overset{(*)}{=}\prod_{i\in[d]}\left(\mathbb{E}_{\omega_i}\left[e^{s|\omega_i|^{\alpha_i}}\right]e^s\right)\leq\prod_{i\in[d]}\left(\mathbb{E}_{\omega_i}\left[e^{s_i|\omega_i|^{\alpha_i}}\right]e^s\right)<\infty,$$

where we used the independence of ω_i-s in $(*)$. We got that $\mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[e^{\tilde{s}\|\boldsymbol{\omega}\|_2^\alpha}\right]<\infty$ which implies that $\boldsymbol{\omega}\in L_{\Psi_\alpha}$.

5.10 Proof of the Properties in Section 4 about the Orlicz norm

- **Properties (i)-(iv)**: These properties are well-known and directly follow from the definition of the Orlicz norm.
- **Property (v)**: The case $\|X\|_\Psi=0$ gives a trivial inequality and can be discarded. Since Ψ is bounded from below by an increasing affine function, $X\in L_\Psi$ implies that X is integrable. Combining (i) with Jensen's inequality gives $\Psi\left(\frac{\mathbb{E}[\|X\|_2]}{\|X\|_\Psi}\right)\leq\mathbb{E}\left[\Psi\left(\frac{\|X\|_2}{\|X\|_\Psi}\right)\right]\leq 1$, and the result follows.

- **Property (vi)**: It is well-known that the usual triangle inequality holds for $\alpha \geq 1$. We now focus on the case $\alpha \in (0,1]$. Set $c := \left(\|X\|_{\Psi_\alpha}^\alpha + \|X'\|_{\Psi_\alpha}^\alpha\right)^{1/\alpha}$, $p := \frac{c^\alpha}{\|X\|_{\Psi_\alpha}^\alpha}$ and $q := \frac{c^\alpha}{\|X'\|_{\Psi_\alpha}^\alpha}$, and notice that $\frac{1}{p} + \frac{1}{q} = 1$. Then, combining (23) with $\gamma = \alpha \in (0,1)$ and the Hölder inequality with the conjugate exponents (p,q) yields

$$\begin{aligned}
\limsup_{m\to+\infty} \mathbb{E}\left[e^{\left(\frac{m\wedge\|X+X'\|_2}{c}\right)^\alpha}\right] &\leq \limsup_{m\to+\infty} \mathbb{E}\left[e^{\left(\frac{m\wedge\|X\|_2}{c}\right)^\alpha} \, e^{\left(\frac{m\wedge\|X'\|_2}{c}\right)^\alpha}\right] \\
&\leq \left(\limsup_{m\to+\infty} \mathbb{E}\left[e^{m\wedge\frac{p\|X\|_2^\alpha}{c^\alpha}}\right]\right)^{1/p} \left(\limsup_{m\to+\infty} \mathbb{E}\left[e^{m\wedge\frac{q\|X'\|_2^\alpha}{c^\alpha}}\right]\right)^{1/q} \\
&\overset{\text{item (i)}}{\leq} 2^{1/p}2^{1/q} = 2.
\end{aligned}$$

  Therefore, $X + X' \in L_{\Psi_\alpha}$ and $\|X + X'\|_{\Psi_\alpha} \leq c$ by the definition of the α-Orlicz norm. Applying (23) with $\gamma = 1/\alpha$ we get $c \leq 2^{\left(\frac{1}{\alpha}-1\right)_+}\left(\|X\|_{\Psi_\alpha} + \|X'\|_{\Psi_\alpha}\right)$ and hence the claimed result is proved.
- **Property (vii)**: This is a direct consequence of the Markov inequality.
- **Property (viii)**: A similar statement appears in (van der Vaart and Wellner, 1996, Lemma 2.2.2), but under the assumption that Ψ_α is convex (which holds only if $\alpha \geq 1$) and without explicit constant. Our statement is valid for any $\alpha > 0$ with explicit control.
  – **A first inequality**: Let $\alpha \in \mathbb{R}^+$. We claim that for any $x_0 > 0$ and any $x, y \geq 1$, we have

$$\Psi_\alpha\left(x_0^{1/\alpha}x\right)\ \Psi_\alpha\left(x_0^{1/\alpha}y\right) \leq \Psi_\alpha\left(x_0^{1/\alpha}\right)\ \Psi_\alpha\left(x_0^{1/\alpha}xy\right). \tag{42}$$

