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 $\alpha$-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 $\alpha$-exponential Orlicz spectrum and any order of derivative.

Keywords: random Fourier features, kernel derivative, polynomial-growth functions, $\alpha$-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 $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ kernel. Kernels extend the classical notion of inner product on $\mathcal{X} = \mathbb{R}^d$ by assuming the existence of a $\phi : \mathcal{X} \rightarrow \mathcal{H}$ 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 \( \{x_n\}_{n=1}^N \) samples and is facing with an optimization problem expressed in terms of function values and derivatives\(^1\)

\[
\min_{f \in \mathcal{H}_k} l \left( \{\partial^p f(x_n)\}_{n \in [N]}, \|f\|_{\mathcal{H}_k}^2 \right),
\]

where \([N] = \{1, \ldots, N\}\), \(\partial^p f(x_n) := \frac{\partial^{p_1 + \cdots + p_d} f(x_n)}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}}\), \(D_n \subset \mathbb{N}^d\), \(N := \{0, 1, \ldots\}\) and the RKHS \(\mathcal{H}_k\) is characterized by \(f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k}\) (\(\forall x \in \mathcal{X}\), \(\forall f \in \mathcal{H}_k\)) and \(k(\cdot, x) \in \mathcal{H}_k\) (\(\forall x \in \mathcal{X}\))\(^2\). 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: \(k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}_k}\) where \(\phi(x) = k(\cdot, x) \in \mathcal{H}_k\).

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

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

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

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

where \(f'(x) = [\partial^{p_1} f(x); \ldots; \partial^{p_d} f(x)] \in \mathbb{R}^d\) is the derivative of \(f\), \(e_j \in \mathbb{R}^d\) is the \(j\)th canonical basis vector, \(\|\cdot\|_2\) is the Euclidean norm and \(D_n = \{0, \{e_j\}_{j=1}^d\} \quad (n \in [N])\). 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 \(\partial^p f(x) = \langle f, \partial^p k(\cdot, x) \rangle_{\mathcal{H}_k}\) derivative-reproducing property of kernels and the representer theorem (Zhou, 2008) guarantee that the solution of (1) has a finite-dimensional parameterization \(f(\cdot) = \sum_{n \in [N]} \sum_{p \in D_n} a_{n,p} \partial^p k(\cdot, x_n)\) \((a_{n,p} \in \mathbb{R})\) and it is sufficient to solve

\[
\min_{a} l \left( \sum_{n \in [N]} \sum_{p \in D_n} \sum_{q \in D_m} a_{m,q} \partial^p q k(x_n, x_m) \right), \quad \sum_{n \in [N]} \sum_{p \in D_n \cup D_m} a_{n,p} \|a_{m,q}\|_{\mathcal{H}_k}^2 k(x_n, x_m)
\]

\(^1\) 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 $\partial^{p,q}k(x,y) := \frac{\partial^{\sum_{i=1}^{d}(p_i+q_i)}k(x,y)}{\partial x^1 \partial y^1 \cdots \partial x^d \partial y^d}$ kernel derivatives; $a = (a_n,p)_{n\in[N],p\in D_n} \in \mathbb{R}^{\sum_{n\in[N]} |D_n|}$ 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 $k: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ can be written as the Fourier transform of a (finite) measure $\Lambda$, called the spectral measure

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

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

$$k(x,x') \approx \langle \lambda(x), \lambda(x') \rangle_{\mathbb{R}^{2M}}, \quad \hat{f}_w(x) = \langle w, \lambda(x) \rangle_{\mathbb{R}^{2M}},$$ (4)

where the integral representation (3) with respect to the measure $\Lambda$ is replaced by an average over random points; hence the random Fourier feature naming. As a result, one can estimate $w$ 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

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

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** ($p = q = 0$): 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 $tc^4 \|k - \hat{k}\|_{S_M} = O_{\text{a.s.}} \left( \frac{\sqrt{\log |S_M| \sqrt{\log M}}}{\sqrt{M}} \right)$ from $\|k - \hat{k}\|_{S_M} = O_p \left( \frac{|S_M| \sqrt{\log M}}{\sqrt{M}} \right)$, 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 $M = o(N) = O \left( \sqrt{N \log N} \right)$ random Fourier features is sufficient to get $O \left( 1/\sqrt{N} \right)$ generalization error. Under additional $\gamma$-capacity ($\gamma \in [0,1]$) and $r$-range space conditions ($r \geq \frac{1}{2}$), the same authors showed that even faster, minimax optimal $O \left( N^{-2 + \frac{2}{3} + \gamma} \right)$ rates are achievable with $M = o(N) = O \left( N^{1 + \frac{2(2r-1)}{2r+2} \log N} \right)$ RFFs. The result improves the originally proved (Rahimi and Recht, 2008) guarantee holding under the pessimistic $M = O(N)$ 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 $M = o(N) = \tilde{O} \left( N^{\frac{2}{2r+2}} \right)$ optimized RFF features—optimized in the sense of Bach (2017)—are sufficient to achieve a learning rate of $\tilde{O} \left( N^{-\frac{c}{2r+2}} \right)$ 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 $\tilde{O} \left( N^{-1} \right)$ with $M = \tilde{O} \left( \ln^d (N) \right)$ 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 $M = O(N^{2/3})$ (polynomial decay) or $M = O(\sqrt{N})$ (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

$$\| \partial^p q_k - \partial^p q_k \|_{S_M} = O_{\text{a.s.}} \left( \frac{\sqrt{\log |S_M| \sqrt{\log M}}}{\sqrt{M}} \right)$$

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


Table 1: Summary of RFF guarantees on kernel values and derivatives. Last line: it includes any measure $\Lambda$ with a finite $\alpha$-exponential moment (for some $\alpha, c > 0$, $\mathbb{E}_{\omega \sim \Lambda} \left(e^{c \|\omega\|_2^\alpha}\right) < +\infty$), 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 $p, q$ or $\Lambda$) are now available for any $p, q$ and any spectral measure $\Lambda$ with $\alpha$-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):

$$
\partial^{p,q}k(x,y) = \int_{\mathbb{R}^d} \left( \prod_{j=1}^d \omega_j^{p_j}(-\omega_j)^{q_j} \right) c_{(\sum_{i=1}^d |p_i+q_i|=n)} \left( \omega^\top (x-y) \right) d\Lambda(\omega),
$$

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

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


function class which controls the error \( \| \partial^{p,q,k} - \partial^{p,q,k} \|_{S} \). We tackle this challenge by imposing the finiteness of the \( \alpha \)-exponential Orlicz norm of the spectral measure (\( \Lambda \)) associated to the kernel, in other words

\[
\exists \alpha > 0, \ c > 0 \ \text{such that} \ E_{\omega \sim \Lambda} \left( e^{c \| \omega \|_{2}^{\alpha}} \right) < +\infty. \tag{5}
\]

Kernels with \( \alpha \)-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 \( X_m \). The distribution of the latter is assumed to have finite \( \alpha \)-exponential Orlicz norm and consequently, the random variables \( f(X_m) \) belong to a \( \gamma \)-exponential Orlicz space with index \( \gamma \) 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 \( \gamma \)-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 \( N = \{0, 1, \ldots\} \), \( \mathbb{R} \), \( \mathbb{C} \), \( \mathbb{Z}^+ = \{1, 2, \ldots\} \), \( \mathbb{R}^+ = (0, \infty) \), \( \mathbb{R}^{\geq} = [0, \infty) \) and \( \mathbb{Z}^{\leq} = \{0, -1, -2, \ldots\} \), 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 \( A \setminus B = \{a \in A : a \notin B \} \). 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 \( \Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx \) for \( t \in \mathbb{C}\setminus\mathbb{Z}_{\leq 0} \). For \( \gamma \in (0, 1] \) and \( x \in \mathbb{R}^{\geq 0} \), let \( I_\gamma(x) = \int_0^x e^{-t^\gamma}dt \) 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 \( J_\nu(z) = \sum_{n\in\mathbb{Z}^+} \frac{z^n}{n!} \) and \( K_\nu(z) = \frac{\pi}{2} \frac{J_{\nu-\frac{1}{2}}(z) - J_{\nu+\frac{1}{2}}(z)}{\sin(\pi \nu)} \) for \( z \in \mathbb{R} \) and non-integer \( \nu \); when \( \nu \) is an integer the limit is taken. The notation \log \ stands for the natural logarithm. The polylogarithm function is \( \text{Li}_n(z) = \sum_{k=1}^{\infty} \frac{z^k}{k^n} \) where \( s, \ z \in \mathbb{R} \) 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 \( \text{sech}(x) = \frac{1}{\cosh(x)} \), respectively. The (imaginary) error function is \( \text{erf}(z) = \sum_{n\in\mathbb{N}} \frac{2(-1)^n}{\sqrt{\pi} n!(2n+1)} z^{2n+1} \) where \( z \in \mathbb{R} \). For \( n \in \mathbb{N} \) let \( a^n = \frac{\Gamma(n+a)}{\Gamma(n)} \) where \( a \in \mathbb{C} \setminus \mathbb{Z}_{\leq 0} \) and \( a + n \in \mathbb{C} \setminus \mathbb{Z}_{\leq 0} \).

The ordinary hyperbolic function is \( \cos(x) = \frac{e^{ix} + e^{-ix}}{2}, \sin(x) = \frac{e^{ix} - e^{-ix}}{2i} \), the rising factorial of \( a \) defined as \( a^n = \frac{\Gamma(a+n)}{\Gamma(a)} \) where \( a \in \mathbb{C}, \ b \in \mathbb{C}, \ c \in \mathbb{C} \setminus \mathbb{Z}_{\leq 0}, \ z \in \mathbb{C} \) and \( |z| < 1 \); for \( |z| \geq 1 \) its analytical continuation is taken. The Kummer’s confluent hypergeometric function is \( _1F_1(a; b; z) = \sum_{n\in\mathbb{N}} \frac{a^n}{b^n n!} \) with \( a \in \mathbb{R}^+, \ b \in \mathbb{R}^+, \ z \in \mathbb{R} \).

