7-2022

p-Adic Statistical Field Theory and Deep Belief Networks

Wilson A. Zuniga-Galindo

p-ADIC STATISTICAL FIELD THEORY AND DEEP BELIEF NETWORKS

W. A. ZÚÑIGA-GALINDO

Abstract. In this work we initiate the study of the correspondence between p-adic statistical field theories (SFTs) and neural networks (NNs). In general quantum field theories over a p-adic spacetime can be formulated in a rigorous way. Nowadays these theories are considered just mathematical toy models for understanding the problems of the true theories. In this work we show these theories are deeply connected with the deep belief networks (DBNs). Hinton et al. constructed DBNs by stacking several restricted Boltzmann machines (RBMs). The purpose of this construction is to obtain a network with a hierarchical structure (a deep learning architecture). An RBM corresponds a certain spin glass, thus a DBN should correspond to an ultrametric (hierarchical) spin glass. A model of such system can be easily constructed by using p-adic numbers. In our approach, a p-adic SFT corresponds to a p-adic continuous DBN, and a discretization of this theory corresponds to a p-adic discrete DBN. We show that these last machines are universal approximators. In the p-adic framework, the correspondence between SFTs and NNs is not fully developed. We point out several open problems.

.

Contents

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

1. Introduction

Recently, it has been proposed the existence of a correspondence between neural networks (NNs) and quantum field theories (QFTs), more precisely with Euclidean QFTs, see e.g. [9], [28], [31], [40], [67], [76], [81], [89], see also [16], [17], [20], and the references therein. This correspondence take several different forms depending on the architecture of the networks involved. This article aims to initiate the study of the mentioned correspondence in the framework of the non-Archimedean statistical field theory (SFT), see e.g. [92], [5], see also [1], [46]-[47], [59], [62], [68], [71]-[73], [97], and the references therein. In this case, the corresponding NNs are new hierarchical generalizations of the classical restricted Boltzmann machines (RBMs), see e.g. [22], [32]. More precisely, they are p-adic counterparts of the convolutional deep belief networks (DBNs), see e.g. [27], [43], [44], and the references therein.

A fundamental problem is the understanding of the structure of space-time at the level of the Planck scale. In the 1930s Bronstein showed that general relativity and quantum mechanics imply that the uncertainty Δx of any length measurement satisfies $\Delta x \geq L_{\text{Planck}} := \sqrt{\frac{\hbar G}{c^3}}$, where L_{Planck} is the Planck length ($L_{\text{Planck}} \approx 10^{-33}$ cm). This inequality implies that space-time is not an infinitely divisible continuum (mathematically speaking, the spacetime must be a completely disconnected topological space at the level of the Planck scale). Bronstein's inequality has motivated the development of several different physical theories. At any rate, this inequality implies the need of using non-Archimedean mathematics in models dealing with the Planck scale. In the 1980s, Volovich proposed the conjecture that the space-time at the Planck scale has a p-adic nature, see e.g. [88]. This conjecture has propelled a wide variety of investigations in cosmology, quantum mechanics, string theory, QFT, etc., and the influence of this conjecture is still relevant nowadays, see e.g. [1], [5]-[7], [11]-[15], [24]-[26], [33]-[34], [37]-[39], [41], [46]-[54], [58]-[59], [62], [68], [71]-[75], [77]-[80], [87]-[88], [90]-[97].

A p-adic number is a series of the form

$$x = x_{-k}p^{-k} + x_{-k+1}p^{-k+1} + \ldots + x_0 + x_1 p + \ldots, \text{ with } x_{-k} \neq 0, \tag{1.1}$$

where p is a fixed prime number, and the x_js are numbers in the set $\{0, 1, \ldots, p-1\}$. The set of all possible series of the form (1.1) constitutes the field of p-adic numbers $\mathbb{Q}_p$. There are natural field operations, sum and multiplication, on series of the form (1.1), see e.g. [57]. There is also a natural norm in $\mathbb{Q}_p$ defined as $|x|_p = p^k$, for a nonzero p-adic number of the form (1.1). The field of p-adic numbers with the distance induced by $|\cdot|_p$ is a complete ultrametric space. The ultrametric (or non-Archimedean) property refers to the fact that $|x-y|_p \leq \max\left\{|x-z|_p, |z-y|_p\right\}$ for any x, y, $z \in \mathbb{Q}_p$. We denote by $\mathbb{Z}_p$ the unit ball, which consists of all series with expansions of the form (1.1) with $-k \geq 0$. The unit ball is an infinite rooted tree, with valence p. The field $\mathbb{Q}_p$ does not have an order compatible with the field operations and also it has a natural hierarchical structure. We extend the $p-$adic norm to $\mathbb{Q}_p^N$ by taking $||x||_p = \max_{1\leq i\leq N} |x_i|_p$, for $x = (x_1, \ldots, x_N) \in \mathbb{Q}_p^N$.

The space $\mathbb{Q}_p^N$ has a very rich mathematical structure. The axiomatic quantum field theory can be extended to $\mathbb{Q}_p^N$, see e.g. [35], [45], [82] for the classical theory. In [68], a family of quantum scalar fields over a $p-$adic spacetime which satisfy $p-$adic analogues of the Gårding–Wightman axioms was constructed. In [5], a large class of interacting Euclidean quantum field theories was constructed by using white noise calculus. These quantum fields fulfill all the Osterwalder-Schrader axioms, except the reflection positivity. In [92], the author constructs, in a rigorous mathematical way, interacting Euclidean quantum field theories on a p-adic spacetime. The main result is the construction of a measure on a function space which allows a rigorous definition of the partition function. The advantage of the approach presented is that all the perturbation calculations can be carried out in the standard way using functional derivatives, but in a mathematically rigorous way. In [1] Abdesselam et al. present the construction of scale invariant non-Gaussian generalized stochastic processes over three dimensional p-adic space. The construction includes that of the associated squared field, this field has a dynamically generated anomalous dimension which rigorously confirms a prediction made more than forty years ago by K. G. Wilson. Traditionally the p-adic QFTs has been considered just mathematical toy models. In this article, we show that these theories are deeply connected with hierarchical versions of RBMs, and then with deep learning.

An Euclidean quantum field theory is a probability measure of the form

$$d\mathbb{P}(\varphi)=\frac{e^{-E(\varphi)}d\mathbb{P}_0(\varphi)}{\int_H e^{-E(\varphi)}d\mathbb{P}_0(\varphi)}$$

on a space H of functions $\varphi:\mathbb{Q}_p^N\to\mathbb{R}$, where $\mathbb{P}_0$ is a Gaussian measure on H. For the sake of simplicity, along this article we assume that $N=1$. By a discretization process, which consists in finding the restriction of $\mathbb{P}$ to a suitable finite dimensional vector subspace H_l of H, one obtains a discrete energy functional E_l and $\mathbb{P}_l$ a finite dimensional Boltzmann distribution for $l\geq l_0$, such that $\mathbb{P}_l\to\mathbb{P}$ in some sense, see [92] and the references therein. The discrete energy functional of a ϕ^4-theory has the form

$$E_l(\phi)=\sum_{i,j\in G_l}\phi_i w_{i,j}^l\phi_j+\sum_{i\in G_l}a_i^l\phi_i+\sum_{i\in G_l}b_i^l\phi_i^2+\sum_{i\in G_l}b_i^l\phi_i^4,$$

where $G_l=\mathbb{Z}_p/p^l\mathbb{Z}_p\simeq\mathbb{Z}/p^l\mathbb{Z}$ is the additive group of the integers modulo p^l, and $\phi=[\phi_i]_{i\in G_l}$.

We identify the elements of G_l with integers of the form $i=i_0+i_1p+\ldots+i_{l-1}p^{l-1}$. The restriction of $|\cdot|_p$ to G_l induces an norm, and thus G_l is a finite ultrametric space. In addition, G_l can be identified with the set of branches (vertices at the top level) of a rooted tree with $l+1$ levels (or layers) and p^l branches.

A p-adic discrete deep belief network is a discrete Euclidean QTF defined by an energy functional of the form

$$E_l(\boldsymbol{v}_l,\boldsymbol{h}_l;\boldsymbol{\theta}_l)=-\sum_{j\in G_l}\sum_{k\in G_l}w_{k,j}^l v_{\boldsymbol{k}}^l h_{\boldsymbol{j}}^l-\sum_{j\in G_l}a_j^l v_j^l-\sum_{j\in G_l}b_j^l h_j^l,$$

where $\boldsymbol{v}_l=\left[v_k^l\right]_{k\in G_l}$ is the state of the visible field, $\boldsymbol{h}_l=\left[h_k^l\right]_{k\in G_l}$ is the state of the hidden field, and $\boldsymbol{\theta}=(w_{k,j},a_j,b_j)$. We assume that the fields $\boldsymbol{v}_l$, $\boldsymbol{h}_l$ are binary-valued, i.e. $\boldsymbol{v}_l$, $\boldsymbol{h}_l\in\{0,1\}^{\#G_l}$, where $\#G_l$ is the cardinality of G_l. The corresponding Boltzmann distribution is given by

$$\boldsymbol{P}_l(\boldsymbol{v}_l,\boldsymbol{h}_l)=\frac{\exp\left(-E_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l\right)\right)}{\sum\limits_{\boldsymbol{v}_l',\boldsymbol{h}_l'}\exp\left(-E_l\left(\boldsymbol{v}_l',\boldsymbol{h}_l'\right)\right)}.$$

If the entries of the matrix $\left[w_{k,j}^l\right]$ do not depend on the topology of G_l neither on the group structure of G_l, then $E_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l;\boldsymbol{\theta}_l\right)$ defines a standard RBM. If the entries of the matrix $[w_{k,j}]$ depend on topology of G_l and on the group structure of G_l, then $[w_{k,j}]$ is a Parisi matrix, see e.g. [23] and the references therein. In this article we assume that $w_{k,j}^l=w\left(|k-j|_p\right)$, in this case $E_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l;\boldsymbol{\theta}_l\right)$ is the energy functional of a new non-Archimedean convolutional deep belief network (DBN), see e.g. [27], [44]. Since G_l is an additive group, the weight $w_{k,j}^l$ depends on one parameter $k-j\in G_l$. More generally, E_l has a translational symmetry: $k\to k+i_0$, $j\to j+i_0$. We denote by $DBN(p,l,\boldsymbol{\theta}_l)$ the p-adic discrete deep belief network attached to $E_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l;\boldsymbol{\theta}_l\right)$.

Two fundamental questions come up immediately: can the p-adic convolutional DBNs perform computations?; does the computational power of the p-adic convolutional DBNs increase as the number of levels of the G_l tree increases?. The answers to both questions is yes, see Theorems 1, 2. These theorems constitute the main result of this article. By identifying G_l with a subset of G_{l+1}, given an $DBN(p,l,\boldsymbol{\theta}_l)$ we construct another $DBN(p,l+1,\boldsymbol{\theta}_l,\boldsymbol{w}_{l+1},b_{j_0}^{l+1})$, here $\boldsymbol{w}_{l+1}\in\mathbb{R}^{\#G_l}$, $b_{j_0}^{l+1}\in\mathbb{R}$, with an extra layer and an extra hidden unit, with the same visible units, whose energy functional $E_{l+1}(\boldsymbol{v}_{l+1},\boldsymbol{h}_{l+1};\boldsymbol{\theta}_l,\boldsymbol{w}_{l+1},b_{j_0}^{l+1})=E_{l+1}(\boldsymbol{v}_l,\boldsymbol{h}_{l+1};\boldsymbol{\theta}_l,\boldsymbol{w}_{l+1},b_{j_0}^{l+1})$ is an extension of $E_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l;\boldsymbol{\theta}_l\right)$, here $\boldsymbol{h}_{l+1}=\left(\boldsymbol{h}_l,h_{j_0}^{l+1}\right)$, and $h_{j_0}^{l+1}$ is the extra hidden unit. This construction allow us to adapt the mathematical techniques introduced by Le Roux and Benigio in [63]. Our Theorems 1, 2 are non-Archimedean counterparts of the Theorems 1, 2 in [63].

An ultrametric space $(\mathcal{M},d)$ is a metric space $\mathcal{M}$ with a distance satisfying the strong triangle inequality $d(A,B)\leq\max\left\{d\left(A,C\right),d\left(B,C\right)\right\}$ for any three points A, B, C in $\mathcal{M}$. The field of p-adic numbers $\mathbb{Q}_p$ constitutes a central example of an ultrametric space. The ultrametricity, which is the emergence of ultrametric spaces in physical models, was discovered in the middle 1980s by Parisi et al. in the context of the spin glass theory, see e.g. [75], [80]. Ultrametric spaces constitute the right framework to formulate models where hierarchy plays a central role. Ultrametric models have been applied in many areas, including, brain and mental states models, relaxation of complex systems, spin glasses, evolutionary dynamics, among other areas, see e.g. [6]-[7], [25], [49]-[54], [77]-[80], [87], [92]-[98], and the references therein.

The Ising models over ultrametric spaces have been studied intensively, see e.g. [29], [39], [53], [62], [74], [77]-[79], [83], [95]-[96] and the references therein. An important motivation comes from the hierarchical Ising model introduced in [29]. The hierarchical Hamiltonian introduced by Dyson in [29] can be naturally studied in p-adic spaces, see e.g. [62], [39]. In [79], see also [54], Parisi and Sourlas presented a p-adic formulation of replica symmetry breaking. In this approach ultrametricity is a natural consequence of the topology of the p-adic numbers.

In [43], see also [44], [27], Hinton et al. introduced the deep belief networks (DBNs), which are multilayer hierarchical generative models constructed by stacking RBMs. A binary RBM is a spin glass, then an DBN must correspond to a hierarchical (ultrametric) spin glass. The simplest way of constructing a hierarchical spin glass is assuming that the space of states has a tree-like structure, such structures can be realized in an easy way by using p-adic numbers. In our view, this is a new approach to understanding deep learning architectures.

The article is organized as follows. In Section 2, we review the basic aspects of the p-adic analysis. In Section 3, we introduce the p-adic RBMs and their discretizations. In Section 4, we show that the discrete p-adic RBMs are universal approximators. Finally, in the last section, we present a discussion of our results compared with other related work, also we propose several open problems.

2. Basic facts on p-adic analysis

In this section we fix the notation and collect some basic results on p-adic analysis that we will use through the article. For a detailed exposition on p-adic analysis the reader may consult [3], [86], [87]. For a quick review of p-adic analysis the reader may consult [12], [61].

2.1. **The field of p-adic numbers.** Throughout this article p will denote a prime number. The field of $p-$adic numbers $\mathbb{Q}_p$ is defined as the completion of the field of rational numbers $\mathbb{Q}$ with respect to the $p-$adic norm $|\cdot|_p$, which is defined as

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

where a and b are integers coprime with p. The integer $\gamma = ord_p(x)$ with $ord_p(0) := +\infty$, is called the *$p-$adic order of x*. The metric space $\left(\mathbb{Q}_p, |\cdot|_p\right)$ is a complete ultrametric space. Ultrametric means that $|x+y|_p \leq \max\left\{|x|_p, |y|_p\right\}$. As a topological space $\mathbb{Q}_p$ is homeomorphic to a Cantor-like subset of the real line, see e.g. [3], [87].

Any $p-$adic number $x \neq 0$ has a unique expansion of the form

$$x = p^{ord_p(x)} \sum_{j=0}^{\infty} x_j p^j, \tag{2.1}$$

where $x_j \in \{0, 1, 2, \dots, p-1\}$ and $x_0 \neq 0$. It follows from (2.1), that any $x \in \mathbb{Q}_p \smallsetminus \{0\}$ can be represented uniquely as $x = p^{ord_p(x)} u(x)$ and $|x|_p = p^{-ord_p(x)}$.

2.2. **Topology of $\mathbb{Q}_p$.** For $r \in \mathbb{Z}$, denote by $B_r(a) = \{x \in \mathbb{Q}_p; |x-a|_p \leq p^r\}$ *the ball of radius p^r with center at $a \in \mathbb{Q}_p$*, and take $B_r(0) := B_r$. The ball B_0 equals *the ring of $p-$adic integers $\mathbb{Z}_p$*. We also denote by $S_r(a) = \{x \in \mathbb{Q}_p; |x-a|_p = p^r\}$ *the sphere of radius p^r with center at $a \in \mathbb{Q}_p$*, and take $S_r(0) := S_r$. We notice that $S_0^1 = \mathbb{Z}_p^{\times}$ (the group of units of $\mathbb{Z}_p$). The balls and spheres are both open and closed subsets in $\mathbb{Q}_p$. In addition, two balls in $\mathbb{Q}_p$ are either disjoint or one is contained in the other.

As a topological space $\left(\mathbb{Q}_p, |\cdot|_p\right)$ is totally disconnected, i.e. the only connected subsets of $\mathbb{Q}_p$ are the empty set and the points. A subset of $\mathbb{Q}_p$ is compact if and only if it is closed and bounded in $\mathbb{Q}_p$, see e.g. [87, Section 1.3], or [3, Section 1.8].