  Because $\Psi_\alpha(x) = \Psi_1(x^\alpha)$ where $\Psi_1(x) =: \Psi(x) = e^x - 1$, the inequality for $\alpha = 1$ clearly implies those for all $\alpha > 0$. To prove the inequality for $\alpha = 1$, let x_0 and x be fixed, and set $H(y) = \Psi(x_0)\Psi(x_0xy) - \Psi(x_0x)\ \Psi(x_0y)$. One has

$$\begin{aligned}
H'(y) &= x_0x\Psi(x_0)e^{x_0xy} - x_0\Psi(x_0x)e^{x_0y} \\
&= x_0^2xe^{x_0}e^{x_0x}\left[\frac{\Psi(x_0)}{x_0e^{x_0}}e^{x_0x(y-1)} - \frac{\Psi(x_0x)}{x_0xe^{x_0x}}e^{x_0(y-1)}\right], \\
\frac{\Psi(x_0)}{x_0e^{x_0}} &= \frac{1-e^{-x_0}}{x_0} = \int_0^1 e^{-ux_0}\mathrm{d}u \geq \int_0^1 e^{-ux_0x}\mathrm{d}u = \frac{\Psi(x_0x)}{x_0xe^{x_0x}}, \\
e^{x_0x(y-1)} &\geq e^{x_0(y-1)},
\end{aligned}$$

  where we used $x_0 > 0$, $x, y \geq 1$ at the two last inequalities. This shows that $H'(y) \geq 0$, and since $H(1) = 0$ we have $H(y) \geq 0$ for any $y \geq 1$. Consequently, (42) is proved.

  – **Final maximal inequality**: We follow the arguments of (van der Vaart and Wellner, 1996, Lemma 2.2.2) with slights modifications. The inequality (42) can be rewritten as

$$\Psi_\alpha(x) \leq \Psi_\alpha\left(x_0^{1/\alpha}\right)\ \Psi_\alpha\left(xy/x_0^{1/\alpha}\right)/\Psi_\alpha(y), \qquad \forall x, y \geq x_0^{1/\alpha}. \tag{43}$$

  Set $c = \max_{m\in[M]} \|X_m\|_{\Psi_\alpha}/x_0^{1/\alpha}$ and let $y \geq x_0^{1/\alpha}$.

- $*$ If $\frac{\max_{m\in[M]}\|X_m\|_2}{cy} \le x_0^{1/\alpha}$, then we have the crude bound $\Psi_\alpha\left(\frac{\max_{m\in[M]}\|X_m\|_2}{cy}\right) \le \Psi_\alpha\left(x_0^{1/\alpha}\right) = \Psi(x_0)$.
- $*$ If $\frac{\max_{m\in[M]}\|X_m\|_2}{cy} \ge x_0^{1/\alpha}$, then (43) yields

$$\begin{aligned}\Psi_\alpha\left(\frac{\max_{m\in[M]}\|X_m\|_2}{cy}\right) &\le \Psi(x_0)\ \Psi_\alpha\left(\frac{\max_{m\in[M]}\|X_m\|_2}{\max_{m\in[M]}\|X_m\|_{\Psi_\alpha}}\right)/\Psi_\alpha(y)\\ &\le \sum_{m\in[M]}\Psi(x_0)\Psi_\alpha\left(\|X_m\|_2/\|X_m\|_{\Psi_\alpha}\right)/\Psi_\alpha(y).\end{aligned}$$