The Fox-Wright generalized hypergeometric function is \( _pF_q((a_1, A_1); (b_1, B_1); z) = \sum_{n\in\mathbb{N}} \frac{(a_1)_n (A_1)_n}{(b_1)_n n!} \) where \( a \in \mathbb{R}^+, \ b \in \mathbb{R}^+, \ z \in \mathbb{R}, \ A \in \mathbb{R}^+, \ B \in \mathbb{R}^+ \) and \( 1 + B > A \). For \( a \) 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 \( N \in \mathbb{N} \) \( \{N\} = \{1, \ldots, N\} \).

Given an \( x \in \mathbb{R}^d \) vector, its transpose is \( x^\top \) and \( \|x\|_p = \left( \sum_{i\in[d]} |x_i|^p \right)^{\frac{1}{p}} \) \((p \in [1, \infty))\) and \( \|x\|_\infty = \max_{i\in[d]} |x_i| \) denotes its \( p \)-norm and maximum norm. Let the \( n \)th derivative of the \( \cos(\cdot) \) function \((n \in \mathbb{N})\) be \( c_n = \cos^{(n)}(\cdot) \). For multi-indices \( p, q \in \mathbb{N}^d \) and \( \omega \in \mathbb{R}^d \) let \( |p| = \sum_{j=1}^d p_j, \omega^p = \prod_{j=1}^d \omega_j^{p_j}, \partial^p f(x) = \partial^{p_1} x_1 \cdots \partial^{p_d} x_d, \partial^p q(x, y) = \partial^{p_1} x_1 \cdots \partial^{p_d} x_d \partial^{q_1} y_1 \cdots \partial^{q_d} y_d \).

The set of \( \mathbb{R}^d \) valued \( \mu \)-integrable functions on \( S (1 \leq r < \infty) \) is \( L^r(S, \mu) \), with \( \|f\|_{L^r(S, \mu)} = \left[ \int_S |f(x)|^r d\mu(x) \right]^\frac{1}{r} \).

We use the shorthand \( \mu f = \int_S f(x) d\mu(x) \) where \( \mu \in \mathcal{M}_+^d(S) \) and \( f \in L^1(S, \mu) \). The product measure of \( \mu_1, \ldots, \mu_M \in \mathcal{M}_+^d(S) \) \((\otimes_{m=1}^M \mu_m)\) specifically when all the components coincide we use the shorthand \( \mu^M = \otimes_{m=1}^M \mu_m \). The empirical measure is \( \mathbb{F}_M = \frac{1}{n} \sum_{k=1}^n \delta_X \) with \( X \) being the Dirac measure concentrated on \( X \) and \( X_1, \ldots, X_M \sim \otimes_{m=1}^M \mu_m \). Let \( (\mathbb{R}^n)_{n\in\mathbb{N}} \) be a positive sequence. The boundedness of \( [X^n_{n\in\mathbb{N}}]_n \) almost surely is denoted by \( (X^n_{n\in\mathbb{N}})_n \). The expectation is \( \mathbb{E} \). Let \( n \in \mathbb{N} \). We say that an \( f : \mathbb{R}^d \to \mathbb{R} \) function is of polynomial growth of order \( n \) (shortly \( f \in \mathcal{F}_P(n) \)) if \( \sup_{x \in \mathbb{R}^d} \frac{|f(x)|}{1 + |x|^2} < \infty \); \( \mathcal{F}_P = \cup_{n \in \mathbb{N}} \mathcal{F}_P(n) \).

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

![Figure 1: $\Psi_\alpha$ for different $\alpha$ values.](image)

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 $\Lambda \in \mathcal{M}_+^1(\mathbb{R}^d)$ spectral measure such that

$$k(x, y) = \int_{\mathbb{R}^d} \cos \left( \langle \omega^\top (x - y) \rangle \right) d\Lambda(\omega)$$

$$= \int_{\mathbb{R}^d} \cos \left( \langle \omega^\top x \rangle \cos \left( \omega^\top y \right) + \sin \left( \omega^\top x \right) \sin \left( \omega^\top y \right) \right) d\Lambda(\omega).$$

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

$$\partial^p q_k(x, y) = \int_{\mathbb{R}^d} \partial^p \cos \left( \langle \omega^\top x \rangle \right) \partial^q \cos \left( \omega^\top y \right) + \partial^p \sin \left( \omega^\top x \right) \partial^q \sin \left( \omega^\top y \right) d\Lambda(\omega).$$

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

$$\overline{\partial^p q_k(x, y)} = \frac{1}{M} \sum_{m=1}^M \partial^p \cos \left( \omega_m^\top x \right) \partial^q \cos \left( \omega_m^\top y \right) + \partial^p \sin \left( \omega_m^\top x \right) \partial^q \sin \left( \omega_m^\top y \right)$$

$$= \langle \lambda_p(x), \lambda_q(y) \rangle_{\mathbb{R}^{2M}}, \quad (6)$$

where $\lambda_p(x) = \frac{1}{\sqrt{M}} \left[ (\partial^p \cos (\omega_m^\top x))_{m \in [M]} ; (\partial^p \sin (\omega_m^\top x))_{m \in [M]} \right] \in \mathbb{R}^{2M}$; this is the RFF feature approximation $\lambda_p$ in (4). For $p = q = 0$, the construction reduces to the traditional RFF technique (Rahimi and Recht, 2007).
This form implies that our target quantity can be written as
\[
\left\| \partial^p q_k - \partial^p q_k \right\|_S = \sup_{z \in S^k} |(\Lambda_M - \Lambda)(f_z)|, \quad f_z(\omega) = \omega^p(-\omega)q_{c[p]}(\omega^T z),
\]
thus the problem boils down to the study of supremum of empirical processes with \( F \subset \mathcal{F}_{\mathcal{P}(n)} \) 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 \( \sup_{f \in \mathcal{F}} \left| \frac{1}{M} \sum_{m=1}^M f(X_m) \right| \) quantity under the following assumptions:

1. **Compact parameterization**: \( \mathcal{F} = \{ f_t : t \in T \} \) where \( f_t : \mathbb{R}^d \rightarrow \mathbb{R} \) is parameterized by a compact set \( T \subset \mathbb{R}^d \).

2. **Lipschitz condition**: There exists \( n \in \mathbb{R}^+ \) and function \( L : \mathbb{R}^d \rightarrow \mathbb{R}^{\geq 0} \), \( L \in \mathcal{F}_{\mathcal{P}(n)} \) such that
   \[
   \begin{align*}
   & (a) \quad |f_{t_0}(x)| \leq L(x) \text{ for some } t_0 \in T, \\
   & (b) \quad |f_{t_1}(x) - f_{t_2}(x)| \leq L(x) \rho(\|t_1 - t_2\|_2) \text{ for all } x \in \mathbb{R}^d, t_1, t_2 \in T, \\
   & (c) \quad \text{with } \rho : [0, |T|] \rightarrow \mathbb{R}^{\geq 0} \text{ continuous strictly increasing mapping with } \rho(0) = 0 \text{ such that } \\
   & \quad I_{\rho}(|T|) := \rho(|T|) \int_0^1 \sqrt{\log \left( 1 + \frac{2|T|}{\rho(u\rho(|T|))} \right)} \, du < \infty.
   \end{align*}
   \]

3. **Independence, finite \( \alpha \)-exponential Orlicz norm**: 
   \[
   \begin{align*}
   & (a) \quad (X_m)_{m \in [M]} \text{ are independent } \mathbb{R}^d \text{-valued random variables; shortly, } (X_m)_{m \in [M]} \sim \otimes_{m \in [M]} \mathbb{P}_m \text{ with } \mathbb{P}_m \in 
   \text{M}^{+}_1 \left( \mathbb{R}^d \right), \\
   & (b) \quad \exists \alpha \in \mathbb{R}^+ \text{ such that } \|X_m\|_{\psi_{\alpha}} < \infty \text{ for all } m \in [M].
   \end{align*}
   \]

4. **Centering**: \( E[f(X_m)] = 0 \) 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 \( (X_m)_{m \in [M]} \) satisfy Assumptions 1-4 and \( \gamma := \frac{\alpha}{n} \leq 1 \). Let \( \mathbb{P} = \otimes_{m \in [M]} \mathbb{P}_m \), and
\[
\|L\|_{L^2(X_{1:M})} := \sqrt{\frac{1}{M} \sum_{m \in [M]} L^2(X_m)}. \quad \text{Let } \Psi^{(l)} \gamma \text{ be the convexification}^5 \text{ of } \Psi, A_{\gamma} := \left( \frac{\Psi^{(l)} \gamma}{\Psi \gamma} \right)^{-1}(1), \\
B_{\gamma} := \left( \Psi^{(l)} \gamma \right)^{-1}(1), \quad C_{\gamma} \text{ and } C_D \text{ be the constants defined in (44) and (45), and } K_{\gamma} := \\
2\left( \frac{1}{\gamma} - 1 \right) \left( C_{\gamma} \left[ 16B_{\gamma} + 2\left( \frac{1}{\gamma} - 1 \right) \left( 1 + A_{\gamma} \right) \right] + 8A_{\gamma} \right). \quad \text{Then for any } \varepsilon > 0 \text{ satisfying}
\]
\[
\varepsilon \geq 6B, \quad B := 2C_D \sqrt{d^d} \frac{E \left[ \|L\|_{L^2(X_{1:M})} \right]}{\sqrt{M}} I_{\rho}(|T|), \tag{9}
\]

5. \( \Psi \gamma \) is not convex for \( \gamma < 1 \). We convexify \( \Psi \gamma \) and use the Section 4(v) based integral control property holding for convex \( \Psi \)-st; for details on \( \Psi^{(l)} \gamma \) see Section 5.3
we have
\[
\mathbb{P} \left( \sup_{t \in T} \frac{1}{M} \sum_{m \in [M]} f_t(X_m) \geq \varepsilon \right) \leq 2e^{-\left( \frac{3K\varepsilon}{\max_{m \in [M]} \sup_{t \in T} \|f_t(X_m)\|_{\Psi,\gamma}} \right)^\gamma} + e^{-\frac{M\varepsilon^2}{2\alpha^2 + 8\varepsilon}} , \quad (10)
\]
where
\[
\sigma^2 := \sup_{t \in T} \frac{1}{M} \sum_{m \in [M]} \mathbb{E} \left[ f_t^2(X_m) \right] ,
\]
\[
c := \max_{m \in [M]} \sup_{t \in T} \|f_t(X_m)\|_{\Psi,\gamma} \left[ \frac{1}{\beta_\gamma} \log \left( \frac{6\Gamma \left( 1 + \frac{1}{\gamma} \right) \max_{m \in [M]} \sup_{t \in T} \|f_t(X_m)\|_{\Psi,\gamma}}{\gamma \varepsilon} \right) \right]^{\frac{1}{\gamma}} \mathbb{E} \left[ \max_{m \in [M]} \sup_{t \in T} \|f_t(X_m)\| \right] \in [0, +\infty).
\]

