UNIVERSAL APPROXIMATION PROPERTY OF BANACH SPACE-VALUED RANDOM FEATURE MODELS INCLUDING RANDOM NEURAL NETWORKS

ARIEL NEUFELD AND PHILIPP SCHMOCKER

ABSTRACT. We introduce a Banach space-valued extension of random feature learning, a data-driven supervised machine learning technique for large-scale kernel approximation. By randomly initializing the feature maps, only the linear readout needs to be trained, which reduces the computational complexity substantially. Viewing random feature models as Banach space-valued random variables, we prove a universal approximation result in the corresponding Bochner space. Moreover, we derive approximation rates and an explicit algorithm to learn an element of the given Banach space by such models.

The framework of this paper includes random trigonometric/Fourier regression and in particular random neural networks which are single-hidden-layer feedforward neural networks whose weights and biases are randomly initialized, whence only the linear readout needs to be trained. For the latter, we can then lift the universal approximation property of deterministic neural networks to random neural networks, even within function spaces over non-compact domains, e.g., weighted spaces, L^p-spaces, and (weighted) Sobolev spaces, where the latter includes the approximation of the (weak) derivatives.

In addition, we analyze when the training costs for approximating a given function grow polynomially in both the input/output dimension and the reciprocal of a pre-specified tolerated approximation error. Furthermore, we demonstrate in a numerical example the empirical advantages of random feature models over their deterministic counterparts.

1. INTRODUCTION

The *random feature model* is an architecture for the data-driven approximation of functions between finite dimensional Euclidean spaces, which was introduced by Rahimi and Recht in [67, 68, 69] building on earlier instances in [10, 60, 79]. It can be seen as one of the simplest supervised machine learning technique: By randomly initializing the inner parameters of the model, only the linear readout needs to be trained, which reduces the computational complexity substantially. In this paper, we introduce a Banach space-valued extension of this architecture, which returns for every random initialization the corresponding model as an element of the given Banach space, allowing us to learn infinite dimensional objects with random features. Some examples include, but are not limited to, random trigonometric/Fourier regression (see [6, 67]), kernel regression tasks (see [5, 8, 67]), Gaussian processes (see [60, 70, 79]), random neural networks (see [31, 32, 43]), operator-valued kernels (see [4, 17, 56, 82]), and random operator learning (see [14, 61, 62]).

Originally, random feature learning was introduced to overcome the computational limitations of traditional kernel methods. These kernel methods map the input data into a high-dimensional feature space to capture the nonlinear input/output relation. Even though the explicit form of this feature map is often unknown, one can still compute the Gram matrix whose entries are given as the inner products of features between all pairs of data points. These inner products can be efficiently calculated using the "kernel trick", even for infinite dimensional feature spaces, but the computational costs increase quadratically in the number of samples. This motivated Rahimi and Recht to explore kernel approximation through random features (see [67]) and to extend this approach to shallow architectures (see [68, 69]). Indeed, by replacing the optimization of the non-linear feature maps with randomization, the explicit calculation of the Gram matrix can be avoided, which reduces the computational complexity and enables the application of kernel-based methods to large-scale datasets.

Subsequently, different works have contribued to the mathematical theory of random feature learning. Rahimi and Recht established in [67, 68, 69] the connection to reproducing kernel Hilbert spaces (RKHS) and proved the approximation rate $\mathcal{O}(1/\sqrt{N})$, where $N \in \mathbb{N}$ denotes the number of features.

Date: October 23, 2024.

Key words and phrases. Random feature learning, random neural networks, machine learning, supervised learning, universal approximation, approximation rates, reservoir computing, law of large numbers.

Financial support by the Nanyang Assistant Professorship Grant (NAP Grant) *Machine Learning based Algorithms in Finance and Insurance* is gratefully acknowledged.

Subsequently, [72] showed that $J := \mathcal{O}(\sqrt{N}\ln(N))$ random features lead to an L^2-generalization error of $\mathcal{O}(1/N^{1/4})$ when approximating functions between Euclidean spaces. Moreover, [18] learned random features with the stochastic gradient descent algorithm instead of least squares, while [45] connected the infinite-width case to neural tangent kernels (see also [20] for an extension to deep linear neural networks). In addition, [53, 54] showed precise asymptotics of the generalization error including the double descent curve as well as a sharp generalization error under a hypercontractivity assumption.

Our first contribution consists of a comprehensive *universal approximation theorem* for Banach space-valued random feature models, presented in Theorem 3.2. In contrast to traditional kernel approximation, this result ensures the convergence of random features to any (random) element of the given Banach space. Indeed, by assuming that the deterministic feature maps are universal (i.e. the linear span of their image can approximate any element of the Banach space), we can apply the strong law of large numbers for Banach space-valued random variables (see [44, Theorem 3.1.10]) to lift the universality to random feature models. We apply this framework to the following three instances of random feature learning: Random trigonometric features, random Fourier regression, and random neural networks.

Random neural networks are single-hidden-layer feed-forward neural networks whose weights and biases inside the activation function are randomly initialized (see the work [43] on extreme learning machines and in particular the work [32] on random neural networks with ReLU activation). By training only the linear readout, one avoids the non-convex optimization problem for training deterministic neural networks (caused by the training of the weights and biases inside the activation function, see [35, p. 282]) and one can replace the computationally expensive backpropagation (see [58, p. 13]) by, e.g., the more efficient least squares method. Using the universal approximation property of deterministic neural networks (first proven in [22, 42], see also [19, 21, 52, 64, 66]), we obtain a universal approximation theorem for random neural networks which significantly generalizes the results in [32] from the case of ReLU activation function and L^2-spaces (resp. C^0-spaces) to more general non-polynomial activation functions and more general function spaces over non-compact domains, e.g., weighted spaces, L^p-spaces, and (weighted) Sobolev spaces over unbounded domains, where the latter *includes the approximation of the (weak) derivatives*.

Our second contribution are *approximation rates* for learning a (possibly infinite dimensional) element of the given Banach space by a random feature model, presented in Theorem 4.5. To this end, we assume that the element belongs to a specific Barron space in order to represent it as expectation of the random features (see also [7, 10, 27, 28, 48, 68]). Then, by using a symmetrization argument with Rademacher averages and the concept of Banach space types, we obtain the desired approximation rates. In L^2-spaces, these rates allow us then to derive a generalization error for learning via the least squares method.

As a corollary, we obtain approximation rates and generalization errors for learning a function by a random neural network, which turn out to be similar to the approximation rates for *deterministic* neural networks (see e.g. [9, 10, 13, 23, 24, 49, 55, 64, 75]). To this end, we use the ridgelet transform (see [16]) and its distributional extension (see [76]) to represent the function to be approximated as expectation of a random neuron. This approach generalizes the approximation rates and generalization errors in [32, Section 4.2] from random neural networks with ReLU activation to more general activation functions and by including the approximation of the (weak) derivatives. In addition, we analyze the situation when random neural networks overcome the curse of dimensionality in the sense that the computational costs (measured as number of neurons) grow polynomially in both the input/output dimensions and the reciprocal of a pre-specified tolerated approximation error.

The theoretical foundations of this paper are also relevant in scientific computing. In particular, random neural networks have been successfully applied for solving partial differential equations (PDEs) in mathematical physics (see [25, 26, 78, 81]), for quantum neural networks and quantum reservoirs (see [34]), for solving the Black-Scholes PDE in mathematical finance (see [31]), for optimal stopping (see [40]), for learning the hedging strategy via Wiener-Ito chaos expansion (see [63]), for solving path-dependent PDEs in the context of rough volatility (see [46]), for pricing American options (see [80]), and for solving non-linear parabolic PDEs in finance by the random deep splitting method (see [65]).

We complement these numerical examples by learning the heat equation, which shows the empirical advantages of random feature learning over their deterministic counterparts.

1.1. **Outline.**

In Section 2, we introduce a Banach space-valued extension of random feature learning. In Section 3, we show a universal approximation result for random feature models, which is applied to random trigonometric/Fourier features and random neural networks, followed by some approximation rates in Section 4. In Section 5, we use the least squares method to learn a random feature model and prove a generalization error. In Section 6, we provide a numerical example, while all proofs are given in Section 7-10.

1.2. **Notation.** As usual, $\mathbb{N} := \{1,2,3,...\}$ and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ denote the sets of natural numbers, while $\mathbb{R}$ and $\mathbb{C}$ (with imaginary unit $\mathbf{i} := \sqrt{-1} \in \mathbb{C}$) represent the sets of real and complex numbers, respectively. In addition, we define $\lceil r \rceil := \min\{k \in \mathbb{N}_0 : k \geqslant r\}$ for all $r \in [0,\infty)$. Furthermore, for any $z \in \mathbb{C}$, we denote its real and imaginary part as $\mathrm{Re}(z)$ and $\mathrm{Im}(z)$, respectively, whereas its complex conjugate is defined as $\overline{z} := \mathrm{Re}(z) - \mathrm{Im}(z)\mathbf{i}$.

Moreover, for any $m \in \mathbb{N}$, we denote by $\mathbb{R}^m$ (and $\mathbb{C}^m$) the m-dimensional (complex) Euclidean space, equipped with the norm $\|u\| = \sqrt{\sum_{i=1}^m |u_i|^2}$, where we define $\mathrm{Re}(z) := (\mathrm{Re}(z_1),...,\mathrm{Re}(z_m))^\top$ for $z := (z_1,...,z_m)^\top \in \mathbb{C}^m$. In addition, for any $m,n \in \mathbb{N}$, we denote by $\mathbb{R}^{m\times n}$ the vector space of matrices $A := (a_{i,j})_{i=1,...,m}^{j=1,...,n} \in \mathbb{R}^{m\times n}$, equipped with the matrix 2-norm $\|A\| = \sup_{x\in\mathbb{R}^n,\, \|x\|\leqslant 1} \|Ax\|$, where $I_m \in \mathbb{R}^{m\times m}$ represents the identity matrix.

In addition, for a metric space (Θ, d_Θ) and a Banach space $(X, \|\cdot\|_X)$, we denote by $C^0(\Theta;X)$ the vector space of continuous maps $g : \Theta \to X$, equipped with the topology of compact convergence (see [59, p. 283]), while $\mathcal{B}(\Theta)$ is the Borel σ-algebra of (Θ, d_Θ). Moreover, $du : \mathcal{L}(U) \to [0,\infty]$ denotes the Lebesgue measure on U, with $\mathcal{L}(U)$ being the σ-algebra of Lebesgue-measurable subsets of $U \in \mathcal{B}(\mathbb{R}^m)$, where a property is said to hold almost everywhere (a.e.) if it holds everywhere except on a set of Lebesgue measure zero.

Furthermore, for every fixed $m,d \in \mathbb{N}$, $k \in \mathbb{N}_0$, $U \subseteq \mathbb{R}^m$ (open, if $k \geqslant 1$), and $p \in [1,\infty)$, we introduce the following function spaces:

- $C^k(U;\mathbb{R}^d)$ denotes the vector space of k-times continuously differentiable functions $f : U \to \mathbb{R}^d$ such that the partial derivative $U \ni u \mapsto \partial_\alpha f(u) := \frac{\partial^{|\alpha|} f}{\partial u_1^{\alpha_1} \cdots \partial u_m^{\alpha_m}}(u) \in \mathbb{R}^d$ is continuous for all $\alpha \in \mathbb{N}_{0,k}^m := \{\alpha := (\alpha_1,...,\alpha_m) \in \mathbb{N}_0^m : |\alpha| := \alpha_1 + ... + \alpha_m \leqslant k\}$. If $m = 1$, we denote the derivatives by $f^{(j)} := \frac{\partial^j f}{\partial u^j} : U \to \mathbb{R}^d$, $j = 0,...,k$.
- $C_b^k(U;\mathbb{R}^d)$ denotes the vector space of functions $f \in C^k(U;\mathbb{R}^d)$ such that $\partial_\alpha f : U \to \mathbb{R}^d$ is bounded for all $\alpha \in \mathbb{N}_{0,k}^m$. Then, the norm $\|f\|_{C_b^k(U;\mathbb{R}^d)} := \max_{\alpha\in\mathbb{N}_{0,k}^m} \sup_{u\in U} \|\partial_\alpha f(u)\|$ turns $C_b^k(U;\mathbb{R}^d)$ into a Banach space. Note that for $k = 0$ and $U \subset \mathbb{R}^m$ compact, we obtain the Banach space of continuous functions $(C^0(U;\mathbb{R}^d), \|\cdot\|_{C^0(U;\mathbb{R}^d)})$ equipped with the supremum norm $\|f\|_{C^0(U;\mathbb{R}^d)} := \|f\|_{C_b^0(U;\mathbb{R}^d)} = \sup_{u\in U} \|f(u)\|$.
- $C_{pol,\gamma}^k(U;\mathbb{R}^d)$, with $\gamma \in [0,\infty)$, denotes the vector space of functions $f \in C^k(U;\mathbb{R})$ such that $\|f\|_{C_{pol,\gamma}^k(U;\mathbb{R}^d)} := \max_{\alpha\in\mathbb{N}_{0,k}^m} \sup_{u\in U} \frac{\|\partial_\alpha f(u)\|}{(1+\|u\|)^\gamma} < \infty$.
- $\overline{C_b^k(U;\mathbb{R}^d)}^\gamma$, with $\gamma \in (0,\infty)$, is defined as the closure of $C_b^k(U;\mathbb{R}^d)$ with respect to $\|\cdot\|_{C_{pol,\gamma}^k(U;\mathbb{R}^d)}$. Then, $(\overline{C_b^k(U;\mathbb{R}^d)}^\gamma, \|\cdot\|_{C_{pol,\gamma}^k(U;\mathbb{R}^d)})$ is by definition a Banach space. If $U \subseteq \mathbb{R}^m$ is bounded, then $\overline{C_b^k(U;\mathbb{R}^d)}^\gamma = C_b^k(U;\mathbb{R}^d)$. Otherwise, $f \in \overline{C_b^k(U;\mathbb{R}^d)}^\gamma$ if and only if $f \in C^k(U;\mathbb{R}^d)$ and $\lim_{r\to\infty} \max_{\alpha\in\mathbb{N}_{0,k}^m} \sup_{u\in U,\, \|u\|\geqslant r} \frac{\|\partial_\alpha f(u)\|}{(1+\|u\|)^\gamma} = 0$ (see [64, Lemma 4.1]).
- $C_c^\infty(U;\mathbb{R}^d)$, with $U \subseteq \mathbb{R}^m$ open, denotes the vector space of smooth functions $f : U \to \mathbb{R}^d$ with $\mathrm{supp}(f) \subseteq U$, where $\mathrm{supp}(f)$ is defined as the closure of $\{u \in U : f(u) \neq 0\}$ in $\mathbb{R}^m$.
- $\mathcal{S}(\mathbb{R}^m;\mathbb{C})$ denotes the Schwartz space consisting of smooth functions $f : \mathbb{R}^m \to \mathbb{C}$ such that $\max_{\alpha\in\mathbb{N}_{0,n}^m} \sup_{u\in\mathbb{R}^m} \left(1+\|u\|^2\right)^n |\partial_\alpha f(u)| < \infty$, for all $n \in \mathbb{N}_0$. Moreover, its dual space $\mathcal{S}'(\mathbb{R}^m;\mathbb{C})$ consists of continuous linear functionals $T : \mathcal{S}(\mathbb{R}^m;\mathbb{C}) \to \mathbb{C}$ called tempered distributions (see [30, p. 332]). For example, every $f \in C_{pol,\gamma}^k(\mathbb{R}^m)$ defines a tempered distribution $\left(g \mapsto T_f(g) := \int_{\mathbb{R}} f(u)g(u)du\right) \in \mathcal{S}'(\mathbb{R}^m;\mathbb{C})$ (see [30, Equation 9.26]).
- $\mathcal{S}_0(\mathbb{R};\mathbb{C}) \subseteq \mathcal{S}(\mathbb{R};\mathbb{C})$ denotes the vector subspace of functions $f \in \mathcal{S}(\mathbb{R};\mathbb{C})$ such that $\int_{\mathbb{R}} u^j f(u)du = 0$ for all $j \in \mathbb{N}_0$ (see [36, Definition 1.1.1]).
- $L^p(U,\Sigma,\mu;\mathbb{R}^d)$, with (possibly non-finite) measure space (U,Σ,μ), denotes the vector space of (equivalence classes of) $\Sigma/\mathcal{B}(\mathbb{R}^d)$-measurable functions $f : U \to \mathbb{R}^d$ such that $\|f\|_{L^p(U,\Sigma,\mu;\mathbb{R}^d)} := \left(\int_U \|f(u)\|^p \mu(du)\right)^{1/p} < \infty$. Then, $(L^p(U,\Sigma,\mu;\mathbb{R}^d), \|\cdot\|_{L^p(U,\Sigma,\mu;\mathbb{R}^d)})$ is a Banach space (see [73, Chapter 3]).

- $W^{k,p}(U,\mathcal{L}(U),du;\mathbb{R}^d)$ denotes the Sobolev space of (equivalence classes of) k-times weakly differentiable functions $f: U \to \mathbb{R}^d$ such that $\partial_\alpha f \in L^p(U,\mathcal{L}(U),du;\mathbb{R}^d)$ for all $\alpha \in \mathbb{N}^m_{0,k}$. Then, the norm $\Vert f \Vert_{W^{k,p}(U,\mathcal{L}(U),du;\mathbb{R}^d)} := \big(\sum_{\alpha\in\mathbb{N}^m_{0,k}} \int_U \Vert \partial_\alpha f(u) \Vert^p du \big)^{1/p}$ turns $W^{k,p}(U,\mathcal{L}(U),du;\mathbb{R}^d)$ into a Banach space (see [2, Chapter 3]).
- $W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d)$, with $\mathcal{L}(U)/\mathcal{B}(\mathbb{R})$-measurable $w: U \to [0,\infty)$, denotes the weighted Sobolev space of (equivalence classes of) k-times weakly differentiable functions $f: U \to \mathbb{R}^d$ such that $\partial_\alpha f \in L^p(U,\mathcal{L}(U),w(u)du;\mathbb{R}^d)$ for all $\alpha \in \mathbb{N}^m_{0,k}$. Hereby, $w: U \to [0,\infty)$ is called a *weight* if it is a.e. strictly positive. In this case, the norm $\Vert f \Vert_{W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d)} := \big(\sum_{\alpha\in\mathbb{N}^m_{0,k}} \int_U \Vert \partial_\alpha f(u) \Vert^p w(u) du \big)^{1/p}$ turns $W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d)$ into a Banach space (see [50, p. 5]). Moreover, we define $L^p(U,\mathcal{L}(U),w;\mathbb{R}^d) := W^{0,p}(U,\mathcal{L}(U),w;\mathbb{R}^d)$.

Moreover, we follow [30, Chapter 7] and define the (multi-dimensional) Fourier transform of any function $f \in L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du;\mathbb{C}^d)$ as

$$\mathbb{R}^m \ni \xi \quad \mapsto \quad \widehat{f}(\xi) := \int_{\mathbb{R}^m} e^{-\mathbf{i}\xi^\top u} f(u) du \in \mathbb{C}^d. \tag{1}$$

Then, by using [44, Proposition 1.2.2], it holds that

$$\sup_{\xi\in\mathbb{R}^m} \left\Vert \widehat{f}(\xi) \right\Vert = \sup_{\xi\in\mathbb{R}^m} \left\Vert \int_{\mathbb{R}^m} e^{-\mathbf{i}\xi^\top u} f(u) du \right\Vert \leqslant \int_{\mathbb{R}^m} \Vert f(u) \Vert du = \Vert f \Vert_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du;\mathbb{R}^d)}. \tag{2}$$

In addition, the Fourier transform of any tempered distribution $T \in \mathcal{S}'(\mathbb{R}^m;\mathbb{C})$ is defined by $\widehat{T}(g) := T(\widehat{g})$, for $g \in \mathcal{S}(\mathbb{R}^m;\mathbb{C})$ (see [30, Equation 9.28]).

Furthermore, for $r \in [1,\infty)$, a probability space $(\Omega,\mathcal{F},\mathbb{P})$, and a Banach space $(X, \Vert\cdot\Vert_X)$, we denote by $L^r(\Omega,\mathcal{F},\mathbb{P};X)$ the Bochner space of (equivalence classes of) strongly $(\mathbb{P},\mathcal{F})$-measurable maps $F: \Omega \to X$ such that $\Vert F \Vert_{L^r(\Omega,\mathcal{F},\mathbb{P};X)} := \mathbb{E}\big[\Vert F \Vert_X^r\big]^{1/r} < \infty$. This norm turns $L^r(\Omega,\mathcal{F},\mathbb{P};X)$ into a Banach space. For more details, we refer to [44, Section 1.2.b].

Moreover, we use the Landau notation: $a_n = \mathcal{O}(b_n)$ (as $n \to \infty$) if $\limsup_{n\to\infty} \frac{|a_n|}{|b_n|} < \infty$.

2. Random feature learning

We now present a Banach space-valued extension of the random feature learning architecture introduced by Rahimi and Recht in [67, 68, 69]. Our approach imposes no specific structure on the random features (e.g. sine/cosine or Fourier), nor does it assume that the Banach space is a particular function space.

To this end, we fix throughout this paper a probability space $(\Omega,\mathcal{F},\mathbb{P})$, a metric space (Θ,d_Θ) representing the parameter space, and a separable Banach space $(X,\Vert\cdot\Vert_X)$ over a field $\mathbb{K}$ (either $\mathbb{K} := \mathbb{R}$ or $\mathbb{K} := \mathbb{C}$), which contains the elements to learn. Moreover, we assume the existence of an independent identically distributed (i.i.d.) sequence of Θ-valued random variables $(\theta_n)_{n\in\mathbb{N}}: \Omega \to \Theta$. Then, by inserting these random initializations $(\theta_n)_{n\in\mathbb{N}}$ into the feature maps taken from a given set $\mathcal{G} \subseteq C^0(\Theta;X)$, we only need to train the linear readout that is assumed to be measurable with respect to the σ-algebra $\mathcal{F}_\theta := \sigma(\{\theta_n : n \in \mathbb{N}\})$.

Definition 2.1. *For given $\mathcal{G} \subseteq C^0(\Theta;X)$, a* random feature model (RF) (with respect to $\mathcal{G}$) *is of the form*

$$\Omega \ni \omega \quad \mapsto \quad G(\omega) := \sum_{n=1}^N y_n(\omega) g_n(\theta_n(\omega)) \in X \tag{3}$$

with respect to some $N \in \mathbb{N}$ denoting the number of features, *where $g_1,\dots,g_N \in \mathcal{G}$ are the* feature maps, *and where the* linear readouts *$y_1,\dots,y_N: \Omega \to \mathbb{K}$ are assumed to be $\mathcal{F}_\theta/\mathcal{B}(\mathbb{K})$-measurable.*

For a given set of feature maps $\mathcal{G} \subseteq C^0(\Theta;X)$, we denote by $\mathcal{RG}$ the set of all random feature models (RFs) of the form (3).

Remark 2.2. *Let us briefly explain how the random feature model $G \in \mathcal{RG}$ in (3) can be implemented. For the random initialization of $(\theta_n)_{n=1,\dots,N}$, we draw some $\omega \in \Omega$ and fix the values of $\theta_1(\omega),\dots,\theta_N(\omega) \in \Theta$. Thus, by using that $y_1,\dots,y_N: \Omega \to \mathbb{K}$ are $\mathcal{F}_\theta/\mathcal{B}(\mathbb{K})$-measurable, the training of $G \in \mathcal{RG}$ consists of finding the optimal $y_1(\omega),\dots,y_N(\omega) \in \mathbb{K}$ given $g_1(\theta_1(\omega)),\dots,g_N(\theta_N(\omega)) \in X$. This can be achieved, e.g., by the least squares method.*

In the following, we give an overview of several applications of this general framework, including random trigonometric/Fourier regression and random neural networks.

2.1. **Random trigonometric features.** Introduced in [67, 68], random trigonometric regression uses trigonometric functions (i.e. sines and cosines) in the feature maps. More precisely, for a compact subset $U \subset \mathbb{R}^m$, we consider the real Banach space $(X, \|\cdot\|_X) := (C^0(U), \|\cdot\|_{C^0(U)})$ and the parameter space $\Theta := \mathbb{R}^m$, where $(\theta_n)_{n\in\mathbb{N}} : \Omega \to \mathbb{R}^m$ denotes the i.i.d. sequence. Then, by choosing $\mathcal{G} := \big\{\mathbb{R}^m \ni \vartheta \mapsto h\big(\vartheta^\top \cdot\big) \in C^0(U) : h \in \{\cos, \sin\}\big\}$, we obtain the following random trigonometric feature model.[1]

Definition 2.3. *A* random trigonometric feature model (RTF) *is of the form*

$$\Omega \ni \omega \quad \mapsto \quad G(\omega) := \sum_{n=1}^N \left(y_n^{(1)}(\omega) \cos\left(\theta_n(\omega)^\top \cdot\right) + y_n^{(2)}(\omega) \sin\left(\theta_n(\omega)^\top \cdot\right) \right) \in C^0(U) \tag{4}$$

with respect to some $N \in \mathbb{N}$ denoting the number of trigonometric features*, where the* linear readouts $y_1^{(1)}, ..., y_N^{(1)}, y_1^{(2)}, ..., y_N^{(2)} : \Omega \to \mathbb{R}$ *are assumed to be $\mathcal{F}_\theta/\mathcal{B}(\mathbb{R})$-measurable.*

We denote by $\mathcal{RT}_{U,1}$ the set of all RTFs of the form (4). Moreover, we could also consider multidimensional extensions $\mathcal{RT}_{U,d}$ with $\mathbb{R}^d$-valued linear readouts.

2.2. **Random Fourier features.** Introduced in [6, 67, 68], random Fourier regression uses the Fourier transform as feature map. For a compact subset $U \subset \mathbb{R}^m$, we consider the complex Banach space $(X, \|\cdot\|_X) := (C^0(U;\mathbb{C}), \|\cdot\|_{C^0(U;\mathbb{C})})$ and the parameter space $\Theta := \mathbb{R}^m$, where $(\theta_n)_{n\in\mathbb{N}} : \Omega \to \mathbb{R}^m$ denotes the i.i.d. sequence. Moreover, we let $\mathcal{G}$ consist of the single feature map $\mathbb{R}^m \ni \vartheta \mapsto \exp\left(\mathbf{i}\vartheta^\top \cdot\right) \in C^0(U;\mathbb{C})$ to obtain random Fourier features.[2]

Definition 2.4. *A* random Fourier feature model (RFF) *is of the form*

$$\Omega \ni \omega \quad \mapsto \quad G(\omega) := \sum_{n=1}^N y_n(\omega) \exp\left(\mathbf{i}\theta_n^\top \cdot\right) \in C^0(U;\mathbb{C}) \tag{5}$$

with respect to some $N \in \mathbb{N}$ denoting the number of Fourier features*, where the* linear readouts $y_1, ..., y_N : \Omega \to \mathbb{C}$ *are assumed to be $\mathcal{F}_\theta/\mathcal{B}(\mathbb{C})$-measurable.*

We denote by $\mathcal{RF}_{U,1}$ the set of RFFs of the form (5). Moreover, we could consider vector-valued versions $\mathcal{RF}_{U,d}$ or Banach spaces containing $C^0(U;\mathbb{C})$ (e.g. certain L^2-spaces).

2.3. **Random neural networks.** As third particular instance of random feature learning, we consider random neural networks that are defined as single-hidden-layer feed-forward neural networks whose weights and biases inside the activation function are randomly initialized. Hence, only the linear readout needs to be trained (see [31, 43]).

To this end, we fix the input and output dimension $m, d \in \mathbb{N}$, the order of differentiability $k \in \mathbb{N}_0$, the domain $U \subseteq \mathbb{R}^m$ (open, if $k \geqslant 1$), and some $\gamma \in (0, \infty)$. Then, we consider the Banach space $(X, \|\cdot\|_X) := (\overline{C_b^k(U;\mathbb{R}^d)}^\gamma, \|\cdot\|_{\overline{C_b^k(U;\mathbb{R}^d)}^\gamma})$ introduced in Section 1.2 and the parameter space $\Theta := \mathbb{R}^m \times \mathbb{R}$, where $(\theta_n)_{n\in\mathbb{N}} := (a_n, b_n)_{n\in\mathbb{N}} : \Omega \to \mathbb{R}^m \times \mathbb{R}$ denotes the i.i.d. sequence of random initializations, which are used for the network weights and biases. Hence, by choosing deterministic (i.e. fully trained) neural networks as feature maps, i.e. by setting $\mathcal{G} := \big\{\mathbb{R}^m \times \mathbb{R} \ni (\vartheta_1, \vartheta_2) \mapsto e_i \rho\big(\vartheta_1^\top \cdot - \vartheta_2\big) \in \overline{C_b^k(U;\mathbb{R}^d)}^\gamma : i = 1, ..., d\big\}$, we obtain random neural networks as particular instance of random feature learning.[3]

[1] The element $h\big(\vartheta^\top \cdot\big)$ denotes the function $U \ni u \mapsto h\big(\vartheta^\top u\big) \in \mathbb{R}$.

[2] The element $\exp\left(\mathbf{i}\vartheta^\top \cdot\right) \in C^0(U;\mathbb{C})$ denotes the function $U \ni u \mapsto \exp\left(\mathbf{i}\vartheta^\top u\right) \in \mathbb{C}$.

[3] The element $y\rho\big(\vartheta_1^\top \cdot - \vartheta_2\big) \in \overline{C_b^k(U;\mathbb{R}^d)}^\gamma$ denotes the function $U \ni u \mapsto y\rho\big(\vartheta_1^\top u - \vartheta_2\big) \in \mathbb{R}^d$, where $y \in \mathbb{R}^d$. Moreover, $e_i \in \mathbb{R}^d$ denotes the i-th unit vector of $\mathbb{R}^d$.

Definition 2.5. *A* random neural network (RN) *is of the form*

$$\Omega \ni \omega \quad \mapsto \quad G(\omega) = \sum_{n=1}^{N} y_n(\omega)\rho\left(a_n(\omega)^\top \cdot - b_n(\omega)\right) \in \overline{C^k_b(U;\mathbb{R}^d)}^\gamma \tag{6}$$

with respect to some $N \in \mathbb{N}$ *denoting the* number of neurons *and* $\rho \in \overline{C^k_b(\mathbb{R})}^\gamma$ *representing the* activation function. *Hereby,* $a_1,...,a_N : \Omega \to \mathbb{R}^m$ *and* $b_1,...,b_N : \Omega \to \mathbb{R}$ *are the* random network weights *and* random network biases, *respectively, and the* linear readouts $y_1,...,y_N : \Omega \to \mathbb{R}^d$ *are assumed to be* $\mathcal{F}_{a,b}/\mathcal{B}(\mathbb{R}^d)$*-measurable, with* $\mathcal{F}_{a,b} := \sigma\left(\{a_n, b_n : n \in \mathbb{N}\}\right)$.

For a given activation function $\rho \in \overline{C^k_b(\mathbb{R})}^\gamma$, we denote by $\mathcal{RN}^\rho_{U,d}$ the set of all random neural networks (RNs) of the form (6). We refer to Remark 2.2 for the implementation and training of such a random neural network.

2.4. **Further random feature models.** Besides the examples in Section 2.1-2.3, the random feature learning model could also be applied, e.g., to kernel regression tasks (see [5, 8, 67]), Gaussian processes (see [60, 70, 79]), and operator learning (see [14, 57, 61]). However, in this paper, we focus on the three particular instances in Section 2.1-2.3.

3. Universal approximation

In this section, we present our universal approximation results for the Banach space-valued random feature models introduced in Definition 2.1. To this end, we consider for every $r \in [1,\infty)$ the Bochner (sub-)space $L^r(\Omega,\mathcal{F}_\theta,\mathbb{P};X) \subseteq L^r(\Omega,\mathcal{F},\mathbb{P};X)$ of $\mathcal{F}_\theta$-strongly measurable maps $F : \Omega \to X$ such that $\mathbb{E}\left[\|F\|_X^r\right]^{1/r} < \infty$. For more details on Bochner spaces, we refer to [44, Section 1.2.b].

Moreover, we impose the following condition on the distribution of the i.i.d. sequence of random initializations $(\theta_n)_{n\in\mathbb{N}} : \Omega \to \Theta$ inserted in the feature maps $\mathcal{G} \subseteq C^0(\Theta;X)$.

Assumption 3.1 (Full support)**.** *Let* $(\theta_n)_{n\in\mathbb{N}} : \Omega \to \Theta$ *be an i.i.d. sequence such that for every* $\vartheta \in \Theta$ *and* $r > 0$ *it holds that* $\mathbb{P}\left[\{\omega \in \Omega : d_\Theta(\theta_1(\omega),\vartheta) < r\}\right] > 0$.

In addition, we assume that the feature maps $\mathcal{G} \subseteq C^0(\Theta;X)$ are universal in the sense that the linear span $\operatorname{span}_{\mathbb{K}}(\mathcal{G}(\Theta)) := \left\{\sum_{n=1}^N y_n g_n(\vartheta_n) : N \in \mathbb{N},\, g_1,...,g_N \in \mathcal{G},\, \vartheta_1,...,\vartheta_N \in \Theta,\, y_1,...,y_N \in \mathbb{K}\right\}$ over a field $\mathbb{K}$ is dense in X. Then, by using the law of large numbers for Banach space-valued random variables (see [44, Theorem 3.3.10]), random feature models inherit the universality from the deterministic feature maps. The proof is given in Section 7.1.

Theorem 3.2 (Universal approximation)**.** *Let Assumption 3.1 hold and let* $\mathcal{G} \subseteq C^0(\Theta;X)$ *such that* $\operatorname{span}_{\mathbb{K}}(\mathcal{G}(\Theta))$ *is dense in* X*. Moreover, let* $F \in L^r(\Omega,\mathcal{F}_\theta,\mathbb{P};X)$ *for some* $r \in [1,\infty)$*. Then, for every* $\varepsilon > 0$ *there exists some* $G \in \mathcal{RG} \cap L^r(\Omega,\mathcal{F}_\theta,\mathbb{P};X)$ *such that*

$$\|F - G\|_{L^r(\Omega,\mathcal{F},\mathbb{P};X)} := \mathbb{E}\left[\|F - G\|_X^r\right]^{\frac{1}{r}} < \varepsilon.$$

In particular, every element $x \in X$ can be approximated arbitrarily well by a random feature model $G \in \mathcal{RG}$ with respect to the Bochner norm $\|\cdot\|_{L^r(\Omega,\mathcal{F},\mathbb{P};X)}$.

Now, we apply Theorem 3.2 to random trigonometric/Fourier regression and random neural networks considered in Section 2.1-2.3. The corresponding proofs are given in Section 7.1.

3.1. **Random trigonometric features.** Assume the setting of Section 2.1 with Banach space $(X,\|\cdot\|_X) := (C^0(U),\|\cdot\|_{C^0(U)})$ and parameter space $\Theta := \mathbb{R}^m$, where $U \subseteq \mathbb{R}^m$ is compact. Since $\operatorname{span}_{\mathbb{R}}(\mathcal{G}(\Theta)) = \operatorname{span}_{\mathbb{R}}\left(\left\{U \ni u \mapsto h\left(\vartheta^\top u\right) \in \mathbb{R} : h \in \{\cos,\sin\},\, \vartheta \in \mathbb{R}^m\right\}\right)$ forms the trigonometric algebra on U which by the Stone-Weierstrass theorem is dense in $C^0(U)$, we obtain the following corollary of Theorem 3.2.

Corollary 3.3 (Universal approximation)**.** *Let Assumption 3.1 hold and let* $F \in L^r(\Omega,\mathcal{F}_\theta,\mathbb{P};C^0(U))$ *for some* $r \in [1,\infty)$*. Then, for every* $\varepsilon > 0$ *there exists a random trigonometric feature model* $G \in \mathcal{RT}_{U,1} \cap L^r(\Omega,\mathcal{F}_\theta,\mathbb{P};C^0(U))$ *such that*

$$\|F - G\|_{L^r(\Omega,\mathcal{F},\mathbb{P};C^0(U))} := \mathbb{E}\left[\|F - G\|_{C^0(U)}^r\right]^{\frac{1}{r}} < \varepsilon.$$

3.2. **Random Fourier features.** Assume the setting of Section 2.1 with Banach space $(X, \|\cdot\|_X) := (C^0(U;\mathbb{C}), \|\cdot\|_{C^0(U;\mathbb{C})})$ and parameter space $\Theta := \mathbb{R}^m$, where $U \subseteq \mathbb{R}^m$ is compact. Then, by using that $\operatorname{span}_{\mathbb{C}}(\mathcal{G}(\Theta)) = \operatorname{span}_{\mathbb{C}}\left(\left\{U \ni u \mapsto \exp\left(\mathbf{i}\vartheta^\top u\right) \in \mathbb{C} : \vartheta \in \mathbb{R}^m\right\}\right)$ is dense in $C^0(U;\mathbb{C})$, we obtain the following corollary.