Consequently, in both cases we have

$$\Psi_\alpha\left(\frac{\max_{m\in[M]}\|X_m\|_2}{cy}\right) \le \Psi(x_0)\left[1+\sum_{m\in[M]}\Psi_\alpha\left(\|X_m\|_2/\|X_m\|_{\Psi_\alpha}\right)/\Psi_\alpha(y)\right].$$

Taking expectation and using Property (i), we arrive at $\mathbb{E}\left[\Psi_\alpha\left(\frac{\max_{m\in[M]}\|X_m\|_2}{cy}\right)\right] \le \Psi(x_0)\left[1+\frac{M}{\Psi_\alpha(y)}\right]$. Let us choose x_0 such that $\Psi(x_0) < 1$. In this case the choice $y = x_0^{1/\alpha} \vee \Psi_\alpha^{-1}\left(\frac{M}{1/\Psi(x_0)-1}\right)$ ensures that the above bound is valid and smaller than 1. Consequently, by the definition of α-Orlicz norm, we get $\left\|\max_{m\in[M]}\|X_m\|_2\right\|_{\Psi_\alpha} \le cy$. The choice $x_0 = \log(3/2)$ satifies the previous requirement: $\Psi(x_0) = \frac{1}{2} < 1$. In this case $y = [\log(3/2) \vee \log(1+M)]^{1/\alpha} = [\log(1+M)]^{1/\alpha}$ since $M \ge 1$. We have obtained the claimed Property (viii).

5.11 External Statements

In this subsection we state external statements which were used to derive our results. Below B stands for a separable Banach space, $L_p(B)$ is the space of B-valued p-integrable functions. The norm $\|\cdot\|_{\Psi_\alpha}$ is defined analogously to $\mathbb{R}^d$ by changing $\|\cdot\|_2$ to $\|\cdot\|_B$.

Theorem 6 ***(Hoffman-Jorgensen inequality, Ledoux and Talagrand (2013), Proposition 6.8)*** *Let $p > 0$, $M \in \mathbb{Z}^+$, $(Y_m)_{m\in[M]}$ be independent random variables in $L_p(B)$, $S_m := \sum_{j=1}^m Y_j$ for $m \in [M]$, $t_0 = \inf\left\{t > 0 : \mathbb{P}\left(\max_{1\le m\le M}\|S_m\|_B > t\right) \le (2\times 4^p)^{-1}\right\}$. Then*

$$\mathbb{E}\left[\max_{m\in[M]}\|S_m\|_B^p\right] \le 2\times 4^p\mathbb{E}\left[\max_{m\in[M]}\|Y_m\|_B^p\right] + 2(4t_0)^p.$$

Theorem 7 ***(Talagrand, 1989, Theorem 3)*** *Let $\gamma \in (0,1]$. Then, there is a constant C_γ such that for all finite sequence $(Y_m)_{m\in[M]}$ of independent, mean zero, integrable random variables in $L_{\Psi_\gamma}(B)$, we have*

$$\left\|\sum_{m\in[M]}Y_m\right\|_{\Psi_\gamma} \le C_\gamma\left(\left\|\sum_{m\in[M]}Y_m\right\|_{L_1(B)} + \left\|\max_{m\in[M]}\|Y_m\|_B\right\|_{\Psi_\gamma}\right). \tag{44}$$

Theorem 8 ***(Klein-Rio inequality for supremum of empirical process - (Klein and Rio, 2005, Theorems 1.1-1.2))*** *Let $M \in \mathbb{Z}^+$, $c \in \mathbb{R}^+$, $(X_m)_{m\in[M]}$ be independent B-valued random variables, and $\mathcal{F}$ a countable set of $\mathbf{f} := (f_1, \ldots, f_M)$ measurable functions from B into $[-c,c]^M$ such that $\mathbb{E}\left[f_m(X_m)\right] = 0$ for all $m \in [M]$. Define $Z := \sup_{\mathbf{f}\in\mathcal{F}} \frac{1}{M}\sum_{m\in[M]} f_m(X_m)$, $\sigma^2 := \frac{1}{M}\sup_{\mathbf{f}\in\mathcal{F}} \mathbb{E}\left[\sum_{m\in[M]} f_m^2(X_m)\right]$. Then, for any $t \geq 0$ the following right and left-hand sided deviation inequalities hold*