**Remark 2**

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

(ii) **Assumption (3):** Assumption (3a) with Assumption (4) is weaker than being i.i.d.: for example \( \mathbb{E} [f_t(X_m)] = 0 \) holds for \( X_m = \mathcal{N} \left( 0, \sigma^2_m \right) \) and \( f_t(x) = c_t x^3 \), 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 \((\alpha', n')\) 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 \( \alpha \) to the value \( \alpha' = n \) using that \( \|X_m\|_{\Psi,\alpha} \leq \infty \) implies \( \|X_m\|_{\Psi,\alpha'} \leq \infty \) 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 \( \forall x \in \mathbb{R}^d \), \( s \) and \( t \in T \)
\[
|T_c f_t(x) - T_c f_s(x)| \leq |f_t(x) - f_s(x)| \leq L(x)\rho(\|t - s\|_2) , \quad (11)
\]
where \( T_c f(x) := f(x) 1_{\|f(x)\| \leq c} + c 1_{f(x) > c} - c 1_{f(x) < -c} \) 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:** \( \|\max_{m \in [M]} \sup_{t \in T} |f_t(X_m)|\|_{\Psi,\gamma} \) and \( \mathbb{E} \left[ \max_{m \in [M]} \sup_{t \in T} |f_t(X_m)| \right] \) are finite (see Section 5.5) in Theorem 1 by the Lipschitz assumption (2a)-(2b), \( \|X_m\|_{\Psi,\alpha} \leq \infty \) (Assumption (3b)) and \( L \in \mathcal{F}_{\mathcal{P}(n)} \) (Assumption (2)).
(v) **RFF specialization:** Assuming that the $\alpha$-exponential Orlicz condition holds for the $\Lambda$ spectral measure associated to $k$ ($\exists \alpha \in \mathbb{R}^+$ such that $\|\omega\|_{\psi_\alpha} < \infty$, $\omega \sim \Lambda$), one can see (Section 5.1) that RFFs are covered by choosing
\[
d' = d, \quad f_t(x) \leftarrow f_z(\omega) - \Lambda f_z, \quad t \leftarrow z, \quad T \leftarrow S_\Delta, \quad X_m \leftarrow \omega_m,
\]
\[
\rho(u) = u^\beta, \quad \beta = \frac{1}{1 + (\log|S_\Delta|)_+} \in (0, 1], \quad n \leftarrow |p + q| + \beta.
\]

While any value of $\beta \in (0, 1]$ would meet the assumptions, allowing $\beta$ to depend on the diameter of $S_\Delta$ 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 $C_{\text{RFF}} \in \mathbb{R}^+$, depending only on $\Lambda$, $|p + q|$, but not on $|S_\Delta|$ and $M$, such that
\[
B \leq C_{\text{RFF}} \sqrt{1 + (\log|S_\Delta|)_+},
\]
\[
s^2 \leq C_{\text{RFF}},
\]
\[
\max_{m \in [M]} \sup_{z \in S_\Delta} \|g_z(\omega_m)\|_{\psi_\gamma} \leq C_{\text{RFF}},
\]
\[
\max_{m \in [M]} \sup_{z \in S_\Delta} |g_z(\omega_m)|_{\psi_\gamma} \leq C_{\text{RFF}} [\log(1 + M)]^{n/\alpha},
\]
\[
E \left[ \max_{m \in [M]} \sup_{z \in S_\Delta} |g_z(\omega_m)| \right] \leq C_{\text{RFF}} [\log(1 + M)]^{n/\alpha}.
\]

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 $\Lambda$. Suppose that $\Lambda$ satisfies the $\alpha$-exponential Orlicz assumption ($\exists \alpha \in \mathbb{R}^+$ such that $\|\omega\|_{\psi_\alpha} < \infty$, $\omega \sim \Lambda$) and let $S \subset \mathbb{R}^d$ be a compact set. Let $\beta = \frac{1}{1 + (\log|S_\Delta|)_+} \in (0, 1]$, let $p, q \in \mathbb{N}^d$, $n := |p + q| + \beta$, and assume that $\gamma := \frac{\alpha}{n} \leq 1$. Let $\hat{\partial^p q_k}$ be the RFF estimate of $\partial^p q_k$ using $(\omega_m)_{m \in [M]}$ i.i.d. $\Lambda$ samples as given in Eq. (6). Then, there exists a constant $C \in \mathbb{R}^+$ (depending only on $\Lambda$, $|p + q|$, but not on $S$ and $M$) such that for any $\epsilon \geq C\sqrt{\frac{1 + (\log|S_\Delta|)_+}{M}}$,
\[
\lambda^M \left( \left\| \hat{\partial^p q_k} - \partial^p q_k \right\|_S \geq \epsilon \right) \leq 2e^{-\frac{(M\epsilon)^n}{C \log(1 + M)}} + e - \frac{M^2}{C \left(1 + \epsilon [\log(C/\epsilon)^{\gamma}/\log(1 + M)]^{1/\gamma}\right)}.
\]

**Corollary 4 (Almost sure convergence for kernel derivative approximation)** Let $p, q \in \mathbb{N}^d$ and $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a continuous, bounded, shift-invariant kernel with spectral measure $\Lambda$ which satisfies the $\alpha$-exponential Orlicz assumption for some $\alpha > 0$. Then, for any
\[
6. This requirement implies that $\int_{\mathbb{R}^d} |\omega^{p + q}| \, d\Lambda(\omega) < \infty$ and thus the existence of $\partial^p q_k$ for any $p, q \in \mathbb{N}^d$. 

11
sequence of compact sets \((S_M)_{M=2}^\infty\) such that \((\log |S_M|)_+ = o(M)\), we have

\[
\left\| \partial_p q_k - \partial_p q_k \right\|_{S_M} = O_{a.s.} \left( \frac{\sqrt{\log |S_M|} + \log M}{\sqrt{M}} \right)
\]  

(14)

Remark 5

(i) **Spectral measure \((\Lambda)\) examples**: Our result assumes the \(\alpha\)-exponential Orlicz property of the spectral measure \(\Lambda\) associated to \(k\). In Table 2 we provide various examples for \(\Lambda\) (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) **\(\alpha\)-exponential Orlicz assumption for tensor product kernels**: Using the \(\alpha\)-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. 

\[
k(x,y) = \prod_{i \in [d]} k_i(x_i,y_i), \Lambda = \otimes_{i \in [d]} \Lambda_i,
\]

and (ii) \(\Lambda_i\), the spectral measure associated to \(k_i\), satisfies the \(\alpha_i\)-exponential Orlicz assumption \((\alpha_i \in \mathbb{R}^+\)). Then \(\omega \sim \Lambda\) is \(\alpha\)-exponential Orlicz with \(\alpha = \min_{i \in [d]} \alpha_i\); see Section 5.9.

(iii) **\(\alpha\)-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 \(|p+q| \leq 2\ell = \alpha\). Indeed, we proved under the more general \(\alpha\)-exponential Orlicz assumption the same \(a.s.\) convergence rates for any arbitrary order (see Corollary 4) kernel derivatives.

<table>
<thead>
<tr>
<th>Spectrum</th>
<th>Spectral density: (f_{\Lambda}(\omega))</th>
<th>Parameters</th>
<th>(\alpha)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>(\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{\omega^2}{2\sigma^2}})</td>
<td>(\sigma &gt; 0)</td>
<td>2</td>
</tr>
<tr>
<td>Laplace</td>
<td>(\frac{\alpha}{2}e^{-\sigma</td>
<td>\omega</td>
<td>})</td>
</tr>
<tr>
<td>generalized Gaussian</td>
<td>(\frac{\alpha}{2\beta\Gamma\left(\frac{1}{\beta}\right)}e^{-\frac{</td>
<td>\omega</td>
<td>^\alpha}{\beta}})</td>
</tr>
<tr>
<td>variance Gamma</td>
<td>(\frac{\sigma^b</td>
<td>\omega</td>
<td>^{b-\frac{1}{2}} K_{b-\frac{1}{2}}(\sigma</td>
</tr>
<tr>
<td>Weibull (S)</td>
<td>(\frac{\alpha}{2a} \left( \frac{</td>
<td>\omega</td>
<td>}{\lambda} \right)^{s-1} e^{-\left( \frac{</td>
</tr>
<tr>
<td>exponentiated exponential (S)</td>
<td>(\frac{\alpha}{2a} \left( 1 - e^{-\frac{</td>
<td>\omega</td>
<td>}{\lambda}} \right)^{\alpha-1} e^{-\frac{</td>
</tr>
<tr>
<td>exponentiated Weibull (S)</td>
<td>(\frac{\alpha s}{2a} \left( \frac{</td>
<td>\omega</td>
<td>}{\lambda} \right)^{s-1} \left[ 1 - e^{-\left( \frac{</td>
</tr>
</tbody>
</table>
## ORLICZ RANDOM FOURIER FEATURES

