2024

Deep Neural Networks: A Formulation Via Non-Archimedean Analysis

Wilson A. Zuniga-Galindo
The University of Texas Rio Grande Valley

DEEP NEURAL NETWORKS: A FORMULATION VIA NON-ARCHIMEDEAN ANALYSIS

W. A. ZÚÑIGA-GALINDO

ABSTRACT. We introduce a new class of deep neural networks (DNNs) with multilayered tree-like architectures. The architectures are codified using numbers from the ring of integers of non-Archimdean local fields. These rings have a natural hierarchical organization as infinite rooted trees. Natural morphisms on these rings allow us to construct finite multilayered architectures. The new DNNs are robust universal approximators of real-valued functions defined on the mentioned rings. We also show that the DNNs are robust universal approximators of real-valued square-integrable functions defined in the unit interval.

1. INTRODUCTION

It is known that neural networks (NN) can learn hierarchically. However, there is currently no theoretical framework that explains how this hierarchical learning occurs; see, e.g., [1], and the references therein. The study of deep learning architectures which can learn data hierarchically organized is an active research area nowadays; see, e.g., [1], [31] and the references therein. In [1], the authors propose to find natural classes of hierarchical functions and to study how regular deep neural networks (DNNs) can learn them. In this article, we follow the converse approach. Using non-Archimedean analysis, we construct a new class of DNNs with tree-like architectures. These DNNs are robust universal approximators of hierarchical functions, and they also are universal approximators for standard functions. The new hierarchical DNNs compute approximations using arithmetic operations in non-Archimedean fields, but they can be trained using the traditional back propagation method.

An non-Archimedean vector space $(M, \|\cdot\|)$ is a normed vector space whose norm satisfies

$$\|x + y\| \leq \max\{\|x\|, \|y\|\},$$

for any two vectors x, y in M. In such a space, the balls are organized in a hierarchical form. This type of space plays a central role in formulating models of complex multi-level systems; in this type of system, the hierarchy plays a significant role; see, e.g., [18], [19], and the references therein. These systems are made up of several subsystems and are characterized by emergent behavior resulting from nonlinear interactions between subsystems for multiple levels of organization. The field of p-adic numbers $\mathbb{Q}_p$ and the field of formal Laurent series $\mathbb{F}_p((T))$ are paramount examples of non-Archimedean vector spaces.

2020 *Mathematics Subject Classification.* Primary 68T07, 65D15; Secondary 41A30, 11S85.

Key words and phrases. Artificial neural networks and deep learning, Algorithms for approximation of functions, non-Archimedean fields, p-adic analysis.

The author was partially supported by the Lokenath Debnath Endowed Professorship.

Nowadays, it is widely accepted that cortical neural networks are arranged in fractal or self-similar patterns and have the small-world property, see, e.g., [14], [25], and the references therein. Methods of non-Archimedean analysis have been successfully used to construct models for this type of network, see, e.g., [20], [33], and the references therein. A restricted Boltzmann machine (RBM) is a generative stochastic artificial neural network based on the dynamics of a spin glass. The p-adic spin glasses constitute a particular case of hierarchical spin glasses. The RBM corresponding to a p-adic spin glass is a hierarchical neural network. These neural networks are universal approximators, [35], and their dynamics can be understood via Euclidean quantum field theory, [34], [36]. p-Adic versions of the cellular networks were developed in [29]-[30]. This work continues our investigation of hierarchical NNs via non-Archimedean analysis.

To discuss our results, here, we restrict to the case of the ring of integers $\mathbb{F}_p[[T]]$ of the field of formal Laurent series with coefficients in a finite field $\mathbb{F}_p = \{0, \ldots, p-1\}$, with p a prime number,

$$\mathbb{F}_p[[T]] = \left\{ T^r \sum_{k=0}^{\infty} a_k T^k ; a_k \in \mathbb{F}_p,\ a_0 \neq 0, r \in \mathbb{Z} \right\} \cup \{0\},$$

where T is an indeterminate. The function $\left| T^r \sum_{k=0}^{\infty} a_k T^k \right| = p^{-r}$, $|0| = 0$ defines a norm. Then $(\mathbb{F}_p[[T]], |\cdot|)$ is a non-Archimedean vector space, which is also a complete space, and $(\mathbb{F}_p[[T]], +)$ is a compact topological group. We denote by dx a Haar measure in $\mathbb{F}_p[[T]]$. The points of $\mathbb{F}_p[[T]]$ are naturally organized in an infinite rooted tree. The points of the form

$$G_l = \left\{ a_0 + \ldots + a_{l-1} T^{l-1} ; a_k \in \mathbb{F}_p \right\} \subset \mathbb{F}_p[[T]],\ l \geq 1,$$

constitute the l-th layer (level) of the tree $\mathbb{F}_p[[T]]$. Geometrically, G_l is a finite rooted tree, but also, G_l is a finite additive group. We assume that every point of G_l there is a neuron. In this way, we will identify $\mathbb{F}_p[[T]]$ with an infinite hierarchical network; G_l is a discrete approximation of $\mathbb{F}_p[[T]]$ obtained by cutting the infinite tree $\mathbb{F}_p[[T]]$ at layer l.

A ball in $\mathbb{F}_p[[T]]$ with center at $a \in G_L$, $L \geq 1$, and radius p^{-L} is the set $B_{-L}(a) = a + T^L \mathbb{F}_p[[T]]$. This ball is an infinite rooted tree, with root a. We denote by $\Omega\left(p^L |x-a|\right)$ the characteristic function of the ball $B_{-L}(a)$. The set of functions

$$\varphi(x) = \sum_{a \in G_l} c_a \Omega\left(p^L |x-a|\right),\ c_a \in \mathbb{R},$$

form an $\mathbb{R}$-vector space, denoted as $\mathcal{D}^l = \mathcal{D}^l(\mathbb{F}_p[[T]])$. The functions from this space are naturally interpreted as hierarchical functions. The $\mathbb{R}$-vector space $\mathcal{D} = \cup_l \mathcal{D}^l$ is called the space of test functions, this space is dense in $L^\rho(\mathbb{F}_p[[T]], dx)$, $\rho \in [1, \infty]$.

We now pick an activation function $\sigma_M : \mathbb{R} \to (-M, M)$, with $M > 0$, a weight function $w \in L^\infty(\mathbb{F}_p[[T]] \times \mathbb{F}_p[[T]], dx \times dx)$, and bias function $\theta \in L^\infty(\mathbb{F}_p[[T]], dx)$; a continuous DNN on $\mathbb{F}_p[[T]]$ is a function of type

$$Y(x) = \sigma_M \left(\int_{\mathbb{F}_p[[T]]} w(x,y) X(y) d^N y + \theta(x) \right), \tag{1.1}$$

where $X\in L^{\infty}\left(\mathbb{F}_{p}\left[\left[T\right]\right],dx\right)$, is the state of the DNN. At first sight, (1.1) seems to be a continuous perceptron, but due to the fact that $\mathbb{F}_{p}\left[\left[T\right]\right]$ has a hierarchical structure, (1.1) is a hierarchical DNN. For practical purposes, only discrete versions of (1.1) are relevant. Here, we consider DNNs where $X\in\mathcal{D}^{l-1}\left(\mathbb{F}_{p}\left[\left[T\right]\right]\right)$, $\theta\in\mathcal{D}^{l}\left(\mathbb{F}_{p}\left[\left[T\right]\right]\right)$, $w\in\mathcal{D}^{l}\left(\mathbb{F}_{p}\left[\left[T\right]\right]\times\mathbb{F}_{p}\left[\left[T\right]\right]\right)$, for $l\geq2$, then $Y\in\mathcal{D}^{l}\left(\mathbb{F}_{p}\left[\left[T\right]\right]\right)$, and

$$Y\left(b\right)=\sigma_{M}\left(\sum_{a\in G_{l}}X\left(\Lambda_{l}\left(a\right)\right)w\left(b,a\right)+\theta\left(b\right)\right)\text{, for }b\in G_{l}, \tag{1.2}$$

where $\Lambda_{l}:G_{l}\rightarrow G_{l-1}$ is the natural group homomorphism. Using (1.2), we construct hierarchical DNNs where the neurons are organized in a tree-like structure having multiple layers. These DNNs can be trained using the standard back propagation method, see Lemma 2. These DNNs are robust universal approximators: given $f\in L^{\rho}\left(\mathbb{F}_{p}\left[\left[T\right]\right],dx\right)$, with $\left\Vert f\right\Vert_{\rho}<M$, $\rho\in\left[1,\infty\right]$, for any input $X\in\mathcal{D}^{L}\left(\mathbb{F}_{p}\left[\left[T\right]\right]\right)$, $L\geq1$, there exist $\sigma_{M},w\in\mathcal{D}^{L+\Delta}\left(\mathbb{F}_{p}\left[\left[T\right]\right]\times\mathbb{F}_{p}\left[\left[T\right]\right]\right),\theta\in\mathcal{D}^{L+\Delta}\left(\mathbb{F}_{p}\left[\left[T\right]\right]\right)$, where Δ is the depth of the network, such that the output of the network $Y\in\mathcal{D}^{L+\Delta}\left(\mathbb{F}_{p}\left[\left[T\right]\right]\right)$ satisfies $\left\Vert Y-f\right\Vert_{\rho}<\epsilon$; furthermore, this approximation remains valid for w and θ belonging to small balls, see Theorem 1. These result can be extended to approximations of a finite set of functions, and functions defined on arbitrary compact subsets from $\mathbb{F}_{p}\left(\left(T\right)\right)$; see Corollary 1 and Theorem 2. It is relevant to mention that the results mentioned are still valid if $w\left(b,a\right)=w\left(b-a\right)$ in (1.2); this case corresponds to a convolutional DNN.

A natural question is if the hierarchical DNNs can approximate non-hierarchical functions, i.e., functions from $L^{\rho}\left(\left[0,1\right],dt\right)$, with $\left\Vert f\right\Vert_{\rho}<M$, $\rho\in\left[1,\infty\right]$, where dt denotes the Lebesgue measure. It turns out that $L^{\rho}\left(\left[0,1\right],dt\right)\simeq L^{\rho}\left(\mathbb{F}_{p}\left[\left[T\right]\right],dx\right)$, where $\simeq$ denotes a linear surjective isometry; see Theorem 3. Therefore, the hierarchical DNNs are robust universal approximators of non-hierarchical functions, see Theorem 4.

In [1], the authors use Fourier-Walsh series to approximate hierarchical functions. The isometry $L^{2}\left(\left[0,1\right],dt\right)\simeq L^{2}\left(\mathbb{F}_{p}\left[\left[T\right]\right],dx\right)$ maps orthonormal bases of $L^{2}\left(\mathbb{F}_{p}\left[\left[T\right]\right],dx\right)$ into orthonormal bases of $L^{2}\left(\left[0,1\right],dt\right)$, cf. Theorem 5. By choosing $p=2$, and a suitable orthonormal basis in $L^{2}\left(\mathbb{F}_{p}\left[\left[T\right]\right],dx\right)$, we recover the classical Fourier-Walsh series. Then, the hierarchical functions used in [1] are elements from $L^{2}\left(\mathbb{F}_{p}\left[\left[T\right]\right],dx\right)$, and also, the notion hierarchy used there has a non-Archimedean origin.

Our approach is inspired by the classical result asserting that the perceptrons are universal approximators; see, e.g., [7, Chapter 9], [9]-[12], [16], [24], among many references. However, (1.2) is not a classical perceptron; see Remark 3. Furthermore, the input of our DNNs is a test function, i.e., the input has a hierarchical representation. We have implemented some p-adic NNs for image processing; see [30], [34]. In this application, the input image is transformed into a test function (a finite weighted tree), and at the end of the calculation, the output, a test function, is transformed into an image.

Finally, several computation models involving p-adic numbers have been studied. In [22], a model of computation over the p-adic numbers, for odd primes p, is defined following the approach of Blum, Shub, and Smale [6]. NNs whose states are p-adic numbers were studied in [3], and p-adic automaton in [4]. In [20], p-adic models for human brain activity were developed.

2. Preliminary results

We quickly review some essential aspects of non-Archimedean analysis required in the article and fix some notations to be used here.

2.1. **Non-Archimedean local fields.**

A non-Archimedean local field $\mathbb{K}$ of arbitrary characteristic is a locally compact topological field with respect to a non-discrete topology, which induced by a norm (or absolute value) $|\cdot|_{\mathbb{K}}$ satisfying

$$|x+y|_{\mathbb{K}} \leq \max\left\{|x|_{\mathbb{K}}, |y|_{\mathbb{K}}\right\} \text{ for } x, y \in \mathbb{K},$$

i.e., $|\cdot|_{\mathbb{K}}$ is a non-Archimedean norm. For an in-depth exposition, the reader may consult [28], see also [2], [21], [26], [27].

The ring of integers $\mathcal{O}_{\mathbb{K}}$ of $\mathbb{K}$ is

$$\mathcal{O}_{\mathbb{K}} = \left\{x \in \mathbb{K} \; ; \; |x|_{\mathbb{K}} \leq 1\right\}.$$

This is a valuation ring with a unique maximal ideal $P_{\mathbb{K}}$. In terms of the absolute value $|\cdot|_{\mathbb{K}}$, $P_{\mathbb{K}}$ can be described as

$$\mathcal{P}_{\mathbb{K}} = \left\{x \in \mathbb{K} \; ; \; |x|_{\mathbb{K}} < 1\right\}.$$

Let $\overline{\mathbb{K}} = \mathcal{O}_{\mathbb{K}}/\mathcal{P}_{\mathbb{K}}$ the residue field of $\mathbb{K}$. Thus $\overline{\mathbb{K}} = \mathbb{F}_q$, the finite field with q elements. Let $\wp$ be a fixed generator of $\mathcal{P}_{\mathbb{K}}$, $\wp$ is called a uniformizing parameter of $\mathbb{K}$, then $\mathcal{P}_{\mathbb{K}} = \wp\mathcal{O}_{\mathbb{K}}$. Furthermore, we assume that $|\wp|_{\mathbb{K}} = q^{-1}$. For $z \in \mathbb{K}$, $ord(z) \in \mathbb{Z} \cup \{+\infty\}$ denotes the valuation of z, and $|z|_{\mathbb{K}} = q^{-ord(z)}$. If $z \in \mathbb{K} \setminus \{0\}$, then $ac(z) = z\pi^{-ord(z)}$ denotes the angular component of z.