The balls and spheres are compact subsets. Thus $\left(\mathbb{Q}_p, |\cdot|_p\right)$ is a locally compact topological space.

Since $(\mathbb{Q}_p, +)$ is a locally compact topological group, there exists a Haar measure dx, which is invariant under translations, i.e. $d(x+a) = dx$. If we normalize this measure by the condition $\int_{\mathbb{Z}_p} dx = 1$, then dx is unique. In a few occasions we use the two-dimensional Haar measure $dxdy$ of the additive group $(\mathbb{Q}_p \times \mathbb{Q}_p, +)$ normalized by the condition $\int_{\mathbb{Z}_p}\int_{\mathbb{Z}_p} dxdy = 1$. For a quick review of the integration in the p-adic framework the reader may consult [12], [61] and the references therein.

Notation 1. *We will use* $\Omega\left(p^{-r}|x-a|_p\right)$ *to denote the characteristic function of the ball* $B_r(a)$.

2.3. **The Bruhat-Schwartz space.** A real-valued function φ defined on $\mathbb{Q}_p$ is *called locally constant* if for any $x \in \mathbb{Q}_p$ there exist an integer $l(x) \in \mathbb{Z}$ such that

$$\varphi(x + x') = \varphi(x) \text{ for any } x' \in B_{l(x)}. \tag{2.2}$$

A function $\varphi : \mathbb{Q}_p \to \mathbb{C}$ 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 $\mathcal{D}(\mathbb{Q}_p)$. For $\varphi \in \mathcal{D}(\mathbb{Q}_p)$, the largest number $l = l(\varphi)$ satisfying (2.2) is called *the exponent of local constancy (or the parameter of constancy) of* φ. Let U be an open subset of $\mathbb{Q}_p$, we denote by $\mathcal{D}(U)$ the $\mathbb{R}$-vector space of all test functions with support in U. For instance $\mathcal{D}(\mathbb{Z}_p)$ is the $\mathbb{R}$-vector space of all test functions with supported in the unit ball $\mathbb{Z}_p$. A function φ in $\mathcal{D}(\mathbb{Z}_p)$ can be written as

$$\varphi(x) = \sum_{j=1}^{M} \varphi(\widetilde{x}_j)\, \Omega\left(p^{r_j}|x - \widetilde{x}_j|_p\right),$$

where the $\widetilde{x}_j$, $j = 1, \ldots, M$, are points in $\mathbb{Z}_p$, the r_j, $j = 1, \ldots, M$, are integers, and $\Omega\left(p^{r_j}|x-\widetilde{x}_j|_p\right)$ denotes the characteristic function of the ball $B_{-r_j}(\widetilde{x}_j) = \widetilde{x}_j + p^{r_j}\mathbb{Z}_p$.

3. A class of non-Archimedean statistical field theories and their discretizations

In this section we introduce a family of p-adic SFTs that we require in this article. A central difference between the classical SFTs and the non-Archimedean counterparts is that the discretization process of the non-Archimedean ones is very simple, and the convergence of the discrete theories to continuous theories can be formulated in rigorous mathematical way, in a large number of cases, see e.g. [92] and the references therein. For a mathematical exposition of the non-Archimedean ϕ^4-QFTs, the reader may consult [92], see also [5], and the references therein.

3.1. **A class of non-Archimedean statistical field theories.** We fix $a(x)$, $b(x) \in \mathcal{D}(\mathbb{Z}_p)$, and a function $w(x,y) : \mathbb{Z}_p \times \mathbb{Z}_p \to \mathbb{R}$. In this article, the function $w(x,y)$ can be a test function or a radial function $w(x,y) = w\left(|x-y|_p\right)$. A *p-adic continuous deep belif network (or a p-adic continuous DBN)* is a statistical field theory $\{\boldsymbol{v}, \boldsymbol{h}\}$ in $\mathcal{D}(\mathbb{Z}_p)$. The function $\boldsymbol{v}(x) \in \mathcal{D}(\mathbb{Z}_p)$ is called the *visible field* and the function $\boldsymbol{h}(x) \in \mathcal{D}(\mathbb{Z}_p)$ is called the *hidden field*. The field $\{\boldsymbol{v}, \boldsymbol{h}\}$ performs thermal fluctuations, assuming that the expectation value of the field is zero, the fluctuations take place around zero. The size of the fluctuations is controlled by an energy functional (or action) of the form

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\iint\limits_{\mathbb{Z}_p \times \mathbb{Z}_p} \boldsymbol{h}(y)\, w(x,y)\, \boldsymbol{v}(x)\, dxdy - \int\limits_{\mathbb{Z}_p} a(x)\boldsymbol{v}(x)\, dx - \int\limits_{\mathbb{Z}_p} b(x)\boldsymbol{h}(x)\, dx. \quad (3.1)$$

Along the article, we assume that the fields $\boldsymbol{v}$, $\boldsymbol{h}$ are binary valued functions, i.e. $\boldsymbol{v}$, $\boldsymbol{h} : \mathbb{Z}_p \rightarrow \{0,1\} \subset \mathbb{R}$, then $E(\boldsymbol{v}, \boldsymbol{h})$ is the energy functional of a p-adic continuous spin glass.

All thermodynamic properties of the system are described by the partition function of the fluctuating field, which is defined as

$$Z^{\text{phys}} = \int d\boldsymbol{v} d\boldsymbol{h}\; e^{-\frac{E(\boldsymbol{v},\boldsymbol{h})}{K_B T}},$$

where K_B is the Boltzmann constant and T is the temperature. We normalize in such a way that $K_B T = 1$. The measure $d\boldsymbol{v}d\boldsymbol{h}$ is ill-defined. It is expected that such measure can be defined rigorously by a limit process.

The statistical field theory corresponding to the energy functional (3.1) is the ill-defined probability measure

$$\boldsymbol{P}^{\text{phys}}(\boldsymbol{v}, \boldsymbol{h}) = d\boldsymbol{v}d\boldsymbol{h} \frac{\exp\left(-E(\boldsymbol{v}, \boldsymbol{h})\right)}{Z^{\text{phys}}}$$

on the space of functions $\mathcal{D}(\mathbb{Z}_p) \times \mathcal{D}(\mathbb{Z}_p)$.

The information about the local properties of the system is contained in the *correlation functions* $G^{(n)}_{\mathbb{I},\mathbb{K}}(x_1, \ldots, x_n)$ of the field $\{\boldsymbol{v}, \boldsymbol{h}\}$: for $n \geq 1$, and two disjoint subsets $\mathbb{I}$, $\mathbb{K} \subset \{1, 2, \ldots, n\}$, with $\mathbb{I}\coprod\mathbb{K} = \{1, 2, \ldots, n\}$, we set

$$\begin{aligned} G^{(n)}_{\mathbb{I},\mathbb{J}}(x_1, \ldots, x_n) &= \left\langle \prod_{i \in \mathbb{I}} \boldsymbol{v}(x_i) \prod_{j \in \mathbb{K}} \boldsymbol{h}(x_j) \right\rangle \\ &= \frac{1}{Z^{\text{phys}}} \int d\boldsymbol{v}d\boldsymbol{h} \prod_{i \in \mathbb{I}} \boldsymbol{v}(x_i) \prod_{j \in \mathbb{K}} \boldsymbol{h}(x_j)\; e^{-E(\boldsymbol{v},\boldsymbol{h})}. \end{aligned}$$

These functions are also called the n-point Green functions. To study of these functions, one introduces two auxiliary external fields $J_0(x)$, $J_1(x) \in \mathcal{D}(\mathbb{Z}_p)$ called *currents,* and adds to the energy functional E as a linear interaction energy of these currents with the field $\{\boldsymbol{v}, \boldsymbol{h}\}$,

$$E_{\text{source}}(\boldsymbol{v}, \boldsymbol{h}, J_0, J_1) = -\int\limits_{\mathbb{Z}_p} J_0(x)\boldsymbol{v}(x)\, dx - \int\limits_{\mathbb{Z}_p} J_1(x)\boldsymbol{h}(x)\, dx,$$

and the energy functional is $E(\boldsymbol{v}, \boldsymbol{h}, J_0, J_1) = E(\boldsymbol{v}, \boldsymbol{h}) + E_{\text{source}}(\boldsymbol{v}, \boldsymbol{h}, J_0, J_1)$. The partition function formed with this energy is

$$Z(J_0, J_1) = \frac{1}{Z^{\text{phys}}} \int d\boldsymbol{v}d\boldsymbol{h}\; e^{-E(\boldsymbol{v},\boldsymbol{h},J_0,J_1)}.$$

The functional derivatives of $Z(J_0,J_1)$ with respect to $J_0(x)$, $J_1(x)$ evaluated at $J_0=0$, $J_1=0$ give the correlation functions of the system:

$$G_{\mathbb{I},\mathbb{K}}^{(n)}\left(x_1,\ldots,x_n\right)=\left[\prod_{i\in\mathbb{I}}\frac{\delta}{\delta J_0\left(x_i\right)}\prod_{j\in\mathbb{K}}\frac{\delta}{\delta J_1\left(x_j\right)}Z(J_0,J_1)\right]_{\substack{J_0=0\\J_1=0}}.$$

The functional $Z(J_0,J_1)$ is called the *generating functional* of the theory.

3.2. **Discretization of the energy functional.** For $l\geq 1$, we set $G_l:=\mathbb{Z}_p/p^l\mathbb{Z}_p$. We use a fixed system of representatives of the form

$$i=i_0+i_1p+\ldots+i_{l-1}p^{l-1},$$

where the i_ks are p-adic digits, for the elements of G_l. We denote by $\mathcal{D}^l(\mathbb{Z}_p)$ the $\mathbb{R}$-vector space of all test functions of the form

$$\varphi\left(x\right)=\sum_{i\in G_l}\varphi\left(i\right)\Omega\left(p^l\left|x-i\right|_p\right),\ \ \varphi\left(i\right)\in\mathbb{R},\tag{3.2}$$

where $\Omega\left(p^l\left|x-i\right|_p\right)$ denotes the characteristic function of the ball $i+p^l\mathbb{Z}_p$. Notice that φ is supported on $\mathbb{Z}_p$ and that $\mathcal{D}^l(\mathbb{Z}_p)$ is a finite dimensional vector space spanned by the basis

$$\left\{\Omega\left(p^l\left|x-i\right|_p\right)\right\}_{i\in G_l}.\tag{3.3}$$

By identifying $\varphi\in\mathcal{D}^l(\mathbb{Z}_p)$ with the column vector $\left[\varphi\left(i\right)\right]_{\boldsymbol{i}\in G_l}\in\mathbb{R}^{\#G_l}$, we get that $\mathcal{D}^l(\mathbb{Z}_p)$ is isomorphic to $\mathbb{R}^{\#G_l}$ endowed with the norm $\left\Vert\left[\varphi\left(i\right)\right]_{\boldsymbol{i}\in G_l^N}\right\Vert=\max_{i\in G_l}\left|\varphi\left(i\right)\right|$. Furthermore,

$$\mathcal{D}^l(\mathbb{Z}_p)\hookrightarrow\mathcal{D}^{l+1}(\mathbb{Z}_p)\hookrightarrow\mathcal{D}(\mathbb{Z}_p),$$

where $\hookrightarrow$ denotes a continuous embedding.

A discretization E_l of the energy functional E is obtained by restricting $\boldsymbol{v},\boldsymbol{h}$ to $\mathcal{D}^l(\mathbb{Z}_p)$, i.e. by taking

$$\boldsymbol{v}\left(x\right)=\sum_{i\in G_l}\boldsymbol{v}\left(i\right)\Omega\left(p^l\left|x-i\right|_p\right),\ \ \boldsymbol{h}\left(x\right)=\sum_{i\in G_l}\boldsymbol{h}\left(i\right)\Omega\left(p^l\left|x-i\right|_p\right).$$

When $\boldsymbol{v},\boldsymbol{h}\in\mathcal{D}^l(\mathbb{Z}_p)$, we use use the following identifications:

$$\boldsymbol{v}_l=\left[\boldsymbol{v}\left(i\right)\right]_{i\in G_l},\ \boldsymbol{h}_l=\left[\boldsymbol{h}\left(i\right)\right]_{i\in G_l}.$$

There are two different types of discrete functionals E_l according if $w(x,y)$ is a test function or a radial function $w(x,y)=w\left(\left|x-y\right|_p\right)$.

3.3. **Standard restricted Boltzmann machines.** Assume that $w(x,y)$ is a test function. Since $w(x,y)$ is locally constant,

$$w(x,y)\Omega\left(p^l\left|x-i\right|_p\right)\Omega\left(p^l\left|x-j\right|_p\right)=w(i,j)\Omega\left(p^l\left|x-i\right|_p\right)\Omega\left(p^l\left|x-j\right|_p\right),$$

for l sufficiently large, and

$$\iint\limits_{\mathbb{Z}_p\times\mathbb{Z}_p} w(x,y)\,\Omega\left(p^l\left|x-i\right|_p\right)\Omega\left(p^l\left|y-j\right|_p\right)dxdy$$
$$=w(i,j)\left(\int\limits_{i+p^l\mathbb{Z}_p} dx\right)\left(\int\limits_{j+p^l\mathbb{Z}_p} dy\right)=p^{-2l}w(i,j).$$

By a similar argument, for l sufficiently large, we have

$$\int\limits_{\mathbb{Z}_p} a(x)\boldsymbol{v}(x)dx=p^{-l}\sum_{\boldsymbol{i}\in G_l}a(i)\boldsymbol{v}\left(i\right),$$
$$\int\limits_{\mathbb{Z}_p} b(x)\boldsymbol{h}(x)dx=p^{-l}\sum_{\boldsymbol{i}\in G_l}b(i)\boldsymbol{h}\left(i\right).$$

Therefore, for l sufficiently large,

$$E_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l\right)=-p^{-2l}\sum_{i,\ j\in G_l}\boldsymbol{v}\left(i\right)w\left(i,j\right)\boldsymbol{h}\left(j\right)-p^{-l}\sum_{i\in G_l}a(i)\boldsymbol{v}\left(i\right)-p^{-l}\sum_{\boldsymbol{i}\in G_l}b(i)\boldsymbol{h}\left(i\right).$$

By taking

$$v_i^l:=\boldsymbol{v}\left(i\right),\ h_i^l:=\boldsymbol{h}\left(i\right),\ w_{i,j}^l:=p^{-2l}w\left(i,j\right),\ \ a_i^l:=p^{-l}a(i),\ b_i^l:=p^{-l}b(i),$$

$$\boldsymbol{v}_l=\left[v_i^l\right]_{i\in G_l},\ \boldsymbol{h}_l=\left[h_i^l\right]_{i\in G_l},\ \boldsymbol{w}_l=\left[w_{i,j}^l\right]_{i,j\in G_l},\ \boldsymbol{a}_l=\left[a_i^l\right]_{\boldsymbol{i}\in G_l},\ \boldsymbol{b}_l=\left[b_i^l\right]_{\boldsymbol{i}\in G_l}$$

we have

$$E_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l\right)=-\sum_{i,\ j\in G_l}v_i^l w_{i,j}^l h_j^l-\sum_{i\in G_l}a_i^l v_i^l-\sum_{i\in G_l}b_i^l h_i^l, \tag{3.4}$$

which is the energy functional of a standard restricted Boltzman machine. Here it is very relevant to notice that the energy functional $E_l\left(\boldsymbol{v},\boldsymbol{h}\right)$ does not depend on the topology of the metric space $\left(G_l,\left|\cdot\right|_p\right)$ neither on the group structure $(G_l,+)$. The Boltzmann distribution is given by

$$\boldsymbol{P}_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l\right)=\frac{\exp\left(-E_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l\right)\right)}{Z_l},$$

where $Z_l=\sum_{\boldsymbol{v}_l,\boldsymbol{h}_l}\exp\left(-E_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l\right)\right)$.

Now, any standard RBM with n visible nodes v_i, $i=1,\ldots,n$ and m hidden nodes $h_{\boldsymbol{i}}$, $i=1,\ldots,m$ can be realized as p-adic discrete RBM of type (3.4), by choosing p and l satisfying $n\leq p^l$, $m\leq p^l$ and taking $v_i^l=v_i$ for $1\leq i\leq n$, $v_i^l=0$ for $n+1\leq i\leq p^l$, and $h_i^l=h_i$ or $1\leq i\leq m$, $h_i^l=0$ for $m+1\leq i\leq p^l$.

The number of the $w_{i,j}^l$ parameters is $\left(\#G_l\right)^2$, the number of the a_i parameters is $\#G_l$, and the number of the h_i parameters is $\#G_l$, and consequently the total number of parameters is

$$\left(\#G_l\right)^2+2\left(\#G_l\right),$$

which is quadratic in the cardinality of G_l.