$$\mathbb{P}\left(Z - \mathbb{E}\left[Z\right] \geq t\right) \leq e^{-\frac{M\,t^2}{2(\sigma^2+2c\;\mathbb{E}[Z])+3c\,t}}, \qquad \mathbb{P}\left(Z - \mathbb{E}\left[Z\right] \leq -t\right) \leq e^{-\frac{M\,t^2}{2(\sigma^2+2c\;\mathbb{E}[Z])+2c\,t}}.$$

Theorem 9 ***(Dudley entropy integral bound)***[9] *Let $\{Z_t : t \in T\}$ be a zero-mean separable stochastic process that is sub-Gaussian w.r.t. a pseudo-metric d on the indexing set T, in other words for every $\lambda \in \mathbb{R}$ $\mathbb{E}\left[e^{\lambda(Z_t - Z_s)}\right] \leq e^{\frac{\lambda^2 d(s,t)^2}{2}}$ $(\forall s, t \in T)$. Then there exists a universal constant C_D such that*

$$\mathbb{E}\left[\sup_{t\in T} Z_t\right] \leq C_D \int_0^\infty \sqrt{\log N(\varepsilon, d, T)}\mathrm{d}\epsilon, \tag{45}$$

where $N(\varepsilon, d, T)$ denotes the covering number.

Theorem 10 ***(Alzer (1997, Theorem 1))***[10] *Let $\gamma \in (0,1]$, $\beta_\gamma := \Gamma\left(1+\frac{1}{\gamma}\right)^{-\gamma}$, $x \in \mathbb{R}^{\geq 0}$, $I_\gamma(x) := \int_0^x e^{-t^\gamma}\mathrm{d}t$. Then $\left(1 - e^{-\beta_\gamma x^\gamma}\right)^{\frac{1}{\gamma}} \leq \frac{I_\gamma(x)}{\Gamma(1+1/\gamma)} \leq \left(1 - e^{-x^\gamma}\right)^{\frac{1}{\gamma}}$.*

Acknowledgments

This research is supported by the *Chair Stress Test, RISK Management and Financial Steering of the Foundation Ecole Polytechnique* and by the *Association Nationale de la Recherche Technique.* This work is part of the first author's doctoral thesis, under the supervision of the two other authors.

The authors wish to thank Radoslaw Adamczak for his detailed comments on his work.

References

Radoslaw Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2008.

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

[9] See (van der Vaart and Wellner, 1996, Corollary 2.2.8) for a general statement. Regarding the numerical value of C_D, van Handel (2016, Corollary 5.25) proves that one can take $C_D = 12$ whereas Bartlett (2013, Lecture 14) suggests a slightly smaller constant $C_D = 8\sqrt{2}$.

[10] The statement here follows by taking the limit of the cited result at $\gamma = 1$ and $x = 0$.

Horst Alzer. On some inequalities for the incomplete Gamma function. *Mathematics of Computation of the American Mathematical Society*, 66(218):771–778, 1997.

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

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18:1–38, 2017.

Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

Ole E. Barndorff-Nielsen, Thomas Mikosch, and Sidney I. Resnick. *Lévy Processess – Theory and Applications*. Springer Science+Business Media, 2001.

Peter Bartlett. Theoretical statistics. Available at `https://www.stat.berkeley.edu/~bartlett/courses/2013spring-stat210b/`, 2013.

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

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

Roxana Ciumara and Vasile Preda. The Weibull-logarithmic distribution in lifetime analysis and its properties. In *International Conference Applied Stochastic Models and Data Analysis (ASMDA)*, pages 395–399. 2009.