The natural map $\mathcal{O}_{\mathbb{K}} \rightarrow \mathcal{O}_{\mathbb{K}}/\mathcal{P}_{\mathbb{K}} \simeq \mathbb{F}_q$ is called the reduction mod $P_{\mathbb{K}}$. We fix $\mathfrak{S} \subset \mathcal{O}_{\mathbb{K}}$ a set of representatives of $\mathbb{F}_q$ in $\mathcal{O}_{\mathbb{K}}$, i.e., $\mathfrak{S}$ is mapped bijectively into $\mathbb{F}_q$ by the reduction mod $P_{\mathbb{K}}$. We assume that $0 \in \mathfrak{S}$. Any non-zero element x of $\mathbb{K}$ can be written as

$$x = \wp^{ord(x)} \sum_{i=0}^{\infty} x_i \wp^i, \; x_i \in \mathfrak{S}, \text{ and } x_0 \neq 0.$$

This series converges in the norm $|\cdot|_{\mathbb{K}}$.

Along this article we assume that $q = p$ a prime number. Then,

$$\mathfrak{S} = \{0, 1, \ldots, p-1\}.$$

Then, the field $\mathbb{K}$ is isomorphic to the field of p-adic numbers $\mathbb{Q}_p$, or to the field of formal Laurent series $\mathbb{F}_p((T))$ with coefficients in a finite field $\mathbb{F}_p$ with p elements.

The field of p-adic numbers $\mathbb{Q}_p$ is can be obtained as the completion of the field of rational numbers $\mathbb{Q}$ with respect to the p-adic norm $|\cdot|_p$, which is defined as

$$|x|_p = \begin{cases} 0 & \text{if} \quad x = 0 \\ \\ p^{-\gamma} & \text{if} \quad x = p^{\gamma}\frac{a}{b}, \end{cases}$$

where a and b are integers coprime with p. The integer $\gamma := ord(x)$, with $ord(0) := +\infty$, is called the p-adic order of x. $\mathbb{Q}_p$ is a field of characteristic zero, while its residue field $\mathbb{F}_p$ has characteristic $p > 0$. The prime p is a local parameter, and the set $\mathfrak{S} = \{0, 1, \ldots, p-1\} \subset \mathbb{Q}_p$ is not a subfield of $\mathbb{Q}_p$.

A series of the form $\sum_{k=k_0}^{\infty} a_k T^k$; $a_k \in \mathbb{F}_p$ is called a formal Laurent series. These series form a field $\mathbb{F}_p((T))$.The standard norm on $\mathbb{F}_p((T))$ is defined as

$$|x| = \begin{cases} 0 & \text{if} \quad x = 0 \\ p^{-k_0} & \text{if} \quad x = \sum_{k=k_0}^{\infty} a_k T^k,\ a_{k_0} \neq 0. \end{cases}$$

Notice that $|T| = p^{-1}$.

The normed vector space $(\mathbb{K}, |\cdot|_K)$ is a non-Archimedean vector space. The non-Archimedean property refers to the fact that $|x+y|_{\mathbb{K}} \leq \max\{|x|_{\mathbb{K}}, |y|_{\mathbb{K}}\}$, and the equality occurs when $|x|_K \neq |y|_{\mathbb{K}}$. As a topological space $\mathbb{K}$ is homeomorphic to a Cantor-like subset of the real line, see, e.g., [2], [27].

2.2. **The space of test functions.** For $r \in \mathbb{Z}$, denote by $B_r(a) = \{x \in \mathbb{K}; |x - a|_{\mathbb{K}} \leq p^r\}$ the ball of radius p^r with center at $a \in \mathbb{K}$, and take $B_r(0) := B_r$. The ball B_0 equals to $\mathcal{O}_{\mathbb{K}}$, the ring of integers of $\mathbb{K}$. The balls are both open and closed subsets in $\mathbb{K}$. In addition, two balls in $\mathbb{K}$ are either disjoint or one is contained in the other. A subset of $\mathbb{K}$ is compact if and only if it is closed and bounded in $\mathbb{K}$, see, e.g., [2, Section 1.8] or [27, Section 1.3]. The balls and spheres are compact subsets. Thus $(\mathbb{K}, |\cdot|_{\mathbb{K}})$ is a locally compact topological space.

Since $(\mathbb{K}, +)$ is a locally compact topological group, there exists a Haar measure dx, which is invariant under translations, i.e., $d(x+a) = dx$, [13]. If we normalize this measure by the condition $\int_{\mathcal{O}_{\mathbb{K}}} dx = 1$, then dx is unique.

Notation 1. *We will use $\Omega(p^{-r}|x-a|_{\mathbb{K}})$ to denote the characteristic function of the ball $B_r(a) = a + p^{-r}\mathcal{O}_{\mathbb{K}}$, where $\mathcal{O}_{\mathbb{K}}$ is the unit ball. For more general sets, we will use the notation 1_A for the characteristic function of set A.*

A real-valued function φ defined on $\mathbb{K}$ is called locally constant if for any $x \in \mathbb{K}$ there exist an integer $l(x) \in \mathbb{Z}$ such that

$$\varphi(x + x') = \varphi(x) \text{ for any } x' \in B_{l(x)}. \tag{2.1}$$

A function $\varphi : \mathbb{K} \to \mathbb{R}$ is called a Bruhat-Schwartz function (or a test function) if it is locally constant with compact support. Any test function can be represented as a linear combination, with real coefficients, of characteristic functions of balls. The $\mathbb{R}$-vector space of Bruhat-Schwartz functions is denoted by $\mathcal{D}(\mathbb{K})$. For $\varphi \in \mathcal{D}(\mathbb{K})$, the largest number $l = l(\varphi)$ satisfying (2.1) is called the exponent of local constancy (or the parameter of constancy) of φ. We denote by $\mathcal{D}(\mathcal{O}_{\mathbb{K}})$ the subspace of the test functions supported in the unit ball.

3. Non-Archimedean deep neural networks

We set $L^\rho(\mathcal{O}_{\mathbb{K}})$, $\rho \in [1, \infty]$, for the $\mathbb{R}$-vector space of functions $f : \mathcal{O}_{\mathbb{K}} \to \mathbb{R}$ satisfying

$$\|f\|_\rho = \begin{cases} \left(\int_{\mathcal{O}_{\mathbb{K}}} |f(x)|^\rho \, d^N x \right)^{\frac{1}{\rho}} < \infty, & \text{for } \rho \in [1, \infty) \\ \|f\|_\infty < \infty & \text{for } \rho = \infty, \end{cases}$$

where $\|f\|_{\infty}$ denotes the essential supremum of $|f(x)|$. The Hölder inequality and $\int_{\mathcal{O}_{\mathbb{K}}} d^N x = 1$ imply that

$$\|f\|_1 \leq \|f\|_{\rho}, \text{ for } 1 \leq \rho \leq \infty,$$

which in turn implies that

$$L^{\rho}(\mathcal{O}_{\mathbb{K}}) \hookrightarrow L^1(\mathcal{O}_{\mathbb{K}}),\ 1 \leq \rho \leq \infty, \tag{3.1}$$

where the arrow denotes a continuous embedding. Since $\mathcal{D}(\mathcal{O}_{\mathbb{K}}) \subset L^{\rho}(\mathcal{O}_{\mathbb{K}})$, for $1 \leq \rho \leq \infty$, and that $\mathcal{D}(\mathcal{O}_{\mathbb{K}})$ is dense in $L^{\rho}(\mathcal{O}_{\mathbb{K}})$, for $1 \leq \rho < \infty$, see [26, Chap. I, Proposition 1.3], $\mathcal{D}(\mathcal{O}_{\mathbb{K}})$ also dense in $L^{\infty}(\mathcal{O}_{\mathbb{K}})$.

We fix a sigmoidal function (or an activation function) $\sigma : \mathbb{R} \to \mathbb{R}$, which is a bounded, differentiable function. We assume that

$$\sigma : \mathbb{R} \to (-1, 1). \tag{3.2}$$

The calculations with neural networks required using scaled versions of this function: $\sigma_M := M\sigma : \mathbb{R} \to (-M, M)$, $M > 1$.

We set $w(x, y) \in L^{\infty}(\mathcal{O}_{\mathbb{K}} \times \mathcal{O}_{\mathbb{K}})$ and assume that

$$w(\cdot, y) \in \mathcal{D}(\mathcal{O}_{\mathbb{K}}),$$

where the index of local constancy of the function $w(\cdot, y)$ is independent of y. We also set $\theta(x) \in L^{\infty}(\mathcal{O}_{\mathbb{K}})$.

Definition 1. *Assuming that σ_M, w, θ are as above, and $X \in L^1(\mathcal{O}_{\mathbb{K}})$, the function*

$$Y(x; \sigma, w, \theta) = \sigma_M \left(\int_{\mathcal{O}_{\mathbb{K}}} w(x, y) X(y) dy + \theta(x) \right) \in L^{\infty}(\mathcal{O}_{\mathbb{K}}). \tag{3.3}$$

is called a continuos non-Archimedean deep neural network (DNN) with activation function σ_M, kernel w, and bias θ. The real-valued function $X(x) \in \mathbb{R}$ describes the state of the neuron $x \in \mathcal{O}_{\mathbb{K}}$. We denote such neural network as $DNN(\mathcal{O}_{\mathbb{K}}, \sigma_M, w, \theta, \infty)$.

The network has infinitely many neurons organized in a rooted tree.

Notice that

$$\left| \int_{\mathcal{O}_{\mathbb{K}}} w(x, y) X(y) dy + \theta(x) \right| \leq \left| \int_{\mathcal{O}_{\mathbb{K}}} w(x, y) X(y) dy \right| + |\theta(x)|$$
$$\leq \|w\|_{\infty} \|X\|_1 + \|\theta\|_{\infty}.$$

3.1. **The spaces $\mathcal{D}^l(\mathcal{O}_{\mathbb{K}})$.** For $l \geq 1$, we set $G_l(\mathcal{O}_{\mathbb{K}}) = G_l := \mathcal{O}_{\mathbb{K}} / \wp^l \mathcal{O}_{\mathbb{K}}$. We use the following system of representatives for the elements of G_l:

$$\boldsymbol{i} = i_0 + i_1 \wp + \ldots + i_{l-1} \wp^{l-1},$$

here the i_j are p-adic digits. The points of G_l are naturally organized in a finite rooted tree with l levels. We denote by $\Omega(p^l |x - \boldsymbol{i}|_{\mathbb{K}})$ the characteristic function of the ball $\boldsymbol{i} + \wp^l \mathcal{O}_{\mathbb{K}}$.

We set $\mathcal{D}^l(\mathcal{O}_{\mathbb{K}})$ to be the $\mathbb{R}$-vector space of all test functions of the form

$$\varphi(x) = \sum_{\boldsymbol{i} \in G_l} \varphi(\boldsymbol{i}) \, \Omega(p^l |x - \boldsymbol{i}|_{\mathbb{K}}), \ \ \varphi(\boldsymbol{i}) \in \mathbb{R}. \tag{3.4}$$

The function φ is supported on $\mathcal{O}_{\mathbb{K}}$ and has parameter of constancy l. The vector space $\mathcal{D}^l(\mathcal{O}_{\mathbb{K}})$ is spanned by the basis

$$\left\{\Omega\left(p^l\left|x-\boldsymbol{i}\right|_{\mathbb{K}}\right)\right\}_{\boldsymbol{i}\in G_l}. \tag{3.5}$$

The identification of $\varphi\in\mathcal{D}^l(\mathcal{O}_{\mathbb{K}})$ with the column vector $\left[\varphi\left(\boldsymbol{i}\right)\right]_{\boldsymbol{i}\in G_l}\in\mathbb{R}^{\#G_l}$, $\#G_l=p^{lN}$, gives rise to an isomorphism between $\mathcal{D}^l(\mathcal{O}_{\mathbb{K}})$ and $\mathbb{R}^{\#G_l}$ endowed with the norm

$$\left\|\left[\varphi\left(\boldsymbol{i}\right)\right]_{\boldsymbol{i}\in G_l}\right\|=\max_{\boldsymbol{i}\in G_l}\left|\varphi\left(\boldsymbol{i}\right)\right|=\left\|\varphi\right\|_{\infty}.$$

Furthermore,

$$\mathcal{D}^l(\mathcal{O}_{\mathbb{K}})\hookrightarrow\mathcal{D}^{l+1}(\mathcal{O}_{\mathbb{K}})\hookrightarrow\mathcal{D}(\mathcal{O}_{\mathbb{K}}), \tag{3.6}$$

where $\hookrightarrow$ denotes a continuous embedding, and

$$\mathcal{D}(\mathcal{O}_{\mathbb{K}})=\cup_l\mathcal{D}^l(\mathcal{O}_{\mathbb{K}}). \tag{3.7}$$

Remark 1. *We warn the reader that through the article, we use* $\left(\mathbb{R}^N,\left|\cdot\right|\right)$*, where* $\left|\cdot\right|$ *is the standard norm, and also* $\left(\mathbb{R}^N,\left\|\cdot\right\|\right)\simeq\left(\mathcal{D}^l(\mathcal{O}_{\mathbb{K}}),\left\|\cdot\right\|_{\infty}\right)$*.*

3.2. **Positive characteristic versus zero characteristic.** For $l\geq 2$, we denote by Λ_l the group homomorphism

$$\begin{array}{cccc}\Lambda_l: & G_l & \rightarrow & G_{l-1}\\ & & & \\ & a_0+a_1\wp+\ldots+a_{l-1}\wp^{l-1} & \rightarrow & a_0+a_1\wp+\ldots+a_{l-2}\wp^{l-2}.\end{array}$$

For a given $\boldsymbol{j}\in G_{l-1}$, the elements in $\boldsymbol{k}\in G_l$ such that $\Lambda_l\left(\boldsymbol{k}\right)=\boldsymbol{j}$ are called the liftings of $\boldsymbol{j}$ to G_l. In positive characteristic,

$$G_l=\left\{q(T)\in\mathbb{F}_p\left[T\right];q(T)=a_0+a_1T+\ldots+a_{l-1}T^{l-1}\right\}$$

is the additive group of polynomials of degree up most $l-1$. Then, G_{l-1} is an additive subgroup of G_l. In characteristic zero,

$$G_l=\left\{a_0+a_1p+\ldots+a_{l-1}p^{l-1};a_i\in\{0,\ldots,p-1\}\right\}.$$

The additive group G_{l-1} is not an additive subgroup of G_l.