3.4. **p-adic discrete deep belief networks.** We now consider the case in which $w(x,y) = w\left(|x-y|_p\right)$ is a radial function. Notice that in this case $w : \mathbb{R}_+ \to \mathbb{R}$. The energy functional $E_l(\boldsymbol{v}_l, \boldsymbol{h}_l)$ depends on the topology of the metric space $\left(G_l, |\cdot|_p\right)$ and on the group structure $(G_l, +)$.

We first notice that

$$I_{i,j}(w) := \iint\limits_{\mathbb{Z}_p\times\mathbb{Z}_p} w\left(|x-y|_p\right)\Omega\left(p^l|x-i|_p\right)\Omega\left(p^l|y-j|_p\right)dxdy \tag{3.5}$$

$$= \begin{cases} p^{-2l}w\left(|i-j|_p\right) & \text{if } \ i\neq j \\ \\ p^{-l}\left(\int\limits_{p^l\mathbb{Z}_p} w\left(|z|_p\right)dz\right) & \text{if } i=j. \end{cases}$$

This identity follows by performing two changes of variables as follows. By changing variables as $x = i + x'$, $y = j + y'$, with x', $y' \in p^l\mathbb{Z}_p$, and $dxdy = dx'dy'$, we have

$$I_{i,j}(w) = \int\limits_{p^l\mathbb{Z}_p}\int\limits_{p^l\mathbb{Z}_p} w\left(|i-j+x'-y'|_p\right)dx'dy'.$$

If $i-j\neq 0$, then $i-j = a_jp^j+\ldots$ with $a_j \neq 0$ and $j<l$, since $x'-y' = b_lp^l+\ldots$ with $b_l \in \{0,\ldots,p-1\}$, it verifies that $|i-j+x'-y'|_p = |i-j|_p$. The first case in formula (3.5) follows from the fact that

$$\int\limits_{p^l\mathbb{Z}_p}\int\limits_{p^l\mathbb{Z}_p} dx'dy' = p^{-2l}.$$

To establish the second case in formula (3.5), we use $i-j=0$ and change variables as $z = x'-y'$, $t = y'$ with z, $t \in p^l\mathbb{Z}_p$. This change of variables preserve the Haar measure $dx'dy'$ of the additive group $p^l\mathbb{Z}_p \times p^l\mathbb{Z}_p$, i.e. $dzdt = dx'dy'$. This follows by using the general formula for changing of variables in p-adic integrals.

Now, by using that $a(x)$, $b(x)$ are test functions supported in the unit ball, and taking l sufficiently large, we have

$$a(x)\,\Omega\left(p^l|x-i|_p\right) = a(i)\,\Omega\left(p^l|x-i|_p\right),$$
$$b(x)\Omega\left(p^l|x-i|_p\right) = b(i)\Omega\left(p^l|x-i|_p\right),$$

and consequently

$$E_l(\boldsymbol{v}_l,\boldsymbol{h}_l) = -p^{-2l}\sum_{\substack{i,\ j\in G_l\\ i\neq j}} \boldsymbol{v}(i)\,w\left(|i-j|_p\right)\boldsymbol{h}(j)$$

$$-p^{-l}\left(\int\limits_{p^l\mathbb{Z}_p} w\left(|z|_p\right)dz\right)\sum_{i\in G_l}\boldsymbol{v}(i)\,\boldsymbol{h}(i) - p^{-l}\sum_{i\in G_l} a(i)\,\boldsymbol{v}(i) - p^{-l}\sum_{i\in G_l} b(i)\,\boldsymbol{h}(i)\,.$$

By taking $v_i^l = \boldsymbol{v}(i)$, $h_i^l = \boldsymbol{h}(i)$,

$$w_{\boldsymbol{i}-\boldsymbol{j}}^l = \begin{cases} p^{-2l} w\left(|i-\boldsymbol{j}|_p\right) & \text{if } \ i \neq j \\ \\ p^{-l}\left(\int\limits_{p^l\mathbb{Z}_p} w\left(|z|_p\right) dz\right) & \text{if } i = j. \end{cases}$$

$a_i^l = p^{-l} a(i)$, $b_i^l = p^{-l} b(i)$, for $i,\ j \in G_l^N$, and $\boldsymbol{\theta}_l = \left\{w_{i,j}^l, a_i^l, b_i^l\right\}$. Then, for l sufficiently large,

$$E_l\left(\boldsymbol{v}_l, \boldsymbol{h}_l; \boldsymbol{\theta}_l\right) = -\sum_{i,\ j \in G_l} w_{i-j}^l v_i^l h_j^l - \sum_{i \in G_l} a_i^l v_i^l - \sum_{i \in G_l} b_i^l h_i^l \tag{3.6}$$

Since $(G_l, +)$ is an additive group and $w_{i-j}^l = w_{j-i}^l$, then

$$\sum_{i,\ j \in G_l} w_{i-j}^l v_i^l h_j^l = \sum_{\boldsymbol{i},\ \boldsymbol{j} \in G_l} w_{j-i}^l v_i^l h_j^l =$$
$$\sum_{j,k \in G_l} w_k^l v_{j+k}^l h_j^l = \sum_{i,k \in G_l} w_k^l v_i^l h_{k+i}^l,$$

and consequently, for l sufficiently large (i.e. for $l \geq L$),

$$E_l\left(\boldsymbol{v}_l, \boldsymbol{h}_l; \boldsymbol{\theta}_l\right) = -\sum_{j \in G_l} \sum_{k \in G_l} w_k^l v_{j+k}^l h_j^l - \sum_{j \in G_l} a_j^l \boldsymbol{v}_j^l - \sum_{j \in G_l} b_j^l h_j^l. \tag{3.7}$$

The total number of parameters of this type of networks is

$$3\left(\#G_l\right),$$

which is linear in the cardinality of G_l.

3.4.1. *Boltzmann probability distributions.* From now on, we set $\boldsymbol{v}_l = \left(v_i^l\right)_{i \in G_l}$, $\boldsymbol{h}_l = \left(h_i^l\right)_{i \in G_l}$, $\boldsymbol{w}_l = \left(w_i^l\right)_{i \in G_l}$, $\boldsymbol{a}_l = \left(a_i^l\right)_{i \in G_l}$, $\boldsymbol{b}_l = \left(b_i^l\right)_{i \in G_l}$, $\boldsymbol{\theta}_l = (\boldsymbol{w}_l, \boldsymbol{a}_l, \boldsymbol{b}_l)$. We warn the reader that, for the sake of simplicity, the dependence of the $\boldsymbol{\theta}_l$ parameters is omitted in most of the formulas. We associate to $E_l\left(\boldsymbol{v}_l, \boldsymbol{h}_l; \boldsymbol{\theta}_l\right)$ the Boltzmann probability distribution

$$\boldsymbol{P}_l(\boldsymbol{v}_l, \boldsymbol{h}_l) = \frac{\exp\left(-E_l\left(\boldsymbol{v}_l, \boldsymbol{h}_l\right)\right)}{Z_l}, \tag{3.8}$$

where

$$Z_l = \sum_{\boldsymbol{v}_l, \boldsymbol{h}_l} \exp\left(-E_l\left(\boldsymbol{v}_l, \boldsymbol{h}_l\right)\right) = \sum_{i,j \in G_l} \exp\left(-E_l\left(\left(v_i^l\right)_{i \in G_l},\ \left(h_j^l\right)_{j \in G_l}\right)\right).$$

It is expected that the limit

$$\frac{e^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z^{\text{phys}}} d\boldsymbol{v} d\boldsymbol{h} \overset{\text{def}}{=} \lim_{l \to \infty} \boldsymbol{P}_l(\boldsymbol{v}_l, \boldsymbol{h}_l)\ d^{\#G_l}\boldsymbol{v}\ d^{\#G_l}\boldsymbol{h}$$

exists in some sense.

The marginal probability distributions are given by

$$\boldsymbol{P}_l(\boldsymbol{v}_l) = \sum_{\boldsymbol{h}_l} \boldsymbol{P}_l(\boldsymbol{v}_l, \boldsymbol{h}_l) = \frac{\sum\limits_{\boldsymbol{h}_l} \exp\left(-E_l\left(\boldsymbol{v}_l, \boldsymbol{h}_l\right)\right)}{\sum\limits_{\boldsymbol{v}_l, \boldsymbol{h}_l} \exp\left(-E_l\left(\boldsymbol{v}_l, \boldsymbol{h}_l\right)\right)}, \tag{3.9}$$

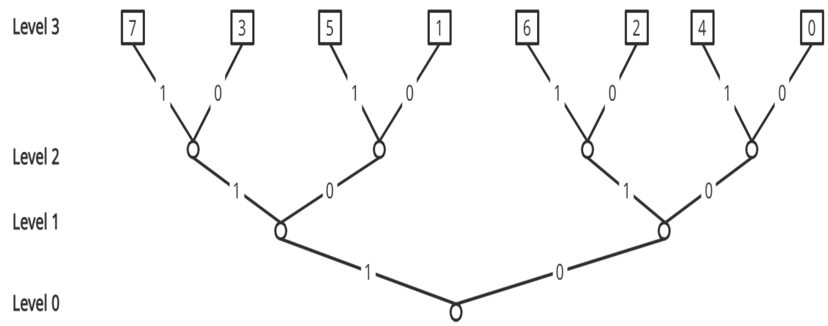

FIGURE 1. *The rooted tree associated with the group $\mathbb{Z}_2/2^3\mathbb{Z}_2$. We identify the elements of $\mathbb{Z}_2/2^3\mathbb{Z}_2$ with the set of integers $\{0,\dots,7\}$ with binary representation $\boldsymbol{i}=\boldsymbol{i}_0+\boldsymbol{i}_1 2+\boldsymbol{i}_3 2^2$, $\boldsymbol{i}_0,\boldsymbol{i}_1,\boldsymbol{i}_2\in\{0,1\}$. Two leaves $\boldsymbol{i},\boldsymbol{j}\in\mathbb{Z}_2/2^3\mathbb{Z}_2$ have a common ancestor at level 2 if and only if $\boldsymbol{i}\equiv\boldsymbol{j} \bmod 2^2$, i.e., $\boldsymbol{i}=\boldsymbol{a}_0+\boldsymbol{a}_1 2+\boldsymbol{i}_2 2^2$ and $\boldsymbol{j}=\boldsymbol{a}_0+\boldsymbol{a}_1 2+\boldsymbol{j}_2 2^2$ with $\boldsymbol{i}_2,\boldsymbol{j}_2\in\{0,1\}$. Now, for $\boldsymbol{i},\boldsymbol{j}\in\mathbb{Z}_2/2^3\mathbb{Z}_2$ have a common ancestor at level 1 if and only if $\boldsymbol{i}\equiv\boldsymbol{j} \bmod 2$. Notice that that the p-adic distance satisfies $-\log_2|\boldsymbol{i}-\boldsymbol{j}|_2=-$(level of the first common ancestor of $\boldsymbol{i}$, $\boldsymbol{j}$). Reprinted from* [91].

$$\boldsymbol{P}_l(\boldsymbol{h}_l)=\sum_{\boldsymbol{v}_l}\boldsymbol{P}_l(\boldsymbol{v}_l,\boldsymbol{h}_l)=\frac{\sum_{\boldsymbol{v}_l}\exp\left(-E_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l\right)\right)}{\sum_{\boldsymbol{v}_l,\boldsymbol{h}_l}\exp\left(-E_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l\right)\right)}.$$

3.4.2. *Tree-like structures, p-adic numbers and DBNs.* The restriction of $|\cdot|_p$ to G_l induces an absolute value and $|G_l|_p=\left\{0,p^{-(l-1)},\cdots,p^{-1},1\right\}$. We endow G_l with the metric induced by $|\cdot|_p$, and thus G_l becomes a finite ultrametric space. In addition, G_l can be identified with the set of branches (vertices at the top level) of a rooted tree with $l+1$ levels and p^l branches. By definition the root of the tree is the only vertex at level 0. There are exactly p vertices at level 1, which correspond with the possible values of the digit i_0 in the p-adic expansion of $i=i_0+i_1p+\ldots+i_{l-1}p^{l-1}$. Each of these vertices is connected to the root by a non-directed edge. At level k, with $1\leq k\leq l+1$, there are exactly p^k vertices, each vertex corresponds to a truncated expansion of i of the form $i_0+\cdots+i_{k-1}p^{k-1}$. The vertex corresponding to $i_0+\cdots+i_{k-1}p^{k-1}$ is connected to a vertex $i_0'+\cdots+i_{k-2}'p^{k-2}$ at the level $k-1$ if and only if $\left(i_0+\cdots+i_{k-1}p^{k-1}\right)-\left(i_0'+\cdots+i_{k-2}'p^{k-2}\right)$ is divisible by p^{k-1}. The unit ball $\mathbb{Z}_p$ is an infinite rooted tree.

We denote by $DBN(p,l,\boldsymbol{\theta}_l)$ the p-adic discrete DBN with energy functional $E_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l;\boldsymbol{\theta}_l\right)$, see (3.7) and marginal distribution $\boldsymbol{P}_l(\boldsymbol{v}_l;\boldsymbol{\theta}_l)$, see (3.9).

We now identify G_l with the set of branches (vertices at the top level) of a rooted tree with $l+1$ levels and p^l branches. Attached to each branch $i\in G_l$ there are two states : v_i^l, h_i^l. The *visible field* is $\boldsymbol{v}_l=\left(v_i^l\right)_{i\in G_l}\in\{0,1\}^{\#G_l}$ and *the hidden field is* $\boldsymbol{h}_l=\left(h_i^l\right)_{i\in G_l}\in\{0,1\}^{\#G_l}$. These values are realizations of two random vectors, that we also called the visible and hidden fields. The values v_i^l, v_j^l , respectively h_i^l,

h_j^l, are statistically independent for $i \neq j$. The $DBN(p,l,\boldsymbol{\theta}_l)$ is a $p-$adic analogue of the convolutional deep belief networks studied in [44]. However, there are several important differences. We discuss these matters in the last section of this article. We denote by $DBN(p,\infty,\boldsymbol{\theta})$ the p-adic deep belief network associated with the energy function $E(\boldsymbol{v},\boldsymbol{h})$, see (3.1). The $DBN(p,l,\boldsymbol{\theta}_l)$ for $l \geq L$ is a discretization of $DBN(p,\infty,\boldsymbol{\theta})$, and for $l' > l \geq L$, $DBN(p,l',\boldsymbol{\theta}_{l'})$ is a larger scaled version of $DBN(p,l,\boldsymbol{\theta}_l)$. In general, the action $E_l(\boldsymbol{v}_l,\boldsymbol{h}_l;\boldsymbol{\theta}_l)$ is non local, which means that $w_{i-j}^l \neq 0$ for any i, $j \in G_l$.

4. The p-adic DBNs are universal approximators

We denote by $\boldsymbol{Q}(\boldsymbol{v})$ an arbitrary probability distribution on a finite set

$$\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_{2^m}\} = \{0,1\}^m$$

with 2^m elements. We fix p, a prime number, and l_0 a positive integer such that $2^m \leq p^{l_0}$, and extend $\boldsymbol{Q}(\boldsymbol{v})$ to the set $\left\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_{2^m},\boldsymbol{v}_{2^m+1},\ldots,\boldsymbol{v}_{p^{l_0}}\right\}$ by taking $\boldsymbol{Q}(\boldsymbol{v}_k)=0$ for $2^m+1 \leq k \leq p^{l_0}$. This observation allows us to extend $\boldsymbol{Q}(\boldsymbol{v})$ to any finite set with cardinality p^l, for any $l \geq l_0$. By identifying $\boldsymbol{v}$ with $\boldsymbol{v}_l = \left(v_j^l\right)_{j\in G_l}$, we can interpret $\boldsymbol{Q}(\boldsymbol{v}_l)$ a probability distributions on the $\boldsymbol{v}_l$s.

In this section, we consider the problem of approximating $\boldsymbol{Q}(\boldsymbol{v})$ by the marginal distribution $\boldsymbol{P}_l(\boldsymbol{v})$ of an $DBN(p,l,\boldsymbol{\theta})$. To measure the "distance" between $\boldsymbol{Q}(\boldsymbol{v})$ and $\boldsymbol{P}_l(\boldsymbol{v})$ we use the Kullback-Leibler (KL) divergence:

$$KL(\boldsymbol{Q} \mid \boldsymbol{P}_l) = \sum_{\boldsymbol{v}} \boldsymbol{Q}(\boldsymbol{v}) \ln \frac{\boldsymbol{Q}(\boldsymbol{v})}{\boldsymbol{P}_l(\boldsymbol{v})} = -H(\boldsymbol{Q}) - \frac{1}{\# G_l} \sum_{j\in G_l} \ln \boldsymbol{P}_l(\boldsymbol{v}_j),$$

where $H(\boldsymbol{Q})$ is the entropy of $\boldsymbol{Q}$. We recall that $KL(\boldsymbol{Q} \mid \boldsymbol{P}_l) = 0$ if and only if $\boldsymbol{Q} = \boldsymbol{P}_l$. We construct an improved version of $DBN(p,l,\boldsymbol{\theta}_l)$ by increasing the the number of levels (or layers) l, and consequently, the number of hidden variables (units), but keeping the number of visible variables fixed.