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

<table>
<thead>
<tr>
<th>Spectrum</th>
<th>Spectral density: ( f_\Lambda(\omega) )</th>
<th>Parameters ( \alpha )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nakagami (S)</td>
<td>( \frac{m^m}{\Gamma(m)1^m}</td>
<td>\omega</td>
</tr>
<tr>
<td>chi-squared (S)</td>
<td>( \frac{1}{2\pi^{d+1}}</td>
<td>\omega</td>
</tr>
<tr>
<td>Erlang (S)</td>
<td>( \frac{\lambda^s</td>
<td>\omega</td>
</tr>
<tr>
<td>Gamma (S)</td>
<td>( \frac{1}{2\pi^{d/2}}</td>
<td>\omega</td>
</tr>
<tr>
<td>generalized Gamma (S)</td>
<td>( \frac{p/a^D}{2\Gamma \left( \frac{D}{1} \right)}</td>
<td>\omega</td>
</tr>
<tr>
<td>Rayleigh (S)</td>
<td>( \frac{</td>
<td>\omega</td>
</tr>
<tr>
<td>Maxwell-Boltzmann (S)</td>
<td>( \frac{1}{2\pi} e^{-\frac{\omega^2}{2a^2}} )</td>
<td>( a &gt; 0 ) 2</td>
</tr>
<tr>
<td>chi (S)</td>
<td>( \frac{1}{2\pi^{1/2}}</td>
<td>\omega</td>
</tr>
<tr>
<td>exponential-logarithmic (S)</td>
<td>( \frac{1}{2\log(p)} \frac{\beta(p-1)\beta</td>
<td>\omega</td>
</tr>
<tr>
<td>Weibull-logarithmic (S)</td>
<td>( \frac{1}{2\log(p)} \frac{\alpha \beta (p-1)</td>
<td>\omega</td>
</tr>
<tr>
<td>Gamma/Gompertz (S)</td>
<td>( \frac{bse</td>
<td>\omega</td>
</tr>
<tr>
<td>hyperbolic secant</td>
<td>( \frac{1}{2} \text{sech} \left( \frac{\pi}{2} \omega \right) )</td>
<td>( s &gt; 0 ) 1</td>
</tr>
<tr>
<td>logistic</td>
<td>( \frac{e^{-\frac{\omega^2}{2s^2}}}{s \left[ 1+e^{-\frac{\omega^2}{2s^2}} \right]^2} )</td>
<td>( s &gt; 0 ) 1</td>
</tr>
<tr>
<td>normal-inverse Gaussian</td>
<td>( \frac{\alpha \beta K_1(\alpha \sqrt{\delta^2+\omega^2}) e^{\delta \omega}}{\pi \sqrt{\delta^2+\omega^2}} )</td>
<td>( \alpha &gt; 0, \delta \in \mathbb{R} ) 1</td>
</tr>
<tr>
<td>hyperbolic</td>
<td>( \frac{1}{2\delta K_1(\delta \omega)} e^{-\alpha \sqrt{\delta^2+\omega^2}} )</td>
<td>( \alpha &gt; 0, \delta \in \mathbb{R} ) 1</td>
</tr>
<tr>
<td>generalized hyperbolic</td>
<td>( \frac{(\alpha/\delta)^{\lambda}}{\sqrt{2\pi} \lambda K_\lambda(\delta \gamma)} \frac{K_\lambda \left( \frac{\alpha \sqrt{\delta^2+\omega^2}}{\lambda} \right)^\frac{1}{2-\lambda}}{\left( \frac{\sqrt{\delta^2+\omega^2}}{\lambda} \right)^{\frac{1}{2-\lambda}}} )</td>
<td>( \alpha &gt; 0, \lambda \in \mathbb{R}, \delta \in \mathbb{R} ) ( 1 )</td>
</tr>
</tbody>
</table>
Figure 2: Relation of the spectral density examples of Table 2. '(S)' stands for symmetrized.
### Table 3: Kernel examples for the spectral densities given in Table 2.

<table>
<thead>
<tr>
<th>Kernel name</th>
<th>Kernel value: (k(x, y))</th>
<th>Spectrum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>(e^{-\frac{\sigma^2(x-y)^2}{2}})</td>
<td>Gaussian</td>
</tr>
<tr>
<td>inverse quadric</td>
<td>(\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{-</td>
<td>\beta(x-y)</td>
</tr>
<tr>
<td>inverse multiquadric</td>
<td>[\frac{\sigma^2}{\pi} \sum_{n=1}^{\infty} (-1)^n \frac{\lambda^n}{n!} \Gamma \left( 1 + \frac{n}{s} \right) \left( \frac{1}{1 + 2i(x-y)} \right)^{-\frac{n}{s}} + \frac{1}{1 - 2i(x-y)}^{-\frac{n}{s}} ]</td>
<td>variance Gamma</td>
</tr>
<tr>
<td></td>
<td>[\frac{1}{\sqrt{2}} \text{erfi} \left( \frac{\sigma(x-y)}{\sqrt{2}} \right)]</td>
<td>Rayleigh (S)&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>[1F_1 \left( \frac{3}{2}; \frac{1}{2}; \frac{(x-y)^2}{\sigma^2} \right)]</td>
<td>chi (S)&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>[\prod_{n=0}^{\infty} \left( -1 \right)^n \frac{\lambda^n}{n!} \Gamma \left( \frac{n+1}{\alpha} \right) \text{Li}_{n+1} \left( 1 - p \right) ]</td>
<td>Weibull-logarithmic (S)&lt;sup&gt;b&lt;/sup&gt; c</td>
</tr>
<tr>
<td></td>
<td>[\frac{1}{2} \left[ c_A(x-y) + c_A(y-x) \right], \text{ with } c_A(t) = \beta^s \frac{\Lambda(t)}{s(t+1)} ]</td>
<td>Gamma/Gompertz (S)&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>(\text{sech}(x-y))</td>
<td>hyperbolic secant</td>
</tr>
<tr>
<td></td>
<td>[\frac{\pi s(x-y)}{\sinh(\pi s(x-y))}]</td>
<td>logistic</td>
</tr>
<tr>
<td></td>
<td>[\delta \left[ \alpha - \sqrt{\sigma^2 + (x-y)^2} \right]^{\lambda} \text{K}<em>\lambda \left( \frac{\delta \sqrt{\sigma^2 + (x-y)^2}}{K</em>\lambda(\delta \alpha)} \right)]</td>
<td>normal-inverse Gaussian</td>
</tr>
<tr>
<td></td>
<td>[\frac{\alpha K_1 \left( \delta \sqrt{\sigma^2 + (x-y)^2} \right)}{\sqrt{\sigma^2 + (x-y)^2} \text{K}_\lambda(\delta \alpha)}]</td>
<td>hyperbolic</td>
</tr>
<tr>
<td></td>
<td>[\left( \frac{\alpha}{\sqrt{\sigma^2 + (x-y)^2}} \right)^\lambda \text{K}_\lambda \left( \delta \sqrt{\sigma^2 + (x-y)^2} \right) ]</td>
<td>generalized hyperbolic</td>
</tr>
</tbody>
</table>

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

<sup>b</sup> In case of symmetrization (S): \(k(x, y) = \frac{1}{2} \left[ c_A(x-y) + c_A(y-x) \right] \) where \(c_A(t) = E_{\omega \sim \Lambda} [e^{it\omega}]\) is the characteristic function of the spectral measure (on \(\mathbb{R}^d\)) before symmetrization; \(i = \sqrt{-1}\).

<sup>c</sup> The characteristic function was obtained by Ciumara and Preda (2009).
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 \( \Psi \) (unless explicitly required).

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

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

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

(iii) Monotonicity in \( \Psi \): \( \Psi \leq \Phi \) 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 \( \Psi \) 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 \( X, X' \in L_\Psi \) and \( \alpha \in \mathbb{R}^+ \). Then \( X + X' \in L_\Psi \) and

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

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

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

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

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

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

Thanks to the \( \alpha \)-exponential Orlicz condition on \( \Lambda \) and the i.i.d. property of \( (\omega_m)_{m=1}^M \) in (6), Assumption 3 is trivially fulfilled. Assumption 4 holds by the definition of \( g_z(\cdot) \) and because the distribution of \( \omega_m \) is \( \Lambda \). Assumption 1 is satisfied since \( S_\Delta \) is a compact set of \( \mathbb{R}^d \). Therefore, it remains to prove Assumption 2, with the existence of \( n \in \mathbb{R}^+ \) and \( L \in \mathcal{F}_{P(n)} \). First, notice that

\[
|f_z(\omega)| \leq \prod_{i \in [d]} |\omega_i|^{p_i+q_i} \leq \|\omega\|_2^{p+q}.
\]

(15)
• **Order:** (15) implies that

\[
|g_2(\omega)| \leq |f_2(\omega)| + \Lambda |f_2| \leq \|\omega\|^2 + \Lambda \left[ \|\cdot\|^2 \right] =: L_1(\omega). \tag{16}
\]

• **Lipschitz condition:** Let \([z_1, z_2] = \{az_1 + (1 - a)z_2 : a \in [0, 1]\}\) denote the segment connecting \(z_1, z_2 \in \mathbb{R}^d\). By using the mean value theorem

\[
|g_{z_1}(\omega) - g_{z_2}(\omega)| \leq \max_{z \in [z_1, z_2]} \left\| \frac{\partial g_z(\omega)}{\partial z} \right\|_2 \|z_1 - z_2\|_2, \tag{17}
\]

\[
\frac{\partial g_z(\omega)}{\partial z} = \frac{\partial f_z(\omega)}{\partial z} - \Lambda \frac{\partial f_z(\omega)}{\partial z} \quad \text{with} \quad \frac{\partial f_z(\omega)}{\partial z} = \omega^p(-\omega)^q, L_z(\omega), \quad \text{and by using similar}
\]

computation as before, one gets

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

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

**Refined \(L\) and \(\rho\):** We now derive refined \(L\) and \(\rho\), by interpolating different bounds. From (16), we can obtain the crude estimate \(|g_{z_1}(\omega) - g_{z_2}(\omega)| \leq 2L_1(\omega)\), which combined with (17)-(18) gives

\[
|g_{z_1}(\omega) - g_{z_2}(\omega)| \leq (2L_1(\omega))^{1 - \beta} \left[ \|z_1 - z_2\|_2 L_2(\omega) \right]^{\beta} \tag{19}
\]

for any \(\beta \in (0, 1]\). Here we have used that if \(0 \leq x \leq \min(x_1, x_2)\) then \(x \leq x_1^{\frac{1 - \beta}{2}} x_2^\beta\). It follows that one can take

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

For \(\beta = 1\), we retrieve the former choice of \(L\) and \(\rho\). Furthermore, we have