4. Discrete non-Archimedean DNNs

Lemma 1. *Let* σ_M *be as before. Assume that* $X\in\mathcal{D}^{l-1}\left(\mathcal{O}_{\mathbb{K}}\right)$*,* $w(\cdot,y)\in\mathcal{D}^l\left(\mathcal{O}_{\mathbb{K}}\right)$*,* $\theta\in\mathcal{D}^l\left(\mathcal{O}_{\mathbb{K}}\right)$*, for some* $l\geq 2$*. Then, all the continuous non-Archimedean DNNs with parameters* σ_M,w,σ,θ *belong to* $\mathcal{D}^l(\mathcal{O}_{\mathbb{K}})$*, and*

$$Y\left(\boldsymbol{i}\right)=\sigma_M\left(\sum_{\boldsymbol{k}\in G_l}X\left(\Lambda_l\left(\boldsymbol{k}\right)\right)w\left(\boldsymbol{i},\boldsymbol{k}\right)+\theta\left(\boldsymbol{i}\right)\right),\text{ for }\boldsymbol{i}\in G_l. \tag{4.1}$$

Proof. The hypothesis $X\in\mathcal{D}^{l-1}\left(\mathcal{O}_{\mathbb{K}}\right)$, $w(\cdot,y)\in\mathcal{D}^l\left(\mathcal{O}_{\mathbb{K}}\right)$, implies that

$$X(y)=\sum_{\boldsymbol{j}\in G_{l-1}}X\left(\boldsymbol{j}\right)\Omega\left(p^{l-1}\left|y-\boldsymbol{j}\right|_{\mathbb{K}}\right),\ w(x,y)=\sum_{\boldsymbol{j}\in G_l}w\left(\boldsymbol{j},y\right)\Omega\left(p^l\left|x-\boldsymbol{j}\right|_p\right).$$

In addition,

$$\theta(x)=\sum_{\boldsymbol{j}\in G_l}\theta\left(\boldsymbol{j}\right)\Omega\left(p^l\left|x-\boldsymbol{j}\right|_p\right)$$

Now, $\Omega\left(p^{l-1}\left|y-\boldsymbol{j}\right|_{p}\right)$ is the characteristic function of the ball $\boldsymbol{j}+\wp^{l-1}\mathcal{O}_{\mathbb{K}}$, this ball is a disjoint union the balls

$$\boldsymbol{j}+\wp^{l-1}\mathcal{O}_{\mathbb{K}}=\bigsqcup_{\boldsymbol{k}\in T_{l}^{N}}\left(\boldsymbol{k}+\wp^{l}\mathcal{O}_{\mathbb{K}}\right),$$

where

$$T_{l}^{N}:=\left\{\boldsymbol{k}\in G_{l};\Lambda_{l}\left(\boldsymbol{k}\right)\in\boldsymbol{j}\right\}.$$

Now, by using that

$$\begin{aligned}X\left(\boldsymbol{j}\right)\Omega\left(p^{l-1}\left|y-\boldsymbol{j}\right|_{\mathbb{K}}\right)&=\sum_{\substack{\boldsymbol{k}\in G_{l}\\ \Lambda_{l}(\boldsymbol{k})=\boldsymbol{j}}}X\left(\boldsymbol{j}\right)\Omega\left(p^{l}\left|y-\boldsymbol{k}\right|_{\mathbb{K}}\right)\\&=\sum_{\substack{\boldsymbol{k}\in G_{l}\\ \Lambda_{l}(\boldsymbol{k})=\boldsymbol{j}}}X\left(\Lambda_{l}\left(\boldsymbol{k}\right)\right)\Omega\left(p^{l}\left|y-\boldsymbol{k}\right|_{\mathbb{K}}\right),\end{aligned}$$

$$X(y)=\sum_{\boldsymbol{j}\in G_{l-1}}X\left(\boldsymbol{j}\right)\Omega\left(p^{l-1}\left|y-\boldsymbol{j}\right|_{\mathbb{K}}\right)=\sum_{\boldsymbol{k}\in G_{l}}X\left(\Lambda_{l}\left(\boldsymbol{k}\right)\right)\Omega\left(p^{l}\left|y-\boldsymbol{k}\right|_{\mathbb{K}}\right),$$

one gets that

$$\begin{aligned}&\int_{\mathcal{O}_{\mathbb{K}}}w(x,y)X(y)dy\\&=\sum_{\boldsymbol{j}\in G_{l}}\left\{\sum_{\boldsymbol{k}\in G_{l}}X\left(\Lambda_{l}\left(\boldsymbol{k}\right)\right)\int_{\mathcal{O}_{\mathbb{K}}}w\left(\boldsymbol{j},y\right)\Omega\left(p^{l}\left|y-\boldsymbol{k}\right|_{\mathbb{K}}\right)d^{N}y\right\}\Omega\left(p^{l}\left|x-\boldsymbol{j}\right|_{\mathbb{K}}\right)\\&=\sum_{\boldsymbol{j}\in G_{l}}\left\{\sum_{\boldsymbol{k}\in G_{l}}X\left(\Lambda_{l}\left(\boldsymbol{k}\right)\right)w\left(\boldsymbol{j},\boldsymbol{k}\right)\right\}\Omega\left(p^{l}\left|x-\boldsymbol{j}\right|_{p}\right),\end{aligned}$$

where

$$w\left(\boldsymbol{j},\boldsymbol{k}\right):=\int_{\mathcal{O}_{\mathbb{K}}}w\left(\boldsymbol{j},y\right)\Omega\left(p^{l}\left|y-\boldsymbol{k}\right|_{\mathbb{K}}\right)dy,$$

and consequently,

$$\begin{aligned}&\sigma\left(\int_{\mathcal{O}_{\mathbb{K}}}w(x,y)X(y)d^{N}y+\theta\left(x\right)\right)\\&=\sum_{\boldsymbol{j}\in G_{l}}\sigma\left(\sum_{\boldsymbol{k}\in G_{l}}X\left(\Lambda_{l}\left(\boldsymbol{k}\right)\right)w\left(\boldsymbol{j},\boldsymbol{k}\right)+\theta\left(\boldsymbol{j}\right)\right)\Omega\left(p^{l}\left|x-\boldsymbol{j}\right|_{\mathbb{K}}\right).\end{aligned}$$

□

Definition 2. *Given σ_{M} an activation function, L,Δ positive integers, $\theta\in\mathcal{D}^{L+\Delta}\left(\mathcal{O}_{\mathbb{K}}\right)$, $w\in\mathcal{D}^{L+\Delta}\left(\mathcal{O}_{\mathbb{K}}\times\mathcal{O}_{\mathbb{K}}\right)$, we define the non-Archimedean discrete $DNN(\mathcal{O}_{\mathbb{K}},\sigma_{M},L,\Delta,w,\theta)$ as a NN with input*

$$X^{in}\left(\boldsymbol{j}\right)=X^{(L)}\left(\boldsymbol{j}\right),\ \textit{for}\ \boldsymbol{j}\in G_{l},$$

states

$$X^{(l)}(\boldsymbol{j}) = \sigma_M \left(\sum_{\boldsymbol{k} \in G_l} X^{(l-1)}(\Lambda_l(\boldsymbol{k}))\, w(\boldsymbol{j}, \boldsymbol{k}) + \theta(\boldsymbol{j}) \right) \tag{4.2}$$

for $l = L+1, \ldots, L+\Delta$, *and* $\boldsymbol{j} \in G_l$, *and output*

$$Y^{out}(\boldsymbol{j}) = X^{(L+\Delta)}(\boldsymbol{j}), \text{ for } \boldsymbol{j} \in G_{L+\Delta}.$$

We say that $DNN(\mathcal{O}_{\mathbb{K}}, \sigma_M, L, \Delta, w, \theta)$ *is a convolutional network, if the states are determined by*

$$X^{(l)}(\boldsymbol{j}) = \sigma_M \left(\sum_{\boldsymbol{k} \in G_l} X^{(l-1)}(\Lambda_l(\boldsymbol{k}))\, w(\boldsymbol{j} - \boldsymbol{k}) + \theta(\boldsymbol{j}) \right)$$

for $l = L+1, \ldots, L+\Delta$.

The input and the output are test functions, more precisely, $X^{in} \in \mathcal{D}^L(\mathcal{O}_{\mathbb{K}})$ and $Y^{out} \in \mathcal{D}^{L+\Delta}(\mathcal{O}_{\mathbb{K}})$. The equation (4.2 establishes that the weights and biases of layer l and the state of layer $l-1$ determine the state of the network at layer l. The functions $X^{(l)}$,.., $X^{(l-1)}$ are different. The network neurons are organized in a tree-like structure with Δ levels. The parameter Δ gives the depth of the network. At the level l, with $L+1 \leq l \leq L+\Delta-1$, there are p^l neurons. The neuron $\boldsymbol{j} \in G_l$ is connected with the neuron $\boldsymbol{k} \in G_{l+1}$ if there exist $\boldsymbol{s} \in G_{l+1}$ such that $\Lambda_{l+1}(\boldsymbol{s}) = \boldsymbol{j}$ and $w(\boldsymbol{j}, \boldsymbol{s}) \neq 0$. Since the cardinality of G_l is p^L, the length of the input should be less than or equal p^L. We pick the smallest possible prime p for practical purposes, and then L is determined by the input size.

5. Implementation of discrete non-Archimedean deep neural networks

5.1. **Matrix form of the DNNs.** From now on, for the sake of simplicity, we use matrix notation. We set

$$\begin{aligned} X^{[l]} &:= [X(\boldsymbol{k})]_{\boldsymbol{k} \in G_l} = \left[X^{[l]}_{\boldsymbol{k}} \right]_{\boldsymbol{k} \in G_l} \in \mathbb{R}^{p^l}, \\ \underline{X}^{[l-1]} &:= \left[X^{(l-1)}(\Lambda_l(\boldsymbol{k})) \right]_{\boldsymbol{k} \in G_l} = \left[\underline{X}^{[l-1]}_{\boldsymbol{k}} \right]_{\boldsymbol{k} \in G_l} \in \mathbb{R}^{p^l}, \\ \theta^{[l]} &:= [\theta(\boldsymbol{k})]_{\boldsymbol{k} \in G_l} = \left[\theta^{[l]}_{\boldsymbol{k}} \right]_{\boldsymbol{k} \in G_l} \in \mathbb{R}^{p^l}, \end{aligned}$$

for $l = L+1, \ldots, L+\Delta$, and the matrix

$$W^{[l]} := [w(\boldsymbol{i}, \boldsymbol{k})]_{\boldsymbol{i}, \boldsymbol{k} \in G_l} = \left[w^{[l]}_{\boldsymbol{i}, \boldsymbol{k}} \right]_{\boldsymbol{i}, \boldsymbol{k} \in G_l}.$$

Notice that $X^{[l-1]} \neq \underline{X}^{[l-1]}$. This is an important difference with the standard case. $\underline{X}^{[l-1]}$ denotes a state at layer l obtained by copying a state $X^{[l-1]}$ at layer $l-1$.

With the above notation, a non-Archimedean DNN can be rewritten as

$$\begin{cases} X^{[L]} \text{ input} \\ X^{[L+\Delta]} \text{ output} \\ Z^{[l]} := W^{[l]}\underline{X}^{[l-1]} + \theta^{[l]} \\ X^{[l]} = \sigma_M\left(Z^{[l]}\right),\ l = L+1,\ldots,L+\Delta, \end{cases} \tag{5.1}$$

where

$$\left(W^{[l]}\underline{X}^{[l-1]} + \theta^{[l]}\right)_{\boldsymbol{j}} = \sigma_M\left(\sum_{\boldsymbol{k}\in G_l} \underline{X}^{[l-1]}_{\boldsymbol{k}} w^{[l]}_{\boldsymbol{j},\boldsymbol{k}} + \theta^{[l]}_{\boldsymbol{j}}\right), \text{ for } \boldsymbol{j} \in G_l,$$

see (6.8).

5.2. **Stochastic gradient.**

Notation 2. *Given two vectors* $A = [A_{\boldsymbol{k}}]_{\boldsymbol{k}\in G_l}$, $B = [B_{\boldsymbol{k}}]_{\boldsymbol{k}\in G_l} \in \mathbb{R}^{p^l}$, *we set*

$$|A - B| = \sqrt{\sum_{\boldsymbol{k}\in G_l} (A_{\boldsymbol{k}} - B_{\boldsymbol{k}})^2}.$$

Assume that $\left\{A^{\{i\}}\right\}_{i=1}^{K}$ are training data for which there are given target outputs $\left\{Y\left(A^{\{i\}}\right)\right\}_{i=1}^{K}$. We use the quadratic cost function

$$C = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{2}\left|Y\left(A^{\{i\}}\right) - X^{[L+\Delta]}\left(A^{\{i\}}\right)\right|^2 = \frac{1}{K}\sum_{i=1}^{K} C_{A^{\{i\}}},$$

to determine the parameters of the network, where $X^{[L+\Delta]}\left(A^{\{i\}}\right)$ denotes the output of the network with input $A^{\{i\}}$, and

$$C_{A^{\{i\}}} = \left|Y\left(A^{\{i\}}\right) - X^{[L+\Delta]}\left(A^{\{i\}}\right)\right|^2.$$

We denote by $\Theta = \left(W^{[L+\Delta]}, \theta^{[L+\Delta]}\right) \in \mathbb{R}^s$, with $s := p^{2(L+\Delta)} + p^{(L+\Delta)}$, the vector constructed with the weights $W^{[L+\Delta]}$ and the biases $\theta^{[L+\Delta]}$ of the network, thus, $C = C(\Theta) : \mathbb{R}^s \to \mathbb{R}$. The minimization of the cost function is obtained by using the gradient descent method:

$$\Theta \to \Theta - \eta\nabla C(\Theta) = \Theta - \frac{\eta}{K}\sum_{i=1}^{K}\nabla C_{A^{\{i\}}}(\Theta), \tag{5.2}$$

where η is the learning constant and $\nabla C(\Theta)$ denotes the gradient of the function $C(\Theta)$. A practical implementation of the updating method (5.2) is obtained by using the stochastic gradient; see, e.g., [15], and the references therein.