Corollary 3.4 (Universal approximation)**.** *Let Assumption 3.1 hold and let $F \in L^r(\Omega, \mathcal{F}_\theta, \mathbb{P}; C^0(U;\mathbb{C}))$ for some $r \in [1,\infty)$. Then, for every $\varepsilon > 0$ there exists a random Fourier feature model $G \in \mathcal{RF}_{U,1} \cap L^r(\Omega, \mathcal{F}_\theta, \mathbb{P}; C^0(U;\mathbb{C}))$ such that*

$$\|F - G\|_{L^r(\Omega,\mathcal{F},\mathbb{P};C^0(U;\mathbb{C}))} := \mathbb{E}\left[\|F - G\|_{C^0(U;\mathbb{C})}^r\right]^{\frac{1}{r}} < \varepsilon.$$

3.3. **Random neural networks.** In view of Theorem 3.2, we obtain a universal approximation result for random neural networks from the universal approximation property of deterministic (i.e. fully trained) neural networks. To this end, we fix the input and output dimension $m, d \in \mathbb{N}$ and consider the following type of function spaces $(X, \|\cdot\|_X)$.

Assumption 3.5. *For $k \in \mathbb{N}_0$, $U \subseteq \mathbb{R}^m$ (open, if $k \geqslant 1$), and $\gamma \in (0,\infty)$, let $(X, \|\cdot\|_X)$ be a Banach space consisting of functions $f : U \to \mathbb{R}^d$ such that the restriction map $(C_b^k(\mathbb{R}^m;\mathbb{R}^d), \|\cdot\|_{C^k_{pol,\gamma}(\mathbb{R}^m;\mathbb{R}^d)}) \ni f \mapsto f|_U \in X$ is a continuous dense embedding.*

Example 3.6 ([64, Example 2.6])**.** *The following function spaces $(X, \|\cdot\|_X)$ satisfy Assumption 3.5:*

(i) *The C_b^k-space $(X, \|\cdot\|_X) := (C_b^k(U;\mathbb{R}^d), \|\cdot\|_{C_b^k(U;\mathbb{R}^d)})$ if $U \subseteq \mathbb{R}^m$ is bounded.*
(ii) *The weighted C^k-space $(X, \|\cdot\|_X) := (\overline{C_b^k(U;\mathbb{R}^d)}^\gamma, \|\cdot\|_{\overline{C_b^k(U;\mathbb{R}^d)}^\gamma})$.*
(iii) *The L^p-space $(X, \|\cdot\|_X) := (L^p(U, \mathcal{B}(U), \mu; \mathbb{R}^d), \|\cdot\|_{L^p(U,\mathcal{B}(U),\mu;\mathbb{R}^d)})$ if $p \in [1,\infty)$ and $\mu : \mathcal{B}(U) \to [0,\infty)$ is a Borel measure with $\int_U (1+\|u\|)^{\gamma p}\mu(du) < \infty$.*
(iv) *The Sobolev space $(X, \|\cdot\|_X) := (W^{k,p}(U, \mathcal{L}(U), du; \mathbb{R}^d), \|\cdot\|_{W^{k,p}(U,\mathcal{L}(U),du;\mathbb{R}^d)})$ if $p \in [1,\infty)$ and $U \subseteq \mathbb{R}^m$ is bounded having the segment property.*[4]
(v) *The weighted Sobolev space $(X, \|\cdot\|_X) := (W^{k,p}(U, \mathcal{L}(U), w; \mathbb{R}^d), \|\cdot\|_{W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d)})$ if $p \in [1,\infty)$, $U \subseteq \mathbb{R}^m$ has the segment property*[4]*, the weight $w : U \to [0,\infty)$ is bounded, $\inf_{u \in B} w(u) > 0$ for all bounded $B \subseteq U$, and $\int_U (1+\|u\|)^{\gamma p} w(u)du < \infty$.*

For the precise definition of these function spaces, we refer to Section 1.2.

Moreover, by using the parameter space $\Theta := \mathbb{R}^m \times \mathbb{R}$ we assume that the random initializations $(\theta_n)_{n\in\mathbb{N}} := (a_n, b_n)_{n\in\mathbb{N}} : \Omega \to \mathbb{R}^m \times \mathbb{R}$ have full support (see also Assumption 3.1).

Assumption 3.7 (Full support)**.** *Let $(a_n, b_n)_{n\in\mathbb{N}} : \Omega \to \mathbb{R}^m \times \mathbb{R}$ be i.i.d. such that for every $(a,b) \in \mathbb{R}^m \times \mathbb{R}$ and $r > 0$ we have $\mathbb{P}[\{\omega \in \Omega : \|(a_1(\omega), b_1(\omega)) - (a,b)\| < r\}] > 0$.*

Then, by using the universal approximation property of deterministic neural networks, i.e. that $\operatorname{span}_{\mathbb{R}}(\mathcal{G}(\Theta)) = \operatorname{span}_{\mathbb{R}}\left(\left\{U \ni u \mapsto e_i\rho\left(\vartheta_1^\top u - \vartheta_2\right) \in \mathbb{R}^d : (\vartheta_1, \vartheta_2) \in \Theta,\ i = 1,\dots,d\right\}\right)$ with non-polynomial activation function $\rho \in \overline{C_b^k(\mathbb{R})}^\gamma$ is dense in X (see [64, Theorem 2.8]), we obtain a universal approximation result for random neural networks.

Corollary 3.8 (Universal approximation)**.** *Let Assumption 3.5+3.7 hold and let $\rho \in \overline{C_b^k(\mathbb{R})}^\gamma$ be non-polynomial. Moreover, let $F \in L^r(\Omega, \mathcal{F}_{a,b}, \mathbb{P}; X)$ for some $r \in [1,\infty)$. Then, for every $\varepsilon > 0$ there exists some random neural network $G \in \mathcal{RN}_{U,d}^\rho \cap L^r(\Omega, \mathcal{F}_{a,b}, \mathbb{P}; X)$ such that*

$$\|F - G\|_{L^r(\Omega,\mathcal{F},\mathbb{P};X)} := \mathbb{E}\left[\|F - G\|_X^r\right]^{\frac{1}{r}} < \varepsilon.$$

In particular, every function $f \in X$ can be approximated arbitrarily well by a random neural network $G \in \mathcal{RG}$ with respect to the Bochner norm $\|\cdot\|_{L^r(\Omega,\mathcal{F},\mathbb{P};X)}$.

Remark 3.9. *Corollary 3.8 extends the universal approximation results in [68, Theorem 3.1], [39, Corollary 2.3], [38, Theorem 2.4.3], and [31, Corollary 3] from particular activation functions and L^2-spaces (resp. C^0-spaces) to more general non-polynomial activation functions and function spaces over non-compact domains, e.g., weighted Sobolev spaces.*

[4] An open subset $U \subseteq \mathbb{R}^m$ is said to have the *segment property* if for every $u \in \partial U := \overline{U} \setminus U$ there exists an open subset $V \subseteq \mathbb{R}^m$ with $u \in V$ and some vector $y \in \mathbb{R}^m \setminus \{0\}$ such that for every $z \in \overline{U} \cap V$ and $t \in (0,1)$ it holds that $z + ty \in U$ (see [2, p. 54]).

4. Approximation rates

In this section, we derive some approximation rates to learn an element $x \in X$ by a random feature model, which relates the number of features needed for a pre-given approximation error. To this end, we assume that the set of feature maps $\mathcal{G} := \{g_1, ..., g_e\}$ consists of finitely many maps $g_1, ..., g_e : \Theta \to X$, where $e \in \mathbb{N}$.

In order to derive the approximation rates, we recall the notion of the *type* of a Banach space $(X, \|\cdot\|_X)$ and refer to [3, Section 6.2] and [51, Chapter 9] for more details.

Definition 4.1. *A Banach space $(X, \|\cdot\|_X)$ is called of* type $t \in [1,2]$ *if there exists a constant $C_X > 0$ such that for every $N \in \mathbb{N}$, $(x_n)_{n=1,...,N} \subseteq X$, and every Rademacher sequence*[5] *$(\epsilon_n)_{n=1,...,N}$ on a (possibly different) probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ it holds that*

$$\widetilde{\mathbb{E}}\left[\left\|\sum_{n=1}^N \epsilon_n x_n\right\|_X^t\right]^{\frac{1}{t}} \leqslant C_X \left(\sum_{n=1}^N \|x_n\|_X^t\right)^{\frac{1}{t}}.$$

Remark 4.2. *Every Banach space $(X, \|\cdot\|_X)$ is of type $t = 1$ (with $C_X = 1$), every Hilbert space $(X, \|\cdot\|_X)$ is of type $t = 2$ (with $C_X = 1$), and every L^p-space as well as (weighted) $W^{k,p}$-Sobolev space (with $p \in [1,\infty)$ and $k \in \mathbb{N}$) are of type $t = \min(2,p)$ (with constant C_X depending only on $p \in [1,\infty)$), see [3, Remark 6.2.11] and [64, Lemma 4.9].*

In addition, we define the Barron space $\mathbb{B}^r_{\mathcal{G},\theta}(X)$ as all elements $x \in X$ having a representation as expectation of the random feature maps, which is similar to the Barron spaces introduced in [7, 10, 27, 28, 48, 68] in the context of neural networks.

Definition 4.3. *For $r \in [1,\infty)$, $e \in \mathbb{N}$, and $\mathcal{G} := \{g_1, ..., g_e\}$, we define the* Barron space $\mathbb{B}^r_{\mathcal{G},\theta}(X) \subseteq X$ *as the subset of all elements $x \in X$ such that*

$$\|x\|_{\mathbb{B}^r_{\mathcal{G},\theta}(X)} := \inf_{y:\Theta\to\mathbb{R}^e} \mathbb{E}\left[\left\|\sum_{i=1}^e y_i(\theta_1) g_i(\theta_1)\right\|_X^r\right]^{\frac{1}{r}} < \infty, \tag{7}$$

where the infimum is taken over all $\mathcal{B}(\Theta)/\mathcal{B}(\mathbb{R}^e)$-measurable maps $y := (y_1, ..., y_e)^\top : \Theta \to \mathbb{R}^e$ satisfying $x = \mathbb{E}\left[\sum_{i=1}^e y_i(\theta_1) g_i(\theta_1)\right]$. Then, we equip the vector space $\mathbb{B}^r_{\mathcal{G},\theta}(X)$ with the Barron norm *$\|\cdot\|_{\mathbb{B}^r_{\mathcal{G},\theta}(X)}$ defined in (7).*

Remark 4.4. *Note that $\|\cdot\|_{\mathbb{B}^r_{\mathcal{G},\theta}(X)}$ satisfies the norm axioms. Moreover, by using Hölder's inequality, we observe that $\mathbb{B}^{r_2}_{\mathcal{G},\theta}(X) \subseteq \mathbb{B}^{r_1}_{\mathcal{G},\theta}(X)$ for all $1 \leqslant r_1 \leqslant r_2 < \infty$.*

Now, we are able to derive the following approximation rates which are based on Rademacher averages and the Banach space type. The proof can be found in Section 8.1.

Theorem 4.5 (Approximation rates)**.** *Let $(X, \|\cdot\|_X)$ be a separable Banach space of type $t \in [1,2]$ (with constant $C_X > 0$), let $(\theta_n)_{n\in\mathbb{N}} : \Omega \to \Theta$ be i.i.d., let $\mathcal{G} := \{g_1, ..., g_e\}$ consist of $\mathcal{B}(\Theta)/\mathcal{B}(X)$-measurable maps $g_1, ..., g_e : \Theta \to X$, and let $r \in [1,\infty)$. Then, there exists a constant $C_{r,t} > 0$ (depending only on $r \in [1,\infty)$ and $t \in [1,2]$) such that for every $x \in \mathbb{B}^r_{\mathcal{G},\theta}(X)$ and $N \in \mathbb{N}$ there exists $G_N \in \mathcal{RG} \cap L^r(\Omega, \mathcal{F}_\theta, \mathbb{P}; X)$ with N features satisfying*

$$\mathbb{E}\left[\|x - G_N\|_X^r\right]^{\frac{1}{r}} \leqslant C_{r,t} C_X \frac{\|x\|_{\mathbb{B}^r_{\mathcal{G},\theta}(X)}}{N^{1-\frac{1}{\min(r,t)}}}. \tag{8}$$

Hence, Theorem 4.5 relates the approximation error (right-hand side of (8)) to the number of features $N \in \mathbb{N}$ needed for the random feature model $G_N \in \mathcal{RG} \cap L^r(\Omega, \mathcal{F}_\theta, \mathbb{P}; X)$.

In the following, we apply Theorem 4.5 to random trigonometric/Fourier regression and random neural networks considered in Section 2.1-2.3. The proofs are given in Section 8.2.

[5] A *Rademacher sequence* $(\epsilon_n)_{n=1,...,N}$ on $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ are i.i.d. random variables with $\widetilde{\mathbb{P}}[\epsilon_1 = \pm 1] = 1/2$.

4.1. **Random trigonometric features.** To obtain rates with random trigonometric regression in a weighted Sobolev space $W^{k,p}(U,\mathcal{L}(U),w)$, we fix the following quantities.

Assumption 4.6. *Let $k \in \mathbb{N}_0$, $p,r \in [1,\infty)$, $U \subseteq \mathbb{R}^m$ (open, if $k \geqslant 1$), and let $w: U \to [0,\infty)$ be a weight satisfying $w(U) := \int_U w(u)du < \infty$.*

Moreover, we let $\Theta := \mathbb{R}^m$ be the parameter space and consider the two features maps $\mathbb{R}^m \ni \vartheta \mapsto g_1(\vartheta) := \cos\left(\vartheta^\top \cdot\right) \in W^{k,p}(U,\mathcal{L}(U),w)$ and $\mathbb{R}^m \ni \vartheta \mapsto g_2(\vartheta) := \sin\left(\vartheta^\top \cdot\right) \in W^{k,p}(U,\mathcal{L}(U),w)$. In addition, we impose the following condition on $(\theta_n)_{n\in\mathbb{N}}$.

Assumption 4.7. *Let $(\theta_n)_{n\in\mathbb{N}} : \Omega \to \mathbb{R}^m$ be an i.i.d. sequence of random variables, each of them having a strictly positive probability density function $p_\theta : \mathbb{R}^m \to (0,\infty)$.*

Then, we use the real and imaginary part of the Fourier transform as linear readouts to obtain a representation of a given $f \in L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du)$ in terms of the random sine and cosine features. This then implies the following result as a corollary of Theorem 4.5.

Corollary 4.8 (Approximation rates)**.** *Let Assumption 4.6+4.7 hold. Then, there exists a constant $C_{p,r} > 0$ (depending only on $p,r \in [1,\infty)$) such that for every $f \in W^{k,p}(U,\mathcal{L}(U),du) \cap L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du)$ with $C_f := \left(\int_{\mathbb{R}^m} \frac{|\widehat{f}(\vartheta)|^r\left(1+\|\vartheta\|^2\right)^{kr/2}}{p_\theta(\vartheta)^{r-1}} d\vartheta\right)^{1/r} < \infty$ and every $N \in \mathbb{N}$ there exists some random trigonometric feature model $G_N \in \mathcal{RT}_{U,1} \cap L^r(\Omega,\mathcal{F}_\theta,\mathbb{P};W^{k,p}(U,\mathcal{L}(U),w))$ with N features satisfying*

$$\mathbb{E}\left[\|f-G_N\|^r_{W^{k,p}(U,\mathcal{L}(U),w)}\right]^{\frac{1}{r}} \leqslant C_{p,r}\frac{m^{\frac{k}{p}}w(U)^{\frac{1}{p}}}{(2\pi)^m}\frac{C_f}{N^{1-\frac{1}{\min(2,p,r)}}}.$$

4.2. **Random Fourier features.** For approximation rates with random Fourier regression in a weighted Sobolev space as above, we let $\Theta := \mathbb{R}^m$ be the parameter space and consider the single feature map $\mathbb{R}^m \ni \vartheta \mapsto g(\vartheta) := \exp\left(\mathbf{i}\vartheta^\top \cdot\right) \in W^{k,p}(U,\mathcal{L}(U),w;\mathbb{C})$.

Corollary 4.9 (Approximation rates)**.** *Let Assumption 4.6+4.7 hold. Then, there exists a constant $C_{p,r} > 0$ (depending only on $p,r \in [1,\infty)$) such that for every $f \in W^{k,p}(U,\mathcal{L}(U),w;\mathbb{C}) \cap L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du;\mathbb{C})$ with $C_f := \left(\int_{\mathbb{R}^m} \frac{|\widehat{f}(\vartheta)|^r\left(1+\|\vartheta\|^2\right)^{kr/2}}{p_\theta(\vartheta)^{r-1}} d\vartheta\right)^{\frac{1}{r}} < \infty$ and every $N \in \mathbb{N}$ there exists some random Fourier feature model $G_N \in \mathcal{RF}_{U,1} \cap L^r(\Omega,\mathcal{F}_\theta,\mathbb{P};W^{k,p}(U,\mathcal{L}(U),w;\mathbb{C}))$ with N features satisfying*

$$\mathbb{E}\left[\|f-G_N\|^r_{W^{k,p}(U,\mathcal{L}(U),w;\mathbb{C})}\right]^{\frac{1}{r}} \leqslant C_{p,r}\frac{m^{\frac{k}{p}}w(U)^{\frac{1}{p}}}{(2\pi)^m}\frac{C_f}{N^{1-\frac{1}{\min(2,p,r)}}}.$$

Remark 4.10. *For $k=0$ and $p=r=2$, the approximation rate in Corollary 4.8+4.9 coincide with the rate $\mathcal{O}\left(1/\sqrt{N}\right)$ proven in [68] (see also [10]).*

4.3. **Random neural networks.** Finally, we apply Theorem 4.5 to obtain some approximation rates for learning a given function $f: U \to \mathbb{R}^d$ by a random neural network in a weighted Sobolev space $W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d)$, where we fix the following quantities.

Assumption 4.11. *Let $k \in \mathbb{N}_0$, $p,r \in [1,\infty)$, $U \subseteq \mathbb{R}^m$ (open, if $k \geqslant 1$), $\gamma \in [0,\infty)$, and let $w: U \to [0,\infty)$ be a weight.*

Moreover, we recall that $\Theta := \mathbb{R}^m \times \mathbb{R}$ and impose the following condition on the random initializations $(\theta_n)_{n\in\mathbb{N}} := (a_n,b_n)_{n\in\mathbb{N}} : \Omega \to \mathbb{R}^m \times \mathbb{R}$ (see also Assumption 4.7).

Assumption 4.12. *Let $(a_n,b_n)_{n\in\mathbb{N}} : \Omega \to \mathbb{R}^m \times \mathbb{R}$ be an i.i.d. sequence, each of the random variables with strictly positive probability density function $p_{a,b} : \mathbb{R}^m \times \mathbb{R} \to (0,\infty)$.*

To obtain rates for random neural networks, we apply the reconstruction formula in [76, Theorem 5.6] to express a given function as expectation of a random neuron. To this end, we consider admissible pairs $(\psi,\rho) \in \mathcal{S}_0(\mathbb{R};\mathbb{C}) \times C^k_{pol,\gamma}(\mathbb{R})$ of a ridgelet function ψ and an activation function ρ, which is a special case of [76, Definition 5.1] (see also [16]).

Definition 4.13. *A pair $(\psi, \rho) \in \mathcal{S}_0(\mathbb{R};\mathbb{C}) \times C^k_{pol,\gamma}(\mathbb{R})$ is called m-*admissible *if the Fourier transform $\widehat{T_\rho} \in \mathcal{S}'(\mathbb{R};\mathbb{C})$ of $\rho \in C^k_{pol,\gamma}(\mathbb{R})$ (in the sense of distribution) coincides*[6] *on $\mathbb{R}\backslash\{0\}$ with a locally integrable function $f_{\widehat{T_\rho}} : \mathbb{R}\backslash\{0\} \to \mathbb{C}$ such that*

$$C_m^{(\psi,\rho)} := (2\pi)^{m-1} \int_{\mathbb{R}\backslash\{0\}} \frac{\overline{\widehat{\psi}(\xi)} f_{\widehat{T_\rho}}(\xi)}{|\xi|^m} d\xi \in \mathbb{C}\backslash\{0\}.$$

Remark 4.14. *If $(\psi, \rho) \in \mathcal{S}_0(\mathbb{R};\mathbb{C}) \times C^k_{pol,\gamma}(\mathbb{R})$ is m-admissible, then $\rho \in C^k_{pol,\gamma}(\mathbb{R})$ has to be non-polynomial (see [64, Remark 3.2]).*

Together with some appropriate $\psi \in \mathcal{S}_0(\mathbb{R};\mathbb{C})$, the most common activation functions such as the sigmoid function and the ReLU function satisfy Definition 4.13.

Example 4.15 ([64, Example 3.3])**.** *Let $\psi \in \mathcal{S}_0(\mathbb{R};\mathbb{C})$ with non-negative $\widehat{\psi} \in C^\infty_c(\mathbb{R})$ and $\operatorname{supp}(\widehat{\psi}) = [\zeta_1, \zeta_2]$ for some $0 < \zeta_1 < \zeta_2 < \infty$. Then, for every $m \in \mathbb{N}$ and the following activation functions $\rho \in C^k_{pol,\gamma}(\mathbb{R})$ the pair $(\psi, \rho) \in \mathcal{S}_0(\mathbb{R};\mathbb{C}) \times C^k_{pol,\gamma}(\mathbb{R})$ is m-admissible:*

(i) *The sigmoid function $\rho(s) := \frac{1}{1+\exp(-s)}$ if $k \in \mathbb{N}_0$ and $\gamma \geqslant 0$.*
(ii) *The tangens hyperbolicus $\rho(s) := \tanh(s)$ if $k \in \mathbb{N}_0$ and $\gamma \geqslant 0$.*
(iii) *The softplus function $\rho(s) := \ln(1 + \exp(s))$ if $k \in \mathbb{N}_0$ and $\gamma \geqslant 1$.*
(iv) *The ReLU function $\rho(s) = \max(s, 0)$ if $k = 0$ and $\gamma \geqslant 1$.*

Moreover, there exists a constant $C_{\psi,\rho} > 0$ (being independent of $m \in \mathbb{N}$) such that for every $m \in \mathbb{N}$ it holds that $\left|C_m^{(\psi,\rho)}\right| \geqslant C_{\psi,\rho}(2\pi/\zeta_2)^m$.

In addition, we follow [16, 76] and define for every $\psi \in \mathcal{S}_0(\mathbb{R};\mathbb{C})$ the (multi-dimensional) *Ridgelet transform* of any function $f \in L^1(\mathbb{R}^m, \mathcal{L}(\mathbb{R}^m), du; \mathbb{R}^d)$ as

$$\mathbb{R}^m \times \mathbb{R} \ni (a, b) \quad \mapsto \quad (\mathfrak{R}_\psi f)(a, b) := \int_{\mathbb{R}^m} \psi\left(a^\top u - b\right) f(u) \|a\| du \in \mathbb{C}^d. \tag{9}$$

Then, we can apply the reconstruction formula in [76, Theorem 5.6] to obtain a representation of any sufficiently integrable function as expectation of a random neuron.

Proposition 4.16 (Reconstruction, [64, Proposition 3.3])**.** *Let Assumption 4.11+4.12 hold, let $(\psi, \rho) \in \mathcal{S}_0(\mathbb{R};\mathbb{C}) \times C^k_{pol,\gamma}(\mathbb{R})$ be m-admissible, and let $f \in L^1(\mathbb{R}^m, \mathcal{L}(\mathbb{R}^m), du; \mathbb{R}^d)$ with $\widehat{f} \in L^1(\mathbb{R}^m, \mathcal{L}(\mathbb{R}^m), du; \mathbb{C}^d)$. Then, for a.e. $u \in \mathbb{R}^m$, it holds that*

$$\mathbb{E}\left[\frac{(\mathfrak{R}_\psi f)(a_1, b_1)}{p_{a,b}(a_1, b_1)} \rho\left(a_1^\top u - b_1\right)\right] = \int_{\mathbb{R}^m} \int_{\mathbb{R}} (\mathfrak{R}_\psi f)(a, b) \rho\left(a^\top u - b\right) db da = C_m^{(\psi,\rho)} f(u).$$

Remark 4.17. *Recall that the set $\mathcal{G}$ consists of the feature maps $\mathbb{R}^m \times \mathbb{R} \ni (a, b) \mapsto g_i(a, b) := e_i \rho\left(a^\top \cdot - b\right) \in W^{k,p}(U, \mathcal{L}(U), w; \mathbb{R}^d)$, $i = 1, ..., d$. Hence, for every function $f \in W^{k,p}(U, \mathcal{L}(U), w; \mathbb{R}^d)$ satisfying the conditions of Proposition 4.16, we choose the linear readout $\mathbb{R}^m \times \mathbb{R} \ni (a, b) \mapsto y(a, b) := (y_i(a, b))_{i=1,...,d}^\top := \operatorname{Re}\left(\frac{(\mathfrak{R}_\psi f)(a,b)}{C_m^{(\psi,\rho)} p_{a,b}(a,b)}\right) \in \mathbb{R}^d$ to obtain $\mathbb{E}\left[\sum_{i=1}^d y_i(a_1, b_1) g_i(a_1, b_1)\right] = f$ a.e. on U, showing that $f \in \mathbb{B}^r_{\mathcal{G},\theta}(W^{k,p}(U, \mathcal{L}(U), w; \mathbb{R}^d))$.*

In order to extend the reconstruction also to other more general functions, we adapt the Barron spaces in Definition 4.3 to this setting with ridgelet transform introduced in (9).

Definition 4.18. *Let Assumption 4.11+4.12 hold and let $\psi \in \mathcal{S}(\mathbb{R};\mathbb{C})$. Then, we define the Barron-ridgelet space $\widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)$ as vector space of $\mathcal{L}(U)/\mathcal{B}(\mathbb{R}^d)$-measurable functions $f : U \to \mathbb{R}^d$ such that*

$$\|f\|_{\widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)} := \inf_h \mathbb{E}\left[\left\|\frac{\left(1 + \|a_1\|^2\right)^{\frac{\gamma+k}{2}} \left(1 + |b_1|^2\right)^{\frac{\gamma}{2}}}{p_{a,b}(a_1, b_1)} (\mathfrak{R}_\psi h)(a_1, b_1)\right\|^r\right]^{\frac{1}{r}} < \infty,$$

where the infimum is taken over all functions $h \in L^1(\mathbb{R}^m, \mathcal{L}(\mathbb{R}^m), du; \mathbb{R}^d)$ satisfying $\widehat{h} \in L^1(\mathbb{R}^m, \mathcal{L}(\mathbb{R}^m), du; \mathbb{C}^d)$ and $h = f$ a.e. on U.

[6]This means that $\widehat{T_\rho}(g) = \int_{\mathbb{R}\backslash\{0\}} f_{\widehat{T_\rho}}(\xi) g(\xi) d\xi$ for all $g \in C^\infty_c(\mathbb{R}\backslash\{0\};\mathbb{C})$.

In the following lemma, we show for the set of feature maps $\mathcal{G}$ defined in Remark 4.17 that $\widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)$ is a subset of $\mathbb{B}^r_{\mathcal{G},\theta}(W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d))$ introduced in Definition 4.3.

Lemma 4.19. *Let Assumption 4.11+4.12 hold, let $\mathcal{G}$ be as in Remark 4.17, let $(\psi,\rho) \in \mathcal{S}_0(\mathbb{R};\mathbb{C}) \times C^k_{pol,\gamma}(\mathbb{R})$, and define $C^{(\gamma,p)}_{U,w} > 0$ as in (10). Then, for every $f \in \widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)$ it holds that $\|f\|_{\mathbb{B}^r_{\mathcal{G},\theta}(W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d))} \leqslant 2^{3+\frac{1}{p}}\|\rho\|_{C^k_{pol,\gamma}(\mathbb{R})}\frac{C^{(\gamma,p)}_{U,w}m^{k/p}}{|C^{(\psi,\rho)}_m|}\|f\|_{\widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)}$. In particular, we have $\widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d) \subseteq \mathbb{B}^r_{\mathcal{G},\theta}(W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d))$.*

Now, as a corollary of Theorem 4.5, we obtain the following rates to approximate any given function $f \in W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d) \cap \widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)$ by a random neural network.

Corollary 4.20 (Approximation rates)**.** *Let Assumption 4.11+4.12 hold such that*

$$C^{(\gamma,p)}_{U,w} := \left(\int_U (1+\|u\|)^{\gamma p} w(u)du\right)^{\frac{1}{p}} < \infty. \tag{10}$$

Moreover, let $(\psi,\rho) \in \mathcal{S}_0(\mathbb{R};\mathbb{C}) \times C^k_{pol,\gamma}(\mathbb{R})$ be m-admissible. Then, there exists $C_{p,r} > 0$ (depending only on $p,r \in [1,\infty)$) such that for every $f \in W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d) \cap \widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)$ there exists a random neural network $G_N \in \mathcal{RN}^\rho_{U,d} \cap L^r(\Omega,\mathcal{F}_{a,b},\mathbb{P};W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d))$ with N neurons satisfying

$$\mathbb{E}\left[\|f-G_N\|^r_{W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d)}\right]^{\frac{1}{r}} \leqslant C_{p,r}\|\rho\|_{C^k_{pol,\gamma}(\mathbb{R})}\frac{C^{(\gamma,p)}_{U,w}m^{\frac{k}{p}}}{\left|C^{(\psi,\rho)}_m\right|}\frac{\|f\|_{\widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)}}{N^{1-\frac{1}{\min(2,p,r)}}}. \tag{11}$$

Remark 4.21. *Theorem 4.20 extends the approximation rates for random neural networks in [32, Section 4.2] and [31, Theorem 1] from ReLU activation functions and L^2-spaces (resp. C^0-spaces) to more general activation functions and weighted Sobolev spaces, where the approximation of the weak derivatives is now included. Moreover, these rates are analogous to the ones for* deterministic *neural networks in [10, 13, 16, 23, 49, 55, 64, 75].*

Next, we give sufficient conditions for a function to belong to a Barron-ridgelet space introduced in Definition 4.18. For example, the solution of the heat equation (with appropriate initial condition) at any fixed time belongs to such a space (see Corollary 6.2.(ii)).

Proposition 4.22. *Let Assumption 4.11 hold, let[7] $(a_n,b_n)_{n\in\mathbb{N}} \sim p_a \otimes t_1$ be i.i.d., and let $\psi \in \mathcal{S}_0(\mathbb{R};\mathbb{C})$ such that $\zeta_1 := \inf\left\{|\zeta| : \zeta \in \operatorname{supp}(\widehat{\psi})\right\} > 0$. Then, there exists a constant $C_1 > 0$ (being independent of $m,d \in \mathbb{N}$) such that for every $f \in L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du;\mathbb{R}^d)$ with $(\lceil\gamma\rceil+2)$-times differentiable Fourier transform it holds that*

$$\|f\|_{\widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)} \leqslant \frac{C_1}{\zeta_1^{\frac{m}{r}}}\sup_{\zeta\in\operatorname{supp}(\widehat{\psi})}\sum_{\beta\in\mathbb{N}^m_{0,\lceil\gamma\rceil+2}}\left(\int_{\mathbb{R}^m}\|\partial_\beta\widehat{f}(\xi)\|^r\frac{\left(1+\|\xi/\zeta\|^2\right)^{\frac{(2\lceil\gamma\rceil+k+3)r}{2}}}{\theta_A(\xi/\zeta)^{r-1}}d\xi\right)^{\frac{1}{r}}. \tag{12}$$

In particular, if $r=2$ and[7] $(a_n,b_n)_{n\in\mathbb{N}} \sim t_m \otimes t_1$ i.i.d., then it holds for every $f \in L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du;\mathbb{R}^d)$ with $(\lceil\gamma\rceil+2)$-times differentiable Fourier transform that

$$\|f\|_{\widetilde{\mathbb{B}}^{k,2,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)} \leqslant \frac{C_1}{\zeta_1^{\frac{m}{2}}}\frac{\pi^{\frac{m+1}{4}}}{\Gamma\left(\frac{m+1}{2}\right)^{\frac{1}{2}}}\sum_{\beta\in\mathbb{N}^m_{0,\lceil\gamma\rceil+2}}\left(\int_{\mathbb{R}^m}\|\partial_\beta\widehat{f}(\xi)\|^2\left(1+\|\xi/\zeta_1\|^2\right)^{2\lceil\gamma\rceil+k+\frac{m+5}{2}}d\xi\right)^{\frac{1}{2}}. \tag{13}$$

Hence, if the right-hand side of (12) or (13) is finite, we obtain that $f \in \widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)$.

Thus, for $r=2$ and $(a_n,b_n)_{n\in\mathbb{N}} \sim t_m \otimes t_1$ i.i.d., we can insert (13) into the right-hand side of (11) to conclude that the approximation rate for random neural networks is the same as the approximation rate for *deterministic* neural networks proven in [64, Theorem 3.6].

Moreover, the following estimate holds true for the constant $C^{(\gamma,p)}_{U,w} > 0$ appearing in (11), while a lower bound for the constant $\left|C^{(\psi,\rho)}_m\right| > 0$ is given below the list of Example 4.15.

[7] For $m \in \mathbb{N}$, t_m denotes the Student's t-distribution with probability density function $\mathbb{R}^m \ni a \mapsto \theta_A(a) = \frac{\Gamma((m+1)/2)}{\pi^{(m+1)/2}}\left(1+\|a\|^2\right)^{-(m+1)/2} \in (0,\infty)$, where Γ is the Gamma function (see [1, Section 6.1]).

Lemma 4.23 ([64, Lemma 3.9]). *Let Assumption 4.11 hold with the weight $U \ni u := (u_1, ..., u_m)^\top \mapsto w(u) := \prod_{l=1}^m w_0(u_l) \in [0, \infty)$, where $w_0 : \mathbb{R} \to [0, \infty)$ satisfies $\int_{\mathbb{R}} w_0(s) ds = 1$ and $C_{\mathbb{R},w_0}^{(\gamma,p)} := \left(\int_{\mathbb{R}} (1 + |s|)^{\gamma p} w_0(s) ds \right)^{1/p} < \infty$. Then, $C_{U,w}^{(\gamma,p)} \leqslant C_{\mathbb{R},w_0}^{(\gamma,p)} m^{\gamma + 1/p}$.*

In addition, we analyze the situation when the approximation of a function by random neural networks overcomes the curse of dimensionality in the sense that the computational costs (measured as the number of neurons $N \in \mathbb{N}$) grow polynomially in both the input/output dimensions $m, d \in \mathbb{N}$ and the reciprocal of a pre-specified tolerated approximation error.

Proposition 4.24. *Let Assumption 4.11 hold with $r = 2$, $p > 1$, and a weight as in Lemma 4.23, let $(a_n, b_n)_{n \in \mathbb{N}} \sim t_m \otimes t_1$ be i.i.d., and let $(\psi, \rho) \in \mathcal{S}_0(\mathbb{R}; \mathbb{C}) \times C^k_{pol,\gamma}(\mathbb{R})$ be a pair as in Example 4.15 (with $0 < \zeta_1 < \zeta_2 < \infty$). In addition, let $f \in W^{k,p}(U, \mathcal{L}(U), w; \mathbb{R}^d)$ satisfy the conditions of Proposition 4.22 such that the right-hand side of (13) satisfies $\mathcal{O}\left(m^s (2/\zeta_2)^m (m+1)^{m/2} \right)$ for some $s \in \mathbb{N}_0$. Then, there exist some constants $C_2, C_3 > 0$ such that for every $m, d \in \mathbb{N}$ and every $\varepsilon > 0$ there exists a random neural network $G_N \in \mathcal{RN}^\rho_{U,d}$ with $N = \left\lceil C_2 m^{C_3} \varepsilon^{-\frac{\min(2,p,r)}{\min(2,p,r)-1}} \right\rceil$ neurons satisfying*

$$\mathbb{E}\left[\| f - G_N \|^2_{W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d)} \right]^{\frac{1}{2}} \leqslant \varepsilon.$$

Hence, in this case, random neural networks overcome the curse of dimensionality.