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

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

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

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

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

René Duijkers, Roland Tóth, Dario Piga, and Vincent Laurain. Shrinking complexity of scheduling dependencies in LS-SVM based LPV system identification. In *IEEE Conference on Decision and Control*, pages 2561–2566, 2014.

Anna Gilbert, Ambuj Tewari, and Yitong Sung. But how does it work in theory? Linear SVM with random features. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3379–3388, 2018.

Wittawat Jitkrittum, Arthur Gretton, Nicolas Heess, Ali Eslami, Balaji Lakshminarayanan, Dino Sejdinovic, and Zoltán Szabó. Kernel-based just-in-time learning for passing expectation propagation messages. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 405–414, 2015.

Nicolas Keriven, Anthony Bourrier, Rémi Gribonval, and Patrick Pérez. Sketching for large-scale learning of mixture models. *Information and Inference: A Journal of the IMA*, 7: 447–508, 2018.

Thierry Klein and Emmanuel Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005.

Fabien Lauer, Van Luong Le, and Gérard Bloch. Learning smooth models of nonsmooth functions via convex optimization. In *International Workshop on Machine Learning for Signal Processing (IEEE-MLSP)*, 2012.

Quoc Le, Tamás Sarlós, and Alexander Smola. Fastfood - computing Hilbert space expansions in loglinear time. In *International Conference on Machine Learning (ICML)*, pages 244–252, 2013.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes.* Springer, 2013.

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

Zhu Li, Jean-François Ton, Dino Oglic, and Dino Sejdinovic. A unified analysis of random Fourier features. In *International Conference on Machine Learning (ICML)*, pages 3905–3914, 2019.

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

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

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

Tibor K. Pogány and Saralees Nadarajah. On the characteristic function of the generalized normal distribution. *Comptes Rendus Mathematique*, 348(3-4):203–206, 2010.

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

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

Lorenzo Rosasco, Matteo Santoro, Sofia Mosci, Alessandro Verri, and Silvia Villa. A regularization approach to nonlinear variable selection. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 653–660, 2010.

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

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

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

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

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

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

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

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57):1–59, 2017.

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

Heiko Strathmann, Dino Sejdinovic, Samuel Livingstone, Zoltán Szabó, and Arthur Gretton. Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families. In *Advances in Neural Information Processing Systems (NIPS)*, pages 955–963, 2015.

Eric V. Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1), 2019. (`https://doi.org/10.1515/jci-2018-0017`).

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

Zoltán Szabó and Bharath K. Sriperumbudur. On kernel derivative approximation with random Fourier features. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 827–836, 2019.

Michel Talagrand. Isoperimetry and integrability of the sum of independent Banach-space valued random variables. *The Annals of Probability*, 17(4):1546–1570, 1989.

John Shawe Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.

Enayat Ullah, Poorya Mianjy, Teodor V. Marinov, and Raman Arora. Streaming kernel PCA with $\tilde{O}(\sqrt{n})$ random features. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7311–7321, 2018.

Sara van de Geer. *Empirical Processes in M-estimation.* Cambridge University Press, 2000.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical processes: with Applications to Statistics.* Springer-Verlag, New-York, 1996.

Ramon van Handel. Probability in high dimension. Available at `https://web.math.princeton.edu/~rvan/APC550.pdf`, 2016.

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

Jiyan Yang, Vikas Sindhwani, Haim Avron, and Michael W. Mahoney. Quasi-Monte Carlo feature maps for shift-invariant kernels. *International Conference on Machine Learning (ICML)*, pages 485–493, 2014.

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

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

Felix X Yu, Ananda Theertha Suresh, Krzysztof M Choromanski, Daniel N Holtmann-Rice, and Sanjiv Kumar. Orthogonal random features. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1975–1983, 2016.

Jian Zhang, Avner May, Tri Dao, and Christopher Ré. Low-precision random Fourier features for memory-constrained kernel approximation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1264–1274, 2019.

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

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