4.1. The key construction.

Remark 1. *Given a positive integer m, and a,b integers, we write*

$$a \equiv b \mod m$$

to mean that m divides $a-b$ in $\mathbb{Z}$, i.e. there exists $n \in \mathbb{Z}$ such that $a-b=nm$.

Notation 2. *From now on, we assume that $p \geq 3$.*

We first recall that $G_l = \mathbb{Z}_p/p^l\mathbb{Z}_p$ is isomorphic to $\mathbb{Z}/p^l\mathbb{Z}$ as Abelian groups. We identify $i_0 + i_1 p + \ldots + i_{l-1}p^{l-1} \in G_l$ with an integer written in base p, and the addition in G_l with the sum of integers mod p^l. There is a natural homomorphism of Abelian groups:

$$\begin{array}{ccc} G_{l+1} & \rightarrow & G_l \\ i & \rightarrow & i \mod p^l. \end{array}$$

But, there are no natural homomorphisms from G_l into G_{l+1}.

We identify G_l with the *subset* of G_{l+1}consisting of integers having the form $i_0 + i_1 p + \ldots + i_{l-1}p^{l-1}$, where the i_ks are p-adic digits. However, with this identification G_l is not a subgroup of the additive group G_{l+1}, because G_l is not closed under the addition in G_{l+1}. Indeed, ap^{l-1}, $(p-a)\,p^{l-1} \in G_l$ for any $a \in \{1,\ldots,p-1\}$, but $ap^{l-1} + (p-a)\,p^{l-1} = p^l \notin G_l$.

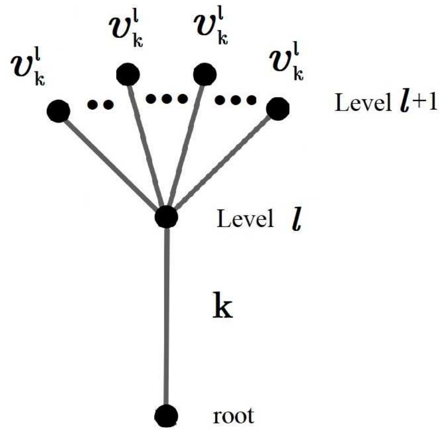

FIGURE 2. Construction of the field $\boldsymbol{v}_{l+1} = \left(\boldsymbol{v}_{\boldsymbol{k}}^{l+1}\right)_{\boldsymbol{k}\in G_{l+1}}$.

We set

$$T_{l+1}^{*} = \left\{ap^{l}; a \in \{1,\ldots,p-1\}\right\} \subset G_{l+1},$$

and $T_{l+1} = T_{l+1}^{*} \cup \{0\}$. Then T_{l+1} is an additive subgroup of G_{l+1}. Furthermore, as sets, it verifies that

$$G_{l+1} = \bigsqcup_{k\in T_{l+1}} (G_l + k),$$

where $\bigsqcup$ denotes the disjoint union and "$+$" denotes the addition in the group $(G_{l+1},+)$.

We now construct a copy $\boldsymbol{v}_{l+1} = \left(v_i^{l+1}\right)_{i\in G_{l+1}}$ in G_{l+1} of the visible field $\boldsymbol{v}_l = \left(v_i^{l}\right)_{i\in G_l}$. We set

$$\boldsymbol{v}_j^{l+1} = \boldsymbol{v}_k^{l} \text{ where } j \equiv k \mod p^{l}, \text{ for } j \in G_{l+1},\ k \in G_l.$$

This construction is illustrated in Figure 2.

We fix $j_0 = \alpha p^{l} \in T_{l+1}^{*}$, $\alpha \in \{2,\ldots,p-1\}$, here we use that $p \geq 3$, and set

$$h_j^{l+1} = \begin{cases} 0 & \text{if} \quad j \in G_{l+1} \smallsetminus (G_l \bigsqcup \{j_0\}) \\ h_{j_0}^{l+1} & \text{if} \quad j = j_0 \in T_{l+1}^{*} \\ h_j^{l} & \text{if} \quad j \in G_l. \end{cases}$$

With this construction the hidden field $\boldsymbol{h}_{l+1} = \left(h_i^{l+1}\right)_{i\in G_{l+1}}$ of the new $DBN(p, l+1, \boldsymbol{\theta}_{l+1})$ consists of the hidden field $\boldsymbol{h}_l = \left(h_i^{l}\right)_{i\in G_l}$ of $DBN(p, l, \boldsymbol{\theta}_l)$ plus an extra hidden unit $h_{j_0}^{l+1}$. This construction is illustrated in Figure 3.

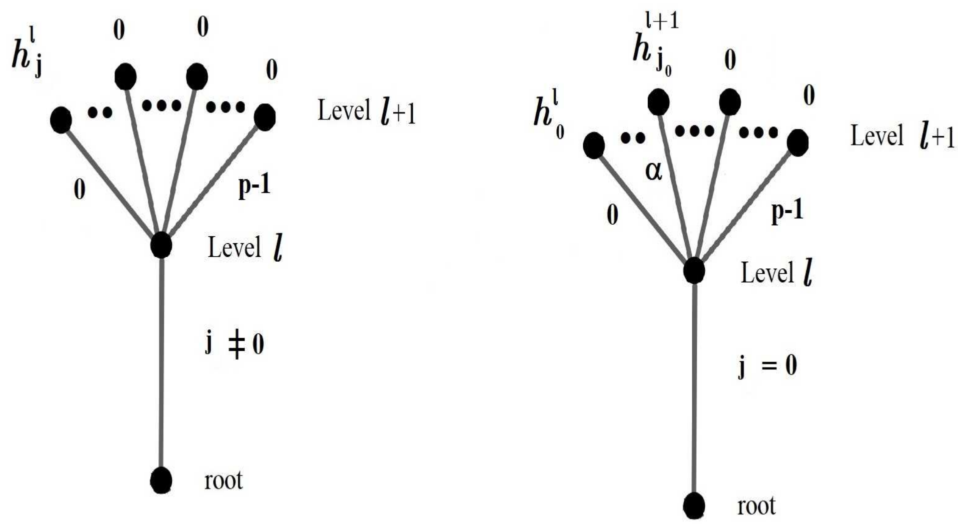

FIGURE 3. Construction of the field $\boldsymbol{h}_{l+1} = \left(h_j^{l+1}\right)_{j \in G_{l+1}}$.

We also set

$$a_j^{l+1} = \begin{cases} 0 & \text{if} \quad j \in G_{l+1} \smallsetminus G_l \\ a_j^l & \text{if} \quad j \in G_l, \end{cases} \qquad b_j^{l+1} = \begin{cases} 0 & \text{if} \quad j \in G_{l+1} \smallsetminus \left(G_l \bigsqcup T_{l+1}^*\right) \\ b_{j_0}^{l+1} & \text{if} \quad j = j_0 \in T_{l+1}^* \\ 0 & \text{if} \quad j \in T_{l+1}^* \smallsetminus \{j_0\} \\ b_j^l & \text{if} \quad j \in G_l, \end{cases}$$

and $\boldsymbol{a}_{l+1} = \left(a_j^{l+1}\right)_{j \in G_{l+1}}$, $\boldsymbol{b}_{l+1} = \left(b_j^{l+1}\right)_{j \in G_{l+1}}$.

We fix $\beta \in \{1, \dots, p-1\}$ and construct a copy $G_l - \beta p^l$ of G_l in G_{l+1}. We now set

$$w_j^{l+1} = \begin{cases} 0 & \text{if} \quad j \in G_{l+1} \smallsetminus \left(G_l \bigsqcup \left(G_l - \beta p^l\right)\right) \\ w_j^{l+1} & \text{if} \quad j \in G_l - \beta p^l \\ w_j^l & \text{if} \quad j \in G_l. \end{cases}$$

Then, the vector $\boldsymbol{w}_{l+1} = \left(w_j^{l+1}\right)_{j \in G_{l+1}} = \left(\boldsymbol{w}_l, \, \boldsymbol{w}_{l+1}^*\right)$ consists of the weigh vector $\boldsymbol{w}_l$ of $DBN(p, l, \boldsymbol{\theta}_l)$ and a new vector $\boldsymbol{w}_{l+1}^* = \left(w_j^{l+1}\right)_{j \in G_l - \beta p^l}$.

Given an $DBN(p, l, \boldsymbol{\theta}_l)$, the key construction allows us to construct a new $DBN(p, l+1, \boldsymbol{\theta}_{l+1})$, $\boldsymbol{\theta}_{l+1} = \left(\boldsymbol{\theta}_l, \boldsymbol{w}_{l+1}^*, b_{j_0}^{l+1}\right)$, with an extra layer, and extra hidden unit $h_{j_0}^{l+1} \in \{0, 1\}$, and $1 + \#G_l$ new parameters: $b_{j_0}^{l+1} \in \mathbb{R}$, $\boldsymbol{w}_{l+1}^* \in \mathbb{R}^{\#G_l}$. The energy functional of the new $DBN(p,l+1,\boldsymbol{\theta}_{l+1})$ is given by

$$E_{l+1}\left(\boldsymbol{v}_{l+1},\boldsymbol{h}_{l+1},\boldsymbol{\theta}_{l+1}\right)=\sum_{j\in G_{l+1}}\sum_{k\in G_{l+1}}w_k^{l+1}v_j^{l+1}h_{j+k}^{l+1}+\sum_{j\in G_{l+1}}a_j^{l+1}v_j^{l+1}+\sum_{j\in G_{l+1}}b_j^{l+1}h_j^{l+1},$$

where $\boldsymbol{\theta}_{l+1}=(\boldsymbol{w}_{l+1},\boldsymbol{a}_{l+1},\boldsymbol{b}_{l+1})$.

Lemma 1. *With the above notation, the following formulas holds true. For* a, $b\in\{0,\dots,p-1\}$, *we set*

$$S(a,b)=\sum_{j\in G_l+ap^l}\sum_{k\in G_l+bp^l}w_k^{l+1}v_j^{l+1}h_{j+k}^{l+1},$$

where "+" denotes the sum in G_{l+1}.
(i) If $a+b\not\equiv 0 \bmod p$, *then*

$$S(a,b)=\begin{cases}h_{j_0}^{l+1}\sum_{j\in G_l}w_{j-\beta p^l}^{l+1}v_j^l & \text{if } (a+b)\,p^l+j+k=\alpha p^l, \text{ for some } j,k\in G_l\\ \\ 0 & \text{otherwise.}\end{cases}$$

(ii) If $a+b\not\equiv 0 \bmod p$, *then*

$$\sum_{a+b\not\equiv 0 \mod p}S(a,b)=\left((p-1)\left(p-2\right)+2\right)h_{j_0}^{l+1}\sum_{j\in G_l}w_{j-\beta p^l}^{l+1}v_j^l.$$

(iii) If $a+b\equiv 0 \bmod p$, *then*

$$S(a,b)=\begin{cases}\sum_{j\in G_l}\sum_{k\in G_l}w_k^l v_j^l h_{j+k}^l & \text{if } a\equiv 0 \bmod p, b\equiv 0 \bmod p\\ \\ 0 & \text{otherwise.}\end{cases}$$

(iv)

$$E_{l+1}\left(\boldsymbol{v}_{l+1},\boldsymbol{h}_{l+1};\boldsymbol{\theta}_{l+1}\right)=E_{l+1}\left(\boldsymbol{v}_{l},\boldsymbol{h}_{l};\boldsymbol{\theta}_{l+1}\right)=E_l\left(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta}_l\right)+\left((p-1)\left(p-2\right)+2\right)h_{j_0}^{l+1}\sum_{j\in G_l}w_{j-\beta p^l}^{l+1}v_j^l+b_{j_0}^{l+1}h_{j_0}^{l+1}.$$

Proof. (i) Notice that

$$S(a,b)=\sum_{j\in G_l}\sum_{k\in G_l}w_{k+bp^l}^{l+1}v_{j+ap^l}^{l+1}h_{(a+b)p^l+j+k}^{l+1}.$$

We now use that $h_{(a+b)p^l+j+k}^{l+1}\neq 0$ if and only if $(a+b)\,p^l+j+k=\alpha p^l$ in G_{l+1}, i.e. if and only if

$$(a+b)\,p^l+j+k\equiv\alpha p^l \mod p^{l+1}. \tag{4.1}$$

The condition (4.1) implies that

$$j+k\equiv 0 \mod p^s \text{ for } 1\leq s\leq l,$$

which in turn implies that $j+k=0$ in G_l, and thus (4.1) becomes

$$a+b\equiv\alpha \mod p.$$

This last congruence has solutions since $a+b\not\equiv 0 \bmod p$. If (4.1) is satisfied, then $k=-j$ in G_l and since $w^{l+1}_{j-bp^l}=w^{l+1}_{-j+bp^l}$, $w^{l+1}_{j-bp^l}\neq 0\Leftrightarrow b=\beta$, and $v^{l+1}_{j+ap^l}=v^l_j$, we have

$$S(a,b)=h^{l+1}_{j_0}\sum_{j\in G_l}w^{l+1}_{j-\beta p^l}v^l_j,$$

otherwise $S(a,b)=0$.

(ii) It follows form the first part by using that

$$\sum_{a+b\not\equiv 0 \bmod p}S(a,b)=\sum_{a\not\equiv 0 \bmod p}\ \sum_{b\not\equiv 0 \bmod p}S(a,b)+\sum_{a\equiv 0 \bmod p}\ \sum_{b\not\equiv 0 \bmod p}S(a,b)$$
$$+\sum_{a\not\equiv 0 \bmod p}\ \sum_{b\equiv 0 \bmod p}S(a,b).$$

(iii) Notice that $b\equiv (p-a) \bmod p$, then

$$S(a,b)=\sum_{j\in G_l}\sum_{k\in G_l}w^{l+1}_{k+(p-a)p^l}v^{l+1}_{j+ap^l}h^{l+1}_{j+k}.$$

If $a\not\equiv 0 \bmod p$, then $v^{l+1}_{j+ap^l}=0$, and $S(a,b)=0$. If $a\equiv 0 \bmod p$, then $b\equiv 0 \bmod p$, and

$$\begin{aligned}S(a,b)&=\sum_{j\in G_l}\sum_{k\in G_l}w^l_kv^l_jh^{l+1}_{j+k}=\sum_{\substack{j,k\in G_l\\ j+k\in G_l}}w^l_kv^l_jh^{l+1}_{j+k}+\sum_{\substack{j,k\in G_l\\ j+k=j_0}}w^l_kv^l_jh^{l+1}_{j+k}\\ &=\sum_{j\in G_l}\sum_{k\in G_l}w^l_kv^l_jh^l_{j+k}+\sum_{\substack{j,k\in G_l\\ j+k=j_0}}w^l_kv^l_jh^{l+1}_{j+k}.\end{aligned}\tag{4.2}$$

A simple inductive argument on l shows that $j+k\equiv 0 \bmod p^l$, with $j,k\in G_l$, is only possible if $j=s+ap^{l-1}\in G_l$ and $k=-s+(p-a)\,p^{l-1}\in G_l$, in this case, $j+k=p^{l+1}\neq j_0$ in G_{l+1} and consequently the last sum in (4.2) is zero.

(iv) We first notice that

$$\begin{aligned}E_{l+1}\left(\boldsymbol{v}_{l+1},\boldsymbol{h}_{l+1};\boldsymbol{\theta}_{l+1}\right)&=\sum_{j,k\in G_{l+1}}w^{l+1}_kv^{l+1}_{j+k}h^{l+1}_j+\sum_{j\in G_l\bigsqcup T_{l+1}}a^{l+1}_jv^{l+1}_j\\ &\quad+\sum_{j\in G_l\bigsqcup T_{l+1}}b^{l+1}_jh^{l+1}_j\\ &=:S^{(0)}_{l+1}\left(\boldsymbol{v}_{l+1},\boldsymbol{h}_{l+1},\boldsymbol{w}_{l+1}\right)+S^{(1)}_{l+1}\left(\boldsymbol{v}_{l+1},\boldsymbol{a}_{l+1}\right)+S^{(2)}_{l+1}\left(\boldsymbol{h}_{l+1},\boldsymbol{b}_{l+1}\right).\end{aligned}\tag{4.3}$$

Now

$$\begin{aligned}S^{(0)}_{l+1}\left(\boldsymbol{v}_{l+1},\boldsymbol{h}_{l+1};\boldsymbol{\theta}_{l+1}\right)&:=\sum_{j\in G_{l+1}}\sum_{k\in G_{l+1}}w^{l+1}_kv^{l+1}_jh^{l+1}_{j+k}=\sum_a\sum_bS(a,b)\\ &=E_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l;\theta_l\right)+\left(\left(p-1\right)\left(p-2\right)+2\right)h^{l+1}_{j_0}\sum_{j\in G_l}w^{l+1}_{j-\beta p^l}v^l_j.\end{aligned}\tag{4.4}$$

It follows immediately that

$$S^{(1)}_{l+1}\left(\boldsymbol{v}_{l+1},\boldsymbol{a}_{l+1}\right)=\sum_{j\in G_l\bigsqcup T_{l+1}}a^{l+1}_jv^{l+1}_j=\sum_{j\in G_l}a^l_jv^l_j,\tag{4.5}$$

and that

$$S^{(2)}_{l+1}\left(\boldsymbol{h}_{l+1},\boldsymbol{b}_{l+1}\right)=\sum_{j\in G_l}b^l_jh^l_j+b^{l+1}_{j_0}h^{l+1}_{j_0}.\tag{4.6}$$

The announce formula follows from formula (4.3), by using (4.4)-(4.6). □

It is relevant to mention that the energy functional of $DBN(p, l+1, \boldsymbol{\theta}_{l+1})$ is an extension of the energy functional of $DBN(p, l, \boldsymbol{\theta}_l)$. Furthermore, the key construction is recursive. Starting with $DBN(p, l+1, \boldsymbol{\theta}_{l+1})$ there exists another $DBN(p, l+2, \boldsymbol{\theta}_{l+2})$ whose energy functional is an extension of the energy functional of $DBN(p, l+1, \boldsymbol{\theta}_{l+1})$.