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

Notice that the advantage of having the additional degree-of-freedom \(\beta\) 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 \(\|f(X)\|_{\psi, \gamma} < \infty\) for \(\|X\|_{\psi, \gamma} < \infty\), \(f \in \mathcal{F}_P(n)\), \(n \in \mathbb{R}^+\), \(\gamma = \frac{q}{n}\). Indeed, by the definition of \(f \in \mathcal{F}_P(n)\), there exists \(C \in \mathbb{R}^+\) such that \(|f(x)| \leq C(1 + \|x\|_2^n)\) for all \(x \in \mathbb{R}^d\). Hence for any \(\gamma > 0\)

\[
|f(x)|^\gamma \leq C^{\gamma} (1 + \|x\|_2^n)^{\gamma} \leq 2^{(\gamma - 1)\gamma} C^{\gamma} (1 + \|x\|_2^{\gamma n}), \tag{22}
\]
where in (**) we used that
\[(a + b)\gamma \leq 2^{(\gamma - 1) + (a\gamma + b\gamma)}, \quad a, b \geq 0, \gamma > 0. \tag{23}\]

Since \(X \in L_{\Psi_\alpha}\) there is some \(s \in \mathbb{R}^+\) for which \(\mathbb{E}\left[e^{s\|X\|^\alpha_2}\right] < \infty\). Combining this property with (22) and recalling that \(n\gamma = \alpha\) yields
\[e^{s'}|f(x)|^{\gamma} \leq e^{s'2(\gamma - 1) + C\gamma(1 + \|x\|^2_2)} \Rightarrow \mathbb{E}\left[e^{s'|f(X)|^{\gamma}}\right] \leq e^{s'2(\gamma - 1) + C\gamma \mathbb{E}\left[e^{s\|X\|^2_2}\right]} < \infty \]
with \(s' = \frac{s}{2(\gamma - 1) + C\gamma}\); this shows that \(f(X) \in L_{\Psi_\gamma}\).

\[5.3 \Psi_{\gamma}^{(l)}, \text{ the Convexification of } \Psi_{\gamma}\]

In the proof of Theorem 1 an integral control with convex \(\Psi\) (see Section 4(v)) is beneficial/applied. However, \(\Psi_{\gamma}\) 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 \(\Psi_{\gamma}\) we get that it is convex iff \(x \geq x_{\gamma} := \left(\frac{1 - \gamma}{\gamma}\right)^{\frac{1}{\gamma}}\).
  Indeed,
  \[
  \Psi_{\gamma}'(x) = \gamma x^{\gamma - 1} e^{x\gamma},
  \Psi_{\gamma}''(x) = \gamma e^{x\gamma} [(\gamma - 1)x^{\gamma - 2} + x^{\gamma - 1}\gamma x^{\gamma - 1}] \Rightarrow
  \Psi_{\gamma}''(x) = 0 \Leftrightarrow x = x_{\gamma}, \quad \Psi_{\gamma}'(x) > 0 \Leftrightarrow x > x_{\gamma}, \quad \Psi_{\gamma}'(x) < 0 \Leftrightarrow x < x_{\gamma}.
  \]

- 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
  \[
  x_{\gamma} := \inf \left\{ x \geq x_{\gamma} : e^{x\gamma} - 1 \leq \gamma x^{\gamma} e^{x\gamma} \right\},
  \]
  and define
  \[
  \Psi_{\gamma}^{(l)}(x) := \begin{cases} \frac{\Psi_{\gamma}(x)}{x_{\gamma}} x & \text{if } x \in [0, x_{\gamma}), \\ \Psi_{\gamma}(x) & \text{if } x \in [x_{\gamma}, \infty). \end{cases}
  \]
  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 $R_c f(x) := f(x) - T_c f(x)$ 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

$$
\sup_{t \in T} \frac{1}{M} \sum_{m \in [M]} f_t(X_m) = \sup_{t \in T} \frac{1}{M} \sum_{m \in [M]} (T_c f_t(X_m) - E[T_c f_t(X_m)] + E[T_c f_t(X_m)] + R_c f_t(X_m))
$$

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

$$
P \left( \sup_{t \in T} \frac{1}{M} \sum_{m \in [M]} f_t(X_m) \geq \varepsilon \right) \leq P \left( Z^{R_c} \geq \varepsilon/3 \right) + P \left( Z^{\mathcal{E}^T} \geq \varepsilon/3 \right). \tag{24}
$$

The structure of our proof is as follows.

1. Unbounded part ($Z^{R_c}$): Based on the Talagrand and the Hoffman-Jorgensen inequalities, for large enough $c$ (referred to as $c_{HJ}$) we will derive an exponential control over $P (Z^{R_c} \geq \varepsilon/3)$ expressed with $\|\max_{m \in [M]} \sup_{t \in T} |f_t(X_m)|\|_{\Phi_\psi}$, which is finite by Section 5.5.

2. Bounded part ($Z^{\mathcal{E}^T}$): 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 $\varepsilon$.

3. Truncation ($\mathcal{E}^T$): As $E[f_t(X_m)] = 0$, $T_c f_t \approx f_t$ and $E[T_c f_t(X_m)] \approx 0$ for large $c$ (called $c_{\text{min}}$). The $\mathcal{E}^T \leq \frac{\varepsilon}{3}$ 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 $Z^{R_c}$, $Z^{\mathcal{E}^T}$ and $\mathcal{E}^T$ quantities is detailed in the following sections. Plugging the (25) and (27) results of the computations into (24) gives the final bound (10). The $\varepsilon$ constraint comes from (32), provided that $c \geq c_{\text{min}} \lor c_{HJ}$. The constants $c_{\text{min}}$ and $c_{HJ}$ are defined in (34) and (37), respectively.

5.4.1 Bounding $Z^{R_c}$

$$
P \left( Z^{R_c} \geq \varepsilon/3 \right) \text{ is bounded as}
$$

$$
P \left( Z^{R_c} \geq \varepsilon/3 \right) \leq P \left( \sup_{t \in T} \sum_{m \in [M]} |R_c f_t(X_m)| \geq M \varepsilon/3 \right) \leq P \left( \sum_{m \in [M]} \sup_{t \in T} |R_c f_t(X_m)| \geq M \varepsilon/3 \right)
$$
\[
(b) \leq 2e^{-\left(\frac{M/3}{\sum_{m \in [M]} \sup_{t \in T} \|T_c f_t(X_m)\|_{\Psi^c}}\right)^\gamma} \\
(c) \leq 2e^{-\left(\frac{M}{\sum_{m \in [M]} \sup_{t \in T} \|T_c f_t(X_m)\|_{\Psi^c}}\right)^\gamma},
\]

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

5.4.2 Bounding \( \mathcal{Z}^T_c \)

Below we will invoke the Klein-Rio inequality and control the expectation \( \mathbb{E} \left[ \mathcal{Z}^T_c \right] \).

- **Klein-Rio inequality:** Let \( g_{m,t} : x \in \mathbb{R}^d \mapsto T_c f_t(x) - \mathbb{E} [T_c f_t(X_m)] \) and let us define the function classes
  \[
  \mathcal{T}_c \mathcal{F}^{[M]} := \{ g_t := (g_{1,t}, \ldots, g_{M,t}) : t \in T \}, \quad \mathcal{T}_c \mathcal{F} := \{ T_c f_t : t \in T \}.
  \]

  - \( g_{m,t} \in [-2c, 2c] \) are measurable and bounded functions.
  - Centering: by construction \( \mathbb{E}[g_{m,t}(X_m)] = 0 \) (\( \forall m \in [M] \)).
  - Countability: Since \( t \mapsto f_t \) is continuous, the \( \sup_{t \in T} \) 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
  \[
  \mathbb{E} \left[ \mathcal{Z}^T_c \right] \leq \varepsilon / 6,
  \]
  then the Klein-Rio inequality (Theorem 8 where the \( \sup_{t \in T} \) and \( \sup_{t \in \mathcal{T}_c \mathcal{F}^{[M]}} \) coincide) implies that
  \[
  \mathbb{P} \left( \mathcal{Z}^T_c \geq \varepsilon / 3 \right) \leq e^{-\frac{M (\varepsilon / 6)^2}{2(\sigma^2 + 4c^2 \mathcal{Z}^T_c) + 6c \varepsilon / 6}} \leq e^{-\frac{M \varepsilon^2}{72 \sigma^2 + 36 c \varepsilon}},
  \]
  where the weak variance \( \bar{\sigma}^2 \) is defined and bounded by
  \[
  \bar{\sigma}^2 := \sup_{t \in T} \frac{1}{M} \sum_{m \in [M]} \mathbb{E} \left[ (T_c f_t(X_m) - \mathbb{E} [T_c f_t(X_m)])^2 \right] \leq \sup_{t \in T} \frac{1}{M} \sum_{m \in [M]} \mathbb{E} \left[ (T_c f_t(X_m))^2 \right] \leq \sup_{t \in T} \frac{1}{M} \sum_{m \in [M]} \mathbb{E} \left[ f_t^2(X_m) \right] =: \sigma^2.
  \]

- **Bounding \( \mathbb{E} \left[ \mathcal{Z}^T_c \right] \):** We control \( \mathbb{E} \left[ \mathcal{Z}^T_c \right] \) 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
    \[
    \mathbb{E} \left[ \mathcal{Z}^T_c \right] \leq 2 \mathbb{E} \left[ \mathcal{R}(X_1:M, \mathcal{T}_c \mathcal{F}) \right],
    \]
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 $\varepsilon$-net are assumed to belong to the set covered.
of a compact set $T \subset \mathbb{R}^{d'}$

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

(29) can be estimated further as

$$\mathcal{R}(x_{1:M}, T_c \mathcal{F}) \leq C_D \sqrt{d} \int_0^{\|L\|_{L^2(\mathbf{x}_{1:M})} \rho(|T|)} \left[ \log \left( \frac{2|T|}{\rho^{-1} \left( \frac{\varepsilon}{\|L\|_{L^2(\mathbf{x}_{1:M})}} \right)} \right) + 1 \right] \, d\varepsilon$$

$$= C_D \sqrt{d} \frac{\|L\|_{L^2(\mathbf{x}_{1:M})} \rho(|T|)}{\sqrt{M}} \int_0^1 \sqrt{\log \left( 1 + \frac{2|T|}{\rho^{-1} \left( \frac{\varepsilon}{\rho(|T|)} \right)} \right)} \, du,$$