5.3. **Back propagation.** We fix a training input $A^{\{i\}}$ and consider $C_{A^{\{i\}}}(\Theta)$ as a function of Θ, and use the notation

$$C_{A^{\{i\}}}(\Theta) := C = \frac{1}{2}\left|Y - X^{[L+\Delta]}\right|^2. \tag{5.3}$$

We set $Z^{[l]} := \left[Z_{\boldsymbol{k}}^{[l]}\right]_{\boldsymbol{k}\in G_l}$, see (5.1). We refer $Z_{\boldsymbol{k}}^{[l]}$ as the weighted input for neuron $\boldsymbol{k} \in G_l$, at layer l. We also set

$$\delta_{\boldsymbol{k}}^{[l]} = \frac{\partial C}{\partial Z_{\boldsymbol{k}}^{[l]}} \text{ for } \boldsymbol{k} \in G_l \text{ and } l = L+1, \ldots, L+\Delta,$$

and $\delta^{[l]} := \left[\delta_{\boldsymbol{k}}^{[l]}\right]_{\boldsymbol{k}\in G_l}$.

Given two vectors $A = [A_{\boldsymbol{k}}]_{\boldsymbol{k}\in G_l}$, $B = [B_{\boldsymbol{k}}]_{\boldsymbol{k}\in G_l} \in \mathbb{R}^{p^{lN}}$, the Hadamard product of them is defined as

$$A \circ B := [(A \circ B)_{\boldsymbol{k}}]_{\boldsymbol{k}\in G_l} = [A_{\boldsymbol{k}} B_{\boldsymbol{k}}]_{\boldsymbol{k}\in G_l}.$$

We also use the notation $\sigma_M\left(Z^{[l]}\right) := \left[\sigma_M\left(Z_{\boldsymbol{k}}^{[l]}\right)\right]_{\boldsymbol{k}\in G_l}$, and

$$\sigma_M'\left(Z^{[l]}\right) = \frac{d}{dt}\sigma_M\left(Z^{[l]}\right) = \left[\frac{d}{dt}\sigma_M\left(Z_{\boldsymbol{k}}^{[l]}\right)\right]_{\boldsymbol{k}\in G_l}.$$

Lemma 2. *With the above notation the following formulae hold:*
(i) $\delta^{[L+\Delta]} = \sigma_M'\left(Z^{[l]}\right) \circ \left(X^{[L+\Delta]} - Y\right)$;
(ii) $\delta^{[l]} = \sigma_M'\left(Z^{[l]}\right) \circ \left(W^{[l+1]}\right)^T \delta^{[l+1]}$ *for* $l = L+1, \ldots, L+\Delta-1$;
(iii) $\frac{\partial C}{\partial \theta_{\boldsymbol{k}}^{[l]}} = \delta_{\boldsymbol{k}}^{[l]}$ *for* $l = L+1, \ldots, L+\Delta$;
(iv) $\frac{\partial C}{\partial w_{\boldsymbol{k},\boldsymbol{j}}^{[l]}} = \delta_{\boldsymbol{k}}^{[l]} \underline{X}_{\boldsymbol{j}}^{[l-1]}$ *for* $l = L+1, \ldots, L+\Delta$.

Proof. The proof is a simple variation of the classical one. See, for instance, the proof of Lemma 5.1 in [15]. □

The output $X^{[L+\Delta]}$ can be computed from the input, $X^{[L]}$, using $Z^{[l]} = W^{[l]}\underline{X}^{[l-1]} + \theta^{[l]}$, $X^{[l]}$, $l = L+1, \ldots, L+\Delta$, by computing $X^{[L]}$, $Z^{[L+1]}$, ..., $Z^{[L+\Delta-1]}$, $X^{[L+\Delta]}$ in order. After this step, $\delta^{[L+\Delta]}$ can be computed, see Lemma 2-(i). Then, by Lemma 2-(ii), $\delta^{[L+\Delta-1]}$, $\delta^{[L+\Delta-2]}$, ..., $\delta^{[L+1]}$ can be computed in a backward pass. Now from Lemma 2-(iii),(iv), we can compute the partial derivatives, and thus the gradients. This a non-Archimedean version of the standard back propagation algorithm.

6. The discrete non-Archimedean DNNs are robust universal approximators

Theorem 1. *Given any* $f \in L^\rho(\mathcal{O}_{\mathbb{K}})$, $1 \le \rho \le \infty$, *with* $\|f\|_\rho < M$, *and any* $\epsilon > 0$; *then, for any input* $X \in \mathcal{D}^L(\mathcal{O}_{\mathbb{K}})$, $L \ge 1$, *there exists a p-adic* $DNN(\mathcal{O}_{\mathbb{K}}, \sigma_M, L, \Delta, w, \theta)$ *such that the output*

$$Y(x; \mathcal{O}_{\mathbb{K}}, \sigma_M, L, \Delta, w, \theta) = \sum_{\boldsymbol{k}\in G_{L+\Delta}} Y(\boldsymbol{k})\,\Omega\left(p^{L+\Delta}|x - \boldsymbol{k}|_{\mathbb{K}}\right)$$

satisfies

$$\|Y(\mathcal{O}_{\mathbb{K}}, \sigma_M, L, \Delta, w, \theta) - f\|_\rho < \epsilon,$$

for any w in a ball in $\mathcal{D}^{L+\Delta}\left(\mathcal{O}_{\mathbb{K}}\times\mathcal{O}_{\mathbb{K}}\right)\approx\mathbb{R}^{2p^{(L+\Delta)}}$, and any θ in a ball in $\mathcal{D}^{L+\Delta}\left(\mathcal{O}_{\mathbb{K}}\right)\approx\mathbb{R}^{p^{(L+\Delta)}}$. Furthermore, this approximation property remains valid if we assume that the output is produced by a convolutional network.

Proof. First, we may assume that $f\in\mathcal{D}\left(\mathcal{O}_{\mathbb{K}}\right)$, with $\left\Vert f\right\Vert_{\rho}<M$. Indeed, since $\mathcal{D}\left(\mathcal{O}_{\mathbb{K}}\right)$ is dense in $L^{\rho}(\mathcal{O}_{\mathbb{K}})$, there exists $\varphi\in\mathcal{D}^{L^{'}}\left(\mathcal{O}_{\mathbb{K}}\right)$ such that $\left\Vert\varphi-f\right\Vert_{\rho}<\frac{\epsilon}{2}$. Since $\mathcal{D}^{L^{'}}\left(\mathcal{O}_{\mathbb{K}}\right)\hookrightarrow\mathcal{D}^{L^{''}}\left(\mathcal{O}_{\mathbb{K}}\right)$ for $L''>L'$, without loss of generality, we may suppose that $L^{'}>L$ so $L^{'}=L+\Delta$, for some positive integer Δ. Then, it is sufficient to construct an DNN with output $Y\in\mathcal{D}^{L+\Delta}\left(\mathcal{O}_{\mathbb{K}}\right)$ such that

$$\left\Vert\varphi-Y\right\Vert_{\infty}<\frac{\epsilon}{2}, \tag{6.1}$$

since $\left\Vert\varphi-Y\right\Vert_{\rho}\leq\left\Vert\varphi-Y\right\Vert_{\infty}$. So, we may replace f by φ, with $\left\Vert\varphi\right\Vert_{\rho}\leq\left\Vert\varphi\right\Vert_{\infty}<M$. Taking

$$\varphi\left(x\right)=\sum_{\boldsymbol{i}\in G_{L+\Delta}}\varphi\left(\boldsymbol{i}\right)\Omega\left(p^{L+\Delta}\left\vert x-\boldsymbol{i}\right\vert_{\mathbb{K}}\right),$$

we look for an DNN with output $Y\in\mathcal{D}^{L+\Delta}\left(\mathcal{O}_{\mathbb{K}}\right)$ such that

$$\max_{\boldsymbol{i}\in G_{L+\Delta}}\left\vert Y\left(\boldsymbol{i}\right)-\varphi\left(\boldsymbol{i}\right)\right\vert<\epsilon, \tag{6.2}$$

with $\varphi\left(\boldsymbol{i}\right)\in\left(-M,M\right)$. By using the activation function σ_{M}, for each $\boldsymbol{i}\in G_{L+\Delta}^{N}$, we fix a preimage $\sigma_{M}^{-1}\left(\varphi\left(\boldsymbol{i}\right)\right)$, now by the continuity of σ_{M}, there exists $\delta\left(\epsilon\right)$ such that

$$\max_{\boldsymbol{i}\in G_{L+\Delta}}\left\vert\sum_{\boldsymbol{k}\in G_{L+\Delta}^{N}}X\left(\Lambda_{L+\Delta}\left(\boldsymbol{k}\right)\right)w\left(\boldsymbol{i},\boldsymbol{k}\right)+\theta\left(\boldsymbol{i}\right)-\sigma_{M}^{-1}\left(\varphi\left(\boldsymbol{i}\right)\right)\right\vert<\delta\left(\epsilon\right)$$

implies (6.2). By taking $\left\vert\theta\left(\boldsymbol{i}\right)-\sigma_{M}^{-1}\left(\varphi\left(\boldsymbol{i}\right)\right)\right\vert<\frac{\delta(\epsilon)}{2}$, for all $\boldsymbol{i}\in G_{L+\Delta}$, i.e.,

$$\max_{\boldsymbol{i}\in G_{L+\Delta}}\left\vert\theta\left(\boldsymbol{i}\right)-\sigma_{M}^{-1}\left(\varphi\left(\boldsymbol{i}\right)\right)\right\vert<\frac{\delta\left(\epsilon\right)}{2}, \tag{6.3}$$

it is sufficient to show that

$$\max_{\boldsymbol{i}\in G_{L+\Delta}}\left\vert\sum_{\boldsymbol{k}\in G_{L+\Delta}}X\left(\Lambda_{L+\Delta}\left(\boldsymbol{k}\right)\right)w\left(\boldsymbol{i},\boldsymbol{k}\right)\right\vert<\frac{\delta\left(\epsilon\right)}{2}. \tag{6.4}$$

For a fix $\boldsymbol{i}\in G_{L+\Delta}^{N}$, the linear mapping

$$\begin{array}{ccc}\mathbb{R}^{p^{(L+\Delta)}} & \rightarrow & \mathbb{R}\\ w\left(\boldsymbol{i},\boldsymbol{k}\right) & \rightarrow & \sum_{\boldsymbol{k}\in G_{L+\Delta}}X\left(\Lambda_{L+\Delta}\left(\boldsymbol{k}\right)\right)w\left(\boldsymbol{i},\boldsymbol{k}\right)\end{array}$$

is continuous at the origin, even if $X\left(\Lambda_{L+\Delta}\left(\boldsymbol{k}\right)\right)=0$ for all $\boldsymbol{k}\in G_{L+\Delta}$, then there exists $\gamma\left(\boldsymbol{i},\delta\right)$ such that

$$\max_{\boldsymbol{k}\in G_{L+\Delta}}\left\vert w\left(\boldsymbol{i},\boldsymbol{k}\right)\right\vert<\gamma\left(\boldsymbol{i},\delta\right)\text{ implies }\left\vert\sum_{\boldsymbol{k}\in G_{L+\Delta}}X\left(\Lambda_{L+\Delta}\left(\boldsymbol{k}\right)\right)w\left(\boldsymbol{i},\boldsymbol{k}\right)\right\vert<\frac{\delta\left(\epsilon\right)}{2}.$$

Consequently,

$$\max_{\boldsymbol{i}\in G_{L+\Delta}}\max_{\boldsymbol{k}\in G_{L+\Delta}}\left|w\left(\boldsymbol{i},\boldsymbol{k}\right)\right|<\min_{\boldsymbol{i}\in G_{L+\Delta}}\gamma\left(\boldsymbol{i},\delta\right) \tag{6.5}$$

implies (6.4).

By the isomorphism of Banach spaces $\left(\mathcal{D}^{L+\Delta}\left(\mathcal{O}_{\mathbb{K}}\right),\|\cdot\|_{\infty}\right)\simeq\left(\mathbb{R}^{p^{L+\Delta}},\|\cdot\|_{\infty}\right)$, condition (6.3) defines a ball in $\mathbb{R}^{p^{(L+\Delta)}}$, while condition (6.5) defines defines a ball on $\mathbb{R}^{2p^{(L+\Delta)}}$, where the approximation (6.2) is valid. Finally, we observe that the above reasoning is valid if we take $w\left(\boldsymbol{i},\boldsymbol{k}\right)=w\left(\boldsymbol{i}-\boldsymbol{k}\right)$.

The prime number p and the integer L are determined by the input $X\in\mathcal{D}^{L}\left(\mathcal{O}_{\mathbb{K}}\right)\simeq\mathbb{R}^{p^{L}}$. Then, p^{L} should be larger than the length of the vector X. Consequently, we may pick p to be any arbitrary prime. □

A good approximation of a function $f\in L^{\rho}(\mathcal{O}_{\mathbb{K}})$ by a test function $\varphi\in\mathcal{D}^{L^{'}}\left(\mathcal{O}_{\mathbb{K}}\right)$ requires a large $L^{'}$, and consequently the computational power of the DNNs increase with Δ. The p-adic discrete DNNs are particular cases of the standard deep neural networks; see, e.g.,[15]. The DNNs can be trained using back propagation, and the convolutional implementation of these networks reduces the number of weights significantly.