4.2. **Better model with increasing number of levels.** In this section we show that the computational power of an $DBN(p, l, \theta)$ increases with the number of levels (or layers). More precisely, we show p-adic counterparts of the main results in [63, Theorems 1, 2].

On the other hand, $\left\{j - \beta p^l; j \in G_l\right\} \subset G_{l+1}$ is a copy (more precisely a fixed lifting) of G_l in G_{l+1}, and since the $w_{j-\beta p^l}^{l+1}$s are new parameters, we rename $w_{j-\beta p^l}^{l+1}$ as w_j^{l+1}, then

$$\sum_{j \in G_l} w_j^{l+1} v_j^l = \sum_{j \in G_l} w_{j-\beta p^l}^{l+1} v_j^l.$$

We rescale $h_{j_0}^{l+1}$ to $\left((p-1)(p-2)+2\right) h_{j_0}^{l+1}$ and $b_{j_0}^{l+1}$ to $\left((p-1)(p-2)+2\right) b_{j_0}^{l+1}$. With this notation the energy functional $E_{l+1}\left(\boldsymbol{v}_l, \boldsymbol{h}_{l+1}; \boldsymbol{\theta}_{l+1}\right)$ becomes

$$E_{l+1}\left(\boldsymbol{v}_l, \boldsymbol{h}_{l+1}; \boldsymbol{\theta}_{l+1}\right) = E_l\left(\boldsymbol{v}_l, \boldsymbol{h}_l; \boldsymbol{\theta}_l\right) + h_{j_0}^{l+1} \sum_{k \in G_l} w_k^{l+1} v_k^l + b_{j_0}^{l+1} h_{j_0}^{l+1}, \tag{4.7}$$

where $\boldsymbol{h}_{l+1} = \left(\boldsymbol{h}_l, h_{\boldsymbol{j}_0}^{l+1}\right)$, $\boldsymbol{\theta}_{l+1} = \left(\boldsymbol{\theta}_l, \boldsymbol{w}_{l+1}, b_{j_0}^{l+1}\right)$. Notice that $DBN(p, l+1, \boldsymbol{\theta}_{l+1}) = DBN(p, l+1, \boldsymbol{\theta}_l, \boldsymbol{w}_{l+1}, b_{j_0}^{l+1})$ has only one additional hidden unit ($h_{j_0}^{l+1}$). The corresponding Boltzmann distribution is given by

$$\boldsymbol{P}_{l+1}\left(\boldsymbol{v}_l, \boldsymbol{h}_{l+1}; \boldsymbol{\theta}_{l+1}\right) = \frac{\exp\left(-E_{l+1}\left(\boldsymbol{v}_l, \boldsymbol{h}_{l+1}; \boldsymbol{\theta}_{l+1}\right)\right)}{Z_{l+1}(\boldsymbol{\theta}_{l+1})},$$

and the marginal distribution is given by

$$\boldsymbol{P}_{l+1}\left(\boldsymbol{v}_l; \boldsymbol{\theta}_{l+1}\right) = \frac{\sum\limits_{\boldsymbol{h}_{l+1}} \exp\left(-E_{l+1}\left(\boldsymbol{v}_l, \boldsymbol{h}_{l+1}; \boldsymbol{\theta}_{l+1}\right)\right)}{Z_{l+1}(\theta_{l+1})}.$$

Lemma 2. *Let $\boldsymbol{P}_l\left(\boldsymbol{v}_l; \boldsymbol{\theta}_l\right)$ be a probability distribution over binary vectors $\{0,1\}^{\#G_l}$ obtained with an $DBN(p, l, \boldsymbol{\theta}_l)$, and let $\boldsymbol{P}_{l+1}\left(\boldsymbol{v}_l; \boldsymbol{\theta}_{l+1}\right) = \boldsymbol{P}_{l+1}\left(\boldsymbol{v}_l; \boldsymbol{\theta}_l, \boldsymbol{w}_{l+1}, b_{j_0}^{l+1}\right)$ be the marginal probability distribution corresponding to $DBN(p, l+1, \boldsymbol{\theta}_l, \boldsymbol{w}_{l+1}, b_{j_0}^{l+1})$, which is obtained from $DBN(p, l, \boldsymbol{\theta}_l)$ by adding one level and one hidden unit. Then $\boldsymbol{P}_{l+1}\left(\boldsymbol{v}_{l+1}; \boldsymbol{\theta}_l, \boldsymbol{w}_{l+1}, b_{j_0}^{l+1}\right)$ is a probability distribution over binary vectors $\{0,1\}^{\#G_l}$ for any $b_{j_0}^{l+1} \in [-\infty, \infty)$, and $\boldsymbol{P}_{l+1}\left(\boldsymbol{v}_l; \boldsymbol{\theta}_l, \boldsymbol{w}_{l+1}, -\infty\right) = \boldsymbol{P}_l\left(\boldsymbol{v}_l; \boldsymbol{\theta}_l\right)$.*

Proof. By using the formula (4.7) and the fact that $\boldsymbol{h}_{l+1} = \left(\boldsymbol{h}_l, h_{j_0}^{l+1}\right)$, we have

$$\sum_{\boldsymbol{h}_{l+1}} \exp\left(-E_{l+1}\left(\boldsymbol{v}_l, \boldsymbol{h}_{l+1}; \boldsymbol{\theta}_{l+1}\right)\right)$$
$$= \sum_{\boldsymbol{h}_l, h_{j_0}^{l+1}} \exp\left(-E_l\left(\boldsymbol{v}_l, \boldsymbol{h}_l; \boldsymbol{\theta}_l\right)\right) \exp\left(h_{j_0}^{l+1} \sum_{k \in G_l} w_k^{l+1} v_k^l + b_{j_0}^{l+1} h_{j_0}^{l+1}\right)$$
$$= \left\{\sum_{\boldsymbol{h}_l} \exp\left(-E_l\left(\boldsymbol{v}_l, \boldsymbol{h}_l; \boldsymbol{\theta}_l\right)\right)\right\} \left\{\sum_{h_{j_0}^{l+1} \in \{0,1\}} \exp\left(h_{j_0}^{l+1} \sum_{k \in G_l} w_k^{l+1} v_k^l + b_{j_0}^{l+1} h_{j_0}^{l+1}\right)\right\}$$
$$= \left\{\sum_{\boldsymbol{h}_l} \exp\left(-E_l\left(\boldsymbol{v}_l, \boldsymbol{h}_l; \boldsymbol{\theta}_l\right)\right)\right\} \left\{1 + \exp\left(\sum_{k \in G_l} w_k^{l+1} v_k^l + b_{j_0}^{l+1}\right)\right\}.$$

Then

$$Z_{l+1}(\boldsymbol{\theta}_l, \boldsymbol{w}_{l+1}, b_{j_0}^{l+1}) = \sum_{\boldsymbol{v}_{l+1}, \boldsymbol{h}_{l+1}} \exp\left(-E_{l+1}\left(\boldsymbol{v}_{l+1}, \boldsymbol{h}_{l+1}; \boldsymbol{\theta}_{l+1}\right)\right) \tag{4.8}$$
$$= \sum_{\boldsymbol{v}_l, \boldsymbol{h}_l} \left\{1 + \exp\left(\sum_{k \in G_l} w_k^{l+1} v_k^l + b_{j_0}^{l+1}\right)\right\} \exp\left(-E_l\left(\boldsymbol{v}_l, \boldsymbol{h}_l; \boldsymbol{\theta}_l\right)\right),$$

and

$$\boldsymbol{P}_{l+1}\left(\boldsymbol{v}_l; \boldsymbol{\theta}_l, \boldsymbol{w}_{l+1}, b_{j_0}^{l+1}\right) = \tag{4.9}$$
$$\frac{\left\{1 + \exp\left(\sum_{k \in G_l} w_k^{l+1} v_k^l + b_{j_0}^{l+1}\right)\right\} \sum_{\boldsymbol{h}_l} \exp\left(-E_l\left(\boldsymbol{v}_l, \boldsymbol{h}_l; \boldsymbol{\theta}_l\right)\right)}{Z_{l+1}(\boldsymbol{\theta}_l, \boldsymbol{w}_{l+1}, b_{j_0}^{l+1})}.$$

Thus $\boldsymbol{P}_{l+1}\left(\boldsymbol{v}_l; \theta_l, w_{l+1}, b_{j_0}^{l+1}\right)$ is a well-defined probability distribution for any $b_{j_0}^{l+1} \in [-\infty, \infty)$, and $\boldsymbol{P}_{l+1}\left(\boldsymbol{v}_l; \theta_l, -\infty\right) = \boldsymbol{P}_l\left(\boldsymbol{v}_l; \theta_l\right)$. $\square$

Lemma 3. *Assume that* $KL(\boldsymbol{Q}\left(\boldsymbol{v}_l\right) \parallel \boldsymbol{P}_l\left(\boldsymbol{v}_l; \theta_l\right)) > 0$. *Then there exists* $\widehat{\boldsymbol{w}}_{l+1} = \left(\widehat{w}_k^{l+1}\right)_{k \in G_l}$ *such that*

$$\sum_{\boldsymbol{v}_l} \exp\left(\sum_{k \in G_l} \widehat{w}_k^{l+1} v_k^l\right) \left(\boldsymbol{P}_l\left(\boldsymbol{v}_l; \boldsymbol{\theta}_l\right) - \boldsymbol{Q}\left(\boldsymbol{v}_l\right)\right) < 0. \tag{4.10}$$

Proof. Take $\widehat{\boldsymbol{v}}_l \neq \mathbf{0}$ such that $\boldsymbol{Q}\left(\widehat{\boldsymbol{v}}_l\right) \in (0,1)$. Then for any $\boldsymbol{w}_l'$, $\boldsymbol{b}_l'$ given there exists $\boldsymbol{a}_l'$ such that

$$\boldsymbol{P}_l\left(\widehat{\boldsymbol{v}}_l; \boldsymbol{\theta}_l'\right) < \boldsymbol{Q}\left(\widehat{\boldsymbol{v}}_l\right), \tag{4.11}$$

where $\boldsymbol{\theta}_l' = \left(\boldsymbol{w}_l', \boldsymbol{a}_l', \boldsymbol{b}_l'\right)$. If such $\widehat{\boldsymbol{v}}_l = \left(\widehat{v}_k^l\right)_{k \in G_l}$ does not exist, then $\boldsymbol{Q}\left(\boldsymbol{v}_l\right)$ is concentrated in one point, i.e. $\boldsymbol{Q}\left(\boldsymbol{v}_0\right) = 1$. In this case $KL(\boldsymbol{Q}\left(\boldsymbol{v}_l\right) \parallel \boldsymbol{P}_l\left(\boldsymbol{v}_l; \boldsymbol{\theta}_l\right)) = 0$. But this case is ruled out by the hypothesis $KL(\boldsymbol{Q}\left(\boldsymbol{v}_l\right) \parallel \boldsymbol{P}_l\left(\boldsymbol{v}_l; \theta_l\right)) > 0$.

We now set

$$\mathbf{1} = \underbrace{(1, \ldots, 1)}_{\#G_l\text{-times}} \text{ and } \widehat{\boldsymbol{w}}_{l+1} = \left(\widehat{w}_k^{l+1}\right)_{k \in G_l} = \alpha(\widehat{\boldsymbol{v}}_l - \frac{1}{2}\mathbf{1}),$$

where α is a positive number. Then, for $\boldsymbol{v}_l \neq \widehat{\boldsymbol{v}}_l$,

$$\lim_{\alpha\to\infty}\frac{\exp\left(\sum_{k\in G_l}\widehat{w}_k^{l+1}v_k^l\right)}{\exp\left(\sum_{k\in G_l}\widehat{w}_k^{l+1}\widehat{v}_k^l\right)}=0. \tag{4.12}$$

A detailed verification of this last inequality can be found in the demonstration of Theorem 1 in [63]. Consequently,

$$\sum_{\boldsymbol{v}_l}\exp\left(\sum_{k\in G_l}\widehat{w}_k^{l+1}v_k^l\right)\left(\boldsymbol{P}_l\left(\boldsymbol{v}_l;\boldsymbol{\theta}_l\right)-\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\right)=$$

$$\exp\left(\sum_{k\in G_l}\widehat{w}_k^{l+1}\widehat{v}_k^l\right)\left\{\boldsymbol{P}_l\left(\widehat{\boldsymbol{v}}_l;\boldsymbol{\theta}_l\right)-\boldsymbol{Q}\left(\widehat{\boldsymbol{v}}_l\right)+\sum_{\boldsymbol{v}_l\neq\widehat{\boldsymbol{v}}_l}\frac{\exp\left(\sum_{k\in G_l}\widehat{w}_k^{l+1}v_k^l\right)}{\exp\left(\sum_{k\in G_l}\widehat{w}_k^{l+1}\widehat{v}_k^l\right)}\right\},$$

and by using (4.12),

$$\sum_{\boldsymbol{v}_l}\exp\left(\sum_{k\in G_l}\widehat{w}_k^{l+1}v_k^l\right)\left(\boldsymbol{P}_l\left(\boldsymbol{v}_l;\boldsymbol{\theta}_l\right)-\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\right)$$
$$\sim\exp\left(\sum_{k\in G_l}\widehat{w}_k^{l+1}\widehat{v}_k^l\right)\left(\boldsymbol{P}_l\left(\widehat{\boldsymbol{v}}_l;\boldsymbol{\theta}_l\right)-\boldsymbol{Q}\left(\widehat{\boldsymbol{v}}_l\right)\right)$$

as $\alpha\to\infty$. Finally, by using (4.11), there exists α_0 such that (4.10) holds true for $\alpha>\alpha_0$. □

Remark 2. *Given positive integers l, l_0, with $l\geq l_0$, we identify G_{l_0} with the subset of G_l consisting of integers having the form $i_0+i_1p+\ldots+i_{l_0-1}p^{l_0-1}$, where the i_ks are p-adic digits.*

Theorem 1. *Assume that $p\geq 3$. Let $\boldsymbol{Q}(\boldsymbol{v})$ be an arbitrary probability distribution on $\{0,1\}^m$. As discussed above, we assume without loss of generality that $m=p^{l_0}$. We identify $\boldsymbol{v}$ with $\boldsymbol{v}_l=\left(v_j^l\right)_{j\in G_l}$, and $\boldsymbol{Q}(\boldsymbol{v}_l)$ with a probability distribution on the $\boldsymbol{v}_l$s. Let $DBN(p,l,\boldsymbol{\theta}_l)$ be a p-adic discrete DBN, with $l\geq l_0$, such that $KL(\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\|\boldsymbol{P}_l\left(\boldsymbol{v}_l;\boldsymbol{\theta}_l\right))>0$. Then the two following assertions hold true.*
(i) There exists an $DBN(p,l+1,\boldsymbol{\theta}_l,\boldsymbol{w}_{l+1},b_{j_0}^{l+1})$ constructed from $DBN(p,l,\boldsymbol{\theta}_l)$ by adding one layer with marginal probability distribution $\boldsymbol{P}_{l+1}\left(\boldsymbol{v}_l;\boldsymbol{\theta}_l,\boldsymbol{w}_{l+1},b_{j_0}^{l+1}\right)$ satisfying

$$KL(\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\|\boldsymbol{P}_{l+1}\left(\boldsymbol{v}_l;\boldsymbol{\theta}_l,\boldsymbol{w}_{l+1},b_{j_0}^{l+1}\right))<KL(\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\|\boldsymbol{P}_l\left(\boldsymbol{v}_l;\boldsymbol{\theta}_l\right)), \tag{4.13}$$

for some $\boldsymbol{\theta}_l,\boldsymbol{w}_{l+1},b_{j_0}^{l+1}$.
(ii) Given $\epsilon>0$ arbitrarily small, there exists an

$$DBN(p,l+k,\boldsymbol{\theta}_l,\boldsymbol{w}_{l+1},\ldots,\boldsymbol{w}_{l+k},b_{j_0}^{l+1},\ldots,b_{j_{k-1}}^{l+k})$$

with marginal probability distribution $\boldsymbol{P}_{l+k}\left(\boldsymbol{v}_l;\boldsymbol{\theta}_l,\boldsymbol{w}_{l+1},\ldots,\boldsymbol{w}_{l+k},b_{j_0}^{l+1},\ldots,b_{j_{k-1}}^{l+k}\right)$ *satisfying*

$$KL(\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\|\boldsymbol{P}_{l+k}\left(\boldsymbol{v}_l;\boldsymbol{\theta}_l,\boldsymbol{w}_{l+1},\ldots,\boldsymbol{w}_{l+k},b_{j_0}^{l+1},\ldots,b_{j_{k-1}}^{l+k}\right))<\epsilon, \tag{4.14}$$

where k is a positive integer depending on ϵ, for some

$$\boldsymbol{\theta}_l,\boldsymbol{w}_{l+1},\ldots,\boldsymbol{w}_{l+k},b_{j_0}^{l+1},\ldots,b_{j_{k-1}}^{l+k}.$$