5. Least squares and generalization error

In this section, we use the least squares method to learn a given function by a random feature model in the Sobolev space $W^{k,2}(U, \mathcal{L}(U), w; \mathbb{R}^d)$, where we fix some $k \in \mathbb{N}_0$, $U \subseteq \mathbb{R}^m$ (open, if $k \geqslant 1$), and a weight $w : U \to [0, \infty)$ that is *normalized*, i.e. $\int_U w(u) du = 1$. To this end, we assume that the set of feature maps $\mathcal{G} := \{g_1, ..., g_e\}$ consists of finitely many $\mathcal{B}(\Theta)/\mathcal{B}(W^{k,2}(U, \mathcal{L}(U), w; \mathbb{R}^d))$-measurable maps $g_1, ..., g_e : \Theta \to W^{k,2}(U, \mathcal{L}(U), w; \mathbb{R}^d)$, where $e \in \mathbb{N}$.

Moreover, we fix an i.i.d. sequence of U-valued random variables $(V_j)_{j \in \mathbb{N}} \sim w$ as training data, which are independent of the random initializations $(\theta_n)_{n \in \mathbb{N}}$. From this, we define the σ-algebra $\mathcal{F}_{\theta,V} := \sigma(\{\theta_n, V_n : n \in \mathbb{N}\})$ satisfying $\mathcal{F}_\theta \subseteq \mathcal{F}_{\theta,V} \subseteq \mathcal{F}$.

In addition, for every fixed $N \in \mathbb{N}$, we denote by $\mathcal{Y}_N$ the vector space of all $\mathbb{R}^{e \times N}$-valued random variables $y := (y_{i,n})_{i=1,...,e}^{n=1,...,N}$, which are $\mathcal{F}_{\theta,V}/\mathcal{B}(\mathbb{R}^{e \times N})$-measurable. Then, for every $y \in \mathcal{Y}_N$, we define the corresponding random feature model as

$$\Omega \ni \omega \quad \mapsto \quad G_N^y(\omega) := \sum_{n=1}^N \sum_{l=1}^e y_{l,n}(\omega) g_l(\theta_n(\omega)) \in W^{k,2}(U, \mathcal{L}(U), w; \mathbb{R}^d). \tag{14}$$

Note that (14) slightly differs from Definition 2.5 as the linear readout $y \in \mathcal{Y}_N$ is now measurable with respect to $\mathcal{F}_{\theta,V}$ (instead of $\mathcal{F}_\theta$) and can therefore only be trained after the training data $(V_j)_{j \in \mathbb{N}}$ has been drawn. Moreover, we denote by $\mathcal{RG}^V$ the set of all random feature models of the form (14).

For some fixed $J \in \mathbb{N}$, we now approximate a given k-times weakly differentiable function $f : U \to \mathbb{R}^d$ by the least squares method on the training data $(V_j)_{j=1,...,J}$. To this end, we aim for the random feature model $G_N^{y^{(J)}} \in \mathcal{RG}^V$ with linear readout $y^{(J)} := \left(y_{i,n}^{(J)} \right)_{i=1,...,e}^{n=1,...,N} \in \mathcal{Y}_N$ that minimizes the empirical (weighted) mean squared error (MSE), i.e. we set

$$y^{(J)}(\omega) = \arg\min_{y \in \mathcal{Y}_N} \left(\frac{1}{J} \sum_{j=1}^J \sum_{\alpha \in \mathbb{N}^m_{0,k}} c_\alpha^2 \left\| \partial_\alpha f(V_j(\omega)) - \partial_\alpha G_N^y(\omega)(V_j(\omega)) \right\|^2 \right) \tag{15}$$

for all $\omega \in \Omega$. Hereby, the constants $\mathbf{c} := (c_\alpha)_{\alpha \in \mathbb{N}^m_{0,k}} \subset (0, \infty)$ control the contribution of the derivatives, e.g., $c_\alpha := m^{-|\alpha|}$, $\alpha \in \mathbb{N}^m_{0,k}$, means equal contribution of each order. Moreover, for $\mathbf{c} := (c_\alpha)_{\alpha \in \mathbb{N}^m_{0,k}}$, we define the number $\kappa(\mathbf{c}) := \frac{\max_{\alpha \in \mathbb{N}^m_{0,k}} c_\alpha}{\min_{\alpha \in \mathbb{N}^m_{0,k}} c_\alpha}$.

To analyze the complexity of Algorithm 1 in the following result, we count every elementary operation, every function evaluation of $g_1, ..., g_e \in \mathcal{G}$, and each generation of a random variable as one unit. Then, we show the following result whose proof is given in Section 9.1.

Algorithm 1: Least squares method to learn a random feature model

Input: $J, N \in \mathbb{N}$ and a k-times weakly differentiable function $f = (f_1, ..., f_d)^\top : U \to \mathbb{R}^d$.

Output: $G_N^{y^{(J)}} \in \mathcal{RG}^V$ with linear readout $y^{(J)} := \big(y_{l,n}^{(J)}\big)_{l=1,...,e}^{n=1,...,N} \in \mathcal{Y}_N$ solving (15).

1 Generate i.i.d. random variables (RVs) $(\theta_n)_{n=1,...,N}$ satisfying Assumption 4.7.

2 Generate i.i.d. random variables (RVs) $(V_j)_{j=1,...,J} \sim w$ independent of $(\theta_n)_{n=1,...,N}$.

3 Define the $\mathbb{R}^{(J \cdot |\mathbb{N}_{0,k}^m| \cdot d) \times (e \cdot N)}$-valued RV $\mathbf{G} = \big(\mathbf{G}_{(j,\alpha,i),(l,n)}\big)_{(j,\alpha,i) \in \{1,...,J\} \times \mathbb{N}_{0,k}^m \times \{1,...,d\}}^{(l,n) \in \{1,...,e\} \times \{1,...,N\}}$ with $\mathbf{G}_{(j,\alpha,i),(l,n)} := c_\alpha \partial_\alpha g_{l,i}(\theta_n)(V_j)$ for $(j, \alpha, i) \in \{1, ..., J\} \times \mathbb{N}_{0,k}^m \times \{1, ..., d\}$ and $(l, n) \in \{1, ..., e\} \times \{1, ..., N\}$, where $g_l(\vartheta)(u) := (g_{l,i}(\vartheta)(u))_{i=1,...,d}^\top \in \mathbb{R}^d$.

4 Define the $\mathbb{R}^{J \cdot |\mathbb{N}_{0,k}^m| \cdot d}$-valued RV $Z := \big(c_\alpha \partial_\alpha f_i(V_j)\big)_{(j,\alpha,i) \in \{1,...,J\} \times \mathbb{N}_{0,k}^m \times \{1,...,d\}}^\top$.

5 Solve the least squares problem $\mathbf{G}^\top \mathbf{G} \vec{y}^{(J)} = \mathbf{G}^\top Z$ for $\vec{y}^{(J)} := \big(y_{(l,n)}^{(J)}\big)_{(l,n) \in \{1,...,e\} \times \{1,...,N\}}^\top$ via Cholesky decomposition and forward/backward substitution (see [12, Section 2.2.2]).

6 Return $\Omega \ni \omega \mapsto G_N^{y^{(J)}}(\omega) := \sum_{n=1}^N \sum_{l=1}^e y_{l,n}^{(J)}(\omega) g_l(\theta_n(\omega)) \in W^{k,2}(U, \mathcal{L}(U), w; \mathbb{R}^d)$.

Proposition 5.1. *For $e \in \mathbb{N}$, let $\mathcal{G} := \{g_1, ..., g_e\}$ consist of maps $g_1, ..., g_e : \Theta \to W^{k,2}(U, \mathcal{L}(U), w; \mathbb{R}^d)$ that are $\mathcal{B}(\Theta)/\mathcal{B}(W^{k,2}(U, \mathcal{L}(U), w; \mathbb{R}^d))$-measurable, let Assumption 4.7 hold, let $J, N \in \mathbb{N}$, $(c_\alpha)_{\alpha \in \mathbb{N}_{0,k}^m} \subset (0, \infty)$, and $f : U \to \mathbb{R}^d$ be k-times weakly differentiable. Then, Algorithm 1 terminates and is correct, i.e. returns $G_N^{y^{(J)}} \in \mathcal{RG}^V$ with $y^{(J)} \in \mathcal{Y}_N$ solving (15). Moreover, the complexity of Algorithm 1 is of order $\mathcal{O}\big(Jm^k d(eN)^2 + (eN)^3\big)$.*

For fixed $k \in \mathbb{N}_0$, Proposition 5.1 shows that the computational costs for learning a given k-times weakly differentiable function by a random feature model including the derivatives up to order k scales polynomially in $J, N, m, d \in \mathbb{N}$.

Now, we bound the generalization error for learning a function by the random feature model $G_N^{y^{(J)}} \in \mathcal{RG}^V$ obtained from Algorithm 1. Since the linear readout $y^{(J)} \in \mathcal{Y}_N$ minimizes the empirical MSE in (15), the random feature model $G_N^{y^{(J)}} \in \mathcal{RG}^V$ is the best choice on the training data $(V_j)_{j=1,...,J}$. In order to bound the error beyond the training data, we combine the approximation rate in Theorem 4.5 with a result on non-parametric function regression (see [37, Theorem 11.3]). Hereby, we introduce the truncation $\mathbb{R}^d \ni z := (z_1, ..., z_d)^\top \mapsto T_L(z) := (\max(\min(z_i, L), -L))_{i=1,...,d}^\top \in \mathbb{R}^d$. The proof can be found in Section 9.2.

Theorem 5.2 (Generalization error). *For $e \in \mathbb{N}$, let $\mathcal{G} := \{g_1, ..., g_e\}$ consist of $\mathcal{B}(\Theta)/\mathcal{B}(W^{k,2}(U, \mathcal{L}(U), w; \mathbb{R}^d))$-measurable maps $g_1, ..., g_e : \Theta \to W^{k,2}(U, \mathcal{L}(U), w; \mathbb{R}^d)$ and let Assumption 4.7 hold. Then, there exists a constant $C_4 > 0$ (being independent of $m, d \in \mathbb{N}$) such that for every $J, N \in \mathbb{N}$, $L > 0$, and $f := (f_1, ..., f_d)^\top \in \mathbb{B}_{\mathcal{G},\theta}^2(W^{k,2}(U, \mathcal{L}(U), w; \mathbb{R}^d))$ satisfying $|\partial_\alpha f_i(u)| \leqslant L$ for all $\alpha \in \mathbb{N}_{0,k}^m$, $i = 1, ..., d$, and $u \in U$, Algorithm 1 returns a random feature model $G_N^{y^{(J)}} \in \mathcal{RG}^V$ with N features being a strongly $(\mathbb{P}, \mathcal{F}_{\theta,V})$-measurable map $G_N^{y^{(J)}} : \Omega \to W^{k,2}(U, \mathcal{L}(U), w; \mathbb{R}^d)$ such that*

$$\mathbb{E}\left[\sum_{\alpha \in \mathbb{N}_{0,k}^m} \int_U \left\| \partial_\alpha f(u) - T_L\left(\partial_\alpha G_N^{y^{(J)}}(\cdot)(u)\right) \right\|^2 w(u) du\right]^{\frac{1}{2}} \leqslant C_4 L m^{\frac{k}{2}} d^{\frac{1}{2}} \sqrt{\frac{(\ln(J) + 1)N}{J}} + C_4 \kappa(\mathbf{c}) \frac{\|f\|_{\mathbb{B}_{\mathcal{G},\theta}^2(W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d))}}{\sqrt{N}}. \tag{16}$$

Hence, the generalization error in (16) can be made arbitrarily small by first choosing the number of random features $N \in \mathbb{N}$ large enough and then by increasing the sample size $J \in \mathbb{N}$.

Remark 5.3. *Choosing $J = \mathcal{O}\big(\sqrt{N} \ln(N)\big)$ random features in (16), we recover the L^2-generalization error of $\mathcal{O}\big(1/N^{1/4}\big)$ in [72, Theorem 1] for random feature models trained via ridge regression. Previously, [69, Theorem 1] showed an L^2-generalization error of $\mathcal{O}\left(1/J^{1/4} + 1/N^{1/4}\right)$ also by using ridge regression.*

5.1. **Random neural networks.** In this section, we now consider random neural networks as particular instance of random feature learning, where $(\theta_n)_{n\in\mathbb{N}} := (a_n, b_n)_{n\in\mathbb{N}} : \Omega \to \mathbb{R}^m \times \mathbb{R}$ is the random initialization. To this end, we fix some $\gamma \in [0,\infty)$ and an activation function $\rho \in C^k_{pol,\gamma}(\mathbb{R})$. Then, for every fixed $N \in \mathbb{N}$, we denote by $\mathcal{Y}_N$ be the vector space of all $\mathcal{F}_{a,b,V}/\mathcal{B}(\mathbb{R}^{d\times N})$-measurable random variables $y := (y_n)^\top_{n=1,\dots,N} := (y_{i,n})^{n=1,\dots,N}_{i=1,\dots,d}$, where $\mathcal{F}_{a,b,V} := \mathcal{F}_{\theta,V}$. Then, for any $y \in \mathcal{Y}_N$, we define the corresponding random neural network as

$$\Omega \ni \omega \quad \mapsto \quad G^y_N(\omega) := \sum_{n=1}^N y_n \rho\left(a_n^\top \cdot - b_n\right) \in C^k_{pol,\gamma}(U;\mathbb{R}^d). \tag{17}$$

Moreover, we denote by $\mathcal{RN}^{\rho,V}_{U,d}$ the set of all random neural networks of the form (17).

For some fixed $J \in \mathbb{N}$, we now approximate a given k-times weakly differentiable function $f : U \to \mathbb{R}^d$ by the least squares method on the training data $(V_j)_{j=1,\dots,J}$, which corresponds to the random neural network $G^{y^{(J)}}_N \in \mathcal{RN}^{\rho,V}_{U,d}$ with linear readout $y^{(J)} \in \mathcal{Y}_N$ solving (15). In this case, we obtain the following version of Algorithm 1 for random neural networks.

Algorithm 2: Least squares method to learn a random neural network

Input: $J, N \in \mathbb{N}$ and k-times weakly differentiable function $f = (f_1, \dots, f_d)^\top : U \to \mathbb{R}^d$.
Output: $G^{y^{(J)}}_N \in \mathcal{RN}^{\rho,V}_{U,d}$ with linear readout $y^{(J)} := \left(y^{(J)}_{i,n}\right)^{n=1,\dots,N}_{i=1,\dots,d} \in \mathcal{Y}_N$ solving (15).

1 Generate i.i.d. random variables (RVs) $(a_n, b_n)_{n=1,\dots,N} \sim p_{a,b}$ (see Assumption 4.12).
2 Generate i.i.d. random variables (RVs) $(V_j)_{j=1,\dots,J} \sim w$ independent of $(a_n, b_n)_{n=1,\dots,N}$.
3 Define the $\mathbb{R}^{(J\cdot|\mathbb{N}^m_{0,k}|)\times N}$-valued RV $\mathbf{G} = (\mathbf{G}_{(j,\alpha),n})^{n=1,\dots,N}_{(j,\alpha)\in\{1,\dots,J\}\times\mathbb{N}^m_{0,k}}$ with
$\mathbf{G}_{(j,\alpha),n} := c_\alpha \rho^{(|\alpha|)}\left(a_n^\top V_j - b_n\right) a_n^\alpha$ for $(j,\alpha) \in \{1,\dots,J\} \times \mathbb{N}^m_{0,k}$ and $n = 1,\dots,N$.
4 **for** $i = 1,\dots,d$ **do**
5 Define the $\mathbb{R}^{J\cdot|\mathbb{N}^m_{0,k}|}$-valued random variable $Z_i := (c_\alpha \partial_\alpha f_i(V_j))^\top_{(j,\alpha)\in\{1,\dots,J\}\times\mathbb{N}^m_{0,k}}$.
6 Solve the least squares problem $\mathbf{G}^\top\mathbf{G} y^{(J)}_i = \mathbf{G}^\top Z_i$ for $y^{(J)}_i := \left(y^{(J)}_{(i,n)}\right)^\top_{n=1,\dots,N}$ via Cholesky decomposition and forward/backward substitution (see [12, Section 2.2.2]).
7 Return $\Omega \ni \omega \mapsto G^{y^{(J)}}_N(\omega) := \sum_{n=1}^N y^{(J)}_n(\omega)\rho\big(a_n(\omega)^\top \cdot - b_n(\omega)\big) \in C^k_{pol,\gamma}(U;\mathbb{R}^d)$.

Corollary 5.4 (Generalization error)**.** *Let $w : U \to [0,\infty)$ be a normalized weight such that the constant $C^{(\gamma,2)}_{U,w} > 0$ defined in (10) is finite. Moreover, let $(\psi,\rho) \in \mathcal{S}_0(\mathbb{R};\mathbb{C}) \times C^k_{pol,\gamma}(\mathbb{R})$ be m-admissible and let Assumption 4.12 hold. Then, there exists some $C_5 > 0$ (being independent of $m, d \in \mathbb{N}$) such that for every $J, N \in \mathbb{N}$, $L > 0$, and $f := (f_1,\dots,f_d)^\top \in W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d) \cap \widetilde{\mathbb{B}}^{k,2,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)$ satisfying $|\partial_\alpha f_i(u)| \leqslant L$ for all $\alpha \in \mathbb{N}^m_{0,k}$, $i = 1,\dots,d$, and $u \in U$, Algorithm 2 returns a random neural network $G^{y^{(J)}}_N \in \mathcal{RN}^\rho_{U,d}$ with N neurons being a strongly $(\mathbb{P},\mathcal{F}_{a,b,V})$-measurable map $G^{y^{(J)}}_N : \Omega \to W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d)$ such that*

$$\begin{aligned}\mathbb{E}&\left[\sum_{\alpha\in\mathbb{N}^m_{0,k}} \int_U \left\|\partial_\alpha f(u) - T_L\left(\partial_\alpha G^{y^{(J)}}_N(\cdot)(u)\right)\right\|^2 w(u)du\right]^{\frac{1}{2}} \\ &\leqslant C_5 L m^{\frac{k}{2}} d^{\frac{1}{2}} \sqrt{\frac{(\ln(J)+1)N}{J}} + C_5 \kappa(\mathbf{c}) \|\rho\|_{C^k_{pol,\gamma}(\mathbb{R})} \frac{C^{(\gamma,2)}_{U,w} m^{\frac{k}{2}}}{\left|C^{(\psi,\rho)}_m\right|} \frac{\|f\|_{\widetilde{\mathbb{B}}^{k,2,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)}}{\sqrt{N}}.\end{aligned} \tag{18}$$

Remark 5.5. *Corollary 5.4 extends the generalization error in [31, Theorem 4.1] for random neural networks with ReLU activation function to more general activation functions and by including the approximation of the weak derivatives. Moreover, (18) coincides up to constants with the generalization error for* deterministic *neural networks obtained in [11].*

6. Numerical example: Heat equation

In this section[8], we follow [29, Section 2.3] and consider the heat equation, which describes the evolution of a given quantity throughout time. More precisely, we consider the partial differential equation (PDE)

$$\frac{\partial f}{\partial t}(t,u) - \lambda \sum_{l=1}^{m} \frac{\partial^2 f}{\partial u_l^2}(t,u) = 0, \qquad (t,u) \in (0,\infty) \times \mathbb{R}^m, \tag{19}$$

with initial condition $g : \mathbb{R}^m \to \mathbb{R}$, i.e. $f(0,u) := \lim_{t \to 0} f(t,u) = g(u)$ for a.e. $u \in \mathbb{R}^m$. The following result guarantees existence and uniqueness of (19), which is a slight generalization of [29, Theorem 2.3.1] to a.e. bounded and a.e. continuous initial conditions $g : \mathbb{R}^m \to \mathbb{R}$. The proofs of the result in this section can be found in Section 10.

Proposition 6.1. *Let $\lambda \in (0,\infty)$ and $g : \mathbb{R}^m \to \mathbb{R}$ be a.e. bounded and a.e. continuous. Then, the function $(0,\infty) \times \mathbb{R}^m \ni (t,u) \mapsto f(t,u) = \frac{1}{(4\pi\lambda t)^{\frac{m}{2}}} \int_{\mathbb{R}^m} e^{-\frac{\|u-v\|^2}{4\lambda t}} g(v) dv \in \mathbb{R}$ is the unique solution of the PDE (19) with initial condition $g : \mathbb{R}^m \to \mathbb{R}$.*

For some fixed $t \in [0,T]$, we now learn $f(t,\cdot)$ by random trigonometric feature models, random neural networks, and their deterministic counterparts. Here, we omit random Fourier regression as it coincides for real-valued functions with random trigonometric features.

In the following, we provide sufficient conditions that the approximation of $f(t,\cdot)$ by trigonometric feature models and random neural networks overcomes the curse of dimensionality in the sense that the computational costs (measured as the number of features/neurons $N \in \mathbb{N}$) grow polynomially in both the input/output dimensions $m, d \in \mathbb{N}$ and the reciprocal of a pre-specified tolerated approximation error. To this end, we apply the approximation rates in Corollary 4.8+4.20 and introduce $\overline{B_r(0)} := \{u \in \mathbb{R}^m : \|u\| \leqslant r\}$, $r \geqslant 0$.

Corollary 6.2. *For $\lambda, t \in (0,\infty)$ and an a.e. bounded and a.e. continuous initial condition $g : \mathbb{R}^m \to \mathbb{R}$, let $f(t,\cdot) : \mathbb{R}^m \to \mathbb{R}$ be the solution of (19) at time t. Moreover, let $p \in [1,\infty)$, $\gamma \in [0,\infty)$, and $w : \mathbb{R}^m \to [0,\infty)$ be as in Lemma 4.23. Then, the following holds:*

(i) *Let $(\theta_n)_{n\in\mathbb{N}} \sim t_m$ be i.i.d. and assume that $g \in L^1(\mathbb{R}^m, \mathcal{L}(\mathbb{R}^m), du)$. Then, there exists $C_6 > 0$ (being independent of $m \in \mathbb{N}$ and $g : \mathbb{R}^m \to \mathbb{R}$) such that for every $N \in \mathbb{N}$ there exists a random trigonometric feature model $G_N \in \mathcal{RT}_{\mathbb{R}^m,1}$ with N features satisfying*

$$\mathbb{E}\left[\|f(t,\cdot) - G_N\|_{L^p(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),w)}^2\right]^{\frac{1}{2}} \leqslant C_6 m^{\frac{1}{4}} \left(\frac{1}{\sqrt{2\lambda t}\pi}\right)^{\frac{m}{2}} \frac{\|g\|_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du)}}{N^{1-\frac{1}{\min(2,p)}}}. \tag{20}$$

In particular, if $p > 1$ and $\mathbb{R}^m \ni u \mapsto g(u) := \mathbb{1}_{B_R(0)}(u) \in \mathbb{R}$ with $R^2 \leqslant \frac{\sqrt{\lambda t}}{\sqrt{2e}}(m+2)$ for all but finitely many $m \in \mathbb{N}$, then there exists $C_7 > 0$ such that for every $m \in \mathbb{N}$ and $\varepsilon > 0$ there exists some $G_N \in \mathcal{RT}_{\mathbb{R}^m,1}$ with $N = \left\lceil C_7 \varepsilon^{-\frac{\min(2,p)}{\min(2,p)-1}} \right\rceil$ features satisfying

$$\mathbb{E}\left[\|f(t,\cdot) - G_N\|_{L^p(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),w)}^2\right]^{\frac{1}{2}} \leqslant \varepsilon. \tag{21}$$

(ii) *Let $(a_n, b_n)_{n\in\mathbb{N}} \sim t_m \otimes t_1$ be i.i.d., let $(\psi,\rho) \in \mathcal{S}_0(\mathbb{R};\mathbb{C}) \times C^0_{pol,\gamma}(\mathbb{R})$ as in Example 4.15 (with $0 < \zeta_1 < \zeta_2 < \infty$), and assume that $g \in L^1(\mathbb{R}^m, \mathcal{L}(\mathbb{R}^m), (1+\|u\|^{\lceil\gamma\rceil+2})du)$. Then, $f(t,\cdot) \in \widetilde{\mathbb{B}}^{0,2,\gamma}_{\psi,a,b}(\mathbb{R}^m)$ and there exist $C_8, C_9 > 0$ (being independent of $m \in \mathbb{N}$ and $g : \mathbb{R}^m \to \mathbb{R}$) such that for every $N \in \mathbb{N}$ there exists a random neural network $G_N \in \mathcal{RN}^\rho_{\mathbb{R}^m,1}$ with N neurons satisfying*

$$\mathbb{E}\left[\|f(t,\cdot) - G_N\|_{L^p(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),w)}^2\right]^{\frac{1}{2}} \leqslant C_8 m^{C_9} \left(\frac{(\zeta_2/\zeta_1)^2}{\sqrt{2\lambda t}\pi}\right)^{\frac{m}{2}} \frac{\|g\|_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),(1+\|u\|^{\lceil\gamma\rceil+2})du)}}{N^{1-\frac{1}{\min(2,p)}}}. \tag{22}$$

In particular, if $p > 1$ and $\mathbb{R}^m \ni u \mapsto g(u) := \mathbb{1}_{B_R(0)}(u) \in \mathbb{R}$ with $R^2 \leqslant \frac{\zeta_1^2\sqrt{\lambda t}}{\zeta_2^2\sqrt{2e}}(m+2)$ for all but finitely many $m \in \mathbb{N}$, then there exist $C_{10}, C_{11} > 0$ such that for every $m \in \mathbb{N}$ and $\varepsilon > 0$ there exists $G_N \in \mathcal{RN}^\rho_{\mathbb{R}^m,1}$ with $N = \left\lceil C_{10} m^{C_{11}} \varepsilon^{-\frac{\min(2,p)}{\min(2,p)-1}} \right\rceil$ neurons satisfying

$$\mathbb{E}\left[\|f(t,\cdot) - G_N\|_{L^p(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),w)}^2\right]^{\frac{1}{2}} \leqslant \varepsilon. \tag{23}$$

Now, we choose $\lambda = 4$, $t = 1$, the initial condition $\mathbb{R}^m \ni u \mapsto g(u) := \mathbb{1}_{\overline{B_R(0)}}(u) \in \mathbb{R}$ with $R := 4m^{0.4}$, and the weight $\mathbb{R}^m \ni u \mapsto w(u) := (2\pi)^{-m/2} \exp\left(-\|u\|^2/2\right) \in [0,\infty)$ satisfying the conditions of Lemma 4.23. Then, we generate $J = 2 \cdot 10^5$ i.i.d. data $(V_j)_{j=1,\dots,J} \sim w$, split up into $80\%/20\%$ for training/testing, and minimize the empirical L^2-error

$$\left(\frac{1}{J} \sum_{j=1}^{J} |f(1,V_j) - G_N(\cdot)(V_j)|^2 \right)^{\frac{1}{2}} \tag{24}$$

over random trigonometric feature models $G_N \in \mathcal{RT}_{\mathbb{R}^m,1}$, random neural networks $G_N \in \mathcal{RN}^{\tanh}_{\mathbb{R}^m,1}$, and their deterministic counterparts, all of them having N features/neurons.

For the training of the random features, we assume that $(\theta_n)_{n\in\mathbb{N}} \sim t_m$ (for $\mathcal{RT}_{\mathbb{R}^m,1}$) and $(a_n,b_n)_{n\in\mathbb{N}} \sim t_m \otimes t_1$ (for $\mathcal{RN}^{\tanh}_{\mathbb{R}^m,1}$). As $R^2 = 16m^{0.8} \leqslant \frac{\sqrt{\lambda t}}{\sqrt{2e}} m$ and $R^2 = 16m^{0.8} \leqslant \frac{\zeta_1^2\sqrt{\lambda t}}{\zeta_2^2\sqrt{2e}} m$ for all but finitely many $m \in \mathbb{N}$, Corollary 6.2 shows that the approximation of $f(t,\cdot)$ by random trigonometric features and random neural networks overcomes the curse of dimensionality. For their deterministic counterparts, we minimize (24) over the deterministic trigonometric feature models $G_N \in \mathcal{T}_{\mathbb{R}^m,1} := \operatorname{span}_{\mathbb{R}}\left(\left\{\mathbb{R}^m \ni u \mapsto h\left(\vartheta^\top u\right) \in \mathbb{R} : h \in \{\cos,\sin\}\right\}\right)$ and the deterministic neural networks $G_N \in \mathcal{N}^{\tanh}_{\mathbb{R}^m,1} := \operatorname{span}_{\mathbb{R}}\left(\left\{\mathbb{R}^m \ni u \mapsto \tanh\left(\vartheta_1^\top u - \vartheta_2\right) \in \mathbb{R} : (\vartheta_1,\vartheta_2) \in \mathbb{R}^m \times \mathbb{R}\right\}\right)$, both of them having N features/neurons. Hereby, we use the Adam algorithm (see [47]) over 3000 epochs with learning rate $\gamma = 10^{-5}$ and batchsize 500.

Figure 1+2 show that random trigonometric feature models as well as random neural networks are able to learn the solution of the heat equation (19) with similar accuracy than their deterministic counterparts. Moreover, in terms of computational efficiency, the random feature models outperform the deterministic models by far (see Table 1).

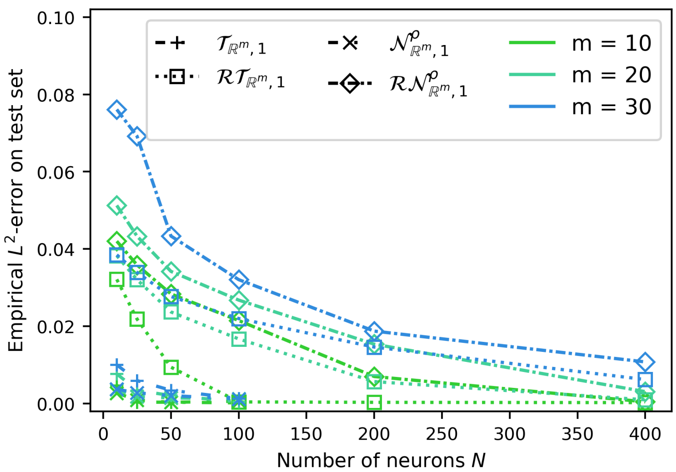

Figure 1. Learning the solution of (19): Empirical L^2-error defined in (24).

[8]The numerical experiment has been implemented in `Python` on an average laptop (Lenovo ThinkPad X13 Gen2a with Processor AMD Ryzen 7 PRO 5850U and Radeon Graphics, 1901 Mhz, 8 Cores, 16 Logical Processors). The code can be found under the following link: https://github.com/psc25/RandomNeuralNetworks

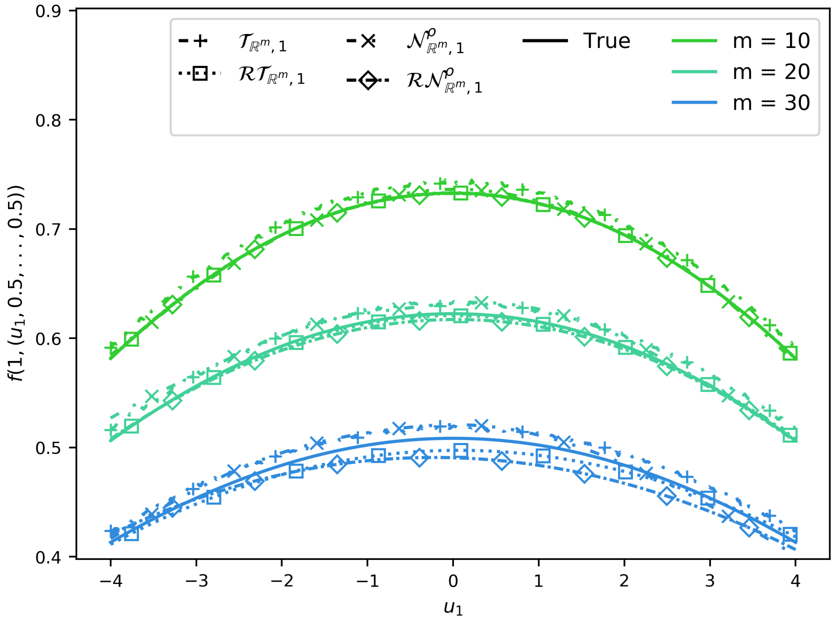

Figure 2. Learning the solution of (19): Approximation of the function $\mathbb{R} \ni u_1 \mapsto f(1,(u_1,0.5,...,0.5)) \in \mathbb{R}$, which is displayed for $N = 200$ features (resp. neurons) for $\mathcal{T}_{\mathbb{R}^m,1}$ and $\mathcal{N}^\rho_{\mathbb{R}^m,1}$ as well as $N = 400$ features (resp. neurons) for $\mathcal{RT}_{\mathbb{R}^m,1}$ and $\mathcal{RN}^\rho_{\mathbb{R}^m,1}$.

		$N=10$	$N=25$	$N=50$	$N=100$	$N=200$	$N=400$
$m=10$	$\mathcal{T}_{\mathbb{R}^m,1}$	*359.38* $1.0 \cdot 10^{10}$	*434.38* $4.8 \cdot 10^{10}$	*546.61* $9.6 \cdot 10^{10}$	*734.56* $1.9 \cdot 10^{11}$		
	$\mathcal{RT}_{\mathbb{R}^m,1}$	*0.13* $8.5 \cdot 10^{6}$	*0.40* $5.0 \cdot 10^{7}$	*0.91* $2.0 \cdot 10^{8}$	*2.60* $8.0 \cdot 10^{8}$	*9.36* $3.2 \cdot 10^{9}$	*30.41* $1.3 \cdot 10^{10}$
	$\mathcal{N}^{\tanh}_{\mathbb{R}^m,1}$	*348.43* $1.0 \cdot 10^{10}$	*382.21* $2.4 \cdot 10^{10}$	*425.88* $4.8 \cdot 10^{10}$	*506.86* $9.6 \cdot 10^{10}$		
	$\mathcal{RN}^{\tanh}_{\mathbb{R}^m,1}$	*0.07* $8.6 \cdot 10^{6}$	*0.21* $5.1 \cdot 10^{7}$	*0.47* $2.0 \cdot 10^{8}$	*1.04* $8.0 \cdot 10^{8}$	*2.59* $3.2 \cdot 10^{9}$	*7.99* $1.3 \cdot 10^{10}$
$m=20$	$\mathcal{T}_{\mathbb{R}^m,1}$	*371.36* $1.0 \cdot 10^{10}$	*495.81* $4.8 \cdot 10^{10}$	*598.58* $9.6 \cdot 10^{10}$	*769.88* $1.9 \cdot 10^{11}$		
	$\mathcal{RT}_{\mathbb{R}^m,1}$	*0.18* $8.5 \cdot 10^{6}$	*0.48* $5.0 \cdot 10^{7}$	*1.04* $2.0 \cdot 10^{8}$	*2.46* $8.0 \cdot 10^{8}$	*7.51* $3.2 \cdot 10^{9}$	*31.62* $1.3 \cdot 10^{10}$
	$\mathcal{N}^{\tanh}_{\mathbb{R}^m,1}$	*344.64* $1.0 \cdot 10^{10}$	*432.72* $2.4 \cdot 10^{10}$	*459.36* $4.8 \cdot 10^{10}$	*563.23* $9.6 \cdot 10^{10}$		
	$\mathcal{RN}^{\tanh}_{\mathbb{R}^m,1}$	*0.09* $6.0 \cdot 10^{8}$	*0.24* $3.5 \cdot 10^{9}$	*0.46* $1.4 \cdot 10^{10}$	*1.00* $3.8 \cdot 10^{10}$	*2.58* $1.4 \cdot 10^{11}$	*9.32* $1.3 \cdot 10^{10}$
$m=30$	$\mathcal{T}_{\mathbb{R}^m,1}$	*439.63* $1.0 \cdot 10^{10}$	*529.50* $4.8 \cdot 10^{10}$	*693.85* $9.6 \cdot 10^{10}$	*819.93* $1.9 \cdot 10^{11}$		
	$\mathcal{RT}_{\mathbb{R}^m,1}$	*0.20* $8.5 \cdot 10^{6}$	*0.48* $5.0 \cdot 10^{7}$	*0.97* $2.0 \cdot 10^{8}$	*2.41* $8.0 \cdot 10^{8}$	*9.09* $3.2 \cdot 10^{9}$	*24.37* $1.3 \cdot 10^{10}$
	$\mathcal{N}^{\tanh}_{\mathbb{R}^m,1}$	*410.55* $1.0 \cdot 10^{10}$	*463.20* $2.4 \cdot 10^{10}$	*494.44* $4.8 \cdot 10^{10}$	*596.33* $9.6 \cdot 10^{10}$		
	$\mathcal{RN}^{\tanh}_{\mathbb{R}^m,1}$	*0.09* $1.0 \cdot 10^{8}$	*0.23* $6.1 \cdot 10^{8}$	*0.50* $2.4 \cdot 10^{9}$	*0.99* $1.6 \cdot 10^{9}$	*2.47* $6.4 \cdot 10^{9}$	*7.12* $2.6 \cdot 10^{10}$

Table 1. Learning the solution of the heat equation (19): Computational time (in italic letters) and complexity (in scientific format $0.0 \cdot 10^0$, see also Proposition 5.1).