Remark 2. *The field $\mathbb{K}$ is invariant under transformations of the form $x\rightarrow b+\wp^{s}x$, with $b\in\mathbb{K}$, $s\in\mathbb{Z}$. For this reason, we will consider that all networks $DNN(\mathcal{O}_{\mathbb{K}},L,\Delta,w,\theta)$ with outputs of the form $Y(b+\wp^{s}x)$ are equivalent. In practical terms, to find an approximation of a function $f:b+\wp^{s}\mathcal{O}_{\mathbb{K}}\rightarrow\mathbb{R}$ is the same as find an approximation of the function $f^{\ast}:\mathcal{O}_{\mathbb{K}}\rightarrow\mathbb{R}$, where $f^{\ast}\left(x\right)=f\left(b+\wp^{s}x\right)$.*

Remark 3. *The natural non-Archimedean version of the perceptron is*

$$G\left(\left[X_{\boldsymbol{k}}\right]_{\boldsymbol{k}\in G_{l}}\right)=\sigma\left(\sum_{\boldsymbol{k}\in G_{l}}W_{\boldsymbol{k}}X_{\boldsymbol{k}}+b\right), \tag{6.6}$$

where $X=\left[X_{\boldsymbol{k}}\right]_{\boldsymbol{k}\in G_{l}},W=\left[W_{\boldsymbol{k}}\right]_{\boldsymbol{k}\in G_{l}}\in\mathbb{R}^{p^{l}}$, and $b\in\mathbb{R}$. Here, the activation function σ is defined as in [7, Definition 2.2.1]. *By using the standard inner product of $\mathbb{R}^{p^{l}}$, we rewrite (6.6) as $G(X)=\sigma\left(W\cdot X+b\right)$. However, (6.6) is radically different from the classical perceptron $\sigma\left(w\cdot x+b\right)$, where $w=\left(w_{1},\ldots,w_{M}\right)$, $x=\left(x_{1},\ldots,x_{M}\right),\in\mathbb{R}^{M}$, because i, x_{i} are real numbers, while $X_{\boldsymbol{k}}$ is a real number, but $\boldsymbol{k}\in\mathbb{K}$. Cybenko's theorem, see* [9], [7, Theorem 9.3.6], *asserts that the finite sums of the form*

$$G(X)=\sum_{r=1}^{R}\alpha_{r}\sigma\left(W_{r}\cdot X+b_{r}\right)$$

are dense in the space of continuos functions defined on $\left[0,1\right]^{M}$ endowed with the norm $\|\cdot\|_{\infty}$. Besides of the apparently similitude, the non-Archimedean DNNs considered here are not non-Archimedean analogues of the classical perceptrons.

6.1. **Parallelization of non-Archimedean DNNs.** A natural problem is the simultaneous approximations of several functions by DNNs. More precisely, given $f_{i}\in L^{\rho}(\mathcal{O}_{\mathbb{K}})$, $i=1,\ldots,R$, by applying Theorem 1, there are DNNs,

$$DNN(\mathcal{O}_{\mathbb{K}},\sigma_{M_{i}},L_{i},\Delta_{i},w_{i},\theta_{i}),\ \ i=1,\ldots,R,$$

whose outputs satisfy

$$\left\| Y_i\left(\mathcal{O}_{\mathbb{K}}, \sigma_{M_i}, L_i, \Delta_i, w_i, \theta_i\right) - f_i \right\|_\rho < \epsilon, , i = 1, \ldots, R.$$

Without loss of generality, we may assume that the parameter p is the same for all the DNNs. For general f_i, $i = 1, \ldots, R$, it is not possible to find a unique DNN that approximates all the f_i, but we can use the $DNN(\mathcal{O}_{\mathbb{K}}, \sigma_{M_i}, L_i, \Delta_i, w_i, \theta_i)$, $i = 1, \ldots, R$, in parallel to construct a simultaneous approximation of all the f_i.

We set

$$\boldsymbol{f} = (f_1, \ldots, f_R) \in L^\rho(\mathcal{O}_{\mathbb{K}}) \bigoplus \cdots \bigoplus L^\rho(\mathcal{O}_{\mathbb{K}}) = \left(L^\rho(\mathcal{O}_{\mathbb{K}})\right)^R,$$

$$\boldsymbol{X} = (X_1, \ldots, X_R) \in \bigoplus_{i=1}^R \mathcal{D}^{L_i}\left(\mathcal{O}_{\mathbb{K}}\right), \boldsymbol{Y} = (Y_1, \ldots, Y_R) \in \bigoplus_{i=1}^R \mathcal{D}^{L_i+\Delta_i}\left(\mathcal{O}_{\mathbb{K}}\right),$$

$$\boldsymbol{L} = (L_1, \ldots, L_R), \boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_R) \in \mathbb{N}^R,$$

$$\boldsymbol{w} = (w_1, \ldots, w_R) \in \bigoplus_{i=1}^R \mathcal{D}^{L_i+\Delta_i}\left(\mathcal{O}_{\mathbb{K}} \times \mathcal{O}_{\mathbb{K}}\right), \boldsymbol{\theta} = (\theta_1, \ldots, \theta_R) \in \bigoplus_{i=1}^R \mathcal{D}^{L_i+\Delta_i}\left(\mathcal{O}_{\mathbb{K}}\right),$$

$$\boldsymbol{\sigma} = \left(\sigma_{M_1}, \ldots, \sigma_{M_R}\right).$$

Given $\boldsymbol{f} \in \left(L^\rho(\mathcal{O}_{\mathbb{K}})\right)^R$, we define the norm

$$\|\boldsymbol{f}\| := \max_{1 \leq i \leq R} \|f_i\|_\rho,$$

and for $\boldsymbol{X} \in \bigoplus_{i=1}^R \mathcal{D}^{L_i}\left(\mathcal{O}_{\mathbb{K}}\right)$,

$$\|\boldsymbol{X}\| = \max_{1 \leq i \leq R} \|X_i\|_\infty.$$

Definition 3. *The direct product of the DNNs $DNN(\mathcal{O}_{\mathbb{K}}, \sigma_{M_i}, L_i, \Delta_i, w_i, \theta_i)$, $i = 1, \ldots, R$, is a DNN, denoted as $DNN(\mathcal{O}_{\mathbb{K}}, \boldsymbol{\sigma}, R, \boldsymbol{L}, \boldsymbol{\Delta}, \boldsymbol{w}, \boldsymbol{\theta})$, with input $\boldsymbol{X} \in \bigoplus_{i=1}^R \mathcal{D}^{L_i}\left(\mathcal{O}_{\mathbb{K}}\right)$, and output $\boldsymbol{Y} \in \bigoplus_{i=1}^R \mathcal{D}^{L_i+\Delta_i}\left(\mathcal{O}_{\mathbb{K}}\right)$.*

By using the above definition and Theorem 1, we obtain the following result.

Corollary 1. *Given any $\boldsymbol{f} \in \left(L^\rho(\mathcal{O}_{\mathbb{K}})\right)^R$, with $\|\boldsymbol{f}\|_\rho < M$, and any $\epsilon > 0$; then, for any input $\boldsymbol{X} \in \bigoplus_{i=1}^R \mathcal{D}^{L_i}\left(\mathcal{O}_{\mathbb{K}}\right)$, with $L_i \geq 1$ for all i,, there exists a p-adic $DNN(\mathcal{O}_{\mathbb{K}}, R, \boldsymbol{L}, \boldsymbol{\Delta}, \boldsymbol{w}, \boldsymbol{\theta})$ such that the output $\boldsymbol{Y}(x; p, \boldsymbol{\sigma}, R, \boldsymbol{L}, \boldsymbol{\Delta}, \boldsymbol{w}, \boldsymbol{\theta}) \in \bigoplus_{i=1}^R \mathcal{D}^{L_i+\Delta_i}\left(\mathcal{O}_{\mathbb{K}}\right)$ satisfies $\left\|\boldsymbol{Y}(x; \mathcal{O}_{\mathbb{K}}, \boldsymbol{\sigma}, R, \boldsymbol{L}, \boldsymbol{\Delta}, \boldsymbol{w}, \boldsymbol{\theta}) - \boldsymbol{f}\right\|_\rho < \epsilon$, for any $\boldsymbol{w}$ in a ball in $\bigoplus_{i=1}^R \mathcal{D}^{L_i+\Delta_i}\left(\mathcal{O}_{\mathbb{K}} \times \mathcal{O}_{\mathbb{K}}\right)$, and any $\boldsymbol{\theta}$ in a ball in $\bigoplus_{i=1}^R \mathcal{D}^{L_i+\Delta_i}\left(\mathcal{O}_{\mathbb{K}}\right)$.*

6.2. **Approximation of functions on open compact subsets.** We now extend Theorem 1 to functions supported in an open compact subset $\mathcal{K}$. A such set is a disjoint union of ball, the, without loss of generality, we assume that

$$\mathcal{K} = \bigsqcup_{i=1}^R \left(a_i + \wp^{N_i}\mathcal{O}_{\mathbb{K}}\right),$$

where $a_i \in \mathbb{K}$, $N_i \in \mathbb{Z}$, for $i = 1, \ldots, R$. Any function $f : \mathcal{K} \to \mathbb{R}$ is completely determined by its restrictions to the balls $B_{-N_i}(a_i) = a_i + \wp^{N_i}\mathcal{O}_{\mathbb{K}}$. More precisely,

$$f(x) = \sum_{i=1}^{R} f_i(x), \text{ with } f_i(x) = f(x)\Omega\left(p^{N_i}\left|x - a_i\right|_{\mathbb{K}}\right). \tag{6.7}$$

We now define the linear transformation

$$\begin{array}{cccc} T_{a_i,N_i}: & a_i + \wp^{N_i}\mathcal{O}_{\mathbb{K}} & \to & \mathcal{O}_{\mathbb{K}} \\ & & & \\ & x & \to & T_{a_i,N_i}(x) = \wp^{-N_i}(x - a_i). \end{array}$$

Then, the mapping

$$\begin{array}{cccc} T^*_{a_i,N_i}: & L^\rho\left(a_i + \wp^{N_i}\mathcal{O}_{\mathbb{K}}\right) & \to & L^\rho\left(\mathcal{O}_{\mathbb{K}}\right) \\ & & & \\ & f & \to & T^*_{a_i,N_i}f, \end{array}$$

with $T^*_{a_i,N_i}f(x) := f\left(T^{-1}_{a_i,N_i}(x)\right)$ defines a linear bounded operator satisfying

$$\left\|T^*_{a_i,N_i}f\right\|_\rho = \begin{cases} p^{\frac{N_i}{\rho}}\left\|f\right\|_\rho & \text{if} \quad \rho \in [1,\infty) \\ & \\ \left\|f\right\|_\rho & \text{if} \quad \rho = \infty. \end{cases}.$$

Consequently, we have a continuous embedding

$$L^\rho\left(a_i + \wp^{N_i}\mathcal{O}_{\mathbb{K}}\right) \hookrightarrow L^\rho\left(\mathcal{O}_{\mathbb{K}}\right),\ 1 \le \rho \le \infty,$$

Since $\mathcal{K}$ has finite Haar measure, $L^\rho(\mathcal{K}) \hookrightarrow L^1(\mathcal{K})$, $1 \le \rho \le \infty$, and $\mathcal{D}(\mathcal{K})$, the space of test functions supported in $\mathcal{K}$, is dense in $L^\rho(\mathcal{K})$, $1 \le \rho \le \infty$.

By (6.7),

$$L^\rho(\mathcal{K}) = \bigoplus_{i=1}^{R} L^\rho(a_i + \wp^{N_i}\mathcal{O}_{\mathbb{K}}).$$

When identifying $f \in L^\rho(\mathcal{K})$ with $(f_1, \ldots, f_R)$, $f_i \in L^\rho(a_i + \wp^{N_i}\mathcal{O}_{\mathbb{K}})$, we use the notation $\boldsymbol{f} = (f_1, \ldots, f_R)$. Notice that

$$\left\|\boldsymbol{f}\right\|_\rho = \left(\sum_{i=1}^{R} \int\limits_{a_i+\wp^{N_i}\mathcal{O}_{\mathbb{K}}} \left|f_i(x)\right|^\rho dx\right)^{\frac{1}{\rho}},\ \rho \in [1,\infty),$$

$$\left\|\boldsymbol{f}\right\|_\infty = \max_{1\le i\le R}\left\{\sup_{x\in a_i+\wp^{N_i}\mathcal{O}_{\mathbb{K}}}\left|f_i(x)\right|\right\}.$$

Theorem 2. *Given $\boldsymbol{f} \in L^\rho(\mathcal{K})$, $1 \le \rho \le \infty$, with $\left\|\boldsymbol{f}\right\|_\rho < M$, and $\epsilon > 0$; then, for any input*

$$\boldsymbol{X} = (X_1, \ldots, X_R) \in \bigoplus_{i=1}^{R} \mathcal{D}^{L_i}\left(a_i + \wp^{N_i}\mathcal{O}_{\mathbb{K}}\right),$$

there exits $DNN(\mathcal{O}_{\mathbb{K}}, \boldsymbol{\sigma}, R, \boldsymbol{L}, \boldsymbol{\Delta}, \boldsymbol{w}, \boldsymbol{\theta})$ with output

$$\boldsymbol{Y} = \left(\left(T^{-1}_{a_1,N_1}\right)^* Y_1, \ldots, \left(T^{-1}_{a_R,N_R}\right)^* Y_R\right) \in \bigoplus_{i=1}^{R} \mathcal{D}^{L_i+\Delta_i}\left(a_i + \wp^{N_i}\mathcal{O}_{\mathbb{K}}\right),$$

such that $\left\|\boldsymbol{f} - \boldsymbol{Y}\right\|_\rho < \epsilon$, for $\boldsymbol{w}$, $\boldsymbol{\theta}$, running in certain balls.