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

$$\mathbb{E} \left[ \mathbb{Z}^{T_c} \right] \leq 2C_D \sqrt{d} \frac{\mathbb{E} \left[ \|L\|_{L^2(\mathbf{x}_{1:M})} \right]}{\sqrt{M}} I_p(|T|) =: B.$$

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

5.4.3 Bounding $\mathcal{E}^{T_c}$

- **Bounding $\mathcal{E}^{T_c}$ by the incomplete Gamma function ($I_\gamma$):**

$$\mathbb{E} \left[ T_c f_t(\mathbf{x}_m) \right] \overset{(a)}{=} -\mathbb{E} \left[ \mathcal{R}_c f_t(\mathbf{x}_m) \right] \leq \mathbb{E} \left[ (-f_t(\mathbf{x}_m) - c) \mathbb{I}_{f_t(\mathbf{x}_m) \leq -c} \right] \overset{(b)}{=} \int_c^\infty \mathbb{P} \left( -f_t(\mathbf{x}_m) \geq y \right) \, dy. \quad (33)$$

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

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

(c) holds by using that for a $Z \geq 0$ random variable, $\mathbb{E} [Z] = \int_0^\infty \mathbb{P} (Z \geq z) \, dz$; we choose $Z = \max (0, -f_t(\mathbf{x}_m) - c)$. Therefore

$$\mathcal{E}^{T_c} = \sup_{t \in T} \mathbb{E} \left[ \frac{1}{M} \sum_{m \in [M]} T_c f_t(\mathbf{x}_m) \right] \overset{(a)}{=} \max_{m \in [M]} \sup_{t \in T} \int_c^\infty \mathbb{P} \left( -f_t(\mathbf{x}_m) \geq y \right) \, dy \overset{(b)}{\leq} 2 \max_{m \in [M]} \sup_{t \in T} \int_c^\infty e^{- \left( \frac{y}{\mathbb{I}_{f_t(\mathbf{x}_m) \leq \Psi}} \right)^\gamma} \, dy \overset{(c)}{=} 2 \max_{m \in [M]} \sup_{t \in T} \left( \|f_t(\mathbf{x}_m)\|_{\Psi} \int_0^\infty \frac{e^{-u^\gamma}}{\|f_t(\mathbf{x}_m)\|_{\Psi}} \, du \right)$$
\[
\begin{align*}
\langle d \rangle & \ 2 \max_{m \in [M]} \sup_{t \in T} \left( \| f_t(X_m) \|_{\Psi_{\gamma}} \left[ \int_0^\infty e^{-u} \, du - \int_0^{\frac{c}{\| f_t(X_m) \|_{\Psi_{\gamma}}}} e^{-u} \, du \right] \right) \\
\langle e \rangle & \ 2 \max_{m \in [M]} \sup_{t \in T} \left( \| f_t(X_m) \|_{\Psi_{\gamma}} \left[ \Gamma \left( 1 + \frac{1}{\gamma} \right) - I_{\gamma} \left( \frac{c}{\| f_t(X_m) \|_{\Psi_{\gamma}}} \right) \right] \right) \\
\langle f \rangle & \ 2 \max_{m \in [M]} \sup_{t \in T} \| f_t(X_m) \|_{\Psi_{\gamma}} \left[ \Gamma \left( 1 + \frac{1}{\gamma} \right) - I_{\gamma} \left( \frac{c}{\max_{m' \in [M]} \sup_{t \in T} \| f_{t'}(X_{m'}) \|_{\Psi_{\gamma}}} \right) \right] \\
\langle g \rangle & \ 2 \Gamma \left( 1 + \frac{1}{\gamma} \right) \left( 1 - \left[ 1 - e^{-\gamma \left( \max_{m \in [M]} \sup_{t \in T} \| f_t(X_m) \|_{\Psi_{\gamma}} \right) \right] \right) \sup_{m \in [M]} \| f_t(X_m) \|_{\Psi_{\gamma}} \\
= & \ \tilde{B},
\end{align*}
\]

where (a) holds by taking maximum over \( m \in [M] \) and using (33), (b) follows from the deviation inequality implied by Section 4(vii). (c) was obtained from a \( \frac{\gamma}{\gamma} \) substitution, and in (d) we decomposed the integral to have the incomplete Gamma function appear. (e) is a consequence of the definition of \( I_{\gamma} \) and the limit

\[
I_{\gamma}(x) = \int_0^x e^{-t} \, dt = \frac{1}{\gamma} \int_0^{x^{\frac{1}{\gamma}}} u^{\frac{1}{\gamma} - 1} e^{-u} \, du \xrightarrow{x \to \infty} \frac{1}{\gamma} \Gamma \left( 1 + \frac{1}{\gamma} \right) = \Gamma \left( 1 + \frac{1}{\gamma} \right),
\]

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

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

\[
1 - \left[ 1 - e^{-\gamma \left( \max_{m \in [M]} \sup_{t \in T} \| f_t(X_m) \|_{\Psi_{\gamma}} \right) \right] \leq \frac{\varepsilon}{6 \Gamma \left( 1 + \frac{1}{\gamma} \right) \max_{m \in [M]} \sup_{t \in T} \| f_t(X_m) \|_{\Psi_{\gamma}}}.
\]

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

\[
x = \frac{c}{\max_{m \in [M]} \sup_{t \in T} \| f_t(X_m) \|_{\Psi_{\gamma}}} \]

it is enough to use \( c \) such that

\[
\frac{1}{\gamma} \left[ 1 - e^{-\gamma \left( \max_{m \in [M]} \sup_{t \in T} \| f_t(X_m) \|_{\Psi_{\gamma}} \right) \right] \leq \frac{\varepsilon}{6 \Gamma \left( 1 + \frac{1}{\gamma} \right) \max_{m \in [M]} \sup_{t \in T} \| f_t(X_m) \|_{\Psi_{\gamma}}}.
\]

Solving this inequality for \( c \) means that

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

(34)
5.5 Proof of $\mathcal{F} \subset \mathcal{F}_P(n)$, $\mathbb{E} \left[ \max_{m \in [M]} \sup_{t \in T} |f_t(X_m)| \right] < \infty$, and $\left\| \max_{m \in [M]} \sup_{t \in T} |f_t(X_m)| \right\|_{\Psi_{\gamma}} < \infty$

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

$$|f_t(x)| \leq |f_t(x) - f_{t_0}(x)| + |f_{t_0}(x)| \leq L(x) |\rho(\|t - t_0\|_2) + 1| \leq L(x) [\rho(|T|) + 1], \quad (35)$$

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

- $\mathcal{F} \subset \mathcal{F}_P(n)$: By $L \in \mathcal{F}_P(n)$ and (35), $f_t \in \mathcal{F}_P(n)$ for all $t \in T$, in other words $\mathcal{F} \subset \mathcal{F}_P(n)$.

- Finiteness of $\left\| \max_{m \in [M]} \sup_{t \in T} |f_t(X_m)| \right\|_{\Psi_{\gamma}}$: Using (35), we get

$$\max_{m \in [M]} \sup_{t \in T} |f_t(X_m)| \leq [\rho(|T|) + 1] \sum_{m \in [M]} L(X_m). \quad (36)$$

Thanks to Section 5.2, each $L(X_m)$ belongs to $L_{\Psi_{\gamma}}$. Combining this with the generalized triangular inequality Section 4(vi) gives the claim.

- Finiteness of $\mathbb{E} \left[ \max_{m \in [M]} \sup_{t \in T} |f_t(X_m)| \right]$: Each $L(X_m)$ is integrable (because $L$ has a polynomial growth and the distribution of $X_m$ satisfies the $\alpha$-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

$$c \geq c_{HJ} := 8 \mathbb{E} \left[ \max_{m \in [M]} \sup_{t \in T} |f_t(X_m)| \right] \quad (37)$$

one has

$$\left\| \sum_{m \in [M]} \sup_{t \in T} |R_c f_t(X_m)| \right\|_{\Psi_{\gamma}} \leq K_\gamma \left\| \max_{m \in [M]} \sup_{t \in T} |f_t(X_m)| \right\|_{\Psi_{\gamma}}. \quad (38)$$

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

$$\left\| \sum_{m \in [M]} \sup_{t \in T} |R_c f_t(X_m)| \right\|_{\Psi_{\gamma}} =$$

$$= \left\| \sum_{m \in [M]} \left( \sup_{t \in T} |R_c f_t(X_m)| - \mathbb{E} \left[ \sup_{t \in T} |R_c f_t(X_m)| \right] + \mathbb{E} \left[ \sup_{t \in T} |R_c f_t(X_m)| \right] \right) \right\|_{\Psi_{\gamma}}$$

$$\overset{(a)}{\leq} 2^{\gamma^{-1}} \left( \left\| \sum_{m \in [M]} \left( \sup_{t \in T} |R_c f_t(X_m)| - \mathbb{E} \left[ \sup_{t \in T} |R_c f_t(X_m)| \right] \right) \right\|_{\Psi_{\gamma}} +$$

$$+ \mathbb{E} \left[ \left\| \sum_{m \in [M]} \sup_{t \in T} |R_c f_t(X_m)| \right\|_{\Psi_{\gamma}} \right] $$
\[
\begin{align*}
\leq 2^{\frac{1}{\gamma} - 1} & \left( C_\gamma \left( \mathbb{E} \left[ \sum_{m \in [M]} \left( \sup_{t \in T} |R_c f_t(X_m)| - \mathbb{E} \left[ \sup_{t \in T} |R_c f_t(X_m)| \right] \right) \right] \right) \\
& \quad + \left\| \max_{m \in [M]} \sup_{t \in T} |R_c f_t(X_m)| - \mathbb{E} \left[ \sup_{t \in T} |R_c f_t(X_m)| \right] \right\|_{\psi_\gamma} + \\
& \quad + \frac{1}{\Psi_\gamma^{-1}(1)} \mathbb{E} \left[ \sum_{m \in [M]} \sup_{t \in T} |R_c f_t(X_m)| \right] \\
& =: 2^{\frac{1}{\gamma} - 1} \left[ C_\gamma (E_1 + E_2) + \frac{1}{\Psi_\gamma^{-1}(1)} E_3 \right].
\end{align*}
\]