Proof. (i) We first compute $KL(\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\|\boldsymbol{P}_{l+1}\left(\boldsymbol{v}_l;\boldsymbol{\theta}_{l+1}\right))$, $\boldsymbol{\theta}_{l+1}=\left(\boldsymbol{\theta}_l,\boldsymbol{w}_{l+1},b_{j_0}^{l+1}\right)$, using formulas (4.9)-(4.8):

$$KL(\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\|\boldsymbol{P}_{l+1}\left(\boldsymbol{v}_l;\boldsymbol{\theta}_{l+1}\right))=\sum_{\boldsymbol{v}_l}\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\ln\boldsymbol{Q}\left(\boldsymbol{v}_l\right)-\sum_{\boldsymbol{v}_l}\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\ln\boldsymbol{P}_{l+1}\left(\boldsymbol{v}_l;\boldsymbol{\theta}_{l+1}\right)=$$

$$-H(\boldsymbol{Q})-\sum_{\boldsymbol{v}_l}\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\ln\frac{\left\{1+\exp\left(\sum_{k\in G_l}w_k^{l+1}v_k^l+b_{\boldsymbol{j}_0}^{l+1}\right)\right\}\sum_{\boldsymbol{h}_l}\exp\left(-E_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l;\boldsymbol{\theta}_l\right)\right)}{\sum_{\widetilde{\boldsymbol{v}}_l,\widetilde{\boldsymbol{h}}_l}\left\{1+\exp\left(\sum_{k\in G_l}w_k^{l+1}\widetilde{v}_k^l+b_{j_0}^{l+1}\right)\right\}\exp\left(-E_l\left(\widetilde{\boldsymbol{v}}_l,\widetilde{\boldsymbol{h}}_l;\boldsymbol{\theta}_l\right)\right)}$$

$$=-H(\boldsymbol{Q})-\sum_{\boldsymbol{v}_l}\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\ln\left(1+\exp\left(\sum_{k\in G_l}w_k^{l+1}v_k^l+b_{j_0}^{l+1}\right)\right)$$

$$-\sum_{\boldsymbol{v}_l}\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\ln\left(\sum_{\boldsymbol{h}_l}\exp\left(-E_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l;\boldsymbol{\theta}_l\right)\right)\right)$$

$$+\left(\sum_{\boldsymbol{v}_l}\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\right)\ln\left(\sum_{\widetilde{\boldsymbol{v}}_l,\widetilde{\boldsymbol{h}}_l}\left\{1+\exp\left(\sum_{\boldsymbol{k}\in G_l}w_k^{l+1}\widetilde{v}_k^l+b_{j_0}^{l+1}\right)\right\}\exp\left(-E_l\left(\widetilde{\boldsymbol{v}}_l,\widetilde{\boldsymbol{h}}_l;\boldsymbol{\theta}_l\right)\right)\right)$$

$$=:-H(\boldsymbol{Q})-KL_1(\boldsymbol{\theta}_l)-KL_2(\boldsymbol{w}_{l+1},b_{j_0}^{l+1})+KL_3\left(\boldsymbol{\theta}_l,\boldsymbol{w}_{l+1},b_{j_0}^{l+1}\right). \tag{4.15}$$

Given any $\boldsymbol{w}_{l+1}$, we may assume that $\exp\left(\sum_{k\in G_l}w_k^{l+1}v_k^l+b_{j_0}^{l+1}\right)$ is very small for any $\boldsymbol{v}$, by taking $-b_{j_0}^{l+1}$ sufficiently large, since $v_k\in\{0,1\}$. Then by using $\ln\left(1+x\right)=x+o(x)$ as $x\to0$, we have

$$\ln\left(1+\exp\left(\sum_{k\in G_l}w_k^{l+1}v_k^l+b_{j_0}^{l+1}\right)\right)=\exp\left(\sum_{k\in G_l}w_k^{l+1}v_k^l+b_{j_0}^{l+1}\right)+o\left(\exp\left(b_{j_0}^{l+1}\right)\right), \tag{4.16}$$

as $b_{j_0}^{l+1}\to-\infty$. Then, the term $KL_1(\boldsymbol{\theta}_l)$ becomes

$$KL_1(\boldsymbol{\theta}_l)=\sum_{\boldsymbol{v}_l}\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\exp\left(\sum_{k\in G_l}w_k^{l+1}v_k^l+b_{j_0}^{l+1}\right)+o\left(\exp\left(b_{\boldsymbol{j}_0}^{l+1}\right)\right)\text{ as }b_{j_0}^{l+1}\to-\infty, \tag{4.17}$$

and the term $KL_3\left(\boldsymbol{\theta}_l, \boldsymbol{w}_{l+1}, b_{j_0}^{l+1}\right)$ becomes

$$KL_3\left(\boldsymbol{\theta}_l, \boldsymbol{w}_{l+1}, b_{j_0}^{l+1}\right) =$$

$$\ln\left(\sum_{\widetilde{\boldsymbol{v}}_l, \widetilde{\boldsymbol{h}}_l}\left\{1+\exp\left(\sum_{k\in G_l} w_k^{l+1}\widetilde{v}_k^l + b_{j_0}^{l+1}\right)\right\}\exp\left(-E_l\left(\widetilde{\boldsymbol{v}}_l, \widetilde{\boldsymbol{h}}_l; \boldsymbol{\theta}_l\right)\right)\right)$$

$$= \ln\left(\sum_{\widetilde{\boldsymbol{v}}_l, \widetilde{\boldsymbol{h}}_l}\exp\left(-E_l\left(\widetilde{\boldsymbol{v}}_l, \widetilde{\boldsymbol{h}}_l; \boldsymbol{\theta}_l\right)\right)\right) +$$

$$\ln\left(1+\frac{\sum_{\widetilde{\boldsymbol{v}}_l, \widetilde{\boldsymbol{h}}_l}\exp\left(\sum_{k\in G_l} w_k^{l+1}\widetilde{v}_k^l + b_{j_0}^{l+1}\right)\exp\left(-E_l\left(\widetilde{\boldsymbol{v}}_l, \widetilde{\boldsymbol{h}}_l; \boldsymbol{\theta}_l\right)\right)}{\sum_{\widetilde{\boldsymbol{v}}_l, \widetilde{\boldsymbol{h}}_l}\exp\left(-E_l\left(\widetilde{\boldsymbol{v}}_l, \widetilde{\boldsymbol{h}}_l; \boldsymbol{\theta}_l\right)\right)}\right).$$

Now, by using (4.16), we have for $b_{j_0}^{l+1} \to -\infty$ that

$$KL_3\left(\boldsymbol{\theta}_l, \boldsymbol{w}_{l+1}, b_{j_0}^{l+1}\right) = \ln\left(\sum_{\widetilde{\boldsymbol{v}}_l, \widetilde{\boldsymbol{h}}_l}\exp\left(-E_l\left(\widetilde{\boldsymbol{v}}_l, \widetilde{\boldsymbol{h}}_l; \boldsymbol{\theta}_l\right)\right)\right) \tag{4.18}$$

$$+\frac{\sum_{\widetilde{\boldsymbol{v}}_l, \widetilde{\boldsymbol{h}}_l}\exp\left(\sum_{k\in G_l} w_k^{l+1}\widetilde{v}_k^l + b_{j_0}^{l+1}\right)\exp\left(-E_l\left(\widetilde{\boldsymbol{v}}_l, \widetilde{\boldsymbol{h}}_l; \boldsymbol{\theta}_l\right)\right)}{\sum_{\widetilde{\boldsymbol{v}}_l, \widetilde{\boldsymbol{h}}_l}\exp\left(-E_l\left(\widetilde{\boldsymbol{v}}_l, \widetilde{\boldsymbol{h}}_l; \boldsymbol{\theta}_l\right)\right)} + o\left(\exp(b_{j_0}^{l+1})\right)$$

$$= \ln\left(\sum_{\widetilde{\boldsymbol{v}}_l, \widetilde{\boldsymbol{h}}_l}\exp\left(-E_l\left(\widetilde{\boldsymbol{v}}_l, \widetilde{\boldsymbol{h}}_l; \boldsymbol{\theta}_l\right)\right)\right) + \sum_{\widetilde{\boldsymbol{v}}_l}\exp\left(\sum_{k\in G_l} w_k^{l+1}\widetilde{v}_k^l + b_{j_0}^{l+1}\right) P_l\left(\widetilde{\boldsymbol{v}}_l; \boldsymbol{\theta}_l\right)$$
$$+o\left(\exp(b_{j_0}^{l+1})\right).$$

Finally, from formulas (4.15)-(4.18), we obtain that

$$KL(\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\|\boldsymbol{P}_{l+1}\left(\boldsymbol{v}_l; \boldsymbol{\theta}_{l+1}\right)) - KL(\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\|\boldsymbol{P}_l\left(\boldsymbol{v}_l; \boldsymbol{\theta}_l\right)) =$$

$$\sum_{\boldsymbol{v}_l}\exp\left(\sum_{k\in G_l} w_k^{l+1} v_k^l + b_{j_0}^{l+1}\right)\left(P_l\left(\boldsymbol{v}_l; \theta_l\right) - \boldsymbol{Q}\left(\boldsymbol{v}_l\right)\right)$$
$$+ o\left(\exp(b_{j_0}^{l+1})\right)$$

as $o\left(\exp(b_{j_0}^{l+1})\right) \to -\infty$. By Applying Lemma 3, there exist $\widetilde{\boldsymbol{w}}_{l+1}$, $\widetilde{b}_{j_0}^{l+1}$ such that

$$KL(\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\|\boldsymbol{P}_{l+1}\left(\boldsymbol{v}_l; \boldsymbol{\theta}_l, \widetilde{\boldsymbol{w}}_{l+1}, \widetilde{b}_{j_0}^{l+1}\right)) - KL(\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\|\boldsymbol{P}_l\left(\boldsymbol{v}_l; \boldsymbol{\theta}_l\right)) < 0.$$

(ii) We proceed recursively. If $KL(\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\|\boldsymbol{P}_{l+1}\left(\boldsymbol{v}_l; \boldsymbol{\theta}_l, \boldsymbol{w}_{l+1}, b_{j_0}^{l+1}\right)) = 0$, for some $\boldsymbol{\theta}_l, \boldsymbol{w}_{l+1}, b_{j_0}^{l+1}$, then the $DBN(p, l+1, \boldsymbol{\theta}_l, \boldsymbol{w}_{l+1}, b_{j_0}^{l+1})$ satisfies the condition required. Otherwise, by using the fact that the key construction can be used in a recursive way, we use the part (i) to construct a Boltzmann machine

$$DBN(p, l+2, \boldsymbol{\theta}_{l+1}, \boldsymbol{w}_{l+1}, \boldsymbol{w}_{l+2}, b_{j_0}^{l+1}, b_{j_1}^{l+2}),$$

which satisfies

$$KL(\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\|\boldsymbol{P}_{l+2}\left(\boldsymbol{v}_l;\boldsymbol{\theta}_{l+1},\boldsymbol{w}_{l+1},\boldsymbol{w}_{l+2},b_{j_0}^{l+1},b_{j_1}^{l+2}\right)) < KL(\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\|\boldsymbol{P}_{l+1}\left(\boldsymbol{v}_l;\boldsymbol{\theta}_l,\boldsymbol{w}_{l+1},b_{j_0}^{l+1}\right)) < KL(\boldsymbol{Q}\left(\boldsymbol{v}_l\right)\|\boldsymbol{P}_l\left(\boldsymbol{v}_l;\boldsymbol{\theta}_l\right)).$$

Therefore there exists $k(\epsilon)$ such that (4.14) holds true. $\square$

Theorem 2. *Assume that $p\geq 3$. Let $\boldsymbol{Q}(\boldsymbol{v})$ be an arbitrary probability distribution on $\{0,1\}^m$. As discussed above, we assume without loss of generality that $m=p^{l_0}$. We identify $\boldsymbol{v}$ with $\boldsymbol{v}_l=\left(v_j^l\right)_{j\in G_l}$, and $\boldsymbol{Q}(\boldsymbol{v}_l)$ with a probability distribution on the $\boldsymbol{v}_l$s.. Then $\boldsymbol{Q}\left(\boldsymbol{v}_l\right)$ can be approximated arbitrarily well, in the sense of the KL divergence, by an $DBN(p,l_0+k,\boldsymbol{\theta}_{l_0+k})$, where k is the number of input vectors whose probability in not zero.*

Proof. The argument is an adaptation of the one given in [63] for Theorem 2. The key is observation is that adding a hidden unit in [63, Theorem 2] corresponds to add a level in our construction. Furthermore, in [63, Theorem 2] the marginal distribution with an extra hidden unit $p_{w,c}(\boldsymbol{v})$ agrees with

$$\boldsymbol{P}_{l+1}\left(\boldsymbol{v}_l;\boldsymbol{\theta}_l,\boldsymbol{w}_{l+1},b_{j_0}^{l+1}\right)= \frac{\left\{1+\exp\left(\sum_{k\in G_l}w_k^{l+1}v_k^l+b_{j_0}^{l+1}\right)\right\}\sum_{\boldsymbol{h}_l}\exp\left(-E_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l;\boldsymbol{\theta}_l\right)\right)}{\sum_{\boldsymbol{v}_l,\boldsymbol{h}_l}\left\{1+\exp\left(\sum_{k\in G_l}w_k^{l+1}v_k^l+b_{j_0}^{l+1}\right)\right\}\exp\left(-E_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l;\boldsymbol{\theta}_l\right)\right)},$$

where $\boldsymbol{\theta}_l=(\boldsymbol{w}_l,\boldsymbol{a}_l,\boldsymbol{b}_l)$, up to the function $E_l\left(\boldsymbol{v}_l,\boldsymbol{h}_l;\boldsymbol{\theta}_l\right)$. Let $\widetilde{\boldsymbol{v}}_l=\left(\widetilde{v}_k^l\right)_{k\in G_l}$ be an arbitrary input vector and let $\widehat{\boldsymbol{w}}_{l+1}$ be the vector defined as the proof of Lemma 3:

$$\widehat{\boldsymbol{w}}_{l+1}=\left[\widehat{w}_k^{l+1}\right]_{k\in G_l},\ \ \widehat{w}_k^{l+1}=\alpha(\widetilde{v}_k^l-\frac{1}{2})\text{ for }k\in G_l,$$

where α is a positive number. We define $\widehat{b}_{j_0}^{l+1}=-\sum_{k\in G_l}\widehat{w}_k^{l+1}\widetilde{v}_k^l+\lambda$, with $\lambda\in\mathbb{R}$. Then

$$\lim_{\alpha\to\infty}1+\exp\left(\sum_{k\in G_l}\widehat{w}_k^{l+1}\widetilde{v}_k^l+\widehat{b}_{j_0}^{l+1}\right)=\begin{cases}1 & \text{if}\quad \boldsymbol{v}_l\neq\widetilde{\boldsymbol{v}}_l\\ 1+\exp\lambda & \text{if}\quad \boldsymbol{v}_l=\widetilde{\boldsymbol{v}}_l,\end{cases}$$

and by using formula (4.9), we have

$$\lim_{\alpha\to\infty}\boldsymbol{P}_{l+1}\left(\boldsymbol{v};\boldsymbol{\theta}_l,\widehat{\boldsymbol{w}}_{l+1},\widehat{b}_{j_0}^{l+1}\right)=\begin{cases}\frac{\boldsymbol{P}_l(\boldsymbol{v}_l;\boldsymbol{\theta}_l)}{1+\exp(\lambda)\boldsymbol{P}_l(\widetilde{\boldsymbol{v}}_l;\boldsymbol{\theta}_l)} & \text{if}\quad \boldsymbol{v}_l\neq\widetilde{\boldsymbol{v}}_l\\ \frac{(1+\exp(\lambda))\boldsymbol{P}_l(\widetilde{\boldsymbol{v}}_l;\boldsymbol{\theta}_l)}{1+\exp(\lambda)\boldsymbol{P}_l(\widetilde{\boldsymbol{v}}_l;\boldsymbol{\theta}_l)} & \text{if}\quad \boldsymbol{v}_l=\widetilde{\boldsymbol{v}}_l.\end{cases} \tag{4.19}$$