7. Proof of results in Section 3

7.1. **Proof of Theorem 3.2.** For the following proofs in this section, we denote the closed ball of radius $r > 0$ around $\vartheta_0 \in \Theta$ by $\overline{B_r(\vartheta_0)} := \{\vartheta \in \Theta : d_\Theta(\vartheta, \vartheta_0) \leqslant r\}$.

Proof of Theorem 3.2. First, by using that $\mathcal{G}(\Theta) := \operatorname{span}_{\mathbb{K}}(\{g(\vartheta) : g \in \mathcal{G}, \vartheta \in \Theta\})$ is by assumption dense in X together with [44, Lemma 1.2.19 (i)], i.e. that $\mathcal{I}_{\mathcal{F}_\theta} \otimes X := \operatorname{span}_{\mathbb{K}}(\{\Omega \ni \omega \mapsto \mathbb{1}_E(\omega)x \in X : E \in \mathcal{F}_\theta, x \in X\})$ is dense in $L^r(\Omega, \mathcal{F}_\theta, \mathbb{P}; X)$, we obtain that $\mathcal{I}_{\mathcal{F}_\theta} \otimes \mathcal{G}(\Theta) := \operatorname{span}_{\mathbb{K}}(\{\Omega \ni \omega \mapsto \mathbb{1}_E(\omega)g(\vartheta) \in X : E \in \mathcal{F}_\theta, g \in \mathcal{G}, \vartheta \in \Theta\})$ is dense in $L^r(\Omega, \mathcal{F}_\theta, \mathbb{P}; X)$. Hence, it suffices to show the approximation of any map of the form $\Omega \ni \omega \mapsto \mathbb{1}_E(\omega)g(\vartheta_0) \in X$, with $E \in \mathcal{F}_\theta$, $g \in \mathcal{G}$, and $\vartheta_0 \in \Theta$, by some $G \in \mathcal{RG} \cap L^r(\Omega, \mathcal{F}_\theta, \mathbb{P}; X)$ with respect to $\|\cdot\|_{L^r(\Omega,\mathcal{F},\mathbb{P};X)}$. To this end, we fix some $E \in \mathcal{F}_\theta$, $g \in \mathcal{G}$, $\vartheta_0 \in \Theta$, and $\varepsilon > 0$. Moreover, for every $M, n \in \mathbb{N}$, we define the map

$$\Omega \ni \omega \quad \mapsto \quad G_{M,n}(\omega) := y_{M,n}(\omega) g(\theta_n(\omega)) \in X$$

with $\mathcal{F}_\theta$-measurable linear readout $\Omega \ni \omega \mapsto y_{M,n}(\omega) := C_M^{-1} \mathbb{1}_{\overline{B_{1/M}(\vartheta_0)}}(\theta_n(\omega)) \in \mathbb{R}$, where $C_M := \mathbb{P}\big[\big\{\omega \in \Omega : \theta_1(\omega) \in \overline{B_{1/M}(\vartheta_0)}\big\}\big] > 0$ due to Assumption 3.1.

Now, we show that the sequence $\big(\mathbb{1}_{\overline{B_{1/M}(\vartheta_0)}}(\theta_n(\omega)) g(\vartheta_0) - C_M G_{M,n}(\omega)\big)_{M\in\mathbb{N}}$ converges uniformly in $\omega \in \Omega$ and $n \in \mathbb{N}$ to $0 \in X$ with respect to $\|\cdot\|_X$. To this end, we fix some $\varepsilon > 0$. Then, by using that $g \in \mathcal{G} \subseteq C^0(\Theta; X)$ is continuous, there exists some $\delta > 0$ such that for every $\vartheta \in B_\delta(\vartheta_0)$ it holds that

$$\|g(\vartheta_0) - g(\vartheta)\|_X < \varepsilon.$$

Hence, by choosing $M_0 \in \mathbb{N}$ large enough such that $M_0 > \delta^{-1}$, it follows for every $M \in \mathbb{N} \cap [M_0, \infty)$ that

$$\begin{aligned}
&\sup_{\omega\in\Omega}\sup_{n\in\mathbb{N}} \left\| \mathbb{1}_{\overline{B_{1/M}(\vartheta_0)}}(\theta_n(\omega)) g(\vartheta_0) - C_M G_{M,n}(\omega) \right\|_X \\
&\quad = \sup_{\omega\in\Omega}\sup_{n\in\mathbb{N}} \left\| \mathbb{1}_{\overline{B_{1/M}(\vartheta_0)}}(\theta_n(\omega)) \big(g(\vartheta_0) - g(\theta_n(\omega))\big) \right\|_X \\
&\quad = \sup_{\vartheta\in\overline{B_{1/M}(\vartheta_0)}} \|g(\vartheta_0) - g(\vartheta)\|_X \\
&\quad \leqslant \sup_{\vartheta\in\overline{B_{1/M_0}(\vartheta_0)}} \|g(\vartheta_0) - g(\vartheta)\|_X < \varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ was chosen arbitrarily, this shows that the sequence $\big(\mathbb{1}_{\overline{B_{1/M}(\vartheta_0)}}(\theta_n(\omega)) g(\vartheta_0) - C_M G_{M,n}(\omega)\big)_{M\in\mathbb{N}}$ converges uniformly in $\omega \in \Omega$ and $n \in \mathbb{N}$ to $0 \in X$ with respect to $\|\cdot\|_X$.

Next, we show for every fixed $M, n \in \mathbb{N}$ that $G_{M,n} \in L^r(\Omega, \mathcal{F}_\theta, \mathbb{P}; X)$. Indeed, by using that $\Omega \ni \omega \mapsto (y_{M,n}(\omega), \theta_n(\omega)) \in \mathbb{R} \times \Theta$ is $\mathcal{F}_\theta/\mathcal{B}(\mathbb{R} \times \Theta)$-measurable and that $g \in \mathcal{G} \subseteq C^0(\Theta; X)$ is continuous, it follows that the concatenation $\Omega \ni \omega \mapsto y_{M,n}(\omega) g(\theta_n(\omega)) \in X$ is $\mathcal{F}_\theta/\mathcal{B}(X)$-measurable. Hence, by using that $(X, \|\cdot\|_X)$ is separable, we can apply [44, Theorem 1.1.6+1.1.20] to conclude that $G_{M,n} : \Omega \to X$ is strongly $(\mathbb{P}, \mathcal{F}_\theta)$-measurable. Moreover, by using Minkowski's inequality and that the sequence $\big(\mathbb{1}_{\overline{B_{1/M}(\vartheta_0)}}(\theta_n(\omega)) g(\vartheta_0) - C_M G_{M,n}(\omega)\big)_{M\in\mathbb{N}}$ is by the previous step uniformly bounded in $\omega \in \Omega$ and $n \in \mathbb{N}$, we have

$$\begin{aligned}
\|G_{M,n}\|_{L^r(\Omega,\mathcal{F},\mathbb{P};X)} &= \mathbb{E}\left[\|G_{M,n}\|_X^r\right]^{\frac{1}{r}} = \frac{1}{C_M} \mathbb{E}\left[\|C_M G_{M,n}\|_X^r\right]^{\frac{1}{r}} \\
&\leqslant \frac{1}{C_M} \mathbb{E}\left[\left\| \mathbb{1}_{\overline{B_{1/M}(\vartheta_0)}}(\theta_n) g(\vartheta_0) \right\|_X^r\right]^{\frac{1}{r}} + \frac{1}{C_M} \mathbb{E}\left[\left\| \mathbb{1}_{\overline{B_{1/M}(\vartheta_0)}}(\theta_n) g(\vartheta_0) - C_M G_{M,n} \right\|_X^r\right]^{\frac{1}{r}} \\
&\leqslant \frac{1}{C_M} \|g(\vartheta_0)\|_X + \frac{1}{C_M} \sup_{\omega\in\Omega} \left\| \mathbb{1}_{\overline{B_{1/M}(\vartheta_0)}}(\theta_n(\omega)) g(\vartheta_0) - C_M G_{M,n} \right\|_X < \infty,
\end{aligned}$$

which shows that $G_{M,n} \in L^r(\Omega, \mathcal{F}_\theta, \mathbb{P}; X)$ for all $M, n \in \mathbb{N}$.

Now, we show that there exists some $M_1 \in \mathbb{N}$ such that the constant maps $(\omega \mapsto g(\vartheta_0)) \in L^r(\Omega, \mathcal{F}_\theta, \mathbb{P}; X)$ and $(\omega \mapsto \mathbb{E}[G_{M_1,1}]) \in L^r(\Omega, \mathcal{F}_\theta, \mathbb{P}; X)$ are $\frac{\varepsilon}{2}$-close to each other with respect to $\|\cdot\|_{L^r(\Omega,\mathcal{F},\mathbb{P};X)}$. Indeed, by using that $\big(\mathbb{1}_{\overline{B_{1/M}(\vartheta_0)}}(\theta_n(\omega)) g(\vartheta_0) - C_M G_{M,n}(\omega)\big)_{M\in\mathbb{N}}$ converges uniformly in $\omega \in \Omega$ and $n \in \mathbb{N}$ to $0 \in X$ with respect to $\|\cdot\|_X$, there exists some $M_1 \in \mathbb{N}$ such that

$$\sup_{n\in\mathbb{N}}\sup_{\omega\in\Omega} \left\| \mathbb{1}_{\overline{B_{1/M_1}(\vartheta_0)}}(\theta_n(\omega)) g(\vartheta_0) - C_{M_1} G_{M_1,n}(\omega) \right\|_X < \frac{\varepsilon}{2}. \tag{25}$$

Hence, by using that $\mathbb{E}\big[\mathbb{1}_{\overline{B_{1/M_1}(\vartheta_0)}}(\theta_1)\big] = \mathbb{P}\big[\big\{\omega \in \Omega : \theta_1(\omega) \in \overline{B_{1/M_1}(\vartheta_0)}\big\}\big] = C_{M_1} > 0$, and [44, Proposition 1.2.2], it follows that

$$\begin{aligned}
\|g(\vartheta_0) - \mathbb{E}[G_{M_1,1}]\|_{L^r(\Omega,\mathcal{F},\mathbb{P};X)} &= \mathbb{E}\left[\|g(\vartheta_0) - \mathbb{E}[G_{M_1,1}]\|_X^r\right]^{\frac{1}{r}} = \|g(\vartheta_0) - \mathbb{E}[G_{M_1,1}]\|_X \\
&= \left\|\mathbb{E}\left[\frac{1}{C_{M_1}}\mathbb{1}_{\overline{B_{1/M_1}(\vartheta_0)}}(\theta_1)g(\vartheta_0) - G_{M_1,1}\right]\right\|_X \\
&\leqslant \mathbb{E}\left[\frac{\mathbb{1}_{\overline{B_{1/M_1}(\vartheta_0)}}(\theta_1)}{C_{M_1}}\left\|\mathbb{1}_{\overline{B_{1/M_1}(\vartheta_0)}}(\theta_1)g(\vartheta_0) - C_{M_1}G_{M_1,1}\right\|_X\right] \\
&\leqslant \underbrace{\frac{\mathbb{E}\left[\mathbb{1}_{\overline{B_{1/M_1}(\vartheta_0)}}(\theta_1)\right]}{C_{M_1}}}_{=1}\sup_{\omega\in\Omega}\left\|\mathbb{1}_{\overline{B_{1/M_1}(\vartheta_0)}}(\theta_1)g(\vartheta_0) - C_{M_1}G_{M_1,1}(\omega)\right\|_X < \frac{\varepsilon}{2}.
\end{aligned} \tag{26}$$

This shows that the constant maps $(\omega \mapsto g(\vartheta_0)) \in L^r(\Omega,\mathcal{F}_\theta,\mathbb{P};X)$ and $(\omega \mapsto \mathbb{E}[G_{M_1,1}]) \in L^r(\Omega,\mathcal{F}_\theta,\mathbb{P};X)$ are $\frac{\varepsilon}{2}$-close to each other with respect to $\|\cdot\|_{L^r(\Omega,\mathcal{F},\mathbb{P};X)}$.

Finally, we approximate the constant random variable $(\omega \mapsto \mathbb{E}[G_{M_1,1}]) \in L^1(\Omega,\mathcal{F}_\theta,\mathbb{P};X)$ by the average of the i.i.d. sequence $(G_{M_1,n})_{n\in\mathbb{N}} \subseteq L^1(\Omega,\mathcal{F}_\theta,\mathbb{P};X)$. Indeed, by applying the strong law of large numbers for Banach space-valued random variables in [44, Theorem 3.3.10] with Banach space $(X,\|\cdot\|_X)$, we conclude that

$$\frac{1}{N}\sum_{n=1}^N G_{M_1,n} \quad \overset{N\to\infty}{\longrightarrow} \quad \mathbb{E}[G_{M_1,1}] \quad \text{in } L^1(\Omega,\mathcal{F}_\theta,\mathbb{P};X) \text{ and } \mathbb{P}\text{-a.s.} \tag{27}$$

Moreover, if $r \in (1,\infty)$, we generalize the convergence in (27) to $L^r(\Omega,\mathcal{F}_\theta,\mathbb{P};X)$. To this end, we define the sequence of real-valued random variables $(Z_N)_{N\in\mathbb{N}}$ by $Z_N(\omega) := \big\|\mathbb{E}\big[G_{M_1,1}\big] - \frac{1}{N}\sum_{n=1}^N G_{M_1,n}\big\|_X^r$, for $\omega \in \Omega$ and $N \in \mathbb{N}$. Then, by using [44, Proposition 1.2.2] and (25), it follows for every $N \in \mathbb{N}$ that

$$\begin{aligned}
\sup_{\omega\in\Omega} Z_N(\omega) &\leqslant \sup_{\omega\in\Omega}\left(\|\mathbb{E}[G_{M_1,1}]\|_X + \frac{1}{N}\sum_{n=1}^N\|G_{M_1,n}(\omega)\|_X\right)^r \\
&\leqslant \sup_{n\in\mathbb{N}}\sup_{\omega\in\Omega}\left(\mathbb{E}\left[\|G_{M_1,1}\|_X\right] + \|G_{M_1,n}(\omega)\|_X\right)^r \\
&\leqslant \frac{2^r}{C_{M_1}^r}\sup_{n\in\mathbb{N}}\sup_{\omega\in\Omega}\|C_{M_1}G_{M_1,n}(\omega)\|_X^r \\
&\leqslant \frac{2^r}{C_{M_1}^r}\sup_{n\in\mathbb{N}}\sup_{\omega\in\Omega}\left(\left\|\mathbb{1}_{\overline{B_{1/M_1}(\vartheta_0)}}(\theta_1(\omega))g(\vartheta_0)\right\|_X + \right. \\
&\qquad\qquad \left. + \left\|\mathbb{1}_{\overline{B_{1/M_1}(\vartheta_0)}}(\theta_1(\omega))g(\vartheta_0) - C_{M_1}G_{M_1,n}(\omega)\right\|_X\right)^r \\
&< \frac{2^r}{C_{M_1}^r}\left(\|g(\vartheta_0)\|_X + \frac{\varepsilon}{2}\right)^r =: C_Z < \infty.
\end{aligned}$$

Hence, $\sup_{N\in\mathbb{N}}\mathbb{E}\left[|Z_N|\mathbb{1}_{\{|Z_N|>C_Z\}}\right] = 0$, which implies that the family of random variables $(Z_N)_{N\in\mathbb{N}}$ is uniformly integrable (see [44, Definition A.3.1]). Thus, by using that $Z_N \to 0$, $\mathbb{P}$-a.s., as $N \to \infty$ (see (27)), and Vitali's convergence theorem (see [44, Proposition A.3.5]), we have

$$\lim_{N\to\infty}\mathbb{E}\left[\left\|\mathbb{E}[G_{M_1,1}] - \frac{1}{N}\sum_{n=1}^N G_{M_1,n}\right\|_X^r\right] = \lim_{N\to\infty}\mathbb{E}[Z_N] = 0. \tag{28}$$

Thus, either by (27) (if $r = 1$) or (28) (if $r \in (1,\infty)$) there exists some $N_0 \in \mathbb{N}$ such that

$$\mathbb{E}\left[\left\|\mathbb{E}[G_{M_1,1}] - \frac{1}{N_0}\sum_{n=1}^{N_0} G_{M_1,n}\right\|_X^r\right]^{\frac{1}{r}} < \frac{\varepsilon}{2}. \tag{29}$$

Finally, we define $G := \left(\omega \mapsto \frac{1}{N_0}\sum_{n=1}^{N_0} \mathbb{1}_E(\omega) G_{M_1,n}(\omega)\right) \in \mathcal{RG} \cap L^r(\Omega,\mathcal{F}_\theta,\mathbb{P};X)$. Hence, by combining (26) and (29) with Minkowski's inequality, it follows that

$$\begin{aligned}
\|\mathbb{1}_E g(\vartheta_0) - G\|_{L^r(\Omega,\mathcal{F},\mathbb{P};X)} &= \mathbb{E}\left[\|\mathbb{1}_E g(\vartheta_0) - G\|^r\right]^{\frac{1}{r}} \\
&= \mathbb{E}\left[\underbrace{\mathbb{1}_E}_{\leqslant 1} \left\| g(\vartheta_0) - \frac{1}{N_0}\sum_{n=1}^{N_0} G_{M_1,n}\right\|^r\right]^{\frac{1}{r}} \\
&\leqslant \mathbb{E}\left[\left\| g(\vartheta_0) - \frac{1}{N_0}\sum_{n=1}^{N_0} G_{M_1,n}\right\|^r\right]^{\frac{1}{r}} \\
&\leqslant \|g(\vartheta_0) - \mathbb{E}[G_{M_1,n}]\|_X + \mathbb{E}\left[\left\| \mathbb{E}[G_{M_1,n}] - \frac{1}{N_0}\sum_{n=1}^{N_0} G_{M_1,n}\right\|^r\right]^{\frac{1}{r}} \\
&< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$, $g \in \mathcal{G}$, and $\vartheta_0 \in \Theta$ were chosen arbitrarily, this shows that $\Omega \ni \omega \mapsto \mathbb{1}_E(\omega) g(\vartheta_0) \in X$ can be approximated by a random feature model $G \in \mathcal{RG} \cap L^r(\Omega,\mathcal{F}_\theta,\mathbb{P};X)$ with respect to $\|\cdot\|_{L^r(\Omega,\mathcal{F},\mathbb{P};X)}$. Combining this together with the first step of the proof, i.e. that $\mathcal{I}_{\mathcal{F}_\theta} \otimes \mathcal{G}(\Theta)$ is dense in $L^r(\Omega,\mathcal{F}_\theta,\mathbb{P};X)$, we obtain the conclusion. ☐

7.2. **Proof of Corollary 3.3+3.4+3.8.**

Proof of Corollary 3.3. We aim to apply Theorem 3.2 with Banach space $(X,\|\cdot\|_X) := (C^0(U),\|\cdot\|_{C^0(U)})$. To this end, we first observe that $(C^0(U),\|\cdot\|_{C^0(U)})$ is by [15, Problem 24] separable. Moreover, we choose $\Theta := \mathbb{R}^m$ and let $(\theta_n)_{n\in\mathbb{N}} : \Omega \to \mathbb{R}^m$ be an i.i.d. sequence satisfying Assumption 3.1. In addition, we define

$$\mathcal{G} := \left\{\mathbb{R}^m \ni \vartheta \mapsto h\left(\vartheta^\top \cdot\right) \in C^0(U) : h \in \{\cos,\sin\}\right\}.$$

Then, for both $h \in \{\cos,\sin\}$, we use that $\mathbb{R}^m \times U \ni (\vartheta,u) \mapsto h(\vartheta^\top u) \in \mathbb{R}$ is continuous to conclude that $K \times U \ni (\vartheta,u) \mapsto h(\vartheta^\top u) \in \mathbb{R}$ is uniformly continuous, for all compact subsets $K \subset \mathbb{R}^m$. Hence, the map $\mathbb{R}^m \ni \vartheta \mapsto h(\vartheta^\top \cdot) \in C^0(U)$ is continuous, which shows that $\mathcal{G} \subseteq C^0(\Theta;X)$. Moreover, by using the trigonometric identities $\cos(s)\cos(t) = (\cos(s-t)+\cos(s+t))/2$, $\sin(s)\sin(t) = (\cos(s-t)-\cos(s+t))/2$, and $\cos(s)\sin(t) = (\sin(s+t)-\sin(s-t))/2$ for any $s,t \in \mathbb{R}$, we observe that

$$\operatorname{span}_{\mathbb{R}}(\mathcal{G}(\Theta)) = \operatorname{span}_{\mathbb{R}}\left(\left\{U \ni u \mapsto h\left(\vartheta^\top u\right) \in \mathbb{R} : h \in \{\cos,\sin\},\ \vartheta \in \mathbb{R}^m\right\}\right)$$

is a subalgebra of $C^0(U)$, i.e. for every $g_1, g_2 \in \operatorname{span}_{\mathbb{R}}(\mathcal{G}(\Theta))$ we have $g_1 + g_2 \in \operatorname{span}_{\mathbb{R}}(\mathcal{G}(\Theta))$ and $g_1 \cdot g_2 \in \operatorname{span}_{\mathbb{R}}(\mathcal{G}(\Theta))$. Moreover, $\operatorname{span}_{\mathbb{R}}(\mathcal{G}(\Theta))$ is point separating, i.e. for any distinct $u_1,u_2 \in U$ there exists some $g \in \operatorname{span}_{\mathbb{R}}(\mathcal{G}(\Theta))$ such that $g(u_1) \neq g(u_2)$. In addition, $\operatorname{span}_{\mathbb{R}}(\mathcal{G}(\Theta))$ vanishes nowhere, i.e. for every $u_0 \in U$ there exists some $g \in \operatorname{span}_{\mathbb{R}}(\mathcal{G}(\Theta))$ such that $g(u_0) \neq 0$. Hence, we can apply the Stone-Weierstrass theorem (see [77]) to obtain that $\operatorname{span}_{\mathbb{R}}(\mathcal{G}(\Theta))$ is dense in $C^0(U)$. Thus, the conclusion follows from Theorem 3.2. ☐

Proof of Corollary 3.4. We aim to apply Theorem 3.2 with Banach space $(X,\|\cdot\|_X) := (C^0(U),\|\cdot\|_{C^0(U)})$. To this end, we first observe that $(C^0(U),\|\cdot\|_{C^0(U)})$ is by [15, Problem 24] separable. Moreover, we choose $\Theta := \mathbb{R}^m$ and let $(\theta_n)_{n\in\mathbb{N}} : \Omega \to \mathbb{R}^m$ be an i.i.d. sequence satisfying Assumption 3.1. In addition, we define the singleton set

$$\mathcal{G} := \left\{\mathbb{R}^m \ni \vartheta \mapsto \exp\left(\mathbf{i}\vartheta^\top \cdot\right) \in C^0(U)\right\}.$$

Then, we follow the proof of Corollary 3.8 to conclude that $\mathbb{R}^m \ni \vartheta \mapsto \exp\left(\mathbf{i}\vartheta^\top \cdot\right) \in C^0(U)$ is continuous, which shows that $\mathcal{G} \subseteq C^0(\Theta;X)$. Moreover, by using the identities $\exp\left(\mathbf{i}\vartheta_1^\top u\right)\exp\left(\mathbf{i}\vartheta_2^\top u\right) = \exp\left(\mathbf{i}(\vartheta_1+\vartheta_2)^\top u\right)$ and $\overline{\exp\left(\mathbf{i}\vartheta_1^\top u\right)} = \exp\left(-\mathbf{i}\vartheta_1^\top u\right) = \exp\left(\mathbf{i}(-\vartheta_1)^\top u\right)$ for any $\vartheta_1,\vartheta_2 \in \mathbb{R}^m$ and $u \in U$, we observe that

$$\operatorname{span}_{\mathbb{C}}(\mathcal{G}(\Theta)) = \operatorname{span}_{\mathbb{C}}\left(\left\{U \ni u \mapsto \exp\left(\mathbf{i}\vartheta^\top u\right) \in \mathbb{C} : \vartheta \in \mathbb{R}^m\right\}\right)$$

is a subalgebra of $C^0(U)$, which is point separating, nowhere vanishing, and self-adjoint, where the latter means that for every $g \in \operatorname{span}_{\mathbb{C}}(\mathcal{G}(\Theta))$ the function $\Theta \ni \vartheta \mapsto \overline{g}(\vartheta) := \overline{g(\vartheta)} \in \mathbb{C}$ satisfies $\overline{g} \in$

$\operatorname{span}_{\mathbb{C}}(\mathcal{G}(\Theta))$. Hence, we can apply the complex-valued Stone-Weierstrass theorem (see e.g. [74, p. 122]) to obtain that $\operatorname{span}_{\mathbb{C}}(\mathcal{G}(\Theta))$ is dense in $C^0(U;\mathbb{C})$. Thus, the conclusion follows from Theorem 3.2. ☐

For the proof of Corollary 3.8, we first show the following auxiliary lemma about neurons and that Banach spaces $(X, \|\cdot\|_X)$ satisfying Assumption 3.5 are separable.

Lemma 7.1. *Let $(X, \|\cdot\|_X)$ satisfy Assumption 3.5. Then, the following holds true:*

(i) *For every $y \in \mathbb{R}^d$, $a \in \mathbb{R}^m$, $b \in \mathbb{R}$, and $\rho \in \overline{C^k_b(\mathbb{R})}^\gamma$ it holds that $y\rho\big(a^\top \cdot - b\big) \in X$.*
(ii) *For every $\rho \in \overline{C^k_b(\mathbb{R})}^\gamma$ the map $\mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R} \ni (y,a,b) \mapsto y\rho\big(a^\top \cdot - b\big) \in X$ is continuous.*
(iii) *The Banach space $(X, \|\cdot\|_X)$ is separable.*

Proof. For Part (i)., we apply [64, Lemma 2.5] to conclude that $y\rho\big(a^\top \cdot - b\big) \in X$ for all $y \in \mathbb{R}^d$, $a \in \mathbb{R}^m$, $b \in \mathbb{R}$, and $\rho \in \overline{C^k_b(\mathbb{R})}^\gamma$.

For Part (ii)., we fix some $\varepsilon > 0$ and a sequence $(y_M, a_M, b_M)_{M \in \mathbb{N}} \subseteq \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}$ converging to $(y,a,b) \in \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}$. Then, by using that $y_M a_M^\alpha$ converges uniformly in $\alpha \in \mathbb{N}^m_{0,k}$ to $y a^\alpha$ (where $a^\alpha := \prod_{l=1}^m a_l^{\alpha_l}$ for $a := (a_1, \ldots, a_m)^\top \in \mathbb{R}^m$ and $\alpha := (\alpha_1, \ldots, \alpha_m) \in \mathbb{N}^m_{0,k}$), the constant $C_{y,a} := 1 + \max_{\alpha \in \mathbb{N}^m_{0,k}} \|y a^\alpha\| + \sup_{M \in \mathbb{N}} \max_{\alpha \in \mathbb{N}^m_{0,k}} \|y_M a_M^\alpha\| > 0$ is finite. Moreover, since $(a_M, b_M)_{M \in \mathbb{N}}^\top \subseteq \mathbb{R}^m \times \mathbb{R}$ converges to $(a,b) \in \mathbb{R}^m \times \mathbb{R}$, the constant $C_{a,b} := 1 + \|(a,b)\| + \sup_{M \in \mathbb{N}} \|(a_M, b_M)\| > 0$ is finite. In addition, there exists by definition of $\overline{C^k_b(\mathbb{R})}^\gamma$ some $\widetilde{\rho} \in C^k_b(\mathbb{R})$ such that

$$\|\rho - \widetilde{\rho}\|_{C^k_{pol,\gamma}(\mathbb{R})} := \max_{\alpha \in \mathbb{N}^m_{0,k}} \sup_{s \in \mathbb{R}} \frac{\left|\rho^{(|\alpha|)}(s) - \widetilde{\rho}^{(|\alpha|)}(s)\right|}{(1+|s|)^\gamma} < \frac{\varepsilon}{6 C_{y,a} C_{a,b}}. \tag{30}$$

Now, we choose some $r > 0$ large enough such that $(1+r)^\gamma \geqslant 6\varepsilon^{-1} C_{y,a} \|\widetilde{\rho}\|_{C^k_b(\mathbb{R})}$. Then, the inequality $1 + \left|a_M^\top u - b_M\right| \leqslant 1 + \|a_M\|\|u\| + |b_M| \leqslant (1 + \|a_M\| + |b_M|)(1 + \|u\|)$ for any $u \in \mathbb{R}^m$ and (30) imply that

$$\begin{aligned}
&\max_{\alpha \in \mathbb{N}^m_{0,k}} \sup_{u \in \mathbb{R}^m \setminus \overline{B_r(0)}} \frac{\left\|y_M \rho^{(|\alpha|)}\left(a_M^\top u - b_M\right) a_M^\alpha\right\|}{(1+\|u\|)^\gamma} \\
&\quad \leqslant \left(\max_{\alpha \in \mathbb{N}^m_{0,k}} \|y_M a_M^\alpha\|\right) \max_{\alpha \in \mathbb{N}^m_{0,k}} \sup_{u \in \mathbb{R}^m \setminus \overline{B_r(0)}} \frac{\left|\rho^{(|\alpha|)}\left(a_M^\top u - b_M\right)\right|}{(1+\|u\|)^\gamma} \\
&\quad \leqslant C_{y,a} \max_{\alpha \in \mathbb{N}^m_{0,k}} \sup_{u \in \mathbb{R}^m \setminus \overline{B_r(0)}} \frac{\left|\rho^{(|\alpha|)}\left(a_M^\top u - b_M\right) - \widetilde{\rho}^{(|\alpha|)}\left(a_M^\top u - b_M\right)\right|}{(1+\|u\|)^\gamma} \\
&\qquad + C_{y,a} \max_{\alpha \in \mathbb{N}^m_{0,k}} \sup_{u \in \mathbb{R}^m \setminus \overline{B_r(0)}} \frac{\left|\widetilde{\rho}^{(|\alpha|)}\left(a_M^\top u - b_M\right)\right|}{(1+\|u\|)^\gamma} \\
&\quad \leqslant C_{y,a}(1 + \|a_M\| + \|b_M\|)^\gamma \max_{\alpha \in \mathbb{N}^m_{0,k}} \sup_{u \in \mathbb{R}^m} \frac{\left|\rho^{(|\alpha|)}\left(a_M^\top u - b_M\right) - \widetilde{\rho}^{(|\alpha|)}\left(a_M^\top u - b_M\right)\right|}{\left(1 + \left|a_M^\top u - b_M\right|\right)^\gamma} \\
&\qquad + C_{y,a} \frac{\|\widetilde{\rho}\|_{C^k_b(\mathbb{R})}}{(1+r)^\gamma} \\
&\quad \leqslant C_{y,a} C_{a,b} \max_{j=0,\ldots,k} \sup_{s \in \mathbb{R}} \frac{\left|\rho^{(j)}(s) - \widetilde{\rho}^{(j)}(s)\right|}{(1+|s|)^\gamma} + C_{y,a} \frac{\varepsilon}{6 C_{y,a} C_{a,b}} \\
&\quad < C_{y,a} C_{a,b} \frac{\varepsilon}{6 C_{y,a} C_{a,b}} + \frac{\varepsilon}{6} = \frac{\varepsilon}{3}.
\end{aligned} \tag{31}$$

Analogously, we conclude that

$$\max_{\alpha \in \mathbb{N}^m_{0,k}} \sup_{u \in \mathbb{R}^m \setminus \overline{B_r(0)}} \frac{\left\|y \rho^{(|\alpha|)}\left(a^\top u - b\right) a^\alpha\right\|}{(1+\|u\|)^\gamma} < \frac{\varepsilon}{3}. \tag{32}$$

Moreover, we define the compact subset $K := \left\{x^\top u - y : u \in \overline{B_r(0)},\ \|x\| + \|y\| \leqslant C_{a,b}\right\} \subseteq \mathbb{R}$. Then, by using that $\rho, \rho', \ldots, \rho^{(k)} \in \overline{C^k_b(\mathbb{R})}^\gamma$ are continuous, thus uniformly continuous on K, there exists some

$\delta > 0$ such that for every $j = 0, ..., k$ and $s_1, s_2 \in K$ with $|s_1 - s_2| < \delta$ it holds that

$$\left|\rho^{(j)}(s_1) - \rho^{(j)}(s_2)\right| < \frac{\varepsilon}{6C_{y,a}}. \tag{33}$$

Now, we define the constant $C_{r,\rho} := 1 + \max_{j=0,...,k} \sup_{u \in \overline{B_r(0)}} \left|\rho^{(j)}\left(a^\top u - b\right)\right| > 0$. Moreover, we choose some $M_2 \in \mathbb{N}$ such that for every $M \in \mathbb{N} \cap [M_2, \infty)$ it holds that $\|(a - a_M, b - b_M)\| < \delta/(1+r)$ and that

$$\max_{\alpha \in \mathbb{N}^m_{0,k}} \|y a^\alpha - y_M a_M^\alpha\| < \frac{\varepsilon}{6C_{r,\rho}}. \tag{34}$$

Then, we conclude for every $M \in \mathbb{N} \cap [M_2, \infty)$ that

$$\begin{aligned} \left|\left(a^\top u - b\right) - \left(a_M^\top u - b_M\right)\right| &\leqslant \left|(a - a_M)^\top u - (b - b_M)\right| \\ &\leqslant \|a - a_M\| \|u\| + |b - b_M| \\ &\leqslant (\|a - a_M\| + |b - b_M|)\,(1 + r) \\ &\leqslant \|(a - a_M, b - b_M)\|(1 + r) < \delta. \end{aligned} \tag{35}$$

Hence, by using (34) and by combining (33) with (35), it follows for every $M \in \mathbb{N} \cap [M_2, \infty)$ that

$$\begin{aligned} &\max_{\alpha \in \mathbb{N}^m_{0,k}} \sup_{u \in \overline{B_r(0)}} \left\| y \rho^{(|\alpha|)}\left(a^\top u - b\right) a^\alpha - y_M \rho^{(|\alpha|)}\left(a_M^\top u - b_M\right) a_M^\alpha \right\| \\ &\leqslant \max_{\alpha \in \mathbb{N}^m_{0,k}} \sup_{u \in \overline{B_r(0)}} \left\| y \rho^{(|\alpha|)}\left(a^\top u - b\right) a^\alpha - y_M \rho^{(|\alpha|)}\left(a^\top u - b\right) a_M^\alpha \right\| \\ &\quad + \max_{\alpha \in \mathbb{N}^m_{0,k}} \sup_{u \in \overline{B_r(0)}} \left\| y_M \rho^{(|\alpha|)}\left(a^\top u - b\right) a_M^\alpha - y_M \rho^{(|\alpha|)}\left(a_M^\top u - b_M\right) a_M^\alpha \right\| \\ &\leqslant \max_{\alpha \in \mathbb{N}^m_{0,k}} \|y a^\alpha - y_M a_M^\alpha\| \max_{j=0,...,k} \sup_{u \in \overline{B_r(0)}} \left|\rho^{(j)}\left(a^\top u - b\right)\right| \\ &\quad + \max_{\alpha \in \mathbb{N}^m_{0,k}} \|y_M a_M^\alpha\| \max_{j=0,...,k} \sup_{u \in \overline{B_r(0)}} \left|\rho^{(j)}\left(a_M^\top u - b_M\right) - \rho^{(j)}\left(a^\top u - b\right)\right| \\ &\leqslant \frac{\varepsilon}{6C_{r,\rho}} C_{r,\rho} + C_{y,a} \frac{\varepsilon}{6C_{y,a}} = \frac{\varepsilon}{3}. \end{aligned} \tag{36}$$

Thus, by using the inequalities (31)+(32)+(36) and that $(1 + \|u\|)^\gamma \geqslant 1$ for any $u \in U$, we have

$$\begin{aligned} &\left\| y\rho\left(a^\top \cdot - b\right) - y_M \rho\left(a_M^\top \cdot - b_M\right)\right\|_{C^k_{pol,\gamma}(\mathbb{R}^m;\mathbb{R}^d)} \\ &= \max_{\alpha \in \mathbb{N}^m_{0,k}} \sup_{u \in \mathbb{R}^m} \frac{\left\| y \rho^{(|\alpha|)}\left(a^\top u - b\right) a^\alpha - y_M \rho^{(|\alpha|)}\left(a_M^\top u - b_M\right) a_M^\alpha \right\|}{(1 + \|u\|)^\gamma} \\ &\leqslant \max_{\alpha \in \mathbb{N}^m_{0,k}} \sup_{u \in \overline{B_r(0)}} \left\| y \rho^{(|\alpha|)}\left(a^\top u - b\right) a^\alpha - y_M \rho^{(|\alpha|)}\left(a_M^\top u - b_M\right) a_M^\alpha \right\| \\ &\quad + \max_{\alpha \in \mathbb{N}^m_{0,k}} \sup_{u \in \mathbb{R}^m \setminus \overline{B_r(0)}} \frac{\left\| y \rho^{(|\alpha|)}\left(a^\top u - b\right) a^\alpha \right\|}{(1 + \|u\|)^\gamma} \\ &\quad + \max_{\alpha \in \mathbb{N}^m_{0,k}} \sup_{u \in \mathbb{R}^m \setminus \overline{B_r(0)}} \frac{\left\| y_M \rho^{(|\alpha|)}\left(a_M^\top u - b_M\right) a_M^\alpha \right\|}{(1 + \|u\|)^\gamma} \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ was chosen arbitrarily, this shows that $\mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R} \ni (y, a, b) \mapsto y\rho\left(a^\top \cdot - b\right) \in \overline{C^k_b(U;\mathbb{R}^d)}^\gamma$ is continuous. Hence, by using that $\left(\overline{C^k_b(\mathbb{R}^m;\mathbb{R}^d)}^\gamma, \|\cdot\|_{C^k_{pol,\gamma}(\mathbb{R}^m;\mathbb{R}^d)}\right) \ni f \mapsto f|_U \in (X, \|\cdot\|_X)$ is by Assumption 3.5 continuous, we obtain the conclusion in Part (ii).