Proof. We first construct a DNN to approximate $f_i \in L^\rho(a_i + \wp^{N_i}\mathcal{O}_{\mathbb{K}})$. Notice that $T^*_{a_i,N_i} f_i \in L^\rho(\mathcal{O}_{\mathbb{K}})$, with $\left\| T^*_{a_i,N_i} f_i \right\|_\rho < M_i(\rho)$, where

$$M_i(\rho) = \begin{cases} p^{\frac{N_i}{\rho}} M & \text{if} \quad 1 \leq \rho < \infty \\ \\ M & \text{if} \quad \rho = \infty. \end{cases}$$

By Theorem 1, there exists a

$$DNN(\mathcal{O}_{\mathbb{K}}, \sigma_{M_i(\rho)}, L_i, \Delta_i, w_i, \theta) \tag{6.8}$$

such that the output $Y_i(x; \mathcal{O}_{\mathbb{K}}, \sigma_{M_i(\rho)}, L, \Delta, w, \theta)$ satisfies

$$\left\| Y_i\left(\mathcal{O}_{\mathbb{K}}, \sigma_{M_i(\rho)}, L, \Delta, w, \theta_i\right) - T^*_{a_i,N_i} f_i \right\|_\rho < \frac{\epsilon}{\gamma_i(\rho)}, \tag{6.9}$$

with

$$\gamma_i(\rho) = \begin{cases} p^{\frac{N_i}{\rho}} & \text{if} \quad 1 \leq \rho < \infty \\ \\ 1 & \text{if} \quad \rho = \infty, \end{cases}$$

for any w_i in a ball in $\mathbb{R}^{2p^{(L_i+\Delta_i)}}$, and any θ in a ball in $\mathbb{R}^{p^{(L_i+\Delta_i)}}$.

The inequality (6.9) means that

$$E_i(x) := Y_i\left(y; \mathcal{O}_{\mathbb{K}}, \sigma_{M_i(\rho)}, L, \Delta, w, \theta\right) - T^*_{a_i,N_i} f_i(x), \text{ with } \|E_i\|_\rho < \frac{\epsilon}{\gamma_i(\rho)},$$

where the parameters $\sigma_{M_i(\rho)}, L_i, \Delta_i, w_i, \theta_i$ are functions of a_i, N_i. Then

$$\left(T^{-1}_{a_i,N_i}\right)^* E_i(x) = \left(T^{-1}_{a_i,N_i}\right)^* Y_i\left(x; \mathcal{O}_{\mathbb{K}}, L, \Delta, w, \theta\right) - f_i(x).$$

Now, for $1 \leq \rho < \infty$, taking $y = p^{-N_i}(x - a_i)$, $dy = p^{N_i} dx$, we have

$$\left\| \left(T^{-1}_{a_i,N_i}\right)^* E_i \right\|_\rho = \left(\int_{\mathcal{O}_{\mathbb{K}}} \left| \left(T^{-1}_{a_i,N_i}\right)^* E_i(x) \right|^\rho dx \right)^{\frac{1}{\rho}} = \left(\int_{\mathcal{O}_{\mathbb{K}}} \left| E_i(a_i + \wp^{N_i} x) \right|^\rho dx \right)^{\frac{1}{\rho}}$$

$$= p^{\frac{L_i}{\rho}} \left(\int_{a_i + \wp^{N_i}\mathcal{O}_{\mathbb{K}}} |E_i(z)|^\rho dz \right)^{\frac{1}{\rho}} < \epsilon.$$

In the case $\rho = \infty$, $\left\| \left(T^*_{a_i,N_i}\right)^{-1} E \right\|_\rho = \|E\|_\rho < \epsilon$. In conclusion, there exits a $DNN(\mathcal{O}_{\mathbb{K}}, \sigma_{M_i(\rho)}, L_i, \Delta_i, w_i, \theta)$, with output satisfying

$$\left\| \left(T^{-1}_{a_i,N_i}\right)^* Y\left(\mathcal{O}_{\mathbb{K}}, \sigma_{M_i(\rho)}, L_i, \Delta_i, w_i, \theta_i\right) - f_i \right\|_\rho < \epsilon,$$

for any w_i in a ball in $\mathbb{R}^{2p^{(L_i+\Delta_i)}}$, and any θ in a ball in $\mathbb{R}^{p^{(L_i+\Delta_i)}}$. We now set

$$\boldsymbol{L} = (L_1, \ldots, L_R), \; \boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_R) \in \mathbb{N}^R,$$

$$\boldsymbol{w} = (w_1, \ldots, w_R) \in \bigoplus_{i=1}^R \mathcal{D}^{L_i+\Delta_i}\left(\left(a_i + \wp^{N_i}\mathcal{O}_{\mathbb{K}}\right) \times \left(a_i + \wp^{N_i}\mathcal{O}_{\mathbb{K}}\right)\right),$$

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_R) \in \bigoplus_{i=1}^R \mathcal{D}^{L_i+\Delta_i}\left(a_i + \wp^{N_i}\mathcal{O}_{\mathbb{K}}\right), \quad \boldsymbol{\sigma} = \left(\sigma_{M_1(\rho)}, \ldots, \sigma_{M_R(\rho)}\right).$$

Then, by taking the direct product of (6.8), we obtain a neural network

$$DNN(\mathcal{O}_{\mathbb{K}},\boldsymbol{\sigma},R,\boldsymbol{L},\boldsymbol{\Delta},\boldsymbol{w},\boldsymbol{\theta})$$

with output $\boldsymbol{Y}=\left(\left(T_{a_1,N_1}^{-1}\right)^{*}Y_1,\ldots,\left(T_{a_R,N_R}^{-1}\right)^{*}Y_R\right)\in\bigoplus_{i=1}^{R}\mathcal{D}^{L_i+\Delta_i}\left(a_i+\wp^{N_i}\mathcal{O}_{\mathbb{K}}\right)$, such that $\left\Vert \boldsymbol{f}-\boldsymbol{Y}\right\Vert_{\rho}<\epsilon$ for any $\boldsymbol{w}$, respectively $\boldsymbol{\theta}$, running in some balls. □

7. Approximation of functions from $L^{\rho}([0,1])$ by p-adic DNNs

The following result is well-known in the mathematical folklore, but we cannot find a suitable bibliographic reference for it.

Lemma 3. *Given $x\in[0,1]$, there exists a unique sequence $\{x_i\}_{i\in\mathbb{N}}$, $x_i\in\{0,\ldots,p-1\}$, of p-adic digits such that*

$$x=\sum_{i=0}^{\infty}x_ip^{-i-1}.$$

The sequence of p-adic digits is computed recursively. If $x=\frac{m}{p^n}$, with n, $m=\sum_{i=0}^{k}m_ip^i\in\mathbb{N}\smallsetminus\{0\}$, $x_i\in\{0,\ldots,p-1\}$, and $0<m<p^n$, then

$$x=\sum_{i=0}^{k}m_ip^{i-n},$$

which means that almost all the digits are zero.

Proof. The proof is a recursive algorithm for computing the p-adic digits of x.

Step 0. The digit $x_0=i$ if

$$\frac{i}{p}\leq x<\frac{i+1}{p}\text{, for }i\in\{0,\ldots,p-1\}. \tag{7.1}$$

Induction step. Assume that the digits x_0,…,x_N, for some $N\geq1$, are already computed. Set $\widetilde{x}_N:=\sum_{i=0}^{N}x_ip^{-i-1}$,

$$0\leq x-\widetilde{x}_N\leq\frac{1}{p^{N+1}}.$$

If $x=\widetilde{x}_N$, then $x_i=0$ for $i\geq N+1$. If $x=\widetilde{x}_N+\frac{1}{p^{N+1}}$, then $x_i=0$ for $i\geq N+2$. If $0<x-\widetilde{x}_N<\frac{1}{p^{N+1}}$, we divide the interval $\left(0,\frac{1}{p^{N+1}}\right)$ into p subintervals of length $\frac{1}{p}$, and set $x_{N+1}=i$ if

$$\frac{i}{p^{N+2}}\leq x-\widetilde{x}_N<\frac{i+1}{p^{N+2}}\text{, for for }i\in\{0,\ldots,p-1\}. \tag{7.2}$$

We now replace x by $\widetilde{x}_{N+1}$ and go back to Step 0. The solutions of the decision problems (7.1)-(7.2) are uniquely determined by the number x. Consequently, the sequence of p-adic digits is uniquely determined by x. □

We set

$$\mathcal{I}_p:=\left\{\sum_{i=0}^{\infty}a_ip^{-i-1};a_i\in\{0,\ldots,p-1\}\right\}.$$

Given $x\in[0,1]$, we set $\varrho_p(x):=\sum_{i=0}^{\infty}x_ip^{-i-1}$. By Lemma 3, $\varrho_p:[0,1]\rightarrow\mathcal{I}_p$ is a one-to-one function. But, this function is not surjective. Indeed, the formula

$$\frac{1}{p^{N+1}}=\sum_{i=N+1}^{\infty}(p-1)\,p^{-i-1}$$

implies that

$$\sum_{i=0}^{N-1}x_ip^{-i-1}+\frac{1}{p^{N+1}}=\sum_{i=0}^{N-1}x_ip^{-i-1}+\sum_{i=N+1}^{\infty}(p-1)\,p^{-i-1}.$$

For $x=\sum_{i=0}^{N-1}x_ip^{-i-1}+\frac{1}{p^{N+1}}$, function ϱ_p gives the sequence $(x_0,\ldots,x_{N-1},1)$; the sequence $(x_0,\ldots,x_{N-1},p-1,p-1,\ldots)$ is not in the range of ϱ_p.

We now construct a generalization of function ϱ_p as follows:

$$\begin{array}{cccc}\varrho_{\mathbb{K}}: & [0,1] & \rightarrow & \mathcal{O}_{\mathbb{K}}\\ & x=\sum_{i=0}^{\infty}x_ip^{-i-1} & \rightarrow & \varrho_{\mathbb{K}}(x)=\sum_{i=0}^{\infty}x_i\wp^i.\end{array}$$

This function is one-to-one, but not surjective. We denote the range of $\varrho_{\mathbb{K}}$ as $\mathcal{O}_{\mathbb{K}}\smallsetminus\mathcal{M}$. The following result follows directly from the definition of the function $\varrho_{\mathbb{K}}$.

Lemma 4. *With the above notation, the following assertions hold:*
(i) $\varrho_{\mathbb{K}}(x)$ *is a continuous function.*
(ii) $\left|\varrho_{\mathbb{K}}^{-1}(x)-\varrho_{\mathbb{K}}^{-1}(y)\right|\leq|x-y|_{\mathbb{K}}$, *for* $x,y\in\mathcal{O}_{\mathbb{K}}\smallsetminus\mathcal{M}$.
(iii) For any $b=\sum_{i=0}^{k}x_i\wp^i\in\mathcal{O}_{\mathbb{K}}$,

$$\varrho_{\mathbb{K}}\left(\sum_{i=0}^{k}x_ip^{-i-1}+p^{-k-1}[0,1]\right)=\left(b+\wp^{k+1}\mathcal{O}_{\mathbb{K}}\right)\smallsetminus\mathcal{M},\ \textit{for } k\geq0.$$

(iv) $\varrho_{\mathbb{K}}([0,1])=\mathcal{O}_{\mathbb{K}}\smallsetminus\mathcal{M}$.

We denote by $\mathcal{B}([0,1])$, respectively $\mathcal{B}(\mathcal{O}_{\mathbb{K}})$, the Borel σ-algebra of $[0,1]$, respectively of $\mathcal{O}_{\mathbb{K}}$. By Lemma 4-(i), the function

$$\varrho_{\mathbb{K}}:([0,1],\mathcal{B}([0,1]))\rightarrow(\mathcal{O}_{\mathbb{K}},\mathcal{B}(\mathcal{O}_{\mathbb{K}}))$$

is measurable. We denote by $\mathfrak{L}:\mathcal{B}([0,1])\rightarrow[0,+\infty]$ the Lebesgue measure in $[0,1]$, and by $\mathfrak{H}$ the pushforward measure of $\mathfrak{L}$ to $\mathcal{B}(\mathcal{O}_{\mathbb{K}})$:

$$\mathfrak{H}(B):=\mathfrak{L}\left(\varrho_{\mathbb{K}}^{-1}(B)\right)\text{ for }B\in\mathcal{B}(\mathcal{O}_{\mathbb{K}}).$$

We denote by $\mathcal{B}_{\mathcal{M}}(\mathcal{O}_{\mathbb{K}})$ the restriction of $\mathcal{B}(\mathcal{O}_{\mathbb{K}})$ to $\mathcal{O}_{\mathbb{K}}\smallsetminus\mathcal{M}$, which is the σ-algebra of subsets of the form

$$B\cap(\mathcal{O}_{\mathbb{K}}\smallsetminus\mathcal{M}),\text{ for }B\in\mathcal{B}(\mathcal{O}_{\mathbb{K}}).$$

The function $\varrho_{\mathbb{K}}^{-1}:([0,1],\mathcal{B}([0,1]))\rightarrow(\mathcal{O}_{\mathbb{K}},\mathcal{B}_{\mathcal{M}}(\mathcal{O}_{\mathbb{K}}))$ is also measurable, cf. Lemma 4-(ii).

Proposition 1. *The measure* $\mathfrak{H}$ *equals the Haar measure* μ_{Haar} *in* $\mathcal{B}(\mathcal{O}_{\mathbb{K}})$. *Furthermore,* $\mu_{Haar}(\mathcal{M})=0$.