By choosing a suitable value of λ, and by adding an extra level to an $DBN(p,l,\boldsymbol{\theta}_l)$, the probability of an arbitrary input $\widetilde{\boldsymbol{v}}_l$ can be increased, while the probability of any other input $\boldsymbol{v}_l\neq\widetilde{\boldsymbol{v}}_l$ can be uniformly decreased by a multiplicative factor. Now, the required DBN can be constructed recursively using the technique given in the proof of Theorem 2 in [63]. We index the input vectors as $\boldsymbol{u}_i$ where i is an integer from 1 to 2^m, $m=p^{l_0}$, and sort them such that

$$0=\boldsymbol{Q}(\boldsymbol{u}_{k+1})=\ldots=\boldsymbol{Q}(\boldsymbol{u}_{2^m})<\boldsymbol{Q}(\boldsymbol{u}_1)\leq\boldsymbol{Q}(\boldsymbol{u}_2)\leq\ldots\leq\boldsymbol{Q}(\boldsymbol{u}_k).$$

We denote by

$$\boldsymbol{P}_{l_0+r}\left(\boldsymbol{v}_{l_0}\right)=\boldsymbol{P}_{l_0+r}\left(\boldsymbol{v}_{l_0};\boldsymbol{\theta}_{l_0+r-1},\boldsymbol{w}_{l_0+1},\ldots,\boldsymbol{w}_{l_0+r},b_{j_0}^{l_0+1},\ldots,b_{j_{r-1}}^{l_0+r}\right),$$

for $r=1,2,\ldots$, the marginal distribution of an RBM constructed from

$$DBN\left(p,l_0,\boldsymbol{\theta}_{l_0},\boldsymbol{w}_{l_0+1},b_{j_0}^{l_0+1}\right)$$

by using the the key construction r times. The $\boldsymbol{P}_{l_0+r}$s are defined inductively as follows. If $r=0$, we take $\boldsymbol{P}_{l_0}\left(\boldsymbol{v}_{l_0}\right)=2^{-m}$, $\boldsymbol{v}\in G_{l_0}$, is the uniform distribution. We now set

$$\widehat{\boldsymbol{w}}_{l_0+1}=\alpha\left(\boldsymbol{u}_1-\frac{1}{2}\right)\text{ and }\widehat{b}_{j_0}^{l_0+1}=-\left\langle\widehat{\boldsymbol{w}}_{l_0+1},\boldsymbol{v}_1\right\rangle+\lambda_1,$$

where

$$\left\langle\boldsymbol{w}_{l_0},\boldsymbol{v}_{l_0}\right\rangle:=\sum_{j\in G_{l_0}}w_j^{l_0}v_j^{l_0}.$$

By (4.19),

$$\lim_{\alpha\to\infty}\boldsymbol{P}_{l_0+1}\left(\boldsymbol{v};\boldsymbol{\theta}_{l_0},\widehat{\boldsymbol{w}}_{l_0+1},\widehat{b}_{j_0}^{l_0+1}\right)=\begin{cases}\frac{(1+\exp(\lambda_1))\boldsymbol{2}^{-m}}{1+\exp(\lambda_1)\boldsymbol{2}^{-m}} & \text{if}\quad \boldsymbol{v}_{l_0}=\boldsymbol{u}_1\\ \\ \frac{\boldsymbol{2}^{-m}}{1+\exp(\lambda_1)\boldsymbol{2}^{-m}} & \text{if}\quad \boldsymbol{v}_{l_0}=\boldsymbol{u}_i,\ i\geq 2.\end{cases}$$

We now add an extra level to $DBN\left(p,l_0,\boldsymbol{\theta}_{l_0},\widehat{\boldsymbol{w}}_{l_0+1},\widehat{b}_{j_0}^{l_0+1}\right)$ using the key construction. By choosing λ_2, $\boldsymbol{P}_{l_0+2}\left(\boldsymbol{v}_{l_0}\right)$ satisfies

$$\frac{\boldsymbol{P}_{l_0+2}\left(\boldsymbol{u}_2\right)}{\boldsymbol{P}_{l_0+2}\left(\boldsymbol{u}_1\right)}=\frac{\boldsymbol{Q}\left(\boldsymbol{u}_2\right)}{\boldsymbol{Q}\left(\boldsymbol{u}_1\right)}.$$

By using this construction recursively, one constructs a probability distribution $\boldsymbol{P}_{l_0+k}\left(\boldsymbol{v}_{l_0}\right)$ satisfying

$$\begin{aligned}&\frac{\boldsymbol{P}_{l_0+k}\left(\boldsymbol{u}_k\right)}{\boldsymbol{P}_{l_0+k}\left(\boldsymbol{u}_{k-1}\right)}=\frac{\boldsymbol{Q}\left(\boldsymbol{u}_k\right)}{\boldsymbol{Q}\left(\boldsymbol{u}_{k-1}\right)},\ldots,\frac{\boldsymbol{P}_{l_0+2}\left(\boldsymbol{u}_2\right)}{\boldsymbol{P}_{l_0+2}\left(\boldsymbol{u}_1\right)}=\frac{\boldsymbol{Q}\left(\boldsymbol{u}_2\right)}{\boldsymbol{Q}\left(\boldsymbol{u}_1\right)},\\ &\boldsymbol{P}_{l_0+k}\left(\boldsymbol{u}_{k+1}\right)=\ldots=\boldsymbol{P}_{l_0+k}\left(\boldsymbol{u}_{2^m}\right).\end{aligned}$$

The solution of the above recursive system is given in the proof of Theorem 2 in [63]:

$$\boldsymbol{P}_{l_0+k}\left(\boldsymbol{u}_i\right)=\begin{cases}\frac{\boldsymbol{Q}(\boldsymbol{u}_1)}{1+\exp(\lambda_1)+(2^m-k)\boldsymbol{Q}(\boldsymbol{v}_1)} & \text{if}\quad i>k\\ \\ \boldsymbol{Q}\left(\boldsymbol{u}_i\right)\frac{1+\exp(\lambda_1)}{1+\exp(\lambda_1)+(2^m-k)\boldsymbol{Q}(\boldsymbol{v}_1)} & \text{if}\quad i\leq k.\end{cases}$$

Finally,

$$KL\boldsymbol{Q}(\parallel\boldsymbol{P}_{l_0+k})=\sum_i\boldsymbol{Q}\left(\boldsymbol{u}_i\right)\frac{\left(2^m-k\right)\boldsymbol{Q}\left(\boldsymbol{u}_i\right)}{1+\exp\left(\lambda_1\right)}+o(\exp\left(-\lambda_1\right))\to 0$$

as $\lambda_1\to\infty$. $\square$

5. Discussion

5.1. **Euclidean QFTs and NNs.** The literature about the connections between QFTs with NNs and brain activity is extremely large. In this section we compare our results and our approach with some recent works. We also propose several new open problems.

In [40], the authors propose a correspondence between QFTs and NNs. Many modern network architectures admits a Gaussian limit as the number of neurons per layer tends to infinity. In the limit, these networks can be described by Gaussian processes which naturally correspond to non-interacting field theories. Moving away from the asymptotic limit yields to non-Gaussian processes which are connected with interacting fields theories. In our approach we work exclusively with interacting field theories: a continuous version and a discrete version. See Table 1.

p-adic discrete DBN with l layers	$\xleftarrow{\text{Discretization}}$	p-adic continuous DBN with infinitely many layers
$\Updownarrow$		$\Updownarrow$
Discrete SFT determined by $E_l(\boldsymbol{v}_l, \boldsymbol{h}_l)$	$\xleftarrow{\text{Discretization}}$	SFT determined by $E(\boldsymbol{v}, \boldsymbol{h})$
$\Updownarrow$		$\Updownarrow$
Probability measure $\boldsymbol{P}_l(\boldsymbol{v}_l, \boldsymbol{h}_l)\ d^{\#G_l}\boldsymbol{v}\ d^{\#G_l}\boldsymbol{h}$ on $\mathcal{D}^l(\mathbb{Z}_p) \times \mathcal{D}^l(\mathbb{Z}_p)$	$\xrightarrow{\text{Limit}}$	Probability measure $\frac{e^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z^{\text{phys}}} d\boldsymbol{v}d\boldsymbol{h}$ on $\mathcal{D}(\mathbb{Z}_p) \times \mathcal{D}(\mathbb{Z}_p)$
$DBN(p, l, \boldsymbol{\theta}_l)$	$\xrightarrow{\text{Scaling, } m > l}$	$DBN(p, m, \boldsymbol{\theta}_m)$

Table 1. The table provides a basic dictionary bewteen Euclidean QFTs and NNs. The last line in the table means that $DBN(p, m, \theta_m)$ is a larger and computationally more powerful version of $DBN(p, l, \theta_l)$.

A rigorous mathematical study of the following problem plays a central role in the understanding the neural networks using statistical field theory:

Problem 1. *Determine all the energy functionals $E(\boldsymbol{v}, \boldsymbol{h})$ such that*

$$\frac{e^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z^{phys}} d\boldsymbol{v}d\boldsymbol{h} \overset{def}{=} \lim_{l\to\infty} \boldsymbol{P}_l(\boldsymbol{v}_l, \boldsymbol{h}_l)\ d^{\#G_l}\boldsymbol{v}\ d^{\#G_l}\boldsymbol{h} \tag{5.1}$$

exists in some sense.

In [92], the author establishes, in a rigorous mathematical way, the existence of ϕ^4-interacting Euclidean quantum field theories on a p-adic spacetime for which the limit (5.1) exists. In a forthcoming publication we plan to expand the results given in [92] to case of two fields and find the energy functionals $E(\boldsymbol{v}, \boldsymbol{h})$ for which the limit (5.1) exists. The mentioned limit suggest that the correlation functions of the continuous STF can be very well-approximated by the correlation functions of the corresponding discrete SFT.

In [9], authors study a generalization of the RBMs associated with energy functionals of type:

$$S(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta}) = -\sum_{j\in\mathcal{G}}\sum_{k\in\mathcal{G}} w_{k,j} v_j h_j + \sum_{j\in\mathcal{G}} a_j v_j + \sum_{j\in\mathcal{G}} b_j h_j \tag{5.2}$$
$$+\sum_{j\in\mathcal{G}} c_j v_j^2 + \sum_{j\in\mathcal{G}} d_j h_j^2 + \sum_{j\in\mathcal{G}} f_j v_j^4 + \sum_{j\in\mathcal{G}} g_j h_j^4,$$

where $\mathcal{G}$ is a square lattice. These generalizations are not DBNs due to the topology of $\mathcal{G}$. Also, the authors assume that and $w_{k,j} \neq 0 \Leftrightarrow i$ and j are connected by one edge. This condition implies that the functional $S(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})$ is local, and then the corresponding Boltzmann distribution is related with a Markov processes on $\mathcal{G}$. Our action $E_l(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})$ is non local, which means that (in general) $w_{i-j} \neq 0$ for any i, $j \in G_l$. In [9], the authors also discussed the implementation of several learning algorithms. In forthcoming article, we will discuss the implementation of p-adic discrete DBNs based on energy functionals of type (5.2) with $\mathcal{G} = G_l$.

In [76], [36] a completely different approach for the correspondence between Euclidean QFTs and NNs is presented. Starting with a stochastic differential equation, which plays the role of a master equation for the neural network, the authors construct an action and a path integral, which provides the QFT attached to the network. The non-Archimedean counterpart of this construction is an open problem. Before considering this problem, it is necessary to study non-Archimedean versions of stochastic recurrent neural networks (SRNNs), see e.g. [65]-[66] and the references therein. Based on [65]-[66], [91], [99], we propose the following non-Archimedean version of the SRNNs.

Problem 2. *Let $t \in [0,T]$ and let $\boldsymbol{v} \in C\left([0,T], L^2(\mathbb{Q}_p, dx)\right)$ be a deterministic input signal. A p-adic temporal and spatially continuous SRNN is described the following state-space model:*

$$\frac{d\boldsymbol{h}(x,t)}{dt} = -a\boldsymbol{h}(x,t) + \alpha\left(\int_{\mathbb{Q}_p} A(x,y)\boldsymbol{h}(y,t)\,dy + \int_{\mathbb{Q}_p} B(x,y)\boldsymbol{v}(y,t)\,dy + B(x)\right)$$
$$+\beta(\boldsymbol{h}(x,t),\boldsymbol{v}(x,t))\dot{W}(x,t) \tag{5.3}$$

$$\boldsymbol{y}(x,t) = \sigma(\boldsymbol{h}(x,t)). \tag{5.4}$$

Where (5.3) is a stochastic equation for the hidden state $\boldsymbol{h} \in C\left([0,T], L^2(\mathbb{Q}_p, dx)\right)$, $a > 0$, $\alpha, \sigma : \mathbb{R} \to \mathbb{R}$ are Lipschitz continuous and bounded functions, A, $B \in L^1\left(\mathbb{Q}_p^2, d^2x\right)$, $\beta : \mathbb{R} \to \mathbb{R}$, and $\dot{W}(x,t)$ is the formal notation for a Gaussian random perturbation defined on some probability space, and (5.4) defines the response of the network, $\boldsymbol{y} \in C\left([0,T], L^2(\mathbb{Q}_p, dx)\right)$. A relevant problem is to study the response of SRNNs.

5.2. **The non-Archimedean counterpart of the Buice-Cowan theory.** Buice and Cowan formulated a theory of fluctuating activity of cortical networks in the language of stochastic fields, see e.g. [17], [18], see also [21]. A relevant observation is that the discrete master equation of the spike model can be formulated on G_l. Consider a network of $N = p^l$ neurons. The configuration of each neuron is given by the number of effective spikes n_i that neuron $i \in G_l$ has emitted. There is a weight function $w_{i,j}$ describing the relative innervation of neuron i by neuron j. We assume that $w_{i,j}$ is a function of $|i-j|_p$, which is exactly the hypothesis used by Buice and Cowan. The probability per unit of time that a neuron will emit another spike is given by $f\left(\sum_{j\in G_l} w_{i,j}n_j\right)$. The state of the system is given by the probability distribution $P_{\boldsymbol{n}}(t)$, which is the probability that the network is in configuration $\boldsymbol{n}=(n_i)_{i\in G_l}$ at the time t. The master equation of the network has the form

$$\frac{\partial}{\partial t}P_{n_i}(t)=-\alpha n_i P_{n_i}(t)+\alpha\left(n_i+1\right)P_{n_i+1}(t) \\ -f\left(\sum_{\substack{j\neq i\\ j,i\in G_l}} w_{i,j}n_j\right)\left(P_{n_i}(t)-P_{n_i+1}(t)\right),$$

for $i\in G_l$. Here α represents a decay rate. It is used to account the fact that spikes are effective only for a time interval of approximately $\frac{1}{\alpha}$. By using Doi techniques, see e.g. [17], [20], [76], involving creation-annihilation operator formalism, the dynamics of the network can be described by a vacuum ket $|0\rangle$ using a pair creation-annihilation operators at each site: $\left[\Phi_i,\Phi_j^{\dagger}\right]=\delta_{i,j}$. The state of the system is described by

$$|\phi\left(t\right)\rangle=\sum_{\boldsymbol{n}}P_{\boldsymbol{n}}(t)\prod_{i\in G_l}\Phi_i^{n_i\dagger}\left|0\right\rangle,$$

where the summation is taken over all configurations $\boldsymbol{n}$. In this operator formalism the master equation takes the form

$$\frac{\partial}{\partial t}\left|\phi\left(t\right)\right\rangle=-\widehat{H}\left|\phi\left(t\right)\right\rangle,$$

where

$$\widehat{H}=\sum_{i\in G_l}\alpha\Phi_i^{\dagger}\Phi_i-\sum_{i\in G_l}\Phi_i^{\dagger}f\left(\sum_{j\in G_l}w_{i,j}\left[\Phi_j^{\dagger}\Phi_j+\Phi_j\right]\right).$$

Now by using the work of Peliti, see e.g. [17], [20], [76], the Hamiltonian $\widehat{H}$ can be studied using a path integral. The corresponding action takes the following form (in the continuous time limit):

$$S\left[\phi_i\left(t\right),\widetilde{\phi}_i\left(t\right)\right]= \tag{5.5}$$
$$\int_0^t dt\left\{\sum_{i\in G_l}\widetilde{\phi}_i\partial_t\phi_i+\alpha\phi_i\widetilde{\phi}_i-\widetilde{\phi}_i f\left(\sum_{j\in G_l}w_{i,j}\left[\widetilde{\phi}_j\phi_j+\phi_j\right]\right)\right\},$$

where dt denotes the Lebesgue measure of the real line. To obtain the continuous limit in the spatial variable $i\to x$ in (5.5), Buice and Cowan assumed that the sum in (5.5) runs over a square lattice and thus by a limit process a Riemann-type integral is obtained. In our case, this sum gives rise to and integral with respect to the Haar measure of $\mathbb{Q}_p$:

$$S\left[\phi\left(x,t\right),\widetilde{\phi}\left(x,t\right)\right]=\int_0^t dt\int_{\mathbb{Q}_p}dx\left\{\sum_{i\in G_l}\widetilde{\phi}\partial_t\phi+\alpha\phi_i\widetilde{\phi}_i-\widetilde{\phi}_i f\left(w*\left[\widetilde{\phi}\phi+\phi\right]\right)\right\}.$$

Problem 3. *To develop a non-Archimedean counterpart of the Buice-Cowan theory.*

5.3. **Final comments.** The connections between statistical mechanics and deep learning has been studied intensively in the last ten years, see e.g. [2], [8], [19], [30], [69], [70], [84]. To the best of our knowledge the theoretical results presented here about the p-adic DBNs are new. Here we should mention that p-adic neural networks have been considered before in [4], [55]-[56], [91], but these computational models are completely different to the ones considered here. We finally mention that Khrennikov et al. have developed hierarchical models (based in p-adic numbers) for brain activity and EEG analysis with applications in the diagnosis of mental diseases see e.g. [48]-[50], [85].