For Part (iii)., we define the subsets

$$\mathcal{N}^{\cos}_{U,d}[\mathbb{A}] := \left\{ U \ni u \mapsto \sum_{n=1}^N y_n \cos\left(a_n^\top \cdot - b_n\right) \in \mathbb{R}^d : \begin{array}{c} N \in \mathbb{N},\ y_1, ..., y_N \in \mathbb{A}^d, \\ a_1, ..., a_N \in \mathbb{A}^m,\ b_1, ..., b_N \in \mathbb{A} \end{array} \right\} \subseteq X,$$

for $\mathbb{A} \in \{\mathbb{Q}, \mathbb{R}\}$. Then, by using that the map $\mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R} \ni (y,a,b) \mapsto y\rho\big(a^\top \cdot - b\big) \in X$ is continuous (see Part 2.), we conclude that $\mathcal{N}_{U,d}^{\cos}[\mathbb{R}]$ is contained in the closure of $\mathcal{N}_{U,d}^{\cos}[\mathbb{Q}]$ with respect to $\|\cdot\|_X$. Moreover, by using that $\cos \in \overline{C_b^k(\mathbb{R})}^\gamma$ is non-polynomial, we can apply [64, Theorem 2.8] to conclude that $\mathcal{N}_{U,d}^{\cos}[\mathbb{R}]$ is dense in X. Hence, by combining these two arguments, we obtain that $\mathcal{N}_{U,d}^{\cos}[\mathbb{Q}]$ is also dense in X. Since $\mathcal{N}_{U,d}^{\cos}[\mathbb{Q}]$ is countable, this shows that $(X, \|\cdot\|_X)$ is separable. □

Proof of Corollary 3.8. We aim to apply Theorem 3.2 with Banach space $(X, \|\cdot\|_X)$ satisfying Assumption 3.5. To this end, we first observe that $(X, \|\cdot\|_X)$ is by Lemma 7.1.(iii) separable. Moreover, we choose $\Theta := \mathbb{R}^m \times \mathbb{R}$ and let $(\theta_n)_{n\in\mathbb{N}} := (a_n, b_n)_{n\in\mathbb{N}} : \Omega \to \mathbb{R}^m \times \mathbb{R}$ be an i.i.d. sequence satisfying Assumption 3.7, which implies Assumption 3.1. In addition, we define

$$\mathcal{G} := \left\{ \mathbb{R}^m \times \mathbb{R} \ni (\vartheta_1, \vartheta_2) \mapsto e_i \rho\left(\vartheta_1^\top \cdot - \vartheta_2\right) \in X : i = 1, ..., d \right\}.$$

Since $\mathbb{R}^m \times \mathbb{R} \ni (\vartheta_1, \vartheta_2) \mapsto e_i \rho\big(\vartheta_1^\top \cdot - \vartheta_2\big) \in X$ is by Lemma 7.1.(ii) continuous, we have $\mathcal{G} \subseteq C^0(\Theta; X)$. Moreover, we observe that

$$\operatorname{span}_{\mathbb{R}}(\mathcal{G}(\Theta)) = \operatorname{span}_{\mathbb{R}}\left(\left\{ U \ni u \mapsto e_i \rho\left(\vartheta_1^\top u - \vartheta_2\right) \in \mathbb{R}^d : (\vartheta_1, \vartheta_2) \in \mathbb{R}^m \times \mathbb{R},\ i = 1, ..., d \right\} \right),$$

forms the set of deterministic (i.e. fully trained) neural networks with activation function $\rho \in \overline{C_b^k(\mathbb{R})}^\gamma$, where $e_i \in \mathbb{R}^d$ denotes the i-th unit vector of $\mathbb{R}^d$. Since $\rho \in \overline{C_b^k(\mathbb{R})}^\gamma$ is non-polynomial, we can apply [64, Theorem 2.8] to conclude that $\operatorname{span}_{\mathbb{R}}(\mathcal{G}(\Theta))$ is dense in X. Hence, the conclusion follows from Theorem 3.2. □

8. Proof of results in Section 4

8.1. **Proof of Theorem 4.5.** For the proof of Theorem 4.5, we first show the following auxiliary lemma about Banach space types.

Lemma 8.1. *Let $(X, \|\cdot\|_X)$ be a Banach space of type $t \in [1,2]$ with constant $C_X > 0$, and let $t' \in [1,t]$. Then, $(X, \|\cdot\|_X)$ is a Banach space of type t' with constant $C_X > 0$.*

Proof. Fix some $N \in \mathbb{N}$, $(x_n)_{n=1,...,N} \subseteq X$, and a Rademacher sequence $(\epsilon_n)_{n=1,...,N}$ defined on a (possibly different) probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$. Then, by using Jensen's inequality and the inequality $\big(\sum_{n=1}^N x_n\big)^{t'/t} \leqslant \sum_{n=1}^N x_n^{t'/t}$ for any $x_1, ..., x_N \geqslant 0$, it follows that

$$\widetilde{\mathbb{E}}\left[\left\| \sum_{n=1}^N \epsilon_n x_n \right\|_X^{t'} \right]^{\frac{1}{t'}} \leqslant \widetilde{\mathbb{E}}\left[\left\| \sum_{n=1}^N \epsilon_n x_n \right\|_X^{t} \right]^{\frac{1}{t}} \leqslant C_X \left(\sum_{n=1}^N \|x_n\|_X^t \right)^{\frac{1}{t}} \leqslant C_X \left(\sum_{n=1}^N \|x_n\|_X^{t'} \right)^{\frac{1}{t'}}.$$

This shows that $(X, \|\cdot\|_X)$ is a Banach space of type $t' \in [1,t]$ with constant $C_X > 0$. □

Proof of Theorem 4.5. Fix some $x \in \mathbb{B}_{\mathcal{G},\theta}^r(X)$ and $N \in \mathbb{N}$. Then, by definition of $\mathbb{B}_{\mathcal{G},\theta}^r(X)$, there exists a $\mathcal{B}(\Theta)/\mathcal{B}(\mathbb{R}^e)$-measurable map $y := (y_1, ..., y_e)^\top : \Theta \to \mathbb{R}^e$ such that $x = \mathbb{E}\left[\sum_{i=1}^e y_i(\theta_1) g_i(\theta_1)\right] \in X$ and

$$\mathbb{E}\left[\left\| \sum_{i=1}^e y(\theta_1) g(\theta_1) \right\|_X^r \right]^{\frac{1}{r}} \leqslant 2\|x\|_{\mathbb{B}_{\mathcal{G},\theta}^r(X)} < \infty. \tag{37}$$

From this, we define for every fixed $n = 1, ..., N$, the map

$$\Omega \ni \omega \quad \mapsto \quad G_n(\omega) := \sum_{i=1}^e y_l(\theta_n(\omega)) g_l(\theta_n(\omega)) \in X.$$

Then, by using that $\theta_n : \Omega \to \Theta$ is by definition $\mathcal{F}_\theta/\mathcal{B}(\Theta)$-measurable and that $y := (y_1, ..., y_e)^\top : \Theta \to \mathbb{R}^e$ is by definition $\mathcal{B}(\Theta)/\mathcal{B}(\mathbb{R}^e)$-measurable, the concatenation $\Omega \ni \omega \mapsto G_n := \sum_{l=1}^e y_l(\theta_n(\omega)) g_l(\theta_n(\omega)) \in X$ is $\mathcal{F}_\theta/\mathcal{B}(X)$-measurable. Hence, by using that $(X, \|\cdot\|_X)$ is separable, we can apply [44, Theorem 1.1.6+1.1.20] to conclude that $G_n : \Omega \to X$ is strongly $(\mathbb{P}, \mathcal{F}_\theta)$-measurable. Thus, (37) and $\theta_n \sim \theta_1$ ensure that $G_n \in L^r(\Omega, \mathcal{F}_\theta, \mathbb{P}; X)$.

Now, by using that $x = \mathbb{E}\left[\sum_{i=1}^e y_i(\theta_1) g_i(\theta_1)\right] = \mathbb{E}\left[\sum_{i=1}^e y_i(\theta_n) g_i(\theta_n)\right] = \mathbb{E}[G_n] \in X$ for any $n = 1, ..., N$, the right-hand side of [51, Lemma 6.3] for the independent mean-zero random variables $(\mathbb{E}[G_n] - G_n)_{n=1,...,N}$ with Rademacher sequence $(\epsilon_n)_{n=1,...,N}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ independent of $(\mathbb{E}[G_n] -$

$G_n)_{n=1,...,N}$, the Kahane-Khintchine inequality in [44, Theorem 3.2.23] with constant $\kappa_{r,\min(r,t)} > 0$ (depending only on $r \in [1,\infty)$ and $\min(r,t) \in [1,2]$), that $(X, \|\cdot\|_X)$ is by assumption a Banach space of type $t \in [1,2]$ (with constant $C_X > 0$), thus by Lemma 8.1 of type $\min(r,t) \in (1,t]$ (with the same constant $C_X > 0$), and that $(\mathbb{E}[G_n] - G_n)_{n=1,...,N} \sim \mathbb{E}[G_1] - G_1$ are identically distributed, it follows for the random feature model $G_N := \frac{1}{N}\sum_{n=1}^N G_n \in \mathcal{RG} \cap L^r(\Omega, \mathcal{F}_\theta, \mathbb{P}; X)$ that

$$\begin{aligned}
\mathbb{E}\left[\|x - G_N\|_X^r\right]^{\frac{1}{r}} &= \frac{1}{N}\mathbb{E}\left[\left\|\sum_{n=1}^N (\mathbb{E}[G_n] - G_n)\right\|_X^r\right]^{\frac{1}{r}} \\
&\leqslant \frac{2}{N}\mathbb{E}\left[\left\|\sum_{n=1}^N \epsilon_n (\mathbb{E}[G_n] - G_n)\right\|_X^r\right]^{\frac{1}{r}} \\
&\leqslant \frac{2\kappa_{r,\min(r,t)}}{N}\mathbb{E}\left[\left\|\sum_{n=1}^N \epsilon_n (\mathbb{E}[G_n] - G_n)\right\|_X^{\min(r,t)}\right]^{\frac{1}{\min(r,t)}} \\
&\leqslant \frac{2C_X\kappa_{r,\min(r,t)}}{N}\left(\sum_{n=1}^N \mathbb{E}\left[\|\mathbb{E}[G_n] - G_n\|_X^{\min(r,t)}\right]\right)^{\frac{1}{\min(r,t)}} \\
&= \frac{2C_X\kappa_{r,\min(r,t)}}{N^{1-\frac{1}{\min(r,t)}}}\mathbb{E}\left[\|\mathbb{E}[G_1] - G_1\|_X^{\min(r,t)}\right]^{\frac{1}{\min(r,t)}}.
\end{aligned}$$

Hence, by using Jensen's inequality, Minkowski's inequality, [44, Proposition 1.2.2], the inequality (37), and the constant $C_{r,t} := 8\kappa_{r,\min(r,t)} > 0$ (depending only on $r \in [1,\infty)$ and $t \in [1,2]$), we conclude for $G_N := \frac{1}{N}\sum_{n=1}^N G_n \in \mathcal{RG} \cap L^r(\Omega, \mathcal{F}_\theta, \mathbb{P}; X)$ that

$$\begin{aligned}
\mathbb{E}\left[\|x - G_N\|_X^r\right]^{\frac{1}{r}} &\leqslant \frac{2C_X\kappa_{r,\min(r,t)}}{N^{1-\frac{1}{\min(r,t)}}}\mathbb{E}\left[\|\mathbb{E}[G_1] - G_1\|_X^r\right]^{\frac{1}{r}} \\
&\leqslant \frac{2C_X\kappa_{r,\min(r,t)}}{N^{1-\frac{1}{\min(r,t)}}}\left(\|\mathbb{E}[G_1]\|_X + \mathbb{E}\left[\|G_1\|_X^r\right]^{\frac{1}{r}}\right) \\
&\leqslant \frac{4C_X\kappa_{r,\min(r,t)}}{N^{1-\frac{1}{\min(r,t)}}}\|G_1\|_{L^r(\Omega,\mathcal{F},\mathbb{P};X)} \\
&\leqslant C_{r,t}C_X\frac{\|x\|_{\mathbb{B}^r_{\mathcal{G},\theta}(X)}}{N^{1-\frac{1}{\min(r,t)}}},
\end{aligned}$$

which completes the proof. ∎

8.2. Proof of Corollary 4.8+4.9+4.20 and Proposition 4.22.

Proof of Corollary 4.8. We aim to apply Theorem 4.5 onto a fixed function $f \in W^{k,p}(U, \mathcal{L}(U), du) \cap L^1(\mathbb{R}^m, \mathcal{L}(\mathbb{R}^m), du)$ with $C_f := \left(\int_{\mathbb{R}^m} \frac{|\widehat{f}(\vartheta)|^r\left(1+\|\vartheta\|^2\right)^{kr/2}}{p_\theta(\vartheta)^{r-1}} d\vartheta\right)^{1/r} < \infty$. To this end, we first observe that $(W^{k,p}(U, \mathcal{L}(U), w; \mathbb{R}^d), \|\cdot\|_{W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d)})$ is a separable Banach space (see [64, Lemma 4.7]). Moreover, we define the linear readouts

$$\begin{aligned}
\mathbb{R}^m \ni \vartheta \quad &\mapsto \quad y_1(\vartheta) := \frac{\operatorname{Re}\left(\widehat{f}(\vartheta)\right)}{(2\pi)^m p_\theta(\vartheta)} \in \mathbb{R}, \qquad \text{and} \\
\mathbb{R}^m \ni \vartheta \quad &\mapsto \quad y_2(\vartheta) := -\frac{\operatorname{Im}\left(\widehat{f}(\vartheta)\right)}{(2\pi)^m p_\theta(\vartheta)} \in \mathbb{R},
\end{aligned} \tag{38}$$

which are $\mathcal{B}(\mathbb{R}^m)/\mathcal{B}(\mathbb{R})$-measurable as composition of the continuous function $\mathbb{R}^m \ni \vartheta \mapsto \widehat{f}(\vartheta) \in \mathbb{C}$ (see [30, p. 214]) and the $\mathcal{B}(\mathbb{C})/\mathcal{B}(\mathbb{R})$-measurable functions returning the real and imaginary part. Then, by using Jensen's inequality, it follows that

$$\begin{aligned}\int_{\mathbb{R}^m} |\widehat{f}(\vartheta)| d\vartheta &\leqslant \int_{\mathbb{R}^m} \frac{|\widehat{f}(\vartheta)|}{p_\theta(\vartheta)} p_\theta(\vartheta) d\vartheta \leqslant \left(\int_{\mathbb{R}^m} \frac{|\widehat{f}(\vartheta)|^r}{p_\theta(\vartheta)^r} p_\theta(\vartheta) d\vartheta \right)^{\frac{1}{r}} \\ &\leqslant \left(\int_{\mathbb{R}^m} \frac{|\widehat{f}(\vartheta)|^r}{p_\theta(\vartheta)^{r-1}} \left(1 + \|\vartheta\|^2\right)^{\frac{kr}{2}} d\vartheta \right)^{\frac{1}{r}} = C_f < \infty,\end{aligned}$$

which shows that $\widehat{f} \in L^1(\mathbb{R}^m, \mathcal{L}(\mathbb{R}^m), du; \mathbb{C})$. Hence, we can apply the Fourier inversion theorem (see [30, Equation 7.14]) and use that the left-hand side is real-valued to conclude for a.e. $u \in \mathbb{R}^m$ that

$$\begin{aligned} x(u) &= \frac{1}{(2\pi)^m} \int_{\mathbb{R}^m} \widehat{f}(\vartheta) e^{\mathbf{i}\vartheta^\top u} d\vartheta = \int_{\mathbb{R}^m} \frac{\widehat{f}(\vartheta)}{(2\pi)^m p_\theta(\vartheta)} e^{\mathbf{i}\vartheta^\top u} p_\theta(\vartheta) d\vartheta \\ &= \mathbb{E}\left[\frac{\widehat{f}(\theta_1)}{(2\pi)^m p_\theta(\theta_1)} \cos\left(\theta_1^\top u\right) + \frac{\widehat{f}(\theta_1)}{(2\pi)^m p_\theta(\theta_1)} \mathbf{i} \sin\left(\theta_1^\top u\right) \right] \\ &= \mathbb{E}\left[\frac{\operatorname{Re}\left(\widehat{f}(\theta_1)\right)}{(2\pi)^m p_\theta(\theta_1)} \cos\left(\theta_1^\top u\right) - \frac{\operatorname{Im}\left(\widehat{f}(\theta_1)\right)}{(2\pi)^m p_\theta(\theta_1)} \sin\left(\theta_1^\top u\right) \right] \\ &= \mathbb{E}\left[y_1(\theta_1) \cos\left(\theta_1^\top u\right) + y_2(\theta_1) \sin\left(\theta_1^\top u\right) \right]. \end{aligned} \tag{39}$$

Thus, by using integration by parts, it holds that

$$\begin{aligned} \int_U \mathbb{E}\left[y_1(\theta_1) \cos\left(\theta_1^\top u\right) + y_2(\theta_1) \sin\left(\theta_1^\top u\right) \right] \partial_\alpha h(u) du &= \int_U x(u) \partial_\alpha h(u) du \\ &= (-1)^{|\alpha|} \int_U \partial_\alpha f(u) h(u) du, \end{aligned} \tag{40}$$

which shows that the weak derivatives of $\mathbb{E}\left[y_1(\theta_1) \cos\left(\theta_1^\top \cdot\right) + y_2(\theta_1) \sin\left(\theta_1^\top \cdot\right) \right]$ and f coincide, and thus implies that $\mathbb{E}\left[y_1(\theta_1) \cos\left(\theta_1^\top \cdot\right) + y_2(\theta_1) \sin\left(\theta_1^\top \cdot\right) \right] = f \in W^{k,p}(U, \mathcal{L}(U), w)$. In addition, by using that $\left|\vartheta^\alpha\right| := \left| \prod_{l=1}^m \vartheta_l^{\alpha_l} \right| = \prod_{l=1}^m |\vartheta_l|^{\alpha_l} \leqslant \prod_{l=1}^m \left(1 + \|\vartheta\|^2\right)^{\alpha_l/2} = \left(1 + \|\vartheta\|^2\right)^{k/2}$ for any $\alpha := (\alpha_1, ..., \alpha_m) \in \mathbb{N}_{0,k}^m$ and $\vartheta := (\vartheta_1, ..., \vartheta_m) \in \mathbb{R}^m$, that $\left|\mathbb{N}_{0,k}^m\right| = \sum_{j=0}^k m^j \leqslant 2m^k$, and that $w(U) := \int_U w(u) du < \infty$, we conclude for every $\vartheta \in \mathbb{R}^m$ that

$$\begin{aligned} \left\|\cos\left(\vartheta^\top \cdot\right)\right\|_{W^{k,p}(U,\mathcal{L}(U),w)} &= \left(\sum_{\alpha \in \mathbb{N}_{0,k}^m} \int_U \left|\partial_\alpha \left(\cos\left(\vartheta^\top u\right)\right)\right|^p w(u) du \right)^{\frac{1}{p}} \\ &= \left(\sum_{\alpha \in \mathbb{N}_{0,k}^m} \int_U \left|\cos^{(|\alpha|)}\left(\vartheta^\top u\right) \vartheta^\alpha\right|^p w(u) du \right)^{\frac{1}{p}} \\ &\leqslant \left|\mathbb{N}_{0,k}^m\right|^{\frac{1}{p}} \left(1 + \|\vartheta\|^2\right)^{\frac{k}{2}} \left(\int_U w(u) du \right)^{\frac{1}{p}} \\ &\leqslant 2^{\frac{1}{p}} m^{\frac{k}{p}} \left(1 + \|\vartheta\|^2\right)^{\frac{k}{2}} w(U)^{\frac{1}{p}}. \end{aligned} \tag{41}$$

Moreover, by using the same arguments as in (41), we also obtain that

$$\left\|\sin\left(\vartheta^\top \cdot\right)\right\|_{W^{k,p}(U,\mathcal{L}(U),w)} \leqslant 2^{\frac{1}{p}} m^{\frac{k}{p}} \left(1 + \|\vartheta\|^2\right)^{\frac{k}{2}} w(U)^{\frac{1}{p}}. \tag{42}$$

Hence, by using that $\mathbb{E}\left[y_1(\theta_1)\cos\left(\theta_1^\top\cdot\right) + y_2(\theta_1)\sin\left(\theta_1^\top\cdot\right)\right] = f \in W^{k,p}(U,\mathcal{L}(U),w)$, the inequalities (41)+(42), and that $|\operatorname{Re}(z)| + |\operatorname{Im}(z)| \leqslant \sqrt{2}|z|$ for any $z \in \mathbb{C}$, it follows that

$$\begin{aligned}
\|f\|_{\mathbb{B}^r_{\mathcal{G},\theta}(W^{k,p}(U,\mathcal{L}(U),w))} &\leqslant \mathbb{E}\left[\left\|y_1(\theta_1)\cos\left(\theta_1^\top\cdot\right) + y_2(\theta_1)\sin\left(\theta_1^\top\cdot\right)\right\|^r_{W^{k,p}(U,\mathcal{L}(U),w)}\right]^{\frac{1}{r}} \\
&\leqslant \mathbb{E}\left[\left(|y_1(\theta_1)|\left\|\cos\left(\theta_1^\top\cdot\right)\right\|_{W^{k,p}(U,\mathcal{L}(U),w)} + |y_2(\theta_1)|\left\|\sin\left(\theta_1^\top\cdot\right)\right\|_{W^{k,p}(U,\mathcal{L}(U),w)}\right)^r\right]^{\frac{1}{r}} \\
&\leqslant \frac{2^{\frac{1}{p}} m^{\frac{k}{p}} w(U)^{\frac{1}{p}}}{(2\pi)^m}\mathbb{E}\left[\frac{\left(\left|\operatorname{Re}\left(\widehat{f}(\theta_1)\right)\right| + \left|\operatorname{Im}\left(\widehat{f}(\theta_1)\right)\right|\right)^r}{p_\theta(\theta_1)^r}\left(1+\|\theta_1\|^2\right)^{\frac{kr}{2}}\right]^{\frac{1}{r}} \\
&\leqslant \frac{2^{\frac{1}{p}+\frac{1}{2}} m^{\frac{k}{p}} w(U)^{\frac{1}{p}}}{(2\pi)^m}\left(\int_{\mathbb{R}^m}\frac{\left|\widehat{f}(\vartheta)\right|^r}{p_\theta(\vartheta)^r}\left(1+\|\vartheta\|^2\right)^{\frac{kr}{2}} p_\theta(\vartheta)d\vartheta\right)^{\frac{1}{r}} \\
&= \frac{2^{\frac{1}{p}+\frac{1}{2}} m^{\frac{k}{p}} w(U)^{\frac{1}{p}}}{(2\pi)^m} C_f < \infty,
\end{aligned} \tag{43}$$

which shows that $f \in \mathbb{B}^r_{\mathcal{G},\theta}(X)$. Thus, by using that $(W^{k,p}(U,\mathcal{L}(U),w), \|\cdot\|_{W^{k,p}(U,\mathcal{L}(U),w)})$ is a Banach space of type $t = \min(2,p)$ with constant $C_{W^{k,p}(U,\mathcal{L}(U),w)} > 0$ depending only on $p \in [1,\infty)$ (see [64, Lemma 4.9]), we can apply Theorem 4.5 (with constant $C_{r,t} > 0$ depending only on $r \in [1,\infty)$ and $t \in [1,2]$), insert the inequality (43), and define the constant $C_{p,r} := 2^{1/p+1/2} C_{r,t} C_{W^{k,p}(U,\mathcal{L}(U),w)} > 0$ (depending only on $p,r \in [1,\infty)$) to conclude that there exists a random trigonometric feature model $G_N \in \mathcal{RT}_{U,1} \cap L^r(\Omega,\mathcal{F}_\theta,\mathbb{P};W^{k,p}(U,\mathcal{L}(U),w))$ with N features satisfying

$$\begin{aligned}
\mathbb{E}\left[\|f - G_N\|^r_{W^{k,p}(U,\mathcal{L}(U),w)}\right]^{\frac{1}{r}} &\leqslant C_{r,t} C_{W^{k,p}(U,\mathcal{L}(U),w)} \frac{\|f\|_{\mathbb{B}^r_{\mathcal{G},\theta}(W^{k,p}(U,\mathcal{L}(U),w))}}{N^{1-\frac{1}{\min(2,p,r)}}} \\
&\leqslant C_{p,r}\frac{m^{\frac{k}{p}} w(U)^{\frac{1}{p}}}{(2\pi)^m}\frac{C_f}{N^{1-\frac{1}{\min(2,p,r)}}},
\end{aligned}$$

which completes the proof. □

Proof of Corollary 4.9. We aim to apply Theorem 4.5 onto a fixed function $f \in W^{k,p}(U,\mathcal{L}(U),du;\mathbb{C}) \cap L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du;\mathbb{C})$ with $\left(\int_{\mathbb{R}^m}\frac{|\widehat{f}(\vartheta)|^r\left(1+\|\vartheta\|^2\right)^{kr/2}}{p_\theta(\vartheta)^{r-1}}d\vartheta\right)^{1/r} < \infty$. To this end, we first observe that $(W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d), \|\cdot\|_{W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d)})$ is a separable Banach space (see [64, Lemma 4.7]). Moreover, we define

$$\mathbb{R}^m \ni \vartheta \quad\mapsto\quad y(\vartheta) := \frac{\widehat{f}(\vartheta)}{(2\pi)^m p_\theta(\vartheta)} \in \mathbb{C},$$

and follow the proof of Corollary 4.9, where the Fourier inversion theorem is applied to conclude for a.e. $u \in \mathbb{R}^m$ that

$$f(u) = \frac{1}{(2\pi)^m}\int_{\mathbb{R}^m}\widehat{f}(\vartheta)e^{\mathbf{i}\vartheta^\top u}d\vartheta = \int_{\mathbb{R}^m}\frac{\widehat{f}(\vartheta)}{(2\pi)^m p_\theta(\vartheta)}e^{\mathbf{i}\vartheta^\top u}p_\theta(\vartheta)d\vartheta = \mathbb{E}\left[y(\theta_1)\exp\left(\mathbf{i}\theta_1^\top u\right)\right].$$

Hence, by using the same steps as in the proof of Corollary 4.9, we obtain the conclusion from Theorem 4.5. □

Proof of Lemma 4.19. Let $\mathcal{G}$ as in Remark 4.17, let $\psi \in \mathcal{S}_0(\mathbb{R};\mathbb{C})$, let $\rho \in C^k_{pol,\gamma}(\mathbb{R})$, and fix some $f \in \widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)$. Then, there exists by definition of $\widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)$ some $h \in L^1(\mathbb{R}^m;\mathcal{L}(\mathbb{R}^m),du;\mathbb{R}^d)$ such that $\widehat{h} \in L^1(\mathbb{R}^m;\mathcal{L}(\mathbb{R}^m),du;\mathbb{C}^d)$, $h = f$ a.e. on U, and

$$\mathbb{E}\left[\left\|\frac{\left(1+\|a_1\|^2\right)^{\frac{\gamma+k}{2}}\left(1+|b_1|^2\right)^{\frac{\gamma}{2}}}{p_{a,b}(a_1,b_1)}(\mathfrak{R}_\psi h)(a_1,b_1)\right\|^r\right]^{\frac{1}{r}} \leqslant 2\|f\|_{\widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)} < \infty. \tag{44}$$

Moreover, we recall that $\mathcal{G}$ consists of the feature maps given by

$$\mathbb{R}^m \times \mathbb{R} \ni (a,b) \quad\mapsto\quad g_i(a,b) := e_i\rho\left(a^\top\cdot - b\right) \in W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d), \qquad i = 1,\ldots,d, \tag{45}$$

where $e_i \in \mathbb{R}^d$ denotes the i-th unit vector. In addition, we define the linear readout

$$\mathbb{R}^m \times \mathbb{R} \ni (a,b) \quad \mapsto \quad y(a,b) := (y_i(a,b))_{i=1,\ldots,d}^\top := \mathrm{Re}\left(\frac{(\mathfrak{R}_\psi f)(a,b)}{C_m^{(\psi,\rho)} p_{a,b}(a,b)} \right) \in \mathbb{R}^d. \tag{46}$$

Then, by using Proposition 4.16 together with the fact that $h : \mathbb{R}^m \to \mathbb{R}^d$ is $\mathbb{R}^d$-valued and that $h = f$ a.e. on U, it follows for a.e. $u \in U$ that

$$\mathbb{E}\left[y(a_1,b_1)\rho\left(a_1^\top u - b_1\right)\right] = h(u) = f(u),$$

which implies by following the arguments of (40) that

$$\mathbb{E}\left[\sum_{i=1}^d y_i(a_1,b_1) e_i \rho\left(a_1^\top \cdot - b_1\right)\right] = \mathbb{E}\left[y(a_1,b_1)\rho\left(a_1^\top \cdot - b_1\right)\right] = f \in W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d).$$

Hence, by using that $|a^\alpha| := \left|\prod_{l=1}^m a_l^{\alpha_l}\right| = \prod_{l=1}^m |a_l|^{\alpha_l} \leqslant \left(1+\|a\|^2\right)^{|\alpha|/2} \leqslant \left(1+\|a\|^2\right)^{k/2}$ for any $\alpha := (\alpha_1,\ldots,\alpha_m) \in \mathbb{N}_{0,k}^m$ and $a := (a_1,\ldots,a_m) \in \mathbb{R}^m$, the inequality [64, Equation 42] (with constant $C_{U,w}^{(\gamma,p)} > 0$ defined in (10)), that $\left|\mathbb{N}_{0,k}^m\right| = \sum_{j=0}^k m^j \leqslant 2m^k$, and (44), we have

$$\begin{aligned}
\|f\|_{\mathbb{B}^r_{\mathcal{G},\theta}(W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d))} &\leqslant \mathbb{E}\left[\left\| \sum_{i=1}^d y_i(a_1,b_1)e_i\rho\left(a_1^\top u - b_1\right)\right\|^r_{W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d)}\right]^{\frac{1}{r}} \\
&= \mathbb{E}\left[\left(\sum_{\alpha\in\mathbb{N}_{0,k}^m} \int_U \left| \partial_\alpha \left(y(a_1,b_1)\rho\left(a_1^\top u - b_1\right)\right)\right|^p w(u)du \right)^{\frac{r}{p}}\right]^{\frac{1}{r}} \\
&\leqslant \mathbb{E}\left[\left(\sum_{\alpha\in\mathbb{N}_{0,k}^m} \left\| \mathrm{Re}\left(\frac{a_1^\alpha (\mathfrak{R}_\psi g)(a_1,b_1)}{C_m^{(\psi,\rho)} p_{a,b}(a_1,b_1)}\right)\right\|^p \int_U \left| \rho^{(|\alpha|)}\left(a_1^\top u - b_1\right)\right|^p du \right)^{\frac{r}{p}}\right]^{\frac{1}{r}} \\
&\leqslant 4\|\rho\|_{C^k_{pol,\gamma}(\mathbb{R})} \frac{C_{U,w}^{(\gamma,p)} \left|\mathbb{N}_{0,k}^m\right|^{\frac{1}{p}}}{\left|C_m^{(\psi,\rho)}\right|} \mathbb{E}\left[\left\| \frac{\left(1+\|a_1\|^2\right)^{\frac{\gamma+k}{2}}\left(1+|b_1|^2\right)^{\frac{\gamma}{2}}}{p_{a,b}(a_1,b_1)} (\mathfrak{R}_\psi g)(a_1,b_1)\right\|^r\right]^{\frac{1}{r}} \\
&\leqslant 2^{3+\frac{1}{p}} \|\rho\|_{C^k_{pol,\gamma}(\mathbb{R})} \frac{C_{U,w}^{(\gamma,p)} m^{\frac{k}{p}}}{\left|C_m^{(\psi,\rho)}\right|} \|f\|_{\widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)} < \infty,
\end{aligned} \tag{47}$$

which shows that $f \in \mathbb{B}^r_{\mathcal{G},\theta}(W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d))$. □