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

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

- **Bounding \( E_1 \):**

\[
\begin{align*}
E_1 &= \mathbb{E} \left[ \sum_{m \in [M]} \left( \sup_{t \in T} |R_c f_t(X_m)| - \mathbb{E} \left[ \sup_{t \in T} |R_c f_t(X_m)| \right] \right) \right] \\
& \overset{(a)}{\leq} 2 \mathbb{E} \left[ \sum_{m \in [M]} \sup_{t \in T} |R_c f_t(X_m)| \right] \overset{(b)}{\leq} 16 \mathbb{E} \left[ \max_{m \in [M]} \sup_{t \in T} |R_c f_t(X_m)| \right] \\
& \overset{(c)}{\leq} 16 \mathbb{E} \left[ \max_{m \in [M]} \sup_{t \in T} |f_t(X_m)| \right] \overset{(d)}{\leq} 16 \left\| \max_{m \in [M]} \sup_{t \in T} |f_t(X_m)| \right\|_{\psi_\gamma^{-1}(1)} \\
& \overset{(e)}{\leq} 16 \left\| \max_{m \in [M]} \sup_{t \in T} |f_t(X_m)| \right\|_{\psi_\gamma^{-1}(1)},
\end{align*}
\]

where in (a) we used the triangle inequality, in (b) we applied the Hoffman-Jorgensen inequality (Theorem 6; \( t_0 = 0, p = 1, B = \mathbb{R}, Y_m = \sup_{t \in T} |R_c f_t(X_m)| \)) with

\[
\mathbb{P} \left( \sum_{m \in [M]} \sup_{t \in T} |R_c f_t(X_m)| > 0 \right) \overset{(f)}{=} \mathbb{P} \left( \max_{j \in [J]} \sum_{m \in [j]} \sup_{t \in T} |R_c f_t(X_m)| > 0 \right) = \mathbb{P} \left( \max_{m \in [M]} \sup_{t \in T} |f_t(X_m)| > c \right) \overset{(g)}{\leq} \mathbb{P} \left( \max_{m \in [M]} \sup_{t \in T} |f_t(X_m)| \geq c_H \right) \overset{(h)}{\leq} \frac{1}{8} = \frac{1}{2 \times 4^p} \text{ with } p = 1.
\]

In (f) the non-negativity of \( Y_m \) was exploited; in (g) \( c \geq c_H \) was used. We applied the Markov inequality and the definition of \( c_H \) in (h). (c) holds by \( |R_c f_t(X_m)| \leq |f_t(X_m)| \).
In (d) and (e) we applied Section 4(iv) with the convex $\Psi_\gamma^{(l)}$ and the monotonicity property Section 4(iii) with $\Psi_\gamma^{(l)} \leq \Psi_\gamma$, respectively.

**Bounding $E_2$:**

\[
E_2 = \left\| \max_{m \in [M]} \left( \sup_{t \in T} |R_c f_t(X_m)| - E \left[ \sup_{t \in T} |R_c f_t(X_m)| \right] \right) \right\|_{\psi_\gamma}
\]

\[
\leq \left\| \max_{m \in [M]} \left( \sup_{t \in T} |R_c f_t(X_m)| + E \left[ \sup_{t \in T} |R_c f_t(X_m)| \right] \right) \right\|_{\psi_\gamma}
\]

\[
= 2^{\frac{1}{2} - 1} \left( \left\| \max_{m \in [M]} \left( \sup_{t \in T} |R_c f_t(X_m)| \right) \right\|_{\psi_\gamma} + \left\| \max_{m \in [M]} \left( \sup_{t \in T} |R_c f_t(X_m)| \right) \right\|_{\psi_\gamma} \right)
\]

\[
\leq 2^{\frac{1}{2} - 1} \left( \left\| \max_{m \in [M]} \left( \sup_{t \in T} |R_c f_t(X_m)| \right) \right\|_{\psi_\gamma} + \frac{1}{\Psi_\gamma^{-1}(1)} \left( \max_{m \in [M]} \left( \sup_{t \in T} |R_c f_t(X_m)| \right) \right) \right)
\]

\[
\leq 2^{\frac{1}{2} - 1} \left( \left\| \max_{m \in [M]} \left( \sup_{t \in T} |f_t(X_m)| \right) \right\|_{\psi_\gamma} + \frac{1}{\Psi_\gamma^{-1}(1)} \left( \max_{m \in [M]} \left( \sup_{t \in T} |f_t(X_m)| \right) \right) \right)
\]

\[
\leq 2^{\frac{1}{2} - 1} \left( 1 + \frac{\Psi_\gamma^{-1}(1)}{\Psi_\gamma^{-1}(1)} \right) \left( \left\| \max_{m \in [M]} \left( \sup_{t \in T} |f_t(X_m)| \right) \right\|_{\psi_\gamma} \right).
\]

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 $\left( \frac{1}{2} - 1 \right)^+ = \frac{1}{2} - 1$ as $\gamma \in (0, 1]$.

(d) holds by Section 4(ii) with the constant $\lambda = E \left[ \max_{m \in [M]} \left( \sup_{t \in T} |R_c f_t(X_m)| \right) \right]$, (e) is by the monotonicity Section 4(iv) as $|R_c f_t(X_m)| \leq |f_t(X_m)|$, and by Section 4(v). (f) follows from $\Psi_\gamma^{(l)} \leq \Psi_\gamma$ combined with the monotonicity Section 4(iii).

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

\[
E_3 = E \left[ \sum_{m \in [M]} \left( \sup_{t \in T} |R_c f_t(X_m)| \right) \right] \leq 8 \left\| \max_{m \in [M]} \left( \sup_{t \in T} |f_t(X_m)| \right) \right\|_{\psi_\gamma} \left( \Psi_\gamma^{(l)} \right)^{-1} (1).
\]

By adding the obtained $E_1$, $E_2$ and $E_3$ bounds, we get (38) with $K_\gamma$ 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 $E \left[ \left\| L \right\|_{L^2(\omega_1, M)} \right]$ by $\sqrt{E \omega \sim L} \left[ L^2(\omega) \right]$ gives

\[
B \leq 2C_D \sqrt{d} \sqrt{\frac{E \omega \sim L}{M} \left[ L^2(\omega) \right]} \left| S_\Delta \right|^\beta \int_0^1 \sqrt{\log \left( 1 + 2u^{-1/\beta} \right)} du.
\]
We now aim at showing a tight bound for \(|S_\Delta|^\beta \int_0^1 \sqrt{\log (1 + 2u^{-1/\beta})} \, du\) 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
\[
I_\beta := \int_0^1 \sqrt{\log (1 + 2u^{-1/\beta})} \, du \leq \frac{4}{\sqrt{\beta}}, \quad \forall \beta \in (0, 1].
\] (40)

By the change of variable \(t = \beta \log \left(1 + 2u^{-1/\beta}\right)\) (i.e. \(u = \left(\frac{e^{t/\beta} - 1}{2}\right)^{-\beta}\)), we get
\[
I_\beta = \frac{2\beta}{\sqrt{\beta}} \int_{\beta \log(3)}^{\infty} \sqrt{t} e^{t/\beta} \, dt = \frac{2\beta}{\sqrt{\beta}} \int_{\beta \log(3)}^{\infty} \frac{\sqrt{t}}{e^{(1 - e^{-t/\beta})\beta + 1}} \, dt.
\]

Using the fact that \(1 - e^{-t/\beta} \geq \frac{2}{3}\) on \([\beta \log(3), +\infty)\), we arrive at
\[
I_\beta \leq \frac{3^{\beta+1}}{2\sqrt{\beta}} \int_{\beta \log(3)}^{\infty} \sqrt{te^{-t/\beta}} \, dt \leq \frac{9}{2\sqrt{\beta}} \sqrt{e} \left(\frac{3}{2}\right) \leq \frac{4}{\sqrt{\beta}},
\]
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
\[
|S_\Delta|^\beta \int_0^1 \sqrt{\log (1 + 2u^{-1/\beta})} \, du \leq 4e^{\frac{\log(|S_\Delta|)_+}{1+\log(|S_\Delta|)_+}} \sqrt{1 + (\log|S_\Delta|)_+} \leq 4e\sqrt{1 + (\log|S_\Delta|)_+},
\]
and therefore
\[
B \leq 8eC_D \sqrt{d} \sqrt{\mathbb{E}_{\omega \sim \Lambda} |L^2(\omega)|} \sqrt{1 + (\log|S_\Delta|)_+}.
\]

- **The term \(\sigma^2\):** It is defined in Theorem 1. Since the variance is bounded by the second moment, \(\mathbb{E} \left[ g_z^2(\omega_m) \right] \leq \mathbb{E} \left[ f_z^2(\omega_m) \right] \). Furthermore, since \((\omega_m)_{m=1}^M\) are i.i.d., the previous expectation can bounded by \(\mathbb{E} \left[ \|\omega\|_2^{2p+q} \right] \) using (15). As a result, we get
\[
\sigma^2 \leq \mathbb{E}_{\omega \sim \Lambda} \left[ \|\omega\|_2^{2p+q} \right].
\]

- **The term \(\max_{m \in [M]} \sup_{z \in S_\Delta} \|g_z(\omega_m)\|_{\Psi_\gamma}\)** with \(\gamma = \alpha/n \leq 1\) and \(n = |p+q| + \beta\): It appears in the definition of \(c\) (in Theorem 1). In view of the bound (16) which is uniform in \(z\) and using property (iv) of Section 4, we get
\[
\max_{m \in [M]} \sup_{z \in S_\Delta} \|g_z(\omega_m)\|_{\Psi_\gamma} \leq \left\| \|\omega\|_2^{p+q} + \Lambda \left\| \cdot \right\|_2^{p+q} \right\|_{\Psi_\alpha/n}.
\]