Proof. By Lemma 4-(iii)-(iv),

$$\mathfrak{H}\left(\left(b+\wp^{n}\mathcal{O}_{\mathbb{K}}\right)\smallsetminus\mathcal{M}\right)=\mu_{\text{Haar}}\left(b+\wp^{n}\mathcal{O}_{\mathbb{K}}\right)=p^{-n}\text{ for }n\in\mathbb{N},\,b\in\mathcal{O}_{\mathbb{K}}.$$

By the Carathéodory extension theorem, see, e.g., [5, Section 1.3.10], the Haar measure is uniquely determined by the condition $\mu_{\text{Haar}}\left(b+\wp^{n}\mathcal{O}_{\mathbb{K}}\right)=p^{-n}$ for $n\in\mathbb{N}$, $b\in\mathcal{O}_{\mathbb{K}}$. Therefore, $\mathfrak{H}$ extends to the Haar measure in $\mathcal{B}\left(\mathcal{O}_{\mathbb{K}}\right)$, and $\mu_{\text{Haar}}\left(\mathcal{M}\right)=0$. □

Given a function $f:\mathcal{O}_{\mathbb{K}}\rightarrow\mathbb{R}$, we define $\varrho_{\mathbb{K}}^{*}f:[0,1]\rightarrow\mathbb{R}$, by $\left(\varrho_{\mathbb{K}}^{*}f\right)(x)=f\left(\varrho_{\mathbb{K}}\left(x\right)\right)$.

Theorem 3. *The map $\varrho_{\mathbb{K}}^{*}:L^{\rho}\left(\mathcal{O}_{\mathbb{K}}\right)\rightarrow L^{\rho}\left([0,1]\right)$, $\rho\in[1,\infty]$, is an isometric isomorphism of normed spaces, i.e., a linear surjective isometry.*

Proof. Changing variables as $y=\varrho_{\mathbb{K}}\left(x\right)$, the Lebesgue measure of $[0,1]$ becomes the Haar measure of $\mathcal{O}_{\mathbb{K}}$, then

$$\left\Vert\varrho_{\mathbb{K}}^{*}f\right\Vert_{\rho}=\left(\int\limits_{[0,1]}\left|f\left(\varrho_{\mathbb{K}}\left(x\right)\right)\right|^{\rho}dx\right)^{\frac{1}{\rho}}=\left(\int\limits_{\mathcal{O}_{\mathbb{K}}}\left|f\left(y\right)\right|^{\rho}dy\right)^{\frac{1}{\rho}}=\left\Vert f\right\Vert_{\rho}.$$

□

Theorems 1 and 3 imply that non-Archimedean DNNs can approximate functions from $L^{\rho}\left([0,1]\right)$ But, the spaces $\mathcal{D}^{L}\left(\mathcal{O}_{\mathbb{K}}\right)$ are naturally isomorphic to spaces of simple functions defined on $[0,1]$; this isomorphism allows us to describe the approximation of functions from $L^{\rho}\left([0,1]\right)$ by DNNs in terms of functions defined on $[0,1]$. The details are as follows.

Remark 4. *Let $(\Omega,\mathcal{F},\mu)$ be a measure space, where $\mathcal{F}$ is a σ-algebra of subsets of Ω, and μ is measure in $\mathcal{F}$. A function $h:\Omega\rightarrow\mathbb{R}$ is called simple if $h(x)=\sum_{i=1}^{n}\alpha_{i}1_{A_{i}}\left(x\right)$, where the α_{i} are real numbers, and the A_{i} are disjoint sets in $\mathcal{F}$. The simple functions are dense in $L^{\rho}\left(\Omega,\mathcal{F},\mu\right)$, for $\rho\in[1,\infty]$; see* [5, Theorem 2.4.13].

Lemma 5. *$\varrho_{\mathbb{K}}$ (considered as a change of variables) gives a bijection between the simple functions defined on $\mathcal{O}_{\mathbb{K}}$ into the simple functions defined on* $[0,1]$.

Proof. Take $B\in\mathcal{B}(\mathcal{O}_{\mathbb{K}})$, since $\mu_{\text{Haar}}\left(\mathcal{M}\right)=0$,

$$1_{B\smallsetminus\mathcal{M}}=1_{B}\text{ in }L^{\rho}\left(\mathcal{O}_{\mathbb{K}}\right).$$

By using that $\varrho_{\mathbb{K}}$ is measurable, cf. Lemma 4-(i),

$$\varrho_{\mathbb{K}}^{-1}\left(B\smallsetminus\mathcal{M}\right)=\varrho_{\mathbb{K}}^{-1}\left(B\right)\smallsetminus\varrho_{\mathbb{K}}^{-1}\left(\mathcal{M}\right)$$

is a Borel subset of $[0,1]$. Now the Lebesgue measure of $\varrho_{\mathbb{K}}^{-1}\left(\mathcal{M}\right)$ is zero, cf. Proposition 1, and thus,

$$1_{\varrho_{\mathbb{K}}^{-1}\left(B\right)\smallsetminus\varrho_{\mathbb{K}}^{-1}\left(\mathcal{M}\right)}=1_{\varrho_{\mathbb{K}}^{-1}\left(B\right)}\text{ in }L^{\rho}\left([0,1]\right).$$

Consequently, $\varrho_{\mathbb{K}}^{-1}$ sends the function 1_{B} into $1_{\varrho_{\mathbb{K}}^{-1}\left(B\right)}$. We now take a Borel subset A of $[0,1]$. By using that $\varrho_{\mathbb{K}}^{-1}$ is measurable, cf. Lemma 4-(ii), $\varrho_{\mathbb{K}}\left(A\right)$ is a Borel subset of $\mathcal{O}_{\mathbb{K}}$. Then, $\varrho_{\mathbb{K}}$ sends the function 1_{A} into $1\varrho_{\mathbb{K}}\left(A\right)$. □

Set

$$\varrho_{\mathbb{K}}^{-1}(G_l) = \left\{a_0p^{-1} + a_1p^{-2} + \ldots + a_{l-1}p^{-l-2}; a_i \in \{0, \ldots, p-1\}\right\}.$$

Notice that $\varrho_{\mathbb{K}}^{-1}(G_l)$ is not an additive group. For instance,

$$1p^{-l-2}, (p-1)\,p^{-l-2} \in \varrho_{\mathbb{K}}^{-1}(G_l)\,, \text{ but } 1p^{-l-2} + (p-1)\,p^{-l-2} = p^{-l-1} \notin \varrho_{\mathbb{K}}^{-1}(G_l)\,.$$

Now, $\phi \in \mathcal{D}^L(\mathcal{O}_{\mathbb{K}})$ can be identified with the function

$$\phi(x) = \sum_{\alpha \in \varrho_{\mathbb{K}}^{-1}(G_l)} \phi(\alpha)\, 1_{\alpha + [0, p^{-l-1}]}(x)\,, \text{ for } x \in [0,1]\,.$$

Now, by Theorems 1 and 3, and Lemma 5, we have the following result.

Theorem 4. *Given any* $f \in L^\rho([0,1])$, $1 \leq \rho \leq \infty$, *with* $\|f\|_\rho < M$, *and any* $\epsilon > 0$*; then, for any input* $X \in \mathcal{D}^L(\mathcal{O}_{\mathbb{K}})$, $L \geq 1$, *there exists a*

$$DNN(\mathcal{O}_{\mathbb{K}}, \sigma_M, L, \Delta, w, \theta)$$

with output

$$Y(x; \mathcal{O}_{\mathbb{K}}, \sigma_M, L, \Delta, w, \theta) = \sum_{\alpha \in \varrho_{\mathbb{K}}^{-1}(G_{L+\Delta})} Y(\alpha)\, 1_{\alpha + [0, p^{-L-\Delta-1}]}(x)$$

satisfying

$$\|Y(\mathcal{O}_{\mathbb{K}}, \sigma_M, L, \Delta, w, \theta) - f\|_\rho < \epsilon,$$

for any w *in a ball in* $\mathbb{R}^{2p^{(L+\Delta)}}$, *and any* θ *in a ball in* $\mathbb{R}^{p^{(L+\Delta)}}$.

8. Fourier-Walsh series

In [1], the authors use Fourier-Walsh series to approximate hierarchical functions. Here, we show that this a pure non-Archimedean technique.

8.1. **Additive characters.** An additive character of $\mathbb{K}$ is a continuous group homomorphism from $(\mathbb{K}, +)$ into $(S, \cdot)$, the complex unit circle considered as a multiplicative group. We denote by χ_{triv} by the trivial additive character of $\mathbb{K}$, i.e., $\chi_{\text{triv}}(x) = 1$ for any $x \in \mathbb{K}$. We say $\chi_{\mathbb{K}}$ is a standard additive character of $\mathbb{K}$, if $\chi_{\mathbb{K}} \neq \chi_{\text{triv}}$, and $\chi_{\mathbb{K}}(x) = 1$ for any $x \in \mathcal{O}_{\mathbb{K}}$. In the case, $\mathbb{K} = \mathbb{F}_p((T))$, for $x = \sum_{k=l}^{\infty} a_k T^k$, $l \in \mathbb{Z}$, we set $Res(x) := a_{-1}$, then, the function

$$\chi_{\mathbb{K}}(x) = \exp\left(\frac{2\pi i}{p} Res(x)\right)$$

is a standard additive character of $\mathbb{F}_p((T))$. In the case $\mathbb{K} = \mathbb{Q}_p$, for $x = \sum_{k=r}^{\infty} a_k p^k$, $r \in \mathbb{Z}$, we set

$$\{x\}_p := \begin{cases} 0 & \text{if} \quad r \geq 0 \\ \\ \sum_{k=r}^{-1} a_k p^k & \text{if} \quad r < 0. \end{cases}$$

Then,

$$\chi_{\mathbb{K}}(x) = \exp\left(2\pi i \{x\}_p\right)$$

is a standard additive character of $\mathbb{Q}_p$.

Any additive character of $\mathbb{K}$ has the form $\chi(x) = \chi_{\mathbb{K}}(ax)$ for $x \in \mathbb{K}$, and some $a \in \mathbb{K}$, see [28, Chap. II, Corollary to Theorem 3]. We denote by $\Omega(\mathcal{O}_{\mathbb{K}})$ the group of additive characters of $\mathcal{O}_{\mathbb{K}}$. Then,

$$\Omega(\mathcal{O}_{\mathbb{K}}) = \left\{\chi_{\mathbb{K}}\left(\wp^{-l}\boldsymbol{a}x\right); l \in \mathbb{N} \smallsetminus \{0\}, \boldsymbol{a} \in G_l(\mathcal{O}_{\mathbb{K}})\right\} \cup \{\chi_{\text{triv}}\},$$

where $G_l(\mathcal{O}_{\mathbb{K}}) = \mathcal{O}_{\mathbb{K}}/\wp^l\mathcal{O}_{\mathbb{K}}$.

8.2. **Orthonormal basis of $L(\mathcal{O}_{\mathbb{K}}) \bigotimes \mathbb{C}$ and $L^2([0,1]) \bigotimes \mathbb{C}$.** Given $f, g : \mathcal{O}_{\mathbb{K}} \to \mathbb{C}$, we set

$$\langle f, g\rangle = \int_{\mathcal{O}_{\mathbb{K}}} f(x)\,\overline{g}(x)\,dx, \text{ and } \|f\|_2^2 = \langle f, f\rangle,$$

where the bar denotes the complex conjugate. We set

$$L^2(\mathcal{O}_{\mathbb{K}}) \bigotimes \mathbb{C} = \{f : \mathcal{O}_{\mathbb{K}} \to \mathbb{C}; \|f\|_2 < \infty\}.$$

This space is the complexification of the space $L^2(\mathcal{O}_{\mathbb{K}})$. We set $L^2([0,1]) \bigotimes \mathbb{C}$ for the complexification of $L^2([0,1])$. We denote by $\mathcal{D}(\mathcal{O}_{\mathbb{K}}) \bigotimes \mathbb{C}$, the $\mathbb{C}$-vector space of test functions, which is the complexification of $\mathcal{D}(\mathcal{O}_{\mathbb{K}})$.

Theorem 5. *With the above notation, the following assertions hold.*
(i) The group of characters $\Omega(\mathcal{O}_{\mathbb{K}})$ is an orthonormal basis of $L^2(\mathcal{O}_{\mathbb{K}}) \bigotimes \mathbb{C}$.
(ii) Set

$$\omega_{\wp,\boldsymbol{a},l}(x) := \chi_{\mathbb{K}}\left(\wp^{-l}\boldsymbol{a}\varrho_{\mathbb{K}}(x)\right), \textit{ for } x \in [0,1], \ l \in \mathbb{N} \smallsetminus \{0\}, \boldsymbol{a} \in G_l(\mathcal{O}_{\mathbb{K}}).$$

Then

$$\left\{\omega_{\wp,\boldsymbol{a},l}\right\}_{\wp,\boldsymbol{a},l} \cup \left\{1_{[0,1]}\right\}$$

is an orthonormal basis of $L^2([0,1]) \bigotimes \mathbb{C}$.

Proof. We start by recalling the following fact: the group of characters $\Omega(\mathcal{O}_{\mathbb{K}})$ is an orthonormal basis of the pre-Hilbert space $(\mathcal{D}(\mathcal{O}_{\mathbb{K}}) \bigotimes \mathbb{C}, \langle\cdot,\cdot\rangle)$, i.e.,

$$\langle \chi, \chi'\rangle = \begin{cases} 0 & \text{if} \quad \chi \neq \chi' \\ \\ 1 & \text{if} \quad \chi = \chi', \end{cases}$$

for any $\chi, \chi' \in \Omega(\mathcal{O}_{\mathbb{K}})$, and any function $\varphi \in \mathcal{D}(\mathcal{O}_{\mathbb{K}}) \bigotimes \mathbb{C}$ can be expressed as a finite sum

$$\varphi(x) = \sum_{\chi \in \Omega(\mathcal{O}_{\mathbb{K}})} C_\chi \chi(x), \text{ where } C_\chi = \langle \varphi, \chi\rangle. \tag{8.1}$$

This assertion is a particular case of a well-known result of harmonic analysis on compact groups, see [17, Proposition 7.2.2].