Proof of Corollary 4.20. We aim to apply Theorem 4.5 onto a fixed function $f \in W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d) \cap \widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)$, where we recall that $(W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d), \|\cdot\|_{W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d)})$ is separable. To this end, we use that there exists by definition of $\widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)$ some $h \in L^1(\mathbb{R}^m;\mathcal{L}(\mathbb{R}^m),du;\mathbb{R}^d)$ such that $\widehat{h} \in L^1(\mathbb{R}^m;\mathcal{L}(\mathbb{R}^m),du;\mathbb{C}^d)$, $h = f$ a.e. on U, and (44) holds true. Moreover, we recall that the feature maps are given by (45) and define the linear $y : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^d$ as in (46). Then, by using that $(W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d), \|\cdot\|_{W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d)})$ is a Banach space of type $t = \min(2,p)$ with constant $C_{W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d)} > 0$ depending only on $p \in [1,\infty)$ (see [64, Lemma 4.9]), we can use Theorem 4.5 (with constant $C_{r,t} > 0$ depending only on $r \in [1,\infty)$ and $t \in [1,2]$), Lemma 4.19, and the constant $C_{p,r} := 2^{3+1/p} C_{r,t} C_{W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d)} > 0$ (depending only on $p,r \in [1,\infty)$) to conclude that there exists a random neural network $G_N \in \mathcal{RN}^\rho_{U,d} \cap L^r(\Omega,\mathcal{F}_{a,b},\mathbb{P};W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d))$ with N neurons satisfying

$$\begin{aligned}
\mathbb{E}\left[\|f - G_N\|^r_{W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d)}\right]^{\frac{1}{r}} &\leqslant C_{r,t} C_{W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d)} \frac{\|f\|_{\mathbb{B}^r_{\mathcal{G},\theta}(W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d))}}{N^{1-\frac{1}{\min(2,p,r)}}} \\
&\leqslant C_{p,r} \|\rho\|_{C^k_{pol,\gamma}(\mathbb{R})} \frac{C_{U,w}^{(\gamma,p)} m^{\frac{k}{p}}}{\left|C_m^{(\psi,\rho)}\right|} \frac{\|f\|_{\widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)}}{N^{1-\frac{1}{\min(2,p,r)}}},
\end{aligned}$$

which completes the proof. □

Proof of Proposition 4.22. For (12), let $f \in L^1(\mathbb{R}^m, \mathcal{L}(\mathbb{R}^m), du; \mathbb{R}^d)$ with $(\lceil\gamma\rceil + 2)$-times differentiable Fourier transform and fix some $c \in \{0, \lceil\gamma\rceil + 2\}$. Then, by using that $b_1 \sim t_1$, the inequality [64, Equation 46], and Minkowski's integral inequality (with measure spaces $(\mathbb{R}^m, \mathcal{L}(\mathbb{R}^m), da)$ and $(\mathbb{N}^m_{0,k} \times \mathbb{R}, \mathcal{P}(\mathbb{N}^m_{0,k}) \otimes \mathcal{B}(\mathbb{R}), d\mu \otimes d\zeta)$, where $\mathcal{P}(\mathbb{N}^m_{0,k})$ denotes the power set of $\mathbb{N}^m_{0,k}$, and where $\mathcal{P}(\mathbb{N}^m_{0,k}) \ni E \mapsto \mu(E) := \sum_{\alpha \in \mathbb{N}^m_{0,k}} \mathbb{1}_E(\alpha) \in [0,\infty)$ is the counting measure) together with the probability disitribution function of $a_1 \sim t_m$, we have

$$
\begin{aligned}
\|f\|_{\widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)} &\leqslant \mathbb{E}\left[\left\| \frac{(1+\|a_1\|^2)^{\frac{\gamma+k}{2}} (1+|b_1|^2)^{\frac{\gamma}{2}}}{p_{a,b}(a_1,b_1)} (\mathfrak{R}_\psi f)(a_1,b_1) \right\|^r\right]^{\frac{1}{r}} \\
&\leqslant \mathbb{E}\left[\sup_{\widetilde{b}\in\mathbb{R}} \left\| \frac{(1+\|a_1\|^2)^{\frac{\lceil\gamma\rceil+k}{2}}}{p_a(a_1)} \left(1+\left|\widetilde{b}\right|^2\right)^{\frac{\lceil\gamma\rceil+2}{2}} (\mathfrak{R}_\psi f)\left(a_1,\widetilde{b}\right) \right\|^r \right]^{\frac{1}{r}} \\
&\leqslant 2^{\frac{\lceil\gamma\rceil}{2}} \frac{(\lceil\gamma\rceil+2)!}{\pi} \mathbb{E}\left[\left\| \frac{(1+\|a_1\|^2)^{\frac{2\lceil\gamma\rceil+k+2}{2}}}{p_a(a_1)} \sum_{\beta\in\mathbb{N}^m_{0,\lceil\gamma\rceil+2}} \int_{\mathbb{R}} \left\|\partial_\beta \widehat{f}(\zeta a)\right\| \left|\widehat{\psi}^{(\lceil\gamma\rceil+2-|\beta|)}(\zeta)\right| d\zeta \right\|^r\right]^{\frac{1}{r}} \\
&\leqslant 2^{\frac{\lceil\gamma\rceil}{2}} \frac{(\lceil\gamma\rceil+2)!}{\pi} \sum_{\beta\in\mathbb{N}^m_{0,\lceil\gamma\rceil+2}} \int_{\mathbb{R}} \left|\widehat{\psi}^{(\lceil\gamma\rceil+2-|\beta|)}(\zeta)\right| \left(\int_{\mathbb{R}^m} \left\|\partial_\beta \widehat{f}(\zeta a)\right\|^r \frac{(1+\|a\|^2)^{\frac{(2\lceil\gamma\rceil+k+2)r}{2}}}{p_a(a)^{r-1}} da \right)^{\frac{1}{r}} d\zeta.
\end{aligned}
$$

Hence, by using the substitution $\xi \mapsto \zeta a$ with Jacobi determinant $d\xi = |\zeta|^m da$, that $\zeta_1 := \inf\{|\zeta| : \zeta \in \operatorname{supp}(\widehat{\psi})\} > 0$, and $C_1 := 2^{\lceil\gamma\rceil/2}(\lceil\gamma\rceil+2)! \max_{j=0,\dots,\lceil\gamma\rceil+2} \int_{\mathbb{R}} \left|\widehat{\psi}^{(j)}(\zeta)\right| d\zeta > 0$ (depending only on $\gamma \in [0,\infty)$ and $\psi \in \mathcal{S}(\mathbb{R};\mathbb{C})$), it follows that

$$
\begin{aligned}
&\|f\|_{\widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)} \\
&\quad \leqslant \frac{(\lceil\gamma\rceil+2)!}{\pi} \sum_{\beta\in\mathbb{N}^m_{0,\lceil\gamma\rceil+2}} \int_{\mathbb{R}} \frac{\left|\widehat{\psi}^{(\lceil\gamma\rceil+2-|\beta|)}(\zeta)\right|}{\zeta^{\frac{m}{2}}} \left(\int_{\mathbb{R}^m} \left\|\partial_\beta \widehat{f}(\xi)\right\|^r \frac{(1+\|\xi/\zeta\|^2)^{\frac{(2\lceil\gamma\rceil+k+2)r}{2}}}{p_a(\xi/\zeta)^{r-1}} d\xi \right)^{\frac{1}{r}} d\zeta \\
&\quad \leqslant \frac{C_1}{\zeta_1^{\frac{m}{2}}} \sup_{\zeta\in\operatorname{supp}(\widehat{\psi})} \sum_{\beta\in\mathbb{N}^m_{0,\lceil\gamma\rceil+2}} \left(\int_{\mathbb{R}^m} \left\|\partial_\beta \widehat{f}(\xi)\right\|^r \frac{(1+\|\xi/\zeta\|^2)^{\frac{(2\lceil\gamma\rceil+k+2)r}{2}}}{p_a(\xi/\zeta)^{r-1}} d\xi \right)^{\frac{1}{r}},
\end{aligned}
$$

which proves (12). For (13), we use (12) and that $a_1 \sim t_m$ to obtain that

$$
\begin{aligned}
\|f\|_{\widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)} &\leqslant \frac{C_1}{\zeta_1^{\frac{m}{2}}} \sup_{\zeta\in\operatorname{supp}(\widehat{\psi})} \sum_{\beta\in\mathbb{N}^m_{0,\lceil\gamma\rceil+2}} \left(\int_{\mathbb{R}^m} \left\|\widehat{f}(\xi)\right\|^r \frac{(1+\|\xi/\zeta\|^2)^{2\lceil\gamma\rceil+k+2}}{p_a(\xi/\zeta)} d\xi \right)^{\frac{1}{2}} \\
&= \frac{C_1}{\zeta_1^{\frac{m}{2}}} \frac{\pi^{\frac{m+1}{4}}}{\Gamma\left(\frac{m+1}{2}\right)^{\frac{1}{2}}} \sum_{\beta\in\mathbb{N}^m_{0,\lceil\gamma\rceil+2}} \left(\int_{\mathbb{R}^m} \left\|\partial_\beta \widehat{f}(\xi)\right\|^2 (1+\|\xi/\zeta_1\|^2)^{2\lceil\gamma\rceil+k+\frac{m+5}{2}} d\xi \right)^{\frac{1}{2}},
\end{aligned}
$$

which completes the proof. □

Proof of Proposition 4.24. The proof is based on [64, Proposition 3.10] which has established a similar result, but with respect to *deterministic* neural networks. To this end, we fix some $m, d \in \mathbb{N}$ and $\varepsilon > 0$. Moreover, let $p > 1$ and $w : U \to [0,\infty)$ be a weight as in Lemma 4.23 (with constant $C^{(\gamma,p)}_{\mathbb{R},w_0} > 0$ being independent of $m, d \in \mathbb{N}$ and $\varepsilon > 0$), let $(\psi, \rho) \in \mathcal{S}_0(\mathbb{R};\mathbb{C}) \times C^k_{pol,\gamma}(\mathbb{R})$ be a pair as in Example 4.15 (with $0 < \zeta_1 < \zeta_2 < \infty$ and constant $C_{\psi,\rho} > 0$ being independent of $m, d \in \mathbb{N}$ and $\varepsilon > 0$), and fix some $f \in W^{k,p}(U, \mathcal{L}(U), w; \mathbb{R}^d)$ satisfying the conditions of Proposition 4.22 such that the right-hand side of (13) satisfies $\mathcal{O}\left(m^s (2/\zeta_2)^m (m+1)^{m/2}\right)$ for some $s \in \mathbb{N}_0$. Then, there exists some constant $C > 0$

(being independent of $m, d \in \mathbb{N}$ and $\varepsilon > 0$) such that for every $m, d \in \mathbb{N}$ it holds that

$$\frac{C_1}{\zeta_1^{\frac{m}{2}}} \sum_{\beta \in \mathbb{N}^m_{0,\lceil\gamma\rceil+2}} \left(\int_{\mathbb{R}^m} \left\| \partial_\beta \widehat{f}(\xi) \right\|^2 \left(1 + \|\xi/\zeta_1\|^2\right)^{2\lceil\gamma\rceil+k+\frac{m+5}{2}} d\xi \right)^{\frac{1}{2}} \leqslant C m^s \left(\frac{2}{\zeta_2}\right)^m (m+1)^{\frac{m}{2}}. \tag{48}$$

Hence, by using the inequality (13) in Proposition 4.22 together with (48), that $\Gamma(x) \geqslant \sqrt{2\pi/x}(x/e)^x$ for any $x \in (0,\infty)$ (see [33, Lemma 2.4]), and that $\frac{\pi^{m/4}(2/\zeta_2)^m}{(2\pi/\zeta_2)^m(1/(2e))^{m/2}} = \left(\frac{2e\sqrt{\pi}}{\pi^2}\right)^{m/2} \leqslant 1$ for any $m \in \mathbb{N}$, we conclude that there exist some constants $C_2, C_3 > 0$ (being independent of $m, d \in \mathbb{N}$ and $\varepsilon > 0$) such that

$$\begin{aligned}
& C_p \|\rho\|_{C^k_{pol,\gamma}(\mathbb{R})} \frac{C^{(\gamma,p)}_{\mathbb{R},w_0} m^{\gamma+\frac{k+1}{p}}}{C_{\psi,\rho}\left(\frac{2\pi}{\zeta_2}\right)^m} \|f\|_{\widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)} \\
& \leqslant C_p \|\rho\|_{C^k_{pol,\gamma}(\mathbb{R})} \frac{C^{(\gamma,p)}_{\mathbb{R},w_0} m^{\gamma+\frac{k+1}{p}} \pi^{\frac{m+1}{4}}}{C_{\psi,\rho}\left(\frac{2\pi}{\zeta_2}\right)^m \Gamma\left(\frac{m+1}{2}\right)^{\frac{1}{2}}} C m^s \left(\frac{2}{\zeta_2}\right)^m (m+1)^{\frac{m}{2}} \\
& \leqslant C_p \|\rho\|_{C^k_{pol,\gamma}(\mathbb{R})} \frac{C^{(\gamma,p)}_{\mathbb{R},w_0} m^{\gamma+\frac{k+1}{p}} \pi^{\frac{m+1}{4}}}{C_{\psi,\rho}\left(\frac{2\pi}{\zeta_2}\right)^m \left(\frac{4\pi}{m+1}\right)^{\frac{1}{4}} \left(\frac{m+1}{2e}\right)^{\frac{m+1}{2}}} C m^s \left(\frac{2}{\zeta_2}\right)^m (m+1)^{\frac{m}{2}} \\
& \leqslant 2C_p \frac{C_p \|\rho\|_{C^k_{pol,\gamma}(\mathbb{R})} C^{(\gamma,p)}_{\mathbb{R},w_0} \pi^{\frac{1}{4}} (2e)^{\frac{1}{2}} C_3}{C_{\psi,\rho}(4\pi)^{\frac{1}{4}}} C m^s \\
& \leqslant \left(C_2 m^{C_3}\right)^{1-\frac{1}{\min(2,p)}}.
\end{aligned} \tag{49}$$

Hence, by using that $f \in \widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)$ (see Proposition 4.22), we can apply Theorem 4.5 with $N = \left\lceil C_2 m^{C_3} \varepsilon^{-\frac{\min(2,p)}{\min(2,p)-1}} \right\rceil$ and insert the inequality (49) to obtain a random neural network $G_N \in \mathcal{RN}^\rho_{U,d}$ with N neurons satisfying

$$\begin{aligned}
\mathbb{E}\left[\|f - G_N\|^r_{W^{k,p}(U,\mathcal{L}(U),w;\mathbb{R}^d)}\right]^{\frac{1}{r}} & \leqslant C_{p,r} \|\rho\|_{C^k_{pol,\gamma}(\mathbb{R})} \frac{C^{(\gamma,p)}_{U,w} m^{\frac{k}{p}}}{\left|C^{(\psi,\rho)}_m\right|} \frac{\|f\|_{\widetilde{\mathbb{B}}^{k,r,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)}}{N^{1-\frac{1}{\min(2,p,r)}}} \\
& \leqslant \frac{\left(C_2 m^{C_3}\right)^{1-\frac{1}{\min(2,p)}}}{N^{1-\frac{1}{\min(2,p,r)}}} \leqslant \varepsilon,
\end{aligned}$$

which completes the proof. □

9. Proof of results in Section 5

9.1. Proof of Proposition 5.1.

Proof of Proposition 5.1. Fix some $J, N \in \mathbb{N}$ and a k-times weakly differentiable function $f := (f_1, ..., f_d)^\top : U \to \mathbb{R}^d$. Moreover, in order to ease notation, we define $\widetilde{m} := J|\mathbb{N}^m_{0,k}|d \in \mathbb{N}$ and $\widetilde{n} := eN \in \mathbb{N}$. Then, by using the definition of the Euclidean norm, we first observe that (15) is equivalent to

$$y^{(J)}(\omega) = \arg\min_{y \in \mathcal{Y}_N} \left(\frac{1}{J} \sum_{j=1}^J \sum_{\alpha \in \mathbb{N}^m_{0,k}} \sum_{i=1}^d c_\alpha^2 \left| \partial_\alpha f_i(V_j(\omega)) - \partial_\alpha G^y_{N,i}(\omega)(V_j(\omega)) \right|^2 \right), \tag{50}$$

where $G^y_N(\omega) := (G^y_{N,1}(\omega), ..., G^y_{N,d}(\omega))^\top \in W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d)$ is defined in (14). Hence, for every fixed $\omega \in \Omega$, the least squares problem (50) is by [12, Theorem 1.1.2] equivalent to the normal equations $\mathbf{G}(\omega)^\top \mathbf{G}(\omega) \vec{y}^{(J)}(\omega) = \mathbf{G}(\omega)^\top Z(\omega)$ stated in Line 5 of Algorithm 1, where $\vec{y}^{(J)}(\omega) := \left(y^{(J)}_{(l,n)}(\omega)\right)^\top_{(l,n)\in\{1,...,e\}\times\{1,...,N\}}$ denotes the vectorized version of $y^{(J)}(\omega) := \left(y^{(J)}_{l,n}(\omega)\right)^{n=1,...,N}_{l=1,...,e}$. Thus, the problem (50) admits by [12, Theorem 1.2.10] a solution $y^{(J)}(\omega) := \left(y^{(J)}_{l,n}(\omega)\right)^{n=1,...,N}_{l=1,...,e} \in \mathbb{R}^{e\times N}$, which proves that Algorithm 1 terminates.

Next, we show that Algorithm 1 is correct. To this end, we first prove that the $\mathbb{R}^{e\times N}$-valued random variable $y^{(J)} := \big(y_{l,n}^{(J)}\big)_{l=1,\dots,e}^{n=1,\dots,N}$ defined in (50) is $\mathcal{F}_{\theta,V}/\mathcal{B}(\mathbb{R}^{\widetilde{n}})$-measurable. Let us define the function

$$(\mathbb{R}^{\widetilde{m}\times\widetilde{n}} \times \mathbb{R}^{\widetilde{m}}) \times \mathbb{R}^{\widetilde{n}} \ni ((A,b),y) \quad \mapsto \quad \|Ay-b\|^2 \in \mathbb{R}, \tag{51}$$

whose epigraphical mapping $(\mathbb{R}^{\widetilde{m}\times\widetilde{n}}\times\mathbb{R}^{\widetilde{m}})\times\mathbb{R}^{\widetilde{n}}\times\mathbb{R} \ni ((A,b),y,t) \mapsto \big\{((A,b),y,t) \in (\mathbb{R}^{\widetilde{m}\times\widetilde{n}}\times\mathbb{R}^{\widetilde{m}})\times \mathbb{R}^{\widetilde{n}}\times\mathbb{R} : \|Ay-b\|^2 \leqslant t\big\}$ is closed-valued and measurable (see [71, Definition 14.1] for the definition of the latter). This shows that (51) is a normal integrand in the sense of [71, Definition 14.27]. Hence, we can apply [71, Theorem 14.36] to conclude that there exists a $\mathcal{B}((\mathbb{R}^{\widetilde{m}\times\widetilde{n}}\times\mathbb{R}^{\widetilde{m}})\times\mathbb{R}^{\widetilde{n}})/\mathcal{B}(\mathbb{R}^{\widetilde{n}})$-measurable map $\Upsilon : \mathbb{R}^{\widetilde{m}\times\widetilde{n}}\times\mathbb{R}^{\widetilde{m}} \to \mathbb{R}^{\widetilde{n}}$ returning a minimizer, i.e. such that for every $(A,b) \in \mathbb{R}^{\widetilde{m}\times\widetilde{n}}\times\mathbb{R}^{\widetilde{m}}$ it holds that

$$\|A\Upsilon(A,b)-b\|^2 = \min_{y\in\mathbb{R}^{\widetilde{n}}} \|Ay-b\|^2.$$

Moreover, by using that $(\theta_n)_{n\in\mathbb{N}} : \Omega \to \Theta$ are by definition $\mathcal{F}_{\theta,V}/\mathcal{B}(\Theta)$-measurable, that $(V_j)_{j\in\mathbb{N}} : \Omega \to U$ are by definition $\mathcal{F}_{\theta,V}/\mathcal{B}(U)$-measurable, and that the feature maps $g_1,\dots,g_e : \Theta \to W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d)$ are by assumption $\mathcal{B}(\Theta)/\mathcal{B}(W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d))$-measurable, the $\mathbb{R}^{\widetilde{m}\times\widetilde{n}}$-valued random variable $\mathbf{G} = \big(\mathbf{G}_{(j,\alpha,i),(l,n)}\big)_{(j,\alpha,i)\in\{1,\dots,J\}\times\mathbb{N}_{0,k}^m\times\{1,\dots,d\}}^{(l,n)\in\{1,\dots,e\}\times\{1,\dots,N\}}$ with $\mathbf{G}_{(j,\alpha,i),(l,n)} := c_\alpha\partial_\alpha g_{l,i}(\theta_n)(V_j)$, for $(j,\alpha,i) \in \{1,\dots,J\}\times\mathbb{N}_{0,k}^m\times\{1,\dots,d\}$ and $(l,n) \in \{1,\dots,e\}\times\{1,\dots,N\}$, is $\mathcal{F}_{\theta,V}/\mathcal{B}(\mathbb{R}^{\widetilde{m}\times\widetilde{n}})$-measurable. In addition, by using that $(V_j)_{j\in\mathbb{N}} : \Omega \to U$ are by definition $\mathcal{F}_{\theta,V}/\mathcal{B}(U)$-measurable and that $f : U \to \mathbb{R}^d$ is k-times weakly differentiable, the $\mathbb{R}^{\widetilde{m}}$-valued random variable $Z := (c_\alpha\partial_\alpha f_i(V_j))_{(j,\alpha,i)\in\{1,\dots,J\}\times\mathbb{N}_{0,k}^m\times\{1,\dots,d\}}$ is $\mathcal{F}_{\theta,V}/\mathcal{B}(\mathbb{R}^{\widetilde{m}})$-measurable. Thus, by combining this with the $\mathcal{B}((\mathbb{R}^{\widetilde{m}\times\widetilde{n}}\times\mathbb{R}^{\widetilde{m}})\times\mathbb{R}^{\widetilde{n}})/\mathcal{B}(\mathbb{R}^{\widetilde{n}})$-measurable map $\Upsilon : \mathbb{R}^{\widetilde{m}\times\widetilde{n}}\times\mathbb{R}^{\widetilde{m}} \to \mathbb{R}^{\widetilde{n}}$, it follows that

$$\Omega \ni \omega \quad \mapsto \quad \vec{y}^{(J)}(\omega) := \Upsilon(\mathbf{G}(\omega), Z(\omega)) \in \mathbb{R}^{e\cdot N}$$

is $\mathcal{F}_{\theta,V}/\mathcal{B}(\mathbb{R}^{\widetilde{n}})$-measurable, which shows that $y^{(J)} \in \mathcal{Y}_N$. Since $\vec{y}^{(J)}(\omega) = \Upsilon(\mathbf{G}(\omega),Z(\omega)) = \min_{y\in\mathbb{R}^{\widetilde{n}}}\|\mathbf{G}(\omega)y - Z(\omega)\|^2$ is by [12, Theorem 1.1.2] equivalent to the normal equations $\mathbf{G}(\omega)^\top\mathbf{G}(\omega)\vec{y}^{(J)}(\omega) = \mathbf{G}(\omega)^\top Z(\omega)$ in Line 5, we obtain that the algorithm is correct.

Finally, we compute the complexity of Algorithm 1. In Line 1, we generate N random variables $(\theta_n)_{n=1,\dots,N}$, which costs N units. In Line 2, we generate J random variables $(V_j)_{j=1,\dots,J} \sim w$, which requires J units. In Line 3, we compute the $\mathbb{R}^{\widetilde{m}\times\widetilde{n}}$-valued random variable $\mathbf{G} = \big(\mathbf{G}_{(j,\alpha,i),(l,n)}\big)_{(j,\alpha,i)\in\{1,\dots,J\}\times\mathbb{N}_{0,k}^m\times\{1,\dots,d\}}^{(l,n)\in\{1,\dots,e\}\times\{1,\dots,N\}}$ with $\mathbf{G}_{(j,\alpha,i),(l,n)} := c_\alpha\partial_\alpha g_l(\theta_n)_i(V_j)$, for $(j,\alpha,i) \in \{1,\dots,J\}\times\mathbb{N}_{0,k}^m\times\{1,\dots,d\}$ and $(l,n) \in \{1,\dots,e\}\times\{1,\dots,N\}$, which needs $2J\big|\mathbb{N}_{0,k}^m\big|deN$ units. In Line 4, we compute the $\mathbb{R}^{\widetilde{m}}$-valued random variable $Z := (c_\alpha\partial_\alpha f_i(V_j))_{(j,\alpha,i)\in\{1,\dots,J\}\times\mathbb{N}_{0,k}^m\times\{1,\dots,d\}}$, which requires $2J\big|\mathbb{N}_{0,k}^m\big|d$ units. In Line 5, we solve the least squares problem via Cholesky decomposition and forward/backward substitution (see [12, Section 2.2.2]), which needs

$$\frac{1}{2}\widetilde{m}\widetilde{n}^2 + \frac{1}{6}\widetilde{n}^3 + \mathcal{O}(\widetilde{m}\widetilde{n}) = \frac{1}{2}\left(J\left|\mathbb{N}_{0,k}^m\right|d\right)(eN)^2 + \frac{1}{6}(eN)^3 + \mathcal{O}\left(J\left|\mathbb{N}_{0,k}^m\right|deN\right)$$

units (see [12, p. 45]). Hence, by summing the computational costs and by using that $\big|\mathbb{N}_{0,k}^m\big| = \sum_{j=0}^k m^j \leqslant 2m^k$, the complexity of Algorithm 1 is of order

$$\begin{aligned}
&N + J + 2J\left|\mathbb{N}_{0,k}^m\right|deN + \frac{1}{2}\left(J\left|\mathbb{N}_{0,k}^m\right|d\right)(eN)^2 + \frac{1}{6}(eN)^3 + \mathcal{O}\left(J\left|\mathbb{N}_{0,k}^m\right|deN\right) \\
&\quad \leqslant N + J + 4Jm^k deN + 2Jm^k d(eN)^2 + \frac{1}{6}(eN)^3 + \mathcal{O}\left(Jm^k deN\right) \\
&\quad = \mathcal{O}\left(Jm^k d(eN)^2 + (eN)^3\right),
\end{aligned}$$

which completes the proof. □

9.2. **Proof of Theorem 5.2 and Corollary 5.4.**

Proof of Theorem 5.2. Fix some $J,N \in \mathbb{N}$, $L > 0$, and a function $f := (f_1,\dots,f_d)^\top \in \mathbb{B}_{\mathcal{G},\theta}^2(W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d))$ satisfying $|\partial_\alpha f_i(u)| \leqslant L$ for all $\alpha \in \mathbb{N}_{0,k}^m$, $i = 1,\dots,d$, and $u \in U$. Then, we apply Algorithm 1 to obtain some $G_N^{y^{(J)}} \in \mathcal{RG}^V$ with $\mathbb{R}^{e\times N}$-valued random variable $y^{(J)} = \big(y_{l,n}^{(J)}\big)_{l=1,\dots,e}^{n=1,\dots,N} \in \mathcal{Y}_N$ solving (15). Moreover, by using that $(\theta_n)_{n\in\mathbb{N}} : \Omega \to \Theta$ are by definition

$\mathcal{F}_{\theta,V}/\mathcal{B}(\Theta)$-measurable, that the feature maps $g_1,...,g_e : \Theta \to W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d)$ are by assumption $\mathcal{B}(\Theta)/\mathcal{B}(W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d))$-measurable, and that $y^{(J)} = \left(y_{l,n}^{(J)}\right)_{l=1,...,e}^{n=1,...,N} \in \mathcal{Y}_N$ is $\mathcal{F}_{\theta,V}/\mathcal{B}(\mathbb{R}^{e\times N})$-measurable, it follows that

$$\begin{aligned}\Omega \ni \omega \quad &\mapsto \quad G_N^{y^{(J)}}(\omega) := \left(u \mapsto G_{N,i}^{y^{(J)}}(\omega)(u)\right)_{i=1,...,d}^\top := \sum_{n=1}^N \sum_{l=1}^e y_{l,N}^{(J)}(\omega) g_l(\theta_n(\omega)) \\ &:= \left(u \mapsto \sum_{n=1}^N \sum_{l=1}^e y_{l,N}^{(J)}(\omega) g_{l,i}(\theta_n(\omega))(u)\right)_{i=1,...,d}^\top \in W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d)\end{aligned}$$

is $\mathcal{F}_{\theta,V}/\mathcal{B}(W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d))$-measurable. Hence, by using [64, Lemma 4.7], i.e. that the Banach space $(W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d), \|\cdot\|_{W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d)})$ is separable, we can apply [44, Theorem 1.1.6+1.1.20] to conclude that $G^{y^{(J)}} : \Omega \to W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d)$ is a $(\mathbb{P},\mathcal{F}_{\theta,V})$-strongly measurable map.

In order to show (16), we adapt the proof of [37, Theorem 11.3]. To this end, we define for every $\alpha \in \mathbb{N}_{0,k}^m$ and $i = 1,...,d$ the $L^2(U,\mathcal{L}(U),w)$-valued random variable

$$\Omega \ni \omega \quad \mapsto \quad \Delta_{\alpha,i,L}^{y^{(J)}}(\omega) := \left(u \mapsto \partial_\alpha f_i(u) - T_L\left(\partial_\alpha G_{N,i}^{y^{(J)}}(\omega)(u)\right)\right) \in L^2(U,\mathcal{L}(U),w).$$

Moreover, we define for every fixed $\alpha \in \mathbb{N}_{0,k}^m$, $i = 1,...,d$, and $\vartheta := (\vartheta_1,...,\vartheta_N) \in \times_{n=1}^N \Theta$, the $L^2(U,\mathcal{L}(U),w)$-valued random variable

$$\Omega \ni \omega \quad \mapsto \quad \Delta_{\alpha,i,L}^{y^{(J)},\vartheta}(\omega) := \left(u \mapsto \partial_\alpha f_i(u) - T_L\left(\partial_\alpha G_{N,i}^{y^{(J)},\vartheta}(\omega)(u)\right)\right) \in L^2(U,\mathcal{L}(U),w),$$

where $\Omega \ni \omega \mapsto G_{N,i}^{y^{(J)},\vartheta}(\omega) := \sum_{n=1}^N \sum_{l=1}^e y_{l,n}^{(J)}(\omega) g_{l,i}(\vartheta_n) \in L^2(U,\mathcal{L}(U),w)$. In addition, we define the corresponding (random) empirical mean squared error $\|\cdot\|_J$ of such $L^2(U,\mathcal{L}(U),w)$-valued random variables as

$$\begin{aligned}\Omega \ni \omega \quad &\mapsto \quad \left\|\Delta_{\alpha,i,L}^{y^{(J)}}(\omega)\right\|_J := \left(\frac{1}{J}\sum_{j=1}^J \left|\Delta_{\alpha,i,L}^{y^{(J)}}(\omega)(V_j(\omega))\right|^2\right)^{\frac{1}{2}} \in \mathbb{R} \qquad \text{and} \\ \Omega \ni \omega \quad &\mapsto \quad \left\|\Delta_{\alpha,i,L}^{y^{(J)},\vartheta}(\omega)\right\|_J := \left(\frac{1}{J}\sum_{j=1}^J \left|\Delta_{\alpha,i,L}^{y^{(J)},\vartheta}(\omega)(V_j(\omega))\right|^2\right)^{\frac{1}{2}} \in \mathbb{R}.\end{aligned}$$

Then, by using the inequality $(x+y)^2 \leqslant 2\left(x^2+y^2\right)$ for any $x,y \geqslant 0$, it follows that

$$\begin{aligned}&\mathbb{E}\left[\sum_{\alpha\in\mathbb{N}_{0,k}^m} \int_U \left\|\partial_\alpha f(u) - T_L\left(\partial_\alpha G_N^{y^{(J)}}(\cdot)(u)\right)\right\|^2 w(u)du\right] \\ &\quad \leqslant \mathbb{E}\left[\sum_{\alpha\in\mathbb{N}_{0,k}^m} \sum_{i=1}^d \int_U \left|\partial_\alpha f_i(u) - T_L\left(\partial_\alpha G_{N,i}^{y^{(J)}}(\cdot)(u)\right)\right|^2 w(u)du\right] \\ &\quad = \sum_{\alpha\in\mathbb{N}_{0,k}^m} \sum_{i=1}^d \mathbb{E}\left[\left(\left\|\Delta_{\alpha,i,L}^{y^{(J)}}\right\|_{L^2(U,\mathcal{L}(U),w)} - 2\left\|\Delta_{\alpha,i,L}^{y^{(J)}}\right\|_J + 2\left\|\Delta_{\alpha,i,L}^{y^{(J)}}\right\|_J\right)^2\right] \\ &\quad \leqslant 2\sum_{\alpha\in\mathbb{N}_{0,k}^m} \sum_{i=1}^d \mathbb{E}\left[\max\left(\left\|\Delta_{\alpha,i,L}^{y^{(J)}}\right\|_{L^2(U,\mathcal{L}(U),w)} - 2\left\|\Delta_{\alpha,i,L}^{y^{(J)}}\right\|_J, 0\right)^2 + 4\left\|\Delta_{\alpha,i,L}^{y^{(J)}}\right\|_J^2\right].\end{aligned}$$

Hence, by conditioning on $\mathcal{F}_\theta$, by using that $|\mathbb{N}^m_{0,k}| = \sum_{j=0}^k m^j \leqslant 2m^k$, that the random variables $(V_j)_{j\in\mathbb{N}}$ are independent of $(\theta_n)_{n\in\mathbb{N}}$, and the notation $\theta := (\theta_n)_{n=1,...,N}$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{\alpha\in\mathbb{N}^m_{0,k}} \int_U \left\|\partial_\alpha f(u) - T_L\left(\partial_\alpha G_N^{y^{(J)}}(\cdot)(u)\right)\right\|^2 w(u)du\right] \\
&\quad\leqslant 2\left|\mathbb{N}^m_{0,k}\right| d \max_{\substack{\alpha\in\mathbb{N}^m_{0,k}\\ i=1,...,d}} \mathbb{E}\left[\mathbb{E}\left[\max\left(\left\|\Delta^{y^{(J)}}_{\alpha,i,L}\right\|_{L^2(U,\mathcal{L}(U),w)} - 2\left\|\Delta^{y^{(J)}}_{\alpha,i,L}\right\|_J, 0\right)^2 \middle| \mathcal{F}_\theta\right]\right] \\
&\qquad + 8\mathbb{E}\left[\sum_{\alpha\in\mathbb{N}^m_{0,k}} \sum_{i=1}^d \frac{1}{J}\sum_{j=1}^J \left\|\Delta^{y^{(J)}}_{\alpha,i,L}(\cdot)(V_j)\right\|^2\right] \\
&\quad\leqslant 4m^k d \max_{\substack{\alpha\in\mathbb{N}^m_{0,k}\\ i=1,...,d}} \mathbb{E}\left[\mathbb{E}\left[\max\left(\left\|\Delta^{y^{(J)},\vartheta}_{\alpha,i,L}\right\|_{L^2(U,\mathcal{L}(U),w)} - 2\left\|\Delta^{y^{(J)},\vartheta}_{\alpha,i,L}\right\|_J, 0\right)^2\right]\Bigg|_{\vartheta=\theta}\right] \\
&\qquad + 8\mathbb{E}\left[\sum_{\alpha\in\mathbb{N}^m_{0,k}} \sum_{i=1}^d \frac{1}{J}\sum_{j=1}^J \left\|\Delta^{y^{(J)}}_{\alpha,i,L}(\cdot)(V_j)\right\|^2\right].
\end{aligned}
\tag{52}
$$

Moreover, we define for every fixed $\alpha \in \mathbb{N}^m_{0,k}$, $i = 1, ..., d$, and $\vartheta := (\vartheta_1, ..., \vartheta_N) \in \times_{n=1}^N \Theta$ the vector space of random functions

$$
\mathcal{G}^\vartheta_{\alpha,i} := \left\{\Omega \ni \omega \mapsto \sum_{n=1}^N \sum_{l=1}^e y_{l,n}\partial_\alpha g_{l,i}(\vartheta)(V_j(\omega)) \in L^2(U,\mathcal{L}(U),w) : y = (y_{l,n})^{n=1,...,N}_{l=1,...,e} \in \mathcal{Y}_N\right\}.
$$

Then, by following [37, p. 193], i.e. by using [37, Theorem 11.2] (with the set $T_L\left(\mathcal{G}^\vartheta_{\alpha,i}\right) := \{\Omega \ni \omega \mapsto (u \mapsto T_L(G(\omega)(u))) \in L^2(U,\mathcal{L}(U),w) : G \in \mathcal{G}^\vartheta_{\alpha,i}\}$ and where $\mathcal{G}^\vartheta_{\alpha,i}$ has for fixed $a \in \mathbb{R}^{N\times m}$, $b \in \mathbb{R}^N$, $\alpha \in \mathbb{N}^m_{0,k}$, and $i = 1, ..., d$ the vector space dimension N in the sense of [37, Theorem 11.1]) together with [37, Lemma 9.2+9.4 & Theorem 9.5], it follows for every $u > 576L^2/J$ that

$$
\begin{aligned}
&\mathbb{P}\left[\max\left(\left\|\Delta^{y^{(J)},\vartheta}_{\alpha,i,L}\right\|_{L^2(U,\mathcal{L}(U),w)} - 2\left\|\Delta^{y^{(J)},\vartheta}_{\alpha,i,L}\right\|_J, 0\right)^2 > u\right] \\
&\quad\leqslant \mathbb{P}\left[\exists g \in T_L\left(\mathcal{G}^\vartheta_{\alpha,i}\right) : \|g\|_{L^2(U,\mathcal{L}(U),w)} - 2\|g\|_J > \frac{\sqrt{u}}{2}\right] \\
&\quad\leqslant 9(12eJ)^{2(N+1)}e^{-\frac{Ju}{2304L^2}}.
\end{aligned}
\tag{53}
$$