6. Conclusions

In this work we initiated the study of the correspondence between p-adic SFTs and NNs. An important advantage of the p-adic SFTS over the classical ones is that discretization process that produces discrete SFTs can be carried out in rigorous mathematical way in many relevant cases, for instant in the ϕ^4-theories [92]. A p-adic discrete SFT corresponds to an energy functional defined on a tree G_l, this functional defines a NN whose neurons are organized hierarchically in a tree-like structure. This type of networks are a new particular class of the deep belief networks introduced by Hinton et al. [43]-[44]. A DBN is constructed by stacking several RBMs, the goal of this construction is to get a network where the neurons are organized hierarchically in large tree-like structure a deep learning architecture). A classical DBN correspond naturally to a certain spin glass, then a DBN should correspond to an ultrametric spin glass, which is a system whose space states is hierarchically organized. Ultrametricity was discovered by Parisi et al. in spin glass theory [23], [75], [80]. Ultrametricity takes a simple form in the p-adic world [79].

A p-adic continuous DBN is a SFT, in this case the neurons correspond to the points of the p-adic unit ball $\mathbb{Z}_p$, and thus, the neurons are organized in an infinite rooted tree. A discrete version of this theory corresponds to a p-adic discrete DBN. Intuitively, the discrete version is obtained by cutting the tree $\mathbb{Z}_p$ at level $l \geq 1$. It is expected that in the limit l tends infinite the discrete theories approach to the continuous ones. This behavior is radically different to the one presented in [40]. In this work, in the limit when the number of neurons tend to infinity the network corresponds to a non-interacting QFT, while in the finite case corresponds a interacting QFT. Here, in both cases we have interacting QFTs. The p-adic discrete DBNs are universal approximators, Theorems 1, 2. To establish this result we adapt the techniques developed by Le Roux and Bennagio in [63].

It is important emphasize that the correspondence between SFTs and NNs takes different forms depending on the architecture of the networks. The case in which the architecture is embedded in a stochastic differential equation or in a discrete mater equation has been studied intensively lately, see e.g. [9], [36], [40], see also [17], [21]. The p-adic counterpart of this correspondence is an open problem. It is widely accepted that the brain activity is organized hierarchically. Here we pointed out that the Buice-Cowan theory of fluctuating activity of cortical networks has a p-adic counterpart, where the neurons are organized in an infinite tree-like structure.

In our view, the fully development of the p-adic counterpart of the Buice-Cowan theory is a relevant matter.

References

[1] A. Abdesselam, A. Chandra, G. Guadagni, *Rigorous quantum field theory functional integrals over the $p-$adics I: Anomalous dimensions,* (2013)*,* https://arxiv.org/abs/1302.5971.

[2] Madhu Advani, Subhaneil Lahiri, Surya Ganguli, *Statistical mechanics of complex neural systems and high dimensional data*, J. Stat. Mech. Theory Exp. 2013, no. 3, P03014, 66 pp., https://doi.org/10.1088/1742-5468/2013/03/P03014.

[3] S. Albeverio, A. Yu. Khrennikov, V. M. Shelkovich, *Theory of p-adicdistributions: linear and nonlinear models.* Cambridge University Press, Cambridge, ISBN 9781139119283 (2010).

[4] Sergio Albeverio, Andrei Khrennikov, Brunello Tirozzi, *p-adic dynamical systems and neural networks*, Math. Models Methods Appl. Sci. **9**, no. 9, 1417–1437, (1999), https://doi.org/10.1142/S0218202599000634.

[5] Edilberto Arroyo-Ortiz, W. A. Zúñiga-Galindo, *Construction of p-adic covariant quantum fields in the framework of white noise analysis*, Rep. Math. Phys. **84**(1), 1–34 (2019), https://doi.org/10.1016/S0034-4877(19)30066-7.

[6] V. A. Avetisov, A. Kh. Bikulov, V. A. Osipov, *p-Adic description of characteristic relaxation in complex systems*, J. Phys. A **36**(15), 4239–4246 (2003), https://doi.org/10.1088/0305-4470/36/15/301.

[7] V. A. Avetisov, A. H. Bikulov, S. V. Kozyrev, V. A. Osipov, *p-Adic models of ultrametric diffusion constrained by hierarchical energy landscapes*, J. Phys. A **35**(2), 177–189 (2002), https://doi.org/10.1088/0305-4470/35/2/301.

[8] Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, Sam S. Schoenholz, Jascha Sohl-Dickstein, Surya Ganguli, *Statistical Mechanics of Deep Learning,* Annual Review of Condensed Matter Physics **11**, 501-528 (2020), https://doi.org/10.1146/annurev-conmatphys-031119-050745.

[9] D. Bachtis, G. Aarts, B. Lucini, *Quantum field-theoretic machine learning*, Physical Review D, **103**(7) (2021). https://doi.org/10.1103/physrevd.103.074510

[10] Bengio Yoshua, *Learning Deep Architectures for AI, Foundations and Trends* in Machine Learning: Vol. 2: No. 1, pp 1-127 (2009). http://dx.doi.org/10.1561/2200000006.

[11] M. Bocardo-Gaspar, Willem Veys, W. A. Zúñiga-Galindo, *Meromorphic continuation of Koba-Nielsen string amplitudes,* J. High Energy Phys. **9**, 138, 43 pp. (2020), https://doi.org/10.1007/JHEP09(2020)138.

[12] Miriam Bocardo-Gaspar, H. García-Compeán, W. A. Zúñiga-Galindo, *Regularization of p-adic string amplitudes, and multivariate local zeta functions*, Lett. Math. Phys. **109**, no. 5, 1167–1204 (2019), https://doi.org/10.1007/s11005-018-1137-1.

[13] M. Bocardo-Gaspar, H. García-Compeán, W. A. Zúñiga-Galindo, *On p-adic string amplitudes in the limit p approaches to one*, J. High Energy Phys. **8**, 043, front matter+22 pp. (2018), https://doi.org/10.1007/JHEP08(2018)043.

[14] L. Brekke, P. G. O. Freund, M. Olson, E. Witten*, Non-Archimedean string dynamics*, Nuclear Phys. B **302**(3), 365–402 (1988), https://doi.org/10.1016/0550-3213(88)90207-6.

[15] L. Brekke, P. G. O. Freund, *$p-$adic numbers in Physics*, Phys. Rep. **233**(1), 1–66 (1993), https://doi.org/10.1016/0370-1573(93)90043-D.

[16] Michael A. Buice, Carson C. Chow, *Beyond mean field theory: statistical field theory for neural networks*, J. Stat. Mech. Theory Exp. **3**, P03003, 21 pp. (2013), https://doi.org/10.1088/1742-5468/2013/03/P03003.

[17] Michael A. Buice, Jack D. Cowan, *Field-theoretic approach to fluctuation effects in neural networks*, Phys. Rev. E **75**, no. 5, 051919, 14 pp. (2007), https://link.aps.org/doi/10.1103/PhysRevE.75.051919.

[18] Michael A. Buice, Jack D. Cowan, Carson C. Chow, *Systematic Fluctuation Expansion for Neural Network Activity Equations*, Neural Comput **22** (2), 377–426 (2010), https://doi.org/10.1162/neco.2009.02-09-960

[19] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld et al., *Machine learning and the physical sciences*, Rev. Mod. Phys. **91**, 045002 (2019), https://link.aps.org/doi/10.1103/RevModPhys.91.045002.

[20] Carson C. Chow, Michael A. Buice, *Path integral methods for stochastic differential equations*, J. Math. Neurosci. **5**, Art. 8, 35 pp. (2015), https://doi.org/10.1186/s13408-015-0018-5.

[21] Stephen Coombes, Peter beim Graben, Roland Potthast, James Wright, Editors, *Neural fields. Theory and applications*. Springer, Heidelberg, (2014).

[22] Aurélien Decelle, Cyril Furtlehner, *Restricted Boltzmann machine: Recent advances and mean-field theory*, Chin. Phys. B, 2021, **30**(4), 040202-040202, DOI: 10.1088/1674-1056/abd160

[23] B. Dragovich, A. Yu. Khrennikov, S. V. Kozyrev, I. V. Volovich, *Giorgio Parisi: the Nobel Prize in physics 2021*, p-Adic Numbers Ultrametric Anal. Appl. **14** , no. 1, 81–83 (2022), https://doi.org/10.1134/S207004662201006X.

[24] B. Dragovich, *On $p-$adic and Adelic generalization of quantum field theory*, Nucl. Phy. B Proc. Suppl. **102-103**, 150-155 (2001), https://doi.org/10.1016/S0920-5632(01)01550-X.

[25] B. Dragovich, A. Yu. Khrennikov, S. V Kozyrev, I. V. Volovich, *On $p-$adic mathematical physics*, $p-$Adic Numbers Ultrametric Anal. Appl. **1** (1), 1–17 (2009), https://doi.org/10.1134/S2070046609010014.

[26] G. S. Djordjević, B. Dragovich, *$p-$Adic Path Integrals for Quadratic Actions*, Mod. Phys. Lett. A **12**, 1455-1463 (1997), https://doi.org/10.1142/S0217732397001485.

[27] Shi Dong, Ping Wang, Khushnood Abbas, *A survey on deep learning and its applications*, Comput. Sci. Rev. **40**, Paper No. 100379, 22 pp. (2021), https://doi.org/10.1016/j.cosrev.2021.100379.

[28] E. Dyer and G. Gur-Ari, *Asymptotics of wide networks from Feynman diagrams*, https://arxiv.org/abs/1909.11304.

[29] F. J. Dyson, *An Ising ferromagnet with discontinuous long-range order*, Comm. Math. Phys. **21**, 269–283 (1971), https://doi.org/10.1007/BF01645749.

[30] Engel A, den Broeck C.V., *Statistical Mechanics of Learning. Cambridge*,UK, Cambridge Univ. Press (2001), https://doi.org/10.1017/CBO9781139164542.

[31] H. Erbin, V. Lahoche, D. Ousmane Samary, *Nonperturbative renormalization for the neural network-QFT correspondence*, Mach. Learn. Sci. Tech. (2022), https://iopscience.iop.org/article/10.1088/2632-2153/ac4f69.

[32] A. Fischer, C. Igel, *An Introduction to Restricted Boltzmann Machines*. In: Alvarez, L., Mejail, M., Gomez, L., Jacobo, J. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2012. Lecture Notes in Computer Science, vol 7441, Springer, Berlin, Heidelberg, (2012), https://doi.org/10.1007/978-3-642-33275-3_2.

[33] A. R. Fuquen-Tibatá, H. García-Compeán, W. A. Zúñiga-Galindo, *Euclidean quantum field formulation of p-adic open string amplitudes*, Nuclear Phys. B **975**, Paper No. 115684, 27 pp. (2022), https://doi.org/10.1016/j.nuclphysb.2022.115684.

[34] H. García-Compeán, Edgar Y. López, W. A. Zúñiga-Galindo, *p-Adic open string amplitudes with Chan-Paton factors coupled to a constant B-field*, Nuclear Phys. B **951**, 114904, 33 pp. (2020), https://doi.org/10.1016/j.nuclphysb.2019.114904.

[35] James Glimm, Arthur Jaffe, *Quantum physics. A functional integral point of view*. Second edition. New York, Springer-Verlag (1987), ISBN: 978-1-4612-4728-9.

[36] Kevin T. Grosvenor, Ro Jefferson, *The edge of chaos: quantum field theory and deep neural networks*, SciPost Phys. **12** (2022), no. 3, Paper No. 081, 65 pp., DOI: 10.21468/SciPost-Phys.12.3.081.

[37] Steven S. Gubser, Christian B. Jepsen, Ziming Ji, Brian Trundy, Amos Yarom, *Non-local non-linear sigma models*, J. High Energy Phys. **9**, 005, 36 pp. (2019), https://doi.org/10.1007/JHEP09(2019)005.

[38] S. S. Gubser, J. Knaute, S. Parikh, A. Samberg, P. Witaszczyk, *p-adic AdS/CFT*, Commun. Math. Phys. **352**(3), 1019 (2017), https://doi.org/10.1007/s00220-016-2813-6.

[39] S. S. Gubser, Ch. Jepsen, Z. Ji and B. Trundy, *Continuum limits of sparse coupling patterns*, Phys. Rev. D **98**(4), 045009, (2018), https://doi.org/10.1103/PhysRevD.98.045009.

[40] J. Halverson, A. Maiti and K. Stoner, *Neural networks and quantum field theory*, Mach. Learn. Sci. Technol. **2**, 035002 (2021), doi:10.1088/2632-2153/abeca3.

[41] D. Harlow, S. H. Shenker, D. Stanford, L. Susskind, *Tree-like structure of eternal inflation: A solvable model*, Physical Review D **85**, 063516 (2012), https://link.aps.org/doi/10.1103/PhysRevD.85.063516.

[42] Takeyuki Hida ,Hui-Hsiung Kuo, Jürgen Potthoff, Ludwig Streit, *White noise. An infinite dimensional calculus*, Dordrecht, Kluwer Academic Publishers (1993), ISBN 9789401736817, 9401736812.

[43] G. E. Hinton, R.R. Salakhutdinov, *Reducing the dimensionality of data with neural networks*, Science, **313**, 5786 (2006), doi: 10.1126/science.112764.

[44] Honglak Lee, Roger Grosse, Rajesh Ranganath, Andrew Y. Ng, *Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks*, Communications of the ACM, **54**, no. 10, pp. 95-103, (2011), https://doi.org/10.1145/2001269.2001295.

[45] Hagen Kleinert, V. Schulte-Frohlinde, C*ritical properties of* ϕ^4*-theories*, Singapore, World Scientific (2001), doi:10.1142/9789812799944.

[46] A. Yu. Khrennikov, *Representation of second quantization over non-Archimedean number fields*, Soviet Phys. Dokl. **35**(10), 867–869 (1990), .

[47] A. Yu Khrennikov, *The Schrödinger and Bargmann-Fock representations in non-Archimedean quantum mechanics*, Sov. Phys., Dokl. **35**(7), 638-640 (1990).

[48] A. Yu. Khrennikov, *Non-Archimedean Analysis: Quantum Paradoxes, Dynamical Systems and Biological Models*, Kluwer Academic Publishers, (1997), ISBN 978-94-009-1483-4.

[49] A. Khrennikov, *Information Dynamics in Cognitive, Psychological, Social and Anomalous Phenomena,* Springer, (2004), ISBN 978-94-017-0479-3.

[50] Andrei Khrennikov, *Probabilistic pathway representation of cognitive information,* Journal of Theoretical Biology, **231**, Issue 4, 2004, Pages 597-613, https://doi.org/10.1016/j.jtbi.2004.07.015.

[51] A. Yu. Khrennikov, *p-Adic Valued Distributions in Mathematical Physics*, Dordrecht, Kluwer Academic Publishers (1994), ISBN 978-94-015-8356-5.

[52] Andrei Khrennikov, Sergei Kozyrev, W. A. Zúñiga-Galindo, *Ultrametric Equations and its Applications. Encyclopedia of Mathematics and its Applications 168*, Cambridge, Cambridge University Press (2018), ISBN 9781107188822.

[53] A. Yu. Khrennikov, F. M. Mukhamedov, J. F. F. Mendes, *On p-adic Gibbs measures of the countable state Potts model on the Cayley tree*, Nonlinearity **20**(12), 2923–2937 (2007), https://doi.org/10.1088/0951-7715/20/12/010.

[54] A. Yu. Khrennikov, S.V. Kozyrev, *Replica symmetry breaking related to a general ultrametric