We denote by *Span* $(\Omega(\mathcal{O}_{\mathbb{K}}))$, the $\mathbb{C}$-subvector space of $L^2(\mathcal{O}_{\mathbb{K}}) \bigotimes \mathbb{C}$ spanned by the elements of $\Omega(\mathcal{O}_{\mathbb{K}})$. To show that $\Omega(\mathcal{O}_{\mathbb{K}})$ is an orthonormal basis of $L^2(\mathcal{O}_{\mathbb{K}}) \bigotimes \mathbb{C}$, it is sufficient to establish that

$$\overline{Span(\Omega(\mathcal{O}_{\mathbb{K}}))} = L^2(\mathcal{O}_{\mathbb{K}}) \bigotimes \mathbb{C},$$

where the bar denotes the topological closure of *Span* $(\Omega(\mathcal{O}_{\mathbb{K}}))$ in $L^2(\mathcal{O}_{\mathbb{K}}) \bigotimes \mathbb{C}$. Let $f \in L^2(\mathcal{O}_{\mathbb{K}}) \bigotimes \mathbb{C}$ and $\epsilon > 0$ given. Then, by the density of $\mathcal{D}(\mathcal{O}_{\mathbb{K}}) \bigotimes \mathbb{C}$ in $L^2(\mathcal{O}_{\mathbb{K}}) \bigotimes \mathbb{C}$ and (8.1), there exists $\varphi \in \mathcal{D}(\mathcal{O}_{\mathbb{K}}) \bigotimes \mathbb{C} \cap Span(\Omega(\mathcal{O}_{\mathbb{K}}))$ such that $\|f - \varphi\|_2 < \epsilon$.

(ii) By Theorem 3, the map $\varrho_{\mathbb{K}}^* : L^2(\mathcal{O}_{\mathbb{K}}) \to L^2([0,1])$, is a linear surjective isometry, i.e., $\|g\|_2^2 = \|\varrho_{\mathbb{K}}^* g\|_2^2$ for any $g \in L^2(\mathcal{O}_{\mathbb{K}})$. Take $f = f_1 + if_2 \in L^2(\mathcal{O}_{\mathbb{K}}) \bigotimes \mathbb{C}$, with $f_i \in L^2(\mathcal{O}_{\mathbb{K}})$, $i = 1, 2$. We now set $\varrho_{\mathbb{K}}^* f = \varrho_{\mathbb{K}}^* f_1 + i\varrho_{\mathbb{K}}^* f_2 \in L^2([0,1]) \bigotimes \mathbb{C}$. Then

$$\|\varrho_{\mathbb{K}}^* f\|_2^2 = \|\varrho_{\mathbb{K}}^* f_1\|_2^2 + \|\varrho_{\mathbb{K}}^* f_2\|_2^2 = \|f_1\|_2^2 + \|f_2\|_2^2 = \|f\|_2^2.$$

Consequently, $\varrho_{\mathbb{K}}^* : L^2(\mathcal{O}_{\mathbb{K}}) \bigotimes \mathbb{C} \to L^2([0,1]) \bigotimes \mathbb{C}$ is an L^2-isometry, and by the polarization identities,

$$\langle \varrho_{\mathbb{K}}^* f, \varrho_{\mathbb{K}}^* g\rangle = \langle f, g\rangle \text{ for any } f, g \in L^2(\mathcal{O}_{\mathbb{K}}) \bigotimes \mathbb{C}$$

Therefore, $\varrho_{\mathbb{K}}^*$ sends orthonormal bases of $L^2(\mathcal{O}_{\mathbb{K}}) \bigotimes \mathbb{C}$ into orthonormal bases of $L^2([0,1]) \bigotimes \mathbb{C}$. □

Besides the rings $\mathbb{F}_p[[T]]$, $\mathbb{Z}_p$ share many common arithmetic and geometric features, their arithmetic operations are quite different. In the case of $\mathbb{F}_p[[T]]$, given $x = \sum_{i=0}^{\infty} x_i T^i$, $y = \sum_{i=0}^{\infty} y_i T^i \in \mathbb{F}_p[[T]]$,

$$x + y = \sum_{i=0}^{\infty} (x_i + y_i) T^i, \text{ and } xy = \sum_{i=0}^{\infty} \left(\sum_{s=0}^{i} x_s y_{i-s} \right) T^i, \tag{8.2}$$

where the digit operations are performed in the field $\mathbb{F}_p$. In particular, the addition in $\mathbb{F}_p[[T]]$ does not require carry digits. Using (8.2), we rewrite the orthonormal basis given in Theorem 5 as follows. The set of functions

$$\theta_{p,\boldsymbol{n},l}(x) := \exp\left(\frac{2\pi i}{p} \sum_{i=0}^{l} n_{l-1-i} x_i \right), \tag{8.3}$$

where

$$x = \sum_{i=0}^{\infty} x_i p^{-i-1} \in [0,1], \boldsymbol{n} = \sum_{i=0}^{l} n_i p^{i-1} \in \mathbb{N}, n_i \in \{0, \ldots, p-1\}, l \geq 1,$$

form an orthonormal basis of $L^2([0,1]) \bigotimes \mathbb{C}$. The digit operations in (8.3) are performed in the field $\mathbb{F}_p$.

In the case of $\mathbb{Z}_p$, given $x = \sum_{i=0}^{\infty} x_i p^i$, $y = \sum_{i=0}^{\infty} y_i p^i \in \mathbb{Z}_p$, there are no closed formulas for the digits of the sum $x + y = \sum_{i=0}^{\infty} a_i p^i$ or the product $xy = \sum_{i=0}^{\infty} b_i p^i$ because the addition in $\mathbb{Z}_p$ requires carry digits. The digits of the sum and product are compute recursively:

$$\sum_{i=0}^{L-1} a_i p^i \equiv \sum_{i=0}^{L-1} x_i p^i + \sum_{i=0}^{L-1} y_i p^i \mod p^L, \text{ for any } L \geq 1, \tag{8.4}$$

and

$$\sum_{i=0}^{L-1} b_i p^i \equiv \left(\sum_{i=0}^{L-1} x_i p^i \right) \left(\sum_{i=0}^{L-1} y_i p^i \right) \mod p^L, \text{ for any } L \geq 1, \tag{8.5}$$

where for $s, t \in \mathbb{Z}$, $s \equiv t \bmod p^L$ means that $s - t$ is divisible by p^L. We now rewrite the orthonormal basis given in Theorem 5 as follows. The set of functions

$$\gamma_{p,\boldsymbol{n},l}(x) := \exp\left(\frac{2\pi i}{p^l} \left(\sum_{i=0}^{l-1} n_i p^i \right) \left(\sum_{i=0}^{l-1} x_i p^i \right) \right), \tag{8.6}$$

where the digit operations indicated in (8.6) are performed in the quotient ring $\mathbb{Z}/p^l\mathbb{Z}$, i.e., they are arithmetic operations $\bmod p^l$, and

$$x=\sum_{i=0}^{\infty}x_i p^{-i-1}\in[0,1]\,,\ \boldsymbol{n}=\sum_{i=0}^{l}n_i p^{i-1}\in\mathbb{N},\ n_i\in\{0,\ldots,p-1\}\,,\ l\geq 1,$$

form an orthonormal basis of $L^2([0,1])\bigotimes\mathbb{C}$.

The above described orthonormal bases of Walsh-Paley type, see, e.g., [8], [11], [23]. In the case $\mathbb{F}_2[[T]]$, the basis (8.3) agrees with Walsh-Paley system [23]; in this framework the connection with the group of characters $\Omega(\mathbb{F}_2[[T]])$ is well-known, [11]. There are generalized Walsh-Paley bases, see, e.g. [8], however, the digit operations used in [8] are different to the ones used here. On the other hand. in the last thirty-five years orthonormal bases on $L^2(\mathbb{Q}_p)\bigotimes\mathbb{C}$ has been studied intensively; see, e.g., [2], [19], [21], [27], and the references therein.

References

[1] Abbe Emmanuel, Boix-Adsera Enric, Brennan Matthew, BreslerGuy, Nagaraj Dheeraj,The staircase property: How hierarchical structure can guide deep learning. Conference on Neural Information Processing Systems (NeurIPS'21).

[2] Albeverio S., Khrennikov A. Yu., Shelkovich V. M., Theory of p-adicdistributions: linear and nonlinear models. Cambridge University Press, 2010.

[3] Albeverio Sergio, Khrennikov Andrei, Tirozzi, Brunello, p -adic dynamical systems and neural networks, Math. Models Methods Appl. Sci. 9 (1999), no. 9, 1417–1437.

[4] Anashin, Vladimir, The p-adic theory of automata functions. Advances in non-Archimedean analysis and applications—the p-adic methodology in STEAM-H, 9–113. Springer, Cham, 2021, 9–113.

[5] Ash, Robert B., Measure, integration, and functional analysis, Academic Press, New York-London, 1972.

[6] Blum Lenore, Shub Mike, Smale Steve, On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions and universal machines, Bull. Amer. Math. Soc. (N.S.) 21 (1989), no.1, 1–46.

[7] Calin Ovidiu, Deep learning architectures—a mathematical approach. Springer Ser. Data Sci. Springer, Cham, 2020.

[8] Chrestenson, H. E., A class of generalized Walsh functions, Pacific J. Math. 5 (1955), 17–31.

[9] Cybenko, G., Approximation by superpositions of a sigmoidal function.Math. Control Signals Systems 2 (1989), no.4, 303–314.

[10] Elbrächter Dennis, Perekrestenko Dmytro, Grohs Philipp, Bölcskei Helmut, Deep neural network approximation theory, IEEE Trans. Inform. Theory 67 (2021), no. 5, 2581–2623.

[11] Fine N. J., On the Walsh functions, Trans. Amer. Math. Soc. 65 (1949), 372–414.

[12] Gribonval Rémi, Kutyniok Gitta, Nielsen Morten, Voigtlaender Felix, Approximation spaces of deep neural networks, Constr. Approx. 55 (2022), no. 1, 259–367.

[13] Halmos P., Measure Theory, D. Van Nostrand Company Inc., New York, 1950.

[14] Hilgetag C.C., Goulas A., Is the brain really a small-world network?. Brain Struct Funct 221, 2361–2366 (2016). https://doi.org/10.1007/s00429-015-1035-6.

[15] Higham Catherine F., Higham Desmond J., Deep learning: an introduction for applied mathematicians, SIAM Rev. 61 (2019), no. 4, 860–891.

[16] Multilayer feedforward networks are universal approximators, Neural Networks 2 (1989) no 5, 359-366.

[17] Igusa J.-I., An introduction to the theory of local zeta functions, in AMS/IP Studies in Advanced Mathematics, 14, American Mathematical Society, Providence, RI; International Press, Cambridge, MA, 2000

[18] Iordache O., Non-Archimedean Analysis. In: Modeling Multi-Level Systems. Understanding Complex Systems, vol 70. Springer, Berlin, Heidelberg, 2011. https://doi.org/10.1007/978-3-642-17946-4_13

[19] Khrennikov Andrei, Kozyrev Sergei, Zúñiga-Galindo W. A., Ultrametric Equations and its Applications. Encyclopedia of Mathematics and its Applications 168. Cambridge, Cambridge University Press, 2018.
[20] Khrennikov Andrei, Information dynamics in cognitive, psychological, social and anomalous phenomena. Fund. Theories Phys., 138. Kluwer Academic Publishers Group, Dordrecht, 2004.
[21] Kochubei, Anatoly N., Pseudo-differential equations and stochastics over non-Archimedean fields. New York, Marcel Dekker, Inc., 2001.
[22] Maller Michael, Whitehead Jennifer, Computational complexity over the p-adic numbers, J. Complexity13 (1997), no.2, 195–207.
[23] Paley, R. E. A. C., A Remarkable Series of Orthogonal Functions (I), Proc. London Math. Soc. (2) 34 (1932), no. 4, 241–264.
[24] Pinkus A., Approximation theory of the MLP model in neural networks, Acta Numerica. 1999; 8:143-195. doi:10.1017/S0962492900002919
[25] Sporns O., Small-world connectivity, motif composition, and complexity of fractal neuronal connections, Biosystems, 85(1), (2006) 55-64.
[26] Taibleson M. H., Fourier analysis on local fields. Princeton University Press, 1975.
[27] Vladimirov V. S., Volovich I. V., Zelenov E. I., p-Adic analysis and mathematical physics. Singapore, World Scientific, 1994.
[28] Weil A., Basic Number Theory, Springer-Verlag, Berlin, 1967.
[29] Zambrano-Luna B. A., Zúñiga-Galindo W. A., p -adic cellular neural networks, J. Nonlinear Math. Phys. 30 (2023), no. 1, 34–70.
[30] Zambrano-Luna, B. A.; Zúñiga-Galindo, W. A., p -adic cellular neural networks: applications to image processing. Phys. D 446 (2023), Paper No. 133668, 11 pp.
[31] Zhang,Y., Han L. (2019). Learning (from) Deep Hierarchical Structure among Features. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 5837-5844. https://doi.org/10.1609/aaai.v33i01.33015837
[32] Zhou Ding-Xuan, Universality of deep convolutional neural networks, Appl. Comput. Harmon. Anal. 48 (2020), no.2, 787–794.
[33] Zúñiga-Galindo W. A., Zambrano-Luna, B. A., Hierarchical Wilson-Cowan models and connection matrices, Entropy 25 (2023), no. 6, Paper No. 949, 20 pp.
[34] Zúñiga-Galindo W. A., He C., Zambrano-Luna B. A., p -adic statistical field theory and convolutional deep Boltzmann machines, PTEP. Prog. Theor. Exp. Phys.(2023), no. 6, Paper No. 063A01, 17 pp.
[35] Zúñiga-Galindo W. A., p -adic statistical field theory and deep belief networks, Phys. A 612 (2023), Paper No. 128492, 23 pp.
[36] Zúñiga-Galindo W. A., A Correspondence Between Deep Boltzmann Machines and p-Adic Statistical Field Theories. To apper in Adv. Theor. Math. Phys. arXiv:2306.03751.

University of Texas Rio Grande Valley, School of Mathematical & Statistical Sciences, One West University Blvd, Brownsville, TX 78520, United States
Email address: `wilson.zunigagalindo@utrgv.edu`