Hence, by using the constant $v := \frac{2304L^2}{J}\ln\left(9(12eJ)^{2(N+1)}\right) > 576L^2/J$, the inequality (53), and that $\ln(108e) \geqslant 1$ together with $2304 \leqslant 9216\ln(108e)$, we conclude that

$$
\begin{aligned}
&\mathbb{E}\left[\max\left(\left\|\Delta^{y^{(J)},\vartheta}_{\alpha,i,L}\right\|_{L^2(U,\mathcal{L}(U),w)} - 2\left\|\Delta^{y^{(J)},\theta}_{\alpha,i,L}\right\|_J, 0\right)^2\right] \\
&\quad = \int_0^\infty \mathbb{P}\left[\max\left(\left\|\Delta^{y^{(J)},\vartheta}_{\alpha,i,L}\right\|_{L^2(U,\mathcal{L}(U),w)} - 2\left\|\Delta^{y^{(J)},\theta}_{\alpha,i,L}\right\|_J, 0\right)^2 > u\right] du \\
&\quad \leqslant v + \int_v^\infty \mathbb{P}\left[\max\left(\left\|\Delta^{y^{(J)},\vartheta}_{\alpha,i,L}\right\|_{L^2(U,\mathcal{L}(U),w)} - 2\left\|\Delta^{y^{(J)},\vartheta}_{\alpha,i,L}\right\|_J, 0\right)^2 > u\right] du \\
&\quad \leqslant v + 9(12eJ)^{2(N+1)} \int_v^\infty e^{-\frac{Ju}{2304L^2}} du \qquad (54) \\
&\quad = \frac{2304L^2}{J} \underbrace{\ln\left(9(12eJ)^{2(N+1)}\right)}_{\leqslant 4N\ln(108eJ)} + \frac{2304L^2}{J} e^{-\frac{Jv}{2304L^2}} \\
&\quad \leqslant \frac{2304L^2}{J} 4N\left(\ln(108e) + \ln(J)\right) + \frac{2304L^2}{J} \\
&\quad \leqslant 9216\ln(108e)L^2 \frac{(\ln(J)+1)N}{J}.
\end{aligned}
$$

On the other hand, for the second term on the right-hand side of (52), we use that $|\partial_\alpha f_i(u)| \leqslant L$ for any $\alpha \in \mathbb{N}^m_{0,k}$, $i = 1,...,d$, and $u \in U$, that $\|T_L(y)\| \leqslant \|y\|$ for any $y \in \mathbb{R}^d$, and that the $\mathbb{R}^{e\times N}$-valued random variable $y^{(J)} = \left(y^{(J)}_{l,n}\right)^{n=1,...,N}_{l=1,...,d}$ solves (15), to obtain that

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{\alpha\in\mathbb{N}^m_{0,k}} \sum_{i=1}^d \frac{1}{J}\sum_{j=1}^J \left\|\Delta^{y^{(J)}}_{\alpha,i,L}(\cdot)(V_j)\right\|^2\right]^{\frac{1}{2}} \\
&\quad = \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^J \sum_{\alpha\in\mathbb{N}^m_{0,k}} \left\|T_L\left(\partial_\alpha f(V_j) - \partial_\alpha G^{y^{(J)}}_N(\cdot)(V_j)\right)\right\|^2\right]^{\frac{1}{2}} \\
&\quad \leqslant \frac{1}{\min_{\alpha\in\mathbb{N}^m_{0,k}} c_\alpha} \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^J \sum_{\alpha\in\mathbb{N}^m_{0,k}} c_\alpha^2 \left\|\partial_\alpha f(V_j) - \partial_\alpha G^{y^{(J)}}_N(\cdot)(V_j)\right\|^2\right]^{\frac{1}{2}} \\
&\quad = \frac{1}{\min_{\alpha\in\mathbb{N}^m_{0,k}} c_\alpha} \mathbb{E}\left[\min_{y\in\mathcal{Y}_N}\left(\frac{1}{J}\sum_{j=1}^J \sum_{\alpha\in\mathbb{N}^m_{0,k}} c_\alpha^2 \left\|\partial_\alpha f(V_j) - \partial_\alpha G^{y}_N(\cdot)(V_j)\right\|^2\right)\right]^{\frac{1}{2}} \\
&\quad \leqslant \frac{1}{\min_{\alpha\in\mathbb{N}^m_{0,k}} c_\alpha} \inf_{y\in\mathcal{Y}_N} \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^J \sum_{\alpha\in\mathbb{N}^m_{0,k}} c_\alpha^2 \left\|\partial_\alpha f(V_j) - \partial_\alpha G^{y}_N(\cdot)(V_j)\right\|^2\right]^{\frac{1}{2}} \\
&\quad \leqslant \frac{\max_{\alpha\in\mathbb{N}^m_{0,k}} c_\alpha}{\min_{\alpha\in\mathbb{N}^m_{0,k}} c_\alpha} \inf_{y\in\mathcal{Y}_N} \mathbb{E}\left[\sum_{\alpha\in\mathbb{N}^m_{0,k}} \int_U \left\|\partial_\alpha f(u) - \partial_\alpha G^{y}_N(\cdot)(u)\right\|^2 w(u)du\right]^{\frac{1}{2}}.
\end{aligned}
$$

Hence, by using Theorem 4.20 (with constants $C_{2,2} > 0$ and $C_{W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d)} > 0$ independent of $f : U \to \mathbb{R}^d$ and $m, d \in \mathbb{N}$, see also [64, Lemma 4.9]) together with $\mathcal{F}_\theta \subseteq \mathcal{F}_{\theta,V}$ (with $G^f_N \in \mathcal{RG} \cap L^2(\Omega,\mathcal{F}_\theta,\mathbb{P};W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d)$ having $\mathcal{F}_\theta/\mathcal{B}(\mathbb{R}^{e\times N})$-measurable linear readout contained in

$\mathcal{Y}_N$ as $\mathcal{F}_\theta \subseteq \mathcal{F}_{\theta,V}$), we conclude that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\alpha\in\mathbb{N}^m_{0,k}}\sum_{i=1}^d \frac{1}{J}\sum_{j=1}^J \left\|\Delta^{y^{(J)}}_{\alpha,i,L}(\cdot)(V_j)\right\|^2\right]^{\frac{1}{2}} \\
&\quad\leqslant \kappa(\mathbf{c}) \inf_{y\in\mathcal{Y}_N} \mathbb{E}\left[\sum_{\alpha\in\mathbb{N}^m_{0,k}}\int_U \left\|\partial_\alpha f(u) - \partial_\alpha G^y_N(\cdot)(u)\right\|^2 w(u)du\right]^{\frac{1}{2}} \\
&\quad\leqslant \kappa(\mathbf{c})\mathbb{E}\left[\left\|f - G^f_N\right\|^2_{W^{k,2}(U,\mathcal{L}(U),w,\mathbb{R}^d)}\right]^{\frac{1}{2}} \\
&\quad\leqslant \kappa(\mathbf{c}) C_{2,2} C_{W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d)} \frac{\|f\|_{\mathbb{B}^2_{\mathcal{G},\theta}(W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d))}}{\sqrt{N}}.
\end{aligned} \tag{55}$$

Hence, by inserting (54)+(55) into (52) with the inequality $\sqrt{x+y} \leqslant \sqrt{x} + \sqrt{y}$ for any $x, y \geqslant 0$, and by using the constant $C_4 := \max\left(2\sqrt{9216\ln(108e)}, \sqrt{8}C_{2,2}C_{W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d)}\right) > 0$ (being independent of $f : U \to \mathbb{R}^d$ and $m, d \in \mathbb{N}$), it follows that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\alpha\in\mathbb{N}^m_{0,k}}\int_U \left\|\partial_\alpha f(u) - T_L\left(\partial_\alpha G^{y^{(J)}}_N(\cdot)(u)\right)\right\|^2 w(u)du\right]^{\frac{1}{2}} \\
&\quad\leqslant 2m^{\frac{k}{2}} d^{\frac{1}{2}} \max_{\substack{\alpha\in\mathbb{N}^m_{0,k} \\ i=1,\dots,d}} \mathbb{E}\left[\mathbb{E}\left[\max\left(\left\|\Delta^{y^{(J)},\vartheta}_{\alpha,i,L}\right\|_{L^2(U,\mathcal{L}(U),w)} - 2\left\|\Delta^{y^{(J)},\vartheta}_{\alpha,i,L}\right\|_J, 0\right)^2\right]\bigg|_{\vartheta=\theta}\right]^{\frac{1}{2}} \\
&\quad\quad + \sqrt{8}\mathbb{E}\left[\sum_{\alpha\in\mathbb{N}^m_{0,k}}\sum_{i=1}^d \frac{1}{J}\sum_{j=1}^J \left\|\Delta^{y^{(J)}}_{\alpha,i,L}(\cdot)(V_j)\right\|^2\right]^{\frac{1}{2}} \\
&\quad\leqslant 2m^{\frac{k}{2}} d^{\frac{1}{2}} \sqrt{9216\ln(108e)} L \sqrt{\frac{(\ln(J)+1)N}{J}} \\
&\quad\quad + \sqrt{8}\kappa(\mathbf{c}) C_{2,2} C_{W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d)} \frac{\|f\|_{\mathbb{B}^2_{\mathcal{G},\theta}(W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d))}}{\sqrt{N}} \\
&\quad\leqslant C_4 L m^{\frac{k}{2}} d^{\frac{1}{2}} \sqrt{\frac{(\ln(J)+1)N}{J}} + C_4\kappa(\mathbf{c}) \frac{\|f\|_{\mathbb{B}^2_{\mathcal{G},\theta}(W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d))}}{\sqrt{N}},
\end{aligned}$$

which completes the proof. □

Proof of Corollary 5.4. Fix some $J, N \in \mathbb{N}$, $L > 0$, and some $f := (f_1, \dots, f_d)^\top \in W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d) \cap \widetilde{\mathbb{B}}^{k,2,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)$ satisfying $|\partial_\alpha f_i(u)| \leqslant L$ for all $\alpha \in \mathbb{N}^m_{0,k}$, $i = 1, \dots, d$, and $u \in U$. Then, we observe that Algorithm 2 is the same as Algorithm 1 for the special case of random neural networks with feature maps $\Theta := \mathbb{R}^m \times \mathbb{R} \ni (\vartheta_1, \vartheta_2) \mapsto e_i\rho\left(\vartheta_1^\top \cdot - \vartheta_2\right) \in W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d)$, $i = 1, \dots, d$, that are continuous and thus $\mathcal{B}(\Theta)/\mathcal{B}(W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d))$-measurable (see [64, Lemma 4.10]), where $e_i \in \mathbb{R}^d$ denotes the i-th unit vector of $\mathbb{R}^d$. Hence, we can apply Theorem 5.2 (with constant $C_4 > 0$ independent of $f : U \to \mathbb{R}^d$ and $m, d \in \mathbb{N}$) to conclude that Algorithm 2 returns a random neural network $G^{y^{(J)}}_N \in \mathcal{RN}^\rho_{U,d}$ with N neurons being a strongly $(\mathbb{P}, \mathcal{F}_{a,b,V})$-measurable map $G^{y^{(J)}}_N : \Omega \to W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d)$ such that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\alpha\in\mathbb{N}^m_{0,k}}\int_U \left\|\partial_\alpha f(u) - T_L\left(\partial_\alpha G^{y^{(J)}}_N(\cdot)(u)\right)\right\|^2 w(u)du\right]^{\frac{1}{2}} \\
&\quad\leqslant C_4 L m^{\frac{k}{2}} d^{\frac{1}{2}} \sqrt{\frac{(\ln(J)+1)N}{J}} + C_4\kappa(\mathbf{c}) \frac{\|f\|_{\mathbb{B}^2_{\mathcal{G},\theta}(W^{k,2}(U,\mathcal{L}(U),w;\mathbb{R}^d))}}{\sqrt{N}}.
\end{aligned}$$

Thus, by using Lemma 4.19, it follows that

$$\mathbb{E}\left[\sum_{\alpha\in\mathbb{N}^m_{0,k}}\int_U \left\|\partial_\alpha f(u) - T_L\left(\partial_\alpha G_N^{y^{(J)}}(\cdot)(u)\right)\right\|^2 w(u)du\right]^{\frac{1}{2}}$$
$$\leqslant C_4 L m^{\frac{k}{2}} d^{\frac{1}{2}}\sqrt{\frac{(\ln(J)+1)N}{J}} + C_4\kappa(\mathbf{c})2^{3+\frac{1}{2}}\|\rho\|_{C^k_{pol,\gamma}(\mathbb{R})}\frac{C^{(\gamma,2)}_{U,w} m^{\frac{k}{2}}}{\left|C^{(\psi,\rho)}_m\right|}\frac{\|f\|_{\widetilde{\mathbb{B}}^{k,2,\gamma}_{\psi,a,b}(U;\mathbb{R}^d)}}{\sqrt{N}}.$$

Therefore, by defining the constant $C_5 := 2^{3+\frac{1}{2}}C_4 > 0$ (being independent of $f : U \to \mathbb{R}^d$ and $m, d \in \mathbb{N}$), we obtain the result. □

10. Proof of results in Section 6

Proof of Proposition 6.1. Fix some $\lambda \in (0,\infty)$ and assume that $g : \mathbb{R}^m \to \mathbb{R}$ is a.e. bounded and a.e. continuous. Then, we first observe that $f(t,\cdot) = \phi_{\lambda,t} * g$ is the convolution of the kernel $\mathbb{R}^m \ni y \mapsto \phi_{\lambda,t}(y) := (4\pi\lambda t)^{-m/2}\exp\left(-\|y\|^2/(4\lambda t)\right) \in \mathbb{R}$ with the initial condition $g : \mathbb{R}^m \to \mathbb{R}$. Moreover, for every $y \in \mathbb{R}^m$, we observe that

$$\frac{\partial\phi_{\lambda,t}}{\partial t}(y) = \left(\frac{\|y\|^2}{4\lambda t^2} - \frac{m}{2t}\right)\frac{e^{-\frac{\|y\|^2}{4\lambda t}}}{(4\pi\lambda t)^{\frac{m}{2}}} = \lambda\sum_{l=1}^m\left(\frac{4y_l^2}{(4\lambda t)^2} - \frac{2}{4\lambda t}\right)\frac{e^{-\frac{\|y\|^2}{4\lambda t}}}{(4\pi\lambda t)^{\frac{m}{2}}} = \lambda\sum_{l=1}^m\frac{\partial^2\phi_{\lambda,t}}{\partial y_l^2}(y). \tag{56}$$

Hence, by applying [41, Theorem 1.3.1], i.e. that $\frac{\partial f}{\partial t}(t,u) = \left(\frac{\partial\phi_{\lambda,t}}{\partial t}*g\right)(u)$ and $\frac{\partial^2 f}{\partial u_l^2}(t,u) = \left(\frac{\partial^2\phi_{\lambda,t}}{\partial^2 y_l}*g\right)(u)$ for any $(t,u) \in (0,\infty)\times\mathbb{R}^m$ and $l = 1,\dots,m$, and by using the identity (56), it follows for every $(t,u) \in (0,\infty)\times\mathbb{R}^m$ that

$$\begin{aligned}\frac{\partial f}{\partial t}(t,u) &= \frac{\partial(\phi_{\lambda,t}*g)}{\partial t}(u) = \int_{\mathbb{R}^m}\frac{\partial\phi_{\lambda,t}}{\partial t}(u-v)g(v)dv = \int_{\mathbb{R}^m}\lambda\sum_{l=1}^m\frac{\partial^2\phi_{\lambda,t}}{\partial y_l^2}(u-v)g(v)dv \\ &= \lambda\sum_{l=1}^m\frac{\partial^2(\phi_{\lambda,t}*g)}{\partial u_l^2}(u) = \lambda\sum_{l=1}^m\frac{\partial^2 f}{\partial u_l^2}(t,u).\end{aligned}$$

In addition, by using the substitution $y \mapsto \frac{u-v}{2\sqrt{\lambda t}}$ and the dominated convergence theorem (with $g : \mathbb{R}^m \to \mathbb{R}$ being a.e. continuous, hence $\lim_{t\to 0} g\left(u + 2\sqrt{\lambda t}y\right) = g(u)$ for a.e. $u, y \in \mathbb{R}^m$, and $g : \mathbb{R}^m \to \mathbb{R}$ being a.e. bounded, thus there exists some $C > 0$ such that for a.e. $u, y \in \mathbb{R}^m$ it holds that $\max\left(\left|g\left(u+2\sqrt{\lambda t}y\right)\right|, |g(u)|\right) \leqslant C$), we conclude for a.e. $u \in \mathbb{R}^m$ that

$$\begin{aligned}\lim_{t\to 0} f(t,u) &= \lim_{t\to 0}\frac{1}{(4\pi\lambda t)^{\frac{m}{2}}}\int_{\mathbb{R}^m} e^{-\frac{\|u-v\|^2}{4\lambda t}}g(v)dv \\ &= \lim_{t\to 0}\frac{1}{(2\pi)^{\frac{m}{2}}}\int_{\mathbb{R}^m} e^{-\frac{\|y\|^2}{2}} g\left(u+2\sqrt{\lambda t}y\right)dy \\ &= \left(\frac{1}{(2\pi)^{\frac{m}{2}}}\int_{\mathbb{R}^m} e^{-\frac{\|y\|^2}{2}}dy\right)g(u) = g(u).\end{aligned}$$

This shows that $f : (0,\infty)\times\mathbb{R}^m \to \mathbb{R}$ indeed solves the PDE (19). □

Lemma 10.1. *For $\lambda, t \in (0,\infty)$ and an a.e. bounded and a.e. continuous initial condition $g : \mathbb{R}^m \to \mathbb{R}$, let $f(t,\cdot) : \mathbb{R}^m \to \mathbb{R}$ be the solution of (19) at time t. Moreover, let $c \in \mathbb{N}_0$, $s \in [0,\infty)$, $0 < \zeta_1 \leqslant \zeta_2 < \infty$, and assume that $g \in L^1(\mathbb{R}^m, \mathcal{L}(\mathbb{R}^m), (1+\|u\|)^c du)$. Then, $\widehat{f(t,\cdot)} : \mathbb{R}^m \to \mathbb{C}$ is c-times weakly differentiable and there exists a constant $C_{12} > 0$ (being independent of $m \in \mathbb{N}$ and $g : \mathbb{R}^m \to \mathbb{R}$) such that*

$$\frac{\zeta_2^m \pi^{\frac{m+1}{4}}}{\zeta_1^{\frac{m}{2}}(2\pi)^m\Gamma\left(\frac{m+1}{2}\right)^{\frac{1}{2}}}\sum_{\beta\in\mathbb{N}^m_{0,c}}\left(\int_{\mathbb{R}^m}\left|\partial_\beta\widehat{f(t,\cdot)}(\xi)\right|^2\left(1+\|\xi/\zeta_1\|^2\right)^{\frac{m+s}{2}}d\xi\right)^{\frac{1}{2}}$$
$$\leqslant C_{12} m^{\frac{5c+s}{4}}\left(\frac{(\zeta_2/\zeta_1)^2}{\sqrt{2\lambda t\pi}}\right)^{\frac{m}{2}}\|g\|_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),(1+\|u\|)^c du)}.$$

Proof. Fix some $\lambda, t \in (0,\infty)$, $c \in \mathbb{N}_0$, $s \in [0,\infty)$, $0 < \zeta_1 < \zeta_2 < \infty$, and an a.e. bounded and a.e. continuous initial condition $g \in L^1(\mathbb{R}^m, \mathcal{L}(\mathbb{R}^m), (1+\|u\|)^c du)$. Then, by using that $f(t,\cdot) = \phi_{\lambda,t} * g$ (see Proposition 6.1) and Young's convolutional inequality, we obtain that

$$\begin{aligned} \|f(t,\cdot)\|_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du)} &= \|\phi_{\lambda,t} * g\|_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du)} \\ &\leqslant \|\phi_{\lambda,t}\|_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du)} \|g\|_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du)} \\ &= \Bigg(\underbrace{\frac{1}{(4\pi\lambda t)^{\frac{m}{2}}} \int_{\mathbb{R}^m} e^{-\frac{\|y\|^2}{4\lambda t}} dy}_{=1} \Bigg) \|g\|_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),(1+\|u\|)^c du)} < \infty, \end{aligned}$$

which shows that $f(t,\cdot) \in L^1(\mathbb{R}^m, \mathcal{L}(\mathbb{R}^m), du)$. Next, we show for every fixed $t \in (0,\infty)$ that the Fourier transform $\widehat{f(t,\cdot)} : \mathbb{R}^m \to \mathbb{C}$ is c-times weakly differentiable. To this end, we use Fubini's theorem, [30, Table 7.2.9], the substitution $\zeta_l \mapsto \sqrt{2\lambda t}\xi_l$, and the Hermite polynomials $(h_n)_{n\in\mathbb{N}}$ in [1, Equation 22.2.15] to conclude for every $\beta := (\beta_1, ..., \beta_m) \in \mathbb{N}_{0,c}^m$ and $\xi \in \mathbb{R}^m$ that

$$\begin{aligned} \partial_\beta \widehat{\phi_{\lambda,t}}(\xi) &= \partial_\beta \left(\prod_{l=1}^m \int_{\mathbb{R}} e^{-i\xi_l u_l} \frac{e^{-\frac{u_l^2}{4\lambda t}}}{\sqrt{4\pi\lambda t}} du_l \right) = \partial_\beta \left(\prod_{l=1}^m e^{-\lambda t \xi_l^2} \right) = \prod_{l=1}^m \frac{\partial^{\beta_l}}{\partial \xi_l^{\beta_l}} \left(e^{-\lambda t \xi_l^2} \right) \\ &= (2\lambda t)^{\frac{|\beta|}{2}} \prod_{l=1}^m \frac{\partial^{\beta_l}}{\partial \zeta_l^{\beta_l}} \left(e^{-\frac{\zeta_l^2}{2}} \right) \Bigg|_{\zeta_l = \sqrt{2\lambda t}\xi_l} = (2\lambda t)^{\frac{|\beta|}{2}} \prod_{l=1}^m (-1)^{\beta_l} h_{\beta_l}(\zeta_l) e^{-\frac{\zeta_l^2}{2}} \Bigg|_{\zeta_l = \sqrt{2\lambda t}\xi_l} \\ &= (-1)^{|\beta|} (2\lambda t)^{\frac{|\beta|}{2}} \left(\prod_{l=1}^m h_{\beta_l}\left(\sqrt{2\lambda t}\xi_l \right) \right) e^{-\lambda t \|\xi\|^2}. \end{aligned} \tag{57}$$

Moreover, we use the notation $\mathbb{R}^m \ni u \mapsto p_\beta(u) := u^\beta := \prod_{l=1}^m u_l^{\beta_l} \in \mathbb{R}$ and the inequality $\left|u^\beta\right| := \left|\prod_{l=1}^m u_l^{\beta_l}\right| = \prod_{l=1}^m |u_l|^{\beta_l} \leqslant \prod_{l=1}^m (1+\|u\|)^{\beta_l} = (1+\|u\|)^{|\beta|} \leqslant (1+\|u\|)^c$ for any $\beta := (\beta_1, ..., \beta_m) \in \mathbb{N}_{0,c}^m$ and $u := (u_1, ..., u_m) \in \mathbb{R}^m$ to obtain for every $\beta \in \mathbb{N}_{0,c}^m$ that

$$\begin{aligned} \|p_\beta \cdot g\|_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du)} &= \int_{\mathbb{R}^m} \left|u^\beta g(u)\right| du \leqslant \int_{\mathbb{R}^m} |g(u)|(1+\|u\|)^c du \\ &= \|g\|_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),(1+\|u\|)^c du)} < \infty. \end{aligned} \tag{58}$$

Hence, by iteratively applying [30, Theorem 7.8. (c)], we conclude that the partial derivatives $\partial_\beta \widehat{g} : \mathbb{R}^m \to \mathbb{C}$ exist, for all $\beta \in \mathbb{N}_{0,c}^m$. Thus, by using [30, Theorem 7.8. (d)] and the Leibniz product rule, we conclude for every $\beta \in \mathbb{N}_{0,c}^m$ and $\xi \in \mathbb{R}^m$ that

$$\partial_\beta \widehat{f(t,\cdot)}(\xi) = \partial_\beta \left(\widehat{\phi_{\lambda,t}}(\xi) \widehat{g}(\xi) \right) = \sum_{\substack{\beta_1,\beta_2 \in \mathbb{N}_0^m \\ \beta_1+\beta_2=\beta}} \frac{\beta!}{\beta_1!\beta_2!} \partial_{\beta_1} \widehat{\phi_{\lambda,t}}(\xi) \partial_{\beta_2} \widehat{g}(\xi), \tag{59}$$

which shows that $\widehat{f(t,\cdot)} : \mathbb{R}^m \to \mathbb{C}$ is c-times weakly differentiable.

Next, we use the explicit expression of the Hermite polynomials $(h_n)_{n\in\mathbb{N}}$ given in [1, Equation 22.3.11], that $|\zeta_l|^{\beta_l - 2j_l} \leqslant \left(1+\|\zeta\|^2\right)^{(\beta_l-2j_l)/2} \leqslant \left(1+\|\zeta\|^2\right)^{\beta_l/2}$ for any $l = 1, ..., m$, $\beta := (\beta_1, ..., \beta_m) \in \mathbb{N}_0^m$, $j_l = 0, ..., \lfloor \beta_l/2 \rfloor$, and $\zeta \in \mathbb{R}^m$, that $\sum_{j_l=1}^{\lfloor \beta_l/2 \rfloor} \frac{\beta_l!}{2^{j_l} j_l! (\beta_l - 2j_l)!} \leqslant \max_{j_l=1,...,\lfloor \beta_l/2 \rfloor} \frac{(2j_l)!}{j_l!} \sum_{j_l=1}^{\lfloor \beta_l/2 \rfloor} \frac{\beta_l!}{(2j_l)!(\beta_l-2j_l)!} \leqslant \beta_l! \sum_{k_l=1}^{\beta_l} \frac{\beta_l!}{k_l!(\beta_l-k_l)!} = 2^{\beta_l}\beta_l!$ for any $l = 1, ..., m$ and $\beta := (\beta_1, ..., \beta_m) \in \mathbb{N}_0^m$, and that $\prod_{l=1}^m \beta_l! = \beta! \leqslant |\beta|! \leqslant c!$ for any $\beta := (\beta_1, ..., \beta_m) \in \mathbb{N}_{0,c}^m$ to conclude for every $\beta := (\beta_1, ..., \beta_m) \in \mathbb{N}_{0,c}^m$ and $\zeta := (\zeta_1, ..., \zeta_m) \in \mathbb{R}^m$ that

$$\begin{aligned} \prod_{l=1}^m |h_{\beta_l}(\zeta_l)| &\leqslant \prod_{l=1}^m \left(\sum_{j_l=1}^{\lfloor \beta_l/2 \rfloor} \frac{\beta_l! |\zeta_l|^{\beta_l - 2j_l}}{2^{j_l} j_l! (\beta_l - 2j_l)!} \right) \leqslant \prod_{l=1}^m \left(\left(1+\|\zeta\|^2\right)^{\frac{\beta_l}{2}} \sum_{j_l=1}^{\lfloor \beta_l/2 \rfloor} \frac{\beta_l!}{2^{j_l} j_l! (\beta_l - 2j_l)!} \right) \\ &\leqslant \left(1+\|\zeta\|^2\right)^{\frac{|\beta|}{2}} \prod_{l=1}^m \left(2^{\beta_l} \beta_l! \right) \leqslant 2^c c! \left(1+\|\zeta\|^2\right)^{\frac{c}{2}}. \end{aligned} \tag{60}$$

Hence, by using the inequality (60) together with (57) and by using the constant $C_{21} := 2^c c! \max(1, 2\lambda t)^c \max(1, \zeta_1)^c > 0$, we conclude for every $\beta \in \mathbb{N}_{0,c}^m$ and $\xi \in \mathbb{R}^m$ that

$$\begin{aligned} \left| \partial_\beta \widehat{\phi_{\lambda,t}}(\xi) \right| &\leqslant (2\lambda t)^{\frac{|\beta|}{2}} 2^c c! \left(1 + \left\| \sqrt{2\lambda t} \xi \right\|^2 \right)^{\frac{c}{2}} e^{-\lambda t \|\xi\|^2} \\ &\leqslant C_{11} \left(1 + \|\xi/\zeta_1\|^2 \right)^{\frac{c}{2}} e^{-\lambda t \|\xi\|^2}. \end{aligned} \tag{61}$$

Moreover, by using that $Y := \|Z\|^2$ of $Z \sim \mathcal{N}_m(0, I_m)$ follows a $\chi^2(m)$-distribution with probability density function $[0,\infty) \ni y \mapsto \frac{y^{m/2-1}\exp(-y/2)}{2^{m/2}\Gamma(m/2)} \in [0,\infty)$, the substitution $x \mapsto y/2$, and the definition of the Gamma function in [1, Equation 6.1.1], we obtain for every $b \in \mathbb{N}_0$ that

$$\begin{aligned} \int_{\mathbb{R}^m} \|z\|^b \frac{e^{-\frac{\|z\|^2}{2}}}{(2\pi)^{\frac{m}{2}}} dz = \mathbb{E}\left[\|Z\|^b \right] = \mathbb{E}\left[Y^{\frac{b}{2}} \right] &= \int_0^\infty y^{\frac{b}{2}} \frac{y^{\frac{m}{2}-1} e^{-\frac{y}{2}}}{2^{\frac{m}{2}} \Gamma\left(\frac{m}{2}\right)} dy \\ &= \frac{2^{\frac{b+m}{2}}}{2^{\frac{m}{2}} \Gamma\left(\frac{m}{2}\right)} \int_0^\infty x^{\frac{b+m}{2}-1} e^{-x} dx = \frac{2^{\frac{b}{2}} \Gamma\left(\frac{m+b}{2}\right)}{\Gamma\left(\frac{m}{2}\right)}. \end{aligned} \tag{62}$$

Now, we use the inequality (61), the exponent $c_s := 2c + 2\lceil s \rceil \in \mathbb{N}_0$, the inequality $(x+y)^r \leqslant 2^r \left(x^r + y^r \right)$ for any $x, y \geqslant 0$ and $r \geqslant 0$, the constant $C_{22} := C_{21}^2 2^{c_s/2} \sqrt{\pi} > 0$, the substitution $z \mapsto \sqrt{4\lambda t}\xi$, the constant $C_{23} := C_{22} \left(\sqrt{4\lambda t} \zeta_1 \right)^{-c_s} > 0$, and the identity (62) with $b := 0$ and $b := m + 2c_s$ to obtain that

$$\begin{aligned} & \frac{\zeta_2^{2m} \pi^{\frac{m+1}{2}}}{\zeta_1^m (2\pi)^{2m} \Gamma\left(\frac{m+1}{2}\right)} \int_{\mathbb{R}^m} \left| \partial_{\beta_1} \widehat{\phi_{\lambda,t}}(\xi) \right|^2 \left(1 + \|\xi/\zeta_1\|^2 \right)^{\frac{m}{2}+s} d\xi \\ & \leqslant C_{11}^2 \frac{\zeta_2^{2m} \pi^{\frac{m+1}{2}}}{\zeta_1^m (2\pi)^{2m} \Gamma\left(\frac{m+1}{2}\right)} \int_{\mathbb{R}^m} \left(1 + \|\xi/\zeta_1\|^2 \right)^{\frac{m+c_s}{2}} e^{-2\lambda t \|\xi\|^2} d\xi \\ & \leqslant C_{22} \frac{\zeta_2^{2m} (2\pi)^{\frac{m}{2}}}{\zeta_1^m (2\pi)^{2m} \Gamma\left(\frac{m+1}{2}\right)} \left(\int_{\mathbb{R}^m} e^{-2\lambda t \|\xi\|^2} d\xi + \int_{\mathbb{R}^m} \|\xi/\zeta_1\|^{m+c_s} e^{-2\lambda t \|\xi\|^2} d\xi \right) \\ & = C_{22} \frac{\zeta_2^{2m}}{\zeta_1^m (2\pi)^m \Gamma\left(\frac{m+1}{2}\right) (4\lambda t)^{\frac{m}{2}}} \int_{\mathbb{R}^m} \frac{e^{-\frac{\|z\|^2}{2}}}{(2\pi)^{\frac{m}{2}}} dz \\ & \quad + C_{23} \frac{\zeta_2^{2m}}{\zeta_1^{2m} (2\pi)^m \Gamma\left(\frac{m+1}{2}\right) (4\lambda t)^{\frac{m}{2}}} \int_{\mathbb{R}^m} \|z\|^{m+c_s} \frac{e^{-\frac{\|z\|^2}{2}}}{(2\pi)^{\frac{m}{2}}} dz \\ & = \frac{C_{22}}{\Gamma\left(\frac{m+1}{2}\right)} \left(\frac{\zeta_2^2/\zeta_1}{4\pi\sqrt{\lambda t}} \right)^m + C_{23} \left(\frac{(\zeta_2/\zeta_1)^2}{4\pi\sqrt{\lambda t}} \right)^m \frac{2^{\frac{m+c_s}{2}} \Gamma\left(\frac{2m+c_s}{2}\right)}{\Gamma\left(\frac{m+1}{2}\right) \Gamma\left(\frac{m}{2}\right)}. \end{aligned} \tag{63}$$

For the first term on the right-hand side of (63), we use that $C_{24} := C_{22} \sup_{m \in \mathbb{N}} \frac{(\zeta_1/(2\sqrt{2}))^m}{\Gamma((m+1)/2)} < \infty$ to conclude that

$$\frac{C_{22}}{\Gamma\left(\frac{m+1}{2}\right)} \left(\frac{\zeta_2^2/\zeta_1}{4\pi\sqrt{\lambda t}} \right)^m \leqslant \frac{C_{22}}{\Gamma\left(\frac{m+1}{2}\right)} \left(\frac{\zeta_1}{2\sqrt{2}} \right)^m \left(\frac{(\zeta_2/\zeta_1)^2}{\sqrt{2\lambda t}\pi} \right)^m \leqslant C_{24} \left(\frac{(\zeta_2/\zeta_1)^2}{\sqrt{2\lambda t}\pi} \right)^m. \tag{64}$$

Moreover, for the second term on the right-hand side of (63), we use that $\sqrt{2\pi/x}(x/e)^x \leqslant \Gamma(x) \leqslant \sqrt{2\pi/x}(x/e)^x e^{1/(12x)} \leqslant \sqrt{4\pi/x}(x/e)^x$ for any $x \in [1/2, \infty)$ (see [33, Lemma 2.4]), that $(2m + c_s)^{c_s/2} \leqslant m^{c_s/2}(2 + c_s)^{c_s/2}$ and $(2 + c_s/m)^m = 2^m(1 + c_s/(2m))^m \leqslant 2^m e^{c_s/2}$ for any $m \in \mathbb{N}$, and the constant $C_{25} := C_{23} 2^{c_s/2} \sqrt{8\pi} (2e)^{(c_s-1)/2} (4\pi)^{-1} (3c_s)^{c_s/2} e^{c_s/2} > 0$, to obtain that

$$\begin{aligned} C_{23} \left(\frac{(\zeta_2/\zeta_1)^2}{4\pi\sqrt{\lambda t}} \right)^m \frac{2^{\frac{m+c_s}{2}} \Gamma\left(\frac{2m+c_s}{2}\right)}{\Gamma\left(\frac{m+1}{2}\right)\Gamma\left(\frac{m}{2}\right)} & \leqslant C_{23} \left(\frac{(\zeta_2/\zeta_1)^2}{4\pi\sqrt{\lambda t}} \right)^m \frac{2^{\frac{m+c_s}{2}} \sqrt{\frac{8\pi}{2m+c_s}} \left(\frac{2m+c_s}{2e}\right)^{\frac{2m+c_s}{2}}}{\sqrt{\frac{4\pi}{m+1}} \left(\frac{m+1}{2e}\right)^{\frac{m+1}{2}} \sqrt{\frac{4\pi}{m}} \left(\frac{m}{2e}\right)^{\frac{m}{2}}} \\ & \leqslant C_{23} \left(\frac{(\zeta_2/\zeta_1)^2}{4\pi\sqrt{\lambda t}} \right)^m \frac{2^{\frac{m+c_s}{2}} \sqrt{8\pi} (2e)^{\frac{c_s-1}{2}}}{4\pi} \underbrace{\frac{\sqrt{m(m+1)} (2m+c_s)^{\frac{c_s}{2}}}{\sqrt{2m+c_s}\sqrt{m+1}}}_{\leqslant (2m+c_s)^{\frac{c_s}{2}}} \Bigg(\underbrace{\frac{2m+c_s}{m}}_{\leqslant 2+c_s/m} \Bigg)^m \\ & \leqslant C_{23} \left(\frac{(\zeta_2/\zeta_1)^2}{2\pi\sqrt{2\lambda t}} \right)^m \frac{2^{\frac{c_s}{2}} \sqrt{8\pi} (2e)^{\frac{c_s-1}{2}}}{4\pi} (2+c_s)^{\frac{c_s}{2}} m^{\frac{c_s}{2}} 2^m e^{\frac{c_s}{2}} = C_{25} \left(\frac{(\zeta_2/\zeta_1)^2}{\sqrt{2\lambda t}\pi} \right)^m m^{\frac{c_s}{2}}. \end{aligned} \tag{65}$$