- **The term \(\max_{m \in [M]} \sup_{z \in S_\Delta} \|g_z(\omega_m)\|_{\Psi_\gamma}\)**: It shows up in the exponential bound (10). We invoke the maximal inequality for the \(\gamma\)-Orlicz norm (item (viii) of Section 4) with the previous estimate to obtain
\[
\max_{m \in [M]} \sup_{z \in S_\Delta} \|g_z(\omega_m)\|_{\Psi_\gamma} \leq \left[ \log(1 + M) \right]^{n/\alpha} \left[ \|\omega\|_2^{p+q} + \Lambda \left\| \cdot \right\|_2^{p+q} \right]_{\Psi_\alpha/n}.
\]
The term $\mathbb{E} \left[ \max_{m \in [M]} \sup_{z \in S_\Delta} |g_z(\omega_m)| \right]$: It appears in the definition of $c$. Using properties (iii) and (v) of Section 4 and the convexification of $\Psi$, we directly get

$$\mathbb{E} \left[ \max_{m \in [M]} \sup_{z \in S_\Delta} |g_z(\omega_m)| \right] \leq \left( \Psi_{\alpha/n}^{(l)} \right)^{-1} \left( 1 \right) \left[ \log(1 + M) \right]^{n/\alpha} \left\| \omega \right\|_{2}^{p+q} + \Lambda \left[ \left\| \cdot \right\|_{2}^{p+q} \right] \left\| \Psi_{\alpha/n} \right\| .$$

Collecting the different bounds we obtain (12) by setting

$$C_{\text{RFF}}(n) := \max \left( 8eC_D \sqrt{d} \sqrt{E_{\omega \sim \Lambda} \left[ L^2(\omega) \right]} , E_{\omega \sim \Lambda} \left[ \left\| \omega \right\|_{2}^{p+q} \right] , \right) \left( 1 \lor \left( \left( \Psi_{\alpha/n}^{(l)} \right)^{-1} \left( 1 \right) \left[ \log(3/2) \right]^{-n/\alpha} \left\| \omega \right\|_{2}^{p+q} \right) \left\| \Psi_{\alpha/n} \right\| \right).$$

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

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 $C_{\text{RFF}}$ does not depend on $S_\Delta$ and $M$, and $K_{\gamma}$ can be bounded uniformly in $n \in [\lceil p + q \rceil , \lceil p + q \rceil + 1]$. The Talagrand constant $C_{\gamma}$ is uniformly bounded w.r.t. $\gamma$ provided that $\gamma$ is bounded away from 0, see the proof of (Talagrand, 1989, Theorem 3).

Now let us prove Corollary 4; set $\varepsilon_M = \frac{\left( \sqrt{6C_{\gamma}} \right) \sqrt{\left( 1 + \left\| (S_\Delta) \right\|_\Lambda \right) \log(1 + M)}}{\sqrt{M}}$. Observe that

(i) $\varepsilon_M$ satisfies the lower bound requirement on $\varepsilon$ in Corollary 3;

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

(iii) therefore

$$1 + \varepsilon_M \left[ \log \left( \frac{\tilde{C}}{\varepsilon_M} \right) \lor \log(1 + M) \right]^{1/\gamma} \leq 1 + \varepsilon_M \left( \left[ \log \left( \frac{\tilde{C}}{\varepsilon_M} \right) \right]^{1/\gamma} + \left[ \log(1 + M) \right]^{1/\gamma} \right) \leq 2 + \varepsilon_M \left[ \log(1 + M) \right]^{1/\gamma}$$

for $M$ large enough;

(iv) $\varepsilon_M \geq \frac{\left( \sqrt{6C_{\gamma}} \right) \sqrt{\log(1 + M)}}{\sqrt{M}}$.

As a consequence of (iii) and (iv), setting $\delta_M := \frac{1 + \left( \log \left( (S_\Delta) \right) \right) \log \left( (S_\Delta) \right) \log(1 + M)}}{\log(1 + M)} \lor 1$, we get (for $M$ large enough)

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

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

On the other hand, using (iv), we easily get \( \frac{(M \varepsilon M)^{\gamma}}{C \log(1+M)} \geq \left[ \frac{(\sqrt{6Ce \log(1+M)^{1/2+1/\gamma}})}{C \log(1+M)} \right]^{\gamma} \geq 2 \log(1+M) \) for \( M \) large enough.

To sum up, in view of (13), we have proved (still for large enough \( M \))

\[
\Lambda^M \left( \| \partial^p q_k - \partial^p q_k \|_{S_M} \geq \varepsilon_M \right) \leq \frac{2}{(1+M)^2} + \frac{1}{(1+M)^2}
\]

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

5.9 Proof of Remark 5(ii)

\( \omega_i \in \mathbb{L}_{\Psi, i} \) 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 \( |\omega|_\alpha := \left( \sum_{i \in [d]} |\omega_i|^{\alpha} \right)^{1/\alpha} \). Then \( |\omega|_2 \leq \sqrt{d} \sup_{i \in [d]} |\omega_i| \leq \sqrt{d} |\omega|_\alpha \) (\( \forall \omega \in \mathbb{R}^d \)). 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 \( |\omega_i|^{\alpha_i} \leq |\omega_i|^{\alpha_i} + 1 \) 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

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

\[
\mathbb{E}_\omega \left[ e^{\tilde{s} |\omega|_2^2} \right] \leq \mathbb{E}_\omega \left[ e^{\tilde{s} \sum_{i \in [d]} (|\omega_i|^{\alpha_i} + 1)} \right] \leq \prod_{i \in [d]} \left( \mathbb{E}_{\omega_i} \left[ e^{e^{s_i |\omega_i|^{\alpha_i}}} \right] \right) \leq \prod_{i \in [d]} \left( \mathbb{E}_{\omega_i} \left[ e^{e^{s_i |\omega_i|^{\alpha_i}}} \right] e^{s} \right) < \infty,
\]

where we used the independence of \( \omega_i \)-s in (*). We got that \( \mathbb{E}_{\omega_i \sim \Lambda} \left[ e^{\tilde{s} |\omega|_2^2} \right] < \infty \) which implies that \( \omega \in \mathbb{L}_{\Psi, \alpha} \).

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 \( \Psi \) is bounded from below by an increasing affine function, \( X \in \mathbb{L}_\Psi \) implies that \( X \) is integrable. Combining (i) with Jensen’s inequality gives \( \Psi \left( \mathbb{E} \left[ \frac{\|X\|_2}{\|X\|_\Psi} \right] \right) \leq \mathbb{E} \left[ \Psi \left( \frac{\|X\|_2}{\|X\|_\Psi} \right) \right] \leq 1 \), 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 $c := (\|X\|_{\Psi, \alpha}^\alpha + \|X'\|_{\Psi, \alpha}^\alpha)^{1/\alpha}$, $p := \frac{\alpha}{\alpha} \|X\|_{\Psi, \alpha}^\alpha$, and $q := \frac{\alpha}{\alpha} \|X'\|_{\Psi, \alpha}^\alpha$, and notice that $\frac{1}{p} + \frac{1}{q} = 1$. Then, combining (23) with $\gamma = \alpha \in (0, 1)$ and the H"older inequality with the conjugate exponents $(p, q)$ yields

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

Therefore, $X + X' \in L_{\Psi, \alpha}$ and $\|X + X'\|_{\Psi, \alpha} \leq c$ by the definition of the $\alpha$-Orlicz norm. Applying (23) with $\gamma = 1/\alpha$ we get $c \leq 2^{(\frac{1}{\alpha}-1)} (\|X\|_{\Psi, \alpha} + \|X'\|_{\Psi, \alpha})$ 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 $\Psi_{\alpha}$ 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

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

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 $H(y) = \Psi(x_0)\Psi(x_0xy) - \Psi(x_0x) \Psi(x_0).$ One has

$$
H'(y) = x_0x\Psi(x_0)e^{x_0xy} - x_0\Psi(x_0x)e^{x_0y} \\
= x_0^2 e^{x_0} e^{x_0x} \left[ \frac{\Psi(x_0)}{x_0 e^{x_0}} e^{x_0(y-1)} - \frac{\Psi(x_0x)}{x_0 e^{x_0}} e^{x_0(y-1)} \right],
$$

$$
\Psi(x_0) \frac{1 - e^{-x_0}}{x_0} = \int_0^1 e^{-ux_0} du \geq \int_0^1 e^{-ux_0} x_0 du = \Psi(x_0x) x_0 e^{x_0x},
$$

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 slight modifications. The inequality (42) can be rewritten as

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

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

30
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 $\|\cdot\|_{\Psi_a}$ is defined analogously to $\mathbb{R}^d$ by changing $\|\cdot\|_2$ to $\|\cdot\|_{B}$.

**Theorem 6 (Hoffman–Jorgensen inequality, Ledoux and Talagrand (2013), Proposition 6.6)** 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 $m \in [M]$, $t_0 = \inf \{ t > 0 : \mathbb{P} (\max_{1 \leq m \leq M} \|S_m\|_B > t) \leq (2 \times 4^p)^{-1} \}$. Then

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

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

$$
\left\| \sum_{m \in [M]} Y_m \right\|_{\Psi_a} \leq C_\gamma \left( \left\| \sum_{m \in [M]} Y_m \right\|_{L_1(B)} + \max_{m \in [M]} \|Y_m\|_B \right). \quad (44)
$$
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 $F$ a countable set of $f := (f_1, \ldots, f_M)$ measurable functions from $B$ into $[-c,c]^M$ such that $\mathbb{E}[f_m(X_m)] = 0$ for all $m \in [M]$. Define $Z := \sup_{f \in F} \frac{1}{M} \sum_{m \in [M]} f_m(X_m)$, $\sigma^2 := \frac{1}{M} \sup_{f \in F} \mathbb{E} \left[ \sum_{m \in [M]} f_m^2(X_m) \right]$. Then, for any $t \geq 0$ the following right and left-hand sided deviation inequalities hold

$$
P(Z - \mathbb{E}[Z] \geq t) \leq e^{-\frac{M^2 t^2}{2(\sigma^2 + 2c \mathbb{E}[Z]) + 2ct}}, \quad P(Z - \mathbb{E}[Z] \leq -t) \leq e^{-\frac{M^2 t^2}{2(\sigma^2 + 2c \mathbb{E}[Z]) + 2ct}}.$$

Theorem 9 (Dudley entropy integral bound) 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

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

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

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

$$(1 - e^{-\beta_\gamma x})^{\frac{1}{\gamma}} \leq \frac{I_\gamma(x)}{\Gamma(1+1/\gamma)} \leq (1 - e^{-x})^{\frac{1}{\gamma}}.$$

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


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 $x = 0$. 