Hence, by inserting (64)+(65) into (63), it follows that

$$\begin{aligned} & \left(\frac{\zeta_2^{2m} \pi^{\frac{m+1}{2}}}{\zeta_1^m (2\pi)^{2m} \Gamma\left(\frac{m+1}{2}\right)} \int_{\mathbb{R}^m} \left| \partial_{\beta_1} \widehat{\phi_{\lambda,t}}(\xi) \right|^2 \left(1 + \|\xi/\zeta_1\|^2\right)^{\frac{m+s}{2}} d\xi \right)^{\frac{1}{2}} \\ & \leqslant \left(\frac{C_{22}}{\Gamma\left(\frac{m+1}{2}\right)} \left(\frac{\zeta_2^2/\zeta_1}{4\pi\sqrt{\lambda t}} \right)^m + C_{23} \left(\frac{(\zeta_2/\zeta_1)^2}{4\pi\sqrt{\lambda t}} \right)^m \frac{2^{\frac{m+c_s}{2}} \Gamma\left(\frac{2m+c_s}{2}\right)}{\Gamma\left(\frac{m+1}{2}\right)\Gamma\left(\frac{m}{2}\right)} \right)^{\frac{1}{2}} \\ & \leqslant \left(C_{24} \left(\frac{(\zeta_2/\zeta_1)^2}{\sqrt{2\lambda t}\pi} \right)^m + C_{25} \left(\frac{(\zeta_2/\zeta_1)^2}{\sqrt{2\lambda t}\pi} \right)^m m^{\frac{c_s}{2}} \right)^{\frac{1}{2}} \leqslant \sqrt{C_{24} + C_{25}} \left(\frac{(\zeta_2/\zeta_1)^2}{\sqrt{2\lambda t}\pi} \right)^{\frac{m}{2}} m^{\frac{c_s}{4}}. \end{aligned} \tag{66}$$

Thus, by using (59), that $\left|\mathbb{N}_{0,c}^m\right| = \sum_{j=0}^c m^j \leqslant 2m^c$, the inequality $\sum_{\beta_1,\beta_2 \in \mathbb{N}_0^m, \beta_1+\beta_2=\beta} \frac{\beta!}{\beta_1!\beta_2!} \leqslant 2^{|\beta|}$ for any $\beta \in \mathbb{N}_0^m$ (see [64, Equation 12]), the inequality (66), [30, Theorem 7.8. (c)] componentwise with (2), the inequality (58), and the constant $C_{12} := 2^{c+1}\sqrt{C_{24}+C_{25}} > 0$ (being independent of $m \in \mathbb{N}$ and $g : \mathbb{R}^m \to \mathbb{R}$), we conclude that

$$\begin{aligned} & \frac{\zeta_2^m \pi^{\frac{m+1}{4}}}{\zeta_1^{\frac{m}{2}} (2\pi)^m \Gamma\left(\frac{m+1}{2}\right)^{\frac{1}{2}}} \sum_{\beta \in \mathbb{N}_{0,c}^m} \left(\int_{\mathbb{R}^m} \left| \partial_\beta \widehat{f(t,\cdot)}(\xi) \right|^2 \left(1 + \|\xi/\zeta_1\|^2\right)^{\frac{m+s}{2}} d\xi \right)^{\frac{1}{2}} \\ & \leqslant \frac{\zeta_2^m \pi^{\frac{m+1}{4}}}{\zeta_1^{\frac{m}{2}} (2\pi)^m \Gamma\left(\frac{m+1}{2}\right)^{\frac{1}{2}}} \sum_{\beta \in \mathbb{N}_{0,c}^m} \sum_{\substack{\beta_1,\beta_2 \in \mathbb{N}_0^m \\ \beta_1+\beta_2=\beta}} \frac{\beta!}{\beta_1!\beta_2!} \left(\int_{\mathbb{R}^m} \left| \partial_{\beta_1} \widehat{\phi_{\lambda,t}}(\xi) \partial_{\beta_2} \widehat{g}(\xi) \right|^2 \left(1 + \|\xi/\zeta_1\|^2\right)^{\frac{m+s}{2}} d\xi \right)^{\frac{1}{2}} \\ & \leqslant \left|\mathbb{N}_{0,c}^m\right| \max_{\beta \in \mathbb{N}_{0,c}^m} \left(\sum_{\substack{\beta_1,\beta_2 \in \mathbb{N}_0^m \\ \beta_1+\beta_2=\beta}} \frac{\beta!}{\beta_1!\beta_2!} \right) \max_{\beta_2 \in \mathbb{N}_{0,c}^m} \sup_{\xi \in \mathbb{R}^m} \left| \partial_{\beta_2} \widehat{g}(\xi) \right| \\ & \quad \cdot \max_{\beta_1 \in \mathbb{N}_{0,c}^m} \left(\frac{\zeta_2^{2m} \pi^{\frac{m+1}{2}}}{\zeta_1^m (2\pi)^{2m} \Gamma\left(\frac{m+1}{2}\right)} \int_{\mathbb{R}^m} \left| \partial_{\beta_1} \widehat{\phi_{\lambda,t}}(\xi) \right|^2 \left(1 + \|\xi/\zeta_1\|^2\right)^{\frac{m+s}{2}} d\xi \right)^{\frac{1}{2}} \\ & \leqslant 2m^c 2^c \max_{\beta_2 \in \mathbb{N}_{0,c}^m} \|p_{\beta_2} \cdot g\|_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du)} \sqrt{C_{24}+C_{25}} \left(\frac{(\zeta_2/\zeta_1)^2}{\sqrt{2\lambda t}\pi} \right)^{\frac{m}{2}} m^{\frac{c_s}{4}} \\ & \leqslant C_9 m^{\frac{5c+s}{4}} \left(\frac{(\zeta_2/\zeta_1)^2}{\sqrt{2\lambda t}\pi} \right)^{\frac{m}{2}} \|g\|_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),(1+\|u\|)^c du)}, \end{aligned}$$

which completes the proof. □

Proof of Corollary 6.2. For $\lambda, t \in (0,\infty)$ and an a.e. bounded and a.e. continuous initial condition $g : \mathbb{R}^m \to \mathbb{R}$, let $f(t,\cdot) : \mathbb{R}^m \to \mathbb{R}$ be the solution of (19) at time t. Moreover, let $p \in [1,\infty)$, $\gamma \in [0,\infty)$, and $w : \mathbb{R}^m \to [0,\infty)$ satisfy the conditions of Lemma 4.23.

For Part (i)., we fix some $N \in \mathbb{N}$ and assume that $g \in L^1(\mathbb{R}^m, \mathcal{L}(\mathbb{R}^m), du)$. Then, by using the probability density function $p_\theta : \mathbb{R}^m \to [0,\infty)$ of the Student's t-distributed i.i.d. sequence $(\theta_n)_{n\in\mathbb{N}} \sim t_m$ and Lemma 10.1 (with $\zeta_1 := \zeta_2 := 1$, $c := 0$, $s := 1$, and constant $C_{12} > 0$ being independent of $m \in \mathbb{N}$

and $g : \mathbb{R}^m \to \mathbb{R}$), we observe that

$$\begin{aligned}
\frac{C_{f(t,\cdot)}}{(2\pi)^m} &:= \frac{1}{(2\pi)^m} \left(\int_{\mathbb{R}^m} \frac{\left| \widehat{f(t,\cdot)}(\vartheta) \right|^2}{p_\theta(\vartheta)} d\vartheta \right)^{\frac{1}{2}} \\
&= \frac{\pi^{\frac{m+1}{4}}}{(2\pi)^m \Gamma\left(\frac{m+1}{2}\right)^{\frac{1}{2}}} \left(\int_{\mathbb{R}^m} \left| \widehat{f(t,\cdot)}(\vartheta) \right|^2 \left(1 + \|\vartheta\|^2\right)^{\frac{m+1}{2}} d\vartheta \right)^{\frac{1}{2}} \\
&\leqslant C_4 m^{\frac{1}{4}} \left(\frac{1}{\sqrt{2\lambda t \pi}} \right)^{\frac{m}{2}} \|g\|_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du)} < \infty.
\end{aligned} \tag{67}$$

Hence, we can apply Corollary 4.8 (with constant $C_{p,2} > 0$ depending only on $p \in (1,\infty)$) to obtain a random trigonometric feature model $G_N \in \mathcal{RT}_{\mathbb{R}^m,1}$ with N features satisfying

$$\mathbb{E}\left[\|f(t,\cdot) - G_N\|^2_{L^p(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),w)} \right]^{\frac{1}{2}} \leqslant C_{p,2} \frac{w(\mathbb{R}^m)^{\frac{1}{p}}}{(2\pi)^m} \frac{C_{f(t,\cdot)}}{N^{1-\frac{1}{\min(2,p)}}}.$$

Thus, by using that $w : \mathbb{R}^m \to [0,\infty)$ is a weight satisfying the conditions of Lemma 4.23, i.e. that $\mathbb{R}^m \ni u \mapsto w(u) := \prod_{l=1}^m w_0(u_l) \in [0,\infty)$ for some $w_0 : \mathbb{R} \to [0,\infty)$ satisfying $\int_{\mathbb{R}} w_0(s) ds = 1$, which implies that $w(\mathbb{R}^m) = \int_{\mathbb{R}^m} w(u) du = \prod_{l=1}^m \int_{\mathbb{R}} w_0(u_l) du_l = 1$ by Fubini's theorem, Lemma 10.1 (with constant $C_{12} > 0$ being independent of $m \in \mathbb{N}$ and $g : \mathbb{R}^m \to \mathbb{R}$), and the constant $C_6 := C_{p,2} C_{12} > 0$ (being independent of $m \in \mathbb{N}$ and $g : \mathbb{R}^m \to \mathbb{R}$), it follows that

$$\begin{aligned}
\mathbb{E}\left[\|f(t,\cdot) - G_N\|^2_{L^p(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),w)} \right]^{\frac{1}{2}} &\leqslant C_{p,2} \frac{w(\mathbb{R}^m)^{\frac{1}{p}}}{(2\pi)^m} \frac{C_{f(t,\cdot)}}{N^{1-\frac{1}{\min(2,p)}}} \\
&\leqslant C_6 m^{\frac{1}{4}} \left(\frac{1}{\sqrt{2\lambda t \pi}} \right)^{\frac{m}{2}} \frac{\|g\|_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du)}}{N^{1-\frac{1}{\min(2,p)}}},
\end{aligned}$$

which proves the inequality (20). For (21), we assume that $p > 1$ and that $\mathbb{R}^m \ni u \mapsto g(u) := \mathbb{1}_{B_R(0)}(u) \in \mathbb{R}$ with $R^2 \leqslant \frac{\sqrt{\lambda t}}{\sqrt{2} e}(m+2)$ for all but finitely many $m \in \mathbb{N}$. Then, there exists a constant $C_7 > 0$ (being independent of $m \in \mathbb{N}$) such that for every $m \in \mathbb{N}$ it holds that

$$\frac{2eC_6}{\sqrt{4\pi}} \left(\frac{\sqrt{2} e R^2}{\sqrt{\lambda t}(m+2)} \right)^{\frac{m}{2}} \leqslant C_7^{1-\frac{1}{\min(2,p)}}. \tag{68}$$

Hence, for every $\varepsilon > 0$ and $m \in \mathbb{N}$, we use the inequality (20) with $N = \left\lceil C_7 \varepsilon^{-\frac{\min(2,p)}{\min(2,p)-1}} \right\rceil$, that $\|g\|_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du)} = \int_{B_R(0)} du = \frac{\pi^{m/2} R^m}{\Gamma(m/2+1)}$, that $\Gamma(x) \geqslant \sqrt{2\pi/x}(x/e)^x$ for any $x \in (0,\infty)$ (see [33, Lemma 2.4]), and (68) to obtain some $G_N \in \mathcal{RT}_{\mathbb{R}^m,1}$ with N features satisfying

$$\begin{aligned}
\mathbb{E}\left[\|f(t,\cdot) - G_N\|^2_{L^p(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),w)} \right]^{\frac{1}{2}} &\leqslant C_6 m^{\frac{1}{4}} \left(\frac{1}{\sqrt{2\lambda t \pi}} \right)^{\frac{m}{2}} \frac{\|g\|_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du)}}{N^{1-\frac{1}{\min(2,p)}}} \\
&= C_6 m^{\frac{1}{4}} \left(\frac{1}{\sqrt{2\lambda t \pi}} \right)^{\frac{m}{2}} \frac{\frac{\pi^{\frac{m}{2}} R^m}{\Gamma(m/2+1)}}{N^{1-\frac{1}{\min(2,p)}}} \\
&\leqslant C_6 m^{\frac{1}{4}} \left(\frac{R^2}{\sqrt{2\lambda t}} \right)^{\frac{m}{2}} \frac{\sqrt{\frac{m+2}{4\pi}} \left(\frac{2e}{m+2} \right)^{\frac{m+2}{2}}}{N^{1-\frac{1}{\min(2,p)}}} \\
&\leqslant \frac{2eC_6}{\sqrt{4\pi}} \left(\frac{\sqrt{2} e R^2}{\sqrt{\lambda t}(m+2)} \right)^{\frac{m}{2}} \frac{1}{N^{1-\frac{1}{\min(2,p)}}} \\
&\leqslant \frac{C_7^{1-\frac{1}{\min(2,p)}}}{N^{1-\frac{1}{\min(2,p)}}} \leqslant \varepsilon,
\end{aligned}$$

which proves the inequality (21).

For Part (ii)., we fix some $N \in \mathbb{N}$ and assume that $g \in L^1(\mathbb{R}^m, \mathcal{L}(\mathbb{R}^m), (1+\|u\|)^{\lceil \gamma \rceil + 2} du)$. Then, by using (13) and Lemma 10.1 (with $0 < \zeta_1 < \zeta_2 < \infty$, $c := \lceil \gamma \rceil + 2$, $s := 4\lceil \gamma \rceil + 5$, and constant $C_{12} > 0$

being independent of $m \in \mathbb{N}$ and $g : \mathbb{R}^m \to \mathbb{R}$), we observe that

$$
\begin{aligned}
&\frac{\zeta_2^m}{(2\pi)^m} \Vert f(t,\cdot) \Vert_{\widetilde{\mathbb{B}}^{0,2,\gamma}_{\psi,a,b}(\mathbb{R}^m)} \\
&\quad\leqslant C_1 \frac{\zeta_2^m \pi^{\frac{m+1}{4}}}{\zeta_1^{\frac{m}{2}} (2\pi)^m \Gamma\left(\frac{m+1}{2}\right)} \sum_{\beta \in \mathbb{N}^m_{0,c}} \left(\int_{\mathbb{R}^m} \left| \widehat{\partial_\beta f(t,\cdot)}(\xi) \right|^2 \left(1 + \Vert \xi/\zeta_1 \Vert^2\right)^{2\lceil\gamma\rceil + \frac{m+5}{2}} d\xi \right)^{\frac{1}{2}} \\
&\quad\leqslant C_{12} m^{\frac{9\lceil\gamma\rceil+15}{4}} \left(\frac{(\zeta_2/\zeta_1)^2}{\sqrt{2\lambda t \pi}} \right)^{\frac{m}{2}} \Vert g \Vert_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),(1+\Vert u \Vert)^{\lceil\gamma\rceil+2} du)} < \infty.
\end{aligned}
\tag{69}
$$

Since $\Vert g \Vert_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du)} \leqslant \Vert g \Vert_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),(1+\Vert u \Vert)^{\lceil\gamma\rceil+2} du)} < \infty$ and $\Vert \phi_{\lambda,t} \Vert_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du)} = 1$, we can apply Young's convolutional inequality on $f(t,\cdot) = \phi_{\lambda,t} * g$ (see Proposition 6.1) to conclude that $f(t,\cdot) \in L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),du)$, which implies that $f(t,\cdot) \in \widetilde{\mathbb{B}}^{0,2,\gamma}_{\psi,a,b}(\mathbb{R}^m)$. Hence, we can use Corollary 4.20 (with constant $C_{p,2} > 0$ depending only on $p \in (1,\infty)$) to obtain a random neural network $G_N \in \mathcal{RN}^\rho_{\mathbb{R}^m,1}$ with N neurons satisfying

$$
\mathbb{E}\left[\Vert f(t,\cdot) - G_N \Vert^2_{L^p(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),w)} \right]^{\frac{1}{2}} \leqslant C_{p,2} \Vert \rho \Vert_{C^0_{pol,\gamma}(\mathbb{R})} \frac{C^{(\gamma,p)}_{U,w}}{\left| C^{(\psi,\rho)}_m \right|} \frac{\Vert f(t,\cdot) \Vert_{\widetilde{\mathbb{B}}^{0,2,\gamma}_{\psi,a,b}(\mathbb{R}^m)}}{N^{1-\frac{1}{\min(2,p)}}}.
$$

Thus, by using Lemma 4.23 (with constant $C^{(\gamma,p)}_{\mathbb{R},w_0}$ depending only on $\gamma \in [0,\infty)$, $p \in (1,\infty)$, and $w_0 : \mathbb{R} \to [0,\infty)$), Example 4.15 (with constant $C_{\psi,\rho} > 0$ depending only on $\psi \in \mathcal{S}_0(\mathbb{R};\mathbb{C})$ and $\rho \in C^0_{pol,\gamma}(\mathbb{R})$), Lemma 10.1 (with constant $C_{12} > 0$ being independent of $m \in \mathbb{N}$ and $g : \mathbb{R}^m \to \mathbb{R}$), and the constants $C_8 := \frac{C_{p,2} C^{(\gamma,p)}_{\mathbb{R},w_0} C_{12}}{C_{\psi,\rho}} > 0$ and $C_9 := \gamma + \frac{1}{p} + \frac{9\lceil\gamma\rceil+15}{4} > 0$ (being independent of $m \in \mathbb{N}$ and $g : \mathbb{R}^m \to \mathbb{R}$), we have

$$
\begin{aligned}
\mathbb{E}\left[\Vert f(t,\cdot) - G_N \Vert^2_{L^p(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),w)} \right]^{\frac{1}{2}} &\leqslant C_{p,2} \Vert \rho \Vert_{C^0_{pol,\gamma}(\mathbb{R})} \frac{C^{(\gamma,p)}_{U,w}}{\left| C^{(\psi,\rho)}_m \right|} \frac{\Vert f(t,\cdot) \Vert_{\widetilde{\mathbb{B}}^{0,2,\gamma}_{\psi,a,b}(\mathbb{R}^m)}}{N^{1-\frac{1}{\min(2,p)}}} \\
&\leqslant \frac{C_{p,2} C^{(\gamma,p)}_{\mathbb{R},w_0}}{C_{\psi,\rho}} m^{\gamma+\frac{1}{p}} \frac{\zeta_2^m}{(2\pi)^m} \frac{\Vert f(t,\cdot) \Vert_{\widetilde{\mathbb{B}}^{0,2,\gamma}_{\psi,a,b}(\mathbb{R}^m)}}{N^{1-\frac{1}{\min(2,p)}}} \\
&\leqslant C_8 m^{C_9} \left(\frac{(\zeta_2/\zeta_1)^2}{\sqrt{2\lambda t \pi}} \right)^{\frac{m}{2}} \frac{\Vert g \Vert_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),(1+\Vert u \Vert)^{\lceil\gamma\rceil+2} du)}}{N^{1-\frac{1}{\min(2,p)}}},
\end{aligned}
$$

which proves the inequality (22). For (23), we assume that $p > 1$ and that $\mathbb{R}^m \ni u \mapsto g(u) := \mathbb{1}_{B_R(0)}(u) \in \mathbb{R}$ with $R^2 \leqslant \frac{\zeta_1^2 \sqrt{\lambda t}}{\zeta_2^2 \sqrt{2e}}(m+2)$ for all but finitely many $m \in \mathbb{N}$. Then, there exist some constants $C_{10}, C_{11} > 0$ (being independent of $m \in \mathbb{N}$) such that for every $m \in \mathbb{N}$ it holds that

$$
\frac{2e C_8}{\sqrt{4\pi}} m^{C_9} (1+R)^{\lceil\gamma\rceil+2} \left(\frac{\sqrt{2} e (\zeta_2/\zeta_1)^2 R^2}{\sqrt{\lambda t}(m+2)} \right)^{\frac{m}{2}} \leqslant \left(C_{10} m^{C_{11}} \right)^{1-\frac{1}{\min(2,p)}}.
\tag{70}
$$

Hence, for every $\varepsilon > 0$ and $m \in \mathbb{N}$, we use (22) with $N = \left\lceil C_{10} m^{C_{11}} \varepsilon^{-\frac{\min(2,p)}{\min(2,p)-1}} \right\rceil$, that $\Vert g \Vert_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),(1+\Vert u \Vert)^{\lceil\gamma\rceil+2} du)} = \int_{B_R(0)} (1+\Vert u \Vert)^{\lceil\gamma\rceil+2} du \leqslant (1+R)^{\lceil\gamma\rceil+2} \int_{B_R(0)} du = (1+

$R)^{\lceil\gamma\rceil+2}\frac{\pi^{m/2}R^m}{\Gamma(m/2+1)}$, and that $\Gamma(x) \geqslant \sqrt{2\pi/x}(x/e)^x$ for any $x \in (0,\infty)$ (see [33, Lemma 2.4]) to obtain some $G_N \in \mathcal{RN}^{\tanh}_{\mathbb{R}^m,1}$ with N neurons satisfying

$$\begin{aligned}
\mathbb{E}\left[\|f(t,\cdot) - G_N\|^2_{L^p(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),w)}\right]^{\frac{1}{2}} &\leqslant C_8 m^{C_9}\left(\frac{(\zeta_2/\zeta_1)^2}{\sqrt{2\lambda t\pi}}\right)^{\frac{m}{2}} \frac{\|g\|_{L^1(\mathbb{R}^m,\mathcal{L}(\mathbb{R}^m),(1+\|u\|)^{\lceil\gamma\rceil+2}du)}}{N^{1-\frac{1}{\min(2,p)}}} \\
&= C_8 m^{C_9}(1+R)^{\lceil\gamma\rceil+2}\left(\frac{(\zeta_2/\zeta_1)^2}{\sqrt{2\lambda t\pi}}\right)^{\frac{m}{2}} \frac{\frac{\pi^{\frac{m}{2}}R^m}{\Gamma(m/2+1)}}{N^{1-\frac{1}{\min(2,p)}}} \\
&\leqslant C_8 m^{C_9}(1+R)^{\lceil\gamma\rceil+2}\left(\frac{(\zeta_2/\zeta_1)^2 R^2}{\sqrt{2\lambda t}}\right)^{\frac{m}{2}} \frac{\sqrt{\frac{m+2}{4\pi}}\left(\frac{2e}{m+2}\right)^{\frac{m+2}{2}}}{N^{1-\frac{1}{\min(2,p)}}} \\
&\leqslant \frac{2eC_8}{\sqrt{4\pi}} m^{C_9}(1+R)^{\lceil\gamma\rceil+2}\left(\frac{\sqrt{2}e(\zeta_2/\zeta_1)^2R^2}{\sqrt{\lambda t}(m+2)}\right)^{\frac{m}{2}} \frac{1}{N^{1-\frac{1}{\min(2,p)}}} \\
&\leqslant \frac{\left(C_{10}m^{C_{11}}\right)^{1-\frac{1}{\min(2,p)}}}{N^{1-\frac{1}{\min(2,p)}}} \leqslant \varepsilon,
\end{aligned}$$

which proves the inequality (23). □

References

[1] Milton Abramowitz and Irene Ann Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Applied mathematics series / National Bureau of Standards 55, Print. 9. Dover, New York, 9th edition, 1970.

[2] Robert A. Adams. *Sobolev Spaces*. Pure and applied mathematics. Academic Press, 1975.

[3] Fernando Albiac and Nigel J. Kalton. *Topics in Banach space theory*. Graduate Texts in Mathematics 233. Springer, New York, 2006.

[4] Mauricio A Alvarez, Lorenzo Rosasco, Neil D Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

[5] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[6] Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 253–262. PMLR, 06-11 Aug 2017.

[7] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

[8] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017.

[9] Andrew R. Barron. Neural net approximation. In *Proc. 7th Yale workshop on adaptive and learning systems*, volume 1, pages 69–72, 1992.

[10] Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

[11] Andrew R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14:115–134, 1994.

[12] Åke Björck. *Numerical methods for least squares problems*. Society for Industrial and Applied Mathematics, Philadelphia, 1996.

[13] Helmut Bölcskei, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. Optimal approximation with sparsely connected deep neural networks. *SIAM Journal on Mathematics of Data Science*, 1:8–45, 2019.

[14] Romain Brault, Markus Heinonen, and Florence Buc. Random fourier features for operator-valued kernels. In Robert J. Durrant and Kee-Eung Kim, editors, *Proceedings of The 8th Asian Conference on Machine Learning*, volume 63 of *Proceedings of Machine Learning Research*, pages 110–125, The University of Waikato, Hamilton, New Zealand, 16-18 Nov 2016. PMLR.

[15] Haïm Brézis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer, New York, 2011.

[16] Emmanuel J. Candès. *Ridgelets: Theory and Applications*. PhD thesis, Stanford University, 1998.

[17] C. Carmeli, E. De Vito, A. Toigo, and V. Umanità. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 08(01):19–61, 2010.

[18] Luigi Carratino, Alessandro Rudi, and Lorenzo Rosasco. Learning with SGD and random features. *arXiv e-prints 2306.03303*, 2018.

[19] Tianping Chen and Hong Chen. Approximation capability to functions of several variables, nonlinear functionals, and operators by radial basis function neural networks. *IEEE Transactions on Neural Networks*, 6(4):904–910, 1995.

[20] Lénač Chizat, Maria Colombo, Xavier Fernńdez-Real, and Alessio Figalli. Infinite-width limit of deep linear neural networks. *Communications on Pure and Applied Mathematics*, 77(10):3958–4007, 2024.

[21] Christa Cuchiero, Philipp Schmocker, and Josef Teichmann. Global universal approximation of functional input maps on weighted spaces. *arXiv e-prints 2306.03303*, 2023.

[22] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[23] Christian Darken, James M. Donahue, Leonid Gurvits, and Eduardo D. Sontag. Rate of approximation results motivated by robust neural network learning. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, COLT '93, pages 303–309, New York, NY, USA, 1993. Association for Computing Machinery.

[24] Christian Darken, James M. Donahue, Leonid Gurvits, and Eduardo D. Sontag. Rates of convex approximation in non-Hilbert spaces. *Constructive Approximation*, 13:187–220, 1997.

[25] Suchuan Dong and Zongwei Li. Local extreme learning machines and domain decomposition for solving linear and nonlinear partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 387:114–129, 2021.

[26] Vikas Dwivedi and Balaji Srinivasan. Physics informed extreme learning machine (PIELM) – a rapid method for the numerical solution of partial differential equations. *Neurocomputing*, 391:96–118, 2020.

[27] Weinan E, Chao Ma, and Lei Wu. The Barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55:369–406, 2022.

[28] Weinan E, Chao Ma, Lei Wu, and Stephan Wojtowytsch. Towards a mathematical understanding of neural network-based machine learning: What we know and what we don't. *CSIAM Transactions on Applied Mathematics*, 1(4):561–615, 2020.

[29] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate studies in mathematics*. American Mathematical Society, Providence, Rhode Island, 2nd edition, 2010.

[30] Gerald B. Folland. *Fourier analysis and its applications*. Brooks/Cole Publishing Company, Belmont, California, 1st edition, 1992.

[31] Lukas Gonon. Random feature neural networks learn Black-Scholes type PDEs without curse of dimensionality. *Journal of Machine Learning Research*, 24(189):1–51, 2023.

[32] Lukas Gonon, Lyudmila Grigoryeva, and Juan-Pablo Ortega. Approximation bounds for random neural networks and reservoir systems. *The Annals of Applied Probability*, 33(1):28–69, 2023.

[33] Lukas Gonon, Philipp Grohs, Arnulf Jentzen, David Kofler, and David Šiška. Uniform error estimates for artificial neural network approximations for heat equations. *IMA Journal of Numerical Analysis*, 42(3):1991–2054, 08 2021.

[34] Lukas Gonon and Antoine Jacquier. Universal approximation theorem and error bounds for quantum neural networks and quantum reservoirs. *arXiv e-prints 2307.12904*, 2023.

[35] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[36] Loukas Grafakos. *Modern Fourier Analysis*. Graduate Texts in Mathematics, 250. Springer New York, New York, NY, 3rd edition, 2014.

[37] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Non-Parametric Regression*. Springer Series in Statistics. Springer, New York, Berlin, Heidelberg, 2002.

[38] Allen Hart, James Hook, and Jonathan Dawes. Embedding and approximation theorems for echo state networks. *Neural Networks*, 128:234–247, 2020.

[39] Jakob Heiss, Josef Teichmann, and Hanna Wutte. How implicit regularization of relu neural networks characterizes the learned function – part i: the 1-d case of two layers with random first layer. *arXiv e-prints 1911.02903*, 2019.

[40] Calypso Herrera, Florian Krach, Pierre Ruyssen, and Josef Teichmann. Optimal stopping via randomized neural networks. *arXiv e-prints 2104.13669*, 2021.

[41] Lars Hörmander. *The Analysis of Linear Partial Differential Operators I: Distribution Theory and Fourier Analysis*. Classics in Mathematics. Springer, Berlin, Heidelberg, 2nd edition, 1990.

[42] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[43] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1):489–501, 2006. Neural Networks.

[44] Tuomas Hytönen, Jan van Neerven, Mark Veraar, and Lutz Weis. *Analysis in Banach Spaces*, volume 63 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge*. Springer, Cham, 2016.

[45] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[46] Antoine Jacquier and Žan Žurič. Random neural networks for rough volatility. *arXiv e-prints 2305.01035*, 2023.

[47] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. 2015.

[48] Jason M. Klusowski and Andrew R. Barron. Risk bounds for high-dimensional ridge function combinations including neural networks. *arXiv e-prints 1607.01434*, 2016.

[49] Věra Kůrková. Complexity estimates based on integral transforms induced by computational units. *Neural Networks*, 33:160–167, 2012.

[50] Alois Kufner. *Weighted Sobolev spaces*. Teubner-Texte zur Mathematik Bd. 31. B.G. Teubner, Leipzig, 1980.

[51] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces: Isoperimetry and Processes*. Ergebnisse der Mathematik und ihrer Grenzgebiete. Folge 3 Bd. 23. Springer, Berlin, 1991.

[52] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

[53] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022. Special Issue on Harmonic Analysis and Machine Learning.

[54] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

[55] Hrushikesh Narhar Mhaskar and Charles A Micchelli. Degree of approximation by neural and translation networks with a single hidden layer. *Advances in Applied Mathematics*, 16(2):151–183, 1995.

[56] Charles A. Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 01 2005.

[57] Ha Quang Minh. Operator-valued bochner theorem, fourier feature maps for operator-valued kernels, and vector-valued learning. *arXiv e-prints 1608.05639*, 2016.

[58] Grégoire Montavon, Geneviève Orr, and Klaus-Robert Müller. *Neural Networks: Tricks of the Trade*. Theoretical Computer Science and General Issues; 7700. Springer, Berlin, Heidelberg, 2nd edition, 2012.

[59] James R. Munkres. *Topology*. Pearson, Harlow, Essex, UK, 2nd, pearson new international edition, 2014.

[60] Radford M. Neal. *Priors for Infinite Networks*, pages 29–53. Springer, New York, NY, 1996.

[61] Nicholas H. Nelsen and Andrew M. Stuart. The random feature model for input-output maps between Banach spaces. *SIAM Journal on Scientific Computing*, 43(5):A3212–A3243, 2021.

[62] Nicholas H. Nelsen and Andrew M. Stuart. Operator learning using random features: A tool for scientific computing. *SIAM Review*, 66(3):535–571, 2024.

[63] Ariel Neufeld and Philipp Schmocker. Chaotic hedging with iterated integrals and neural networks. *arXiv e-prints 2209.10166*, 2022.

[64] Ariel Neufeld and Philipp Schmocker. Universal approximation results for neural networks with non-polynomial activation function over non-compact domains. *arXiv e-prints 2410.14759*, 2024.

[65] Ariel Neufeld, Philipp Schmocker, and Sizhou Wu. Full error analysis of the random deep splitting method for nonlinear parabolic pdes and pides with infinite activity. *arXiv e-prints 2405.05192*, 2024.

[66] Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999.

[67] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, pages 1177–1184, Red Hook, NY, USA, 2007. Curran Associates Inc.

[68] Ali Rahimi and Benjamin Recht. Uniform approximation of functions with random bases. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 555–561, 2008.

[69] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

[70] Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, MS, 1st edition, 2006.

[71] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational analysis*. Grundlehren der mathematischen Wissenschaften; 317. Springer, Berlin, 1st edition, 1997.

[72] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 3218–3228, Red Hook, NY, USA, 2017. Curran Associates Inc.

[73] Walter Rudin. *Real and complex analysis*. McGraw-Hill series in higher mathematics. WCB/McGraw-Hill, Boston, Massachusetts, 3rd edition, 1987.

[74] Walter Rudin. *Functional analysis*. International series in pure and applied mathematics. McGraw-Hill, Boston, Mass, 2nd edition, 1991.

[75] Jonathan W. Siegel and Jinchao Xu. Approximation rates for neural networks with general activation functions. *Neural Networks*, 128:313–321, 2020.

[76] Sho Sonoda and Noboru Murata. Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, 43(2):233–268, 2017.

[77] Marshall Harvey Stone. The generalized Weierstrass approximation theorem. *Mathematics Magazine*, 21(4):167–184, 1948.

[78] Yiran Wang and Suchuan Dong. An extreme learning machine-based method for computational PDEs in higher dimensions. *arXiv e-prints 2309.07049*, 2023.

[79] Christopher Williams. Computing with infinite networks. In M.C. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996.

[80] Xuwei Yang, Anastasis Kratsios, Florian Krach, Matheus Grasselli, and Aurelien Lucchi. Regret-optimal federated transfer learning for kernel regression – with applications in American option pricing. *arXiv e-prints 2309.04557*, 2023.

[81] Yunlei Yang, Muzhou Hou, and Jianshu Luo. A novel improved extreme learning machine algorithm in solving ordinary differential equations by Legendre neural network methods. *Advances in Difference Equations*, 2018:469, 2018.

[82] Haizhang Zhang, Yuesheng Xu, and Jun Zhang. Reproducing kernel banach spaces for machine learning. *Journal of Machine Learning Research*, 10(95):2741–2775, 2009.

NANYANG TECHNOLOGICAL UNIVERSITY, DIVISION OF MATHEMATICAL SCIENCES, 21 NANYANG LINK, SINGAPORE
Email address: ariel.neufeld@ntu.edu.sg

NANYANG TECHNOLOGICAL UNIVERSITY, DIVISION OF MATHEMATICAL SCIENCES, 21 NANYANG LINK, SINGAPORE
Email address: philippt001@e.ntu.edu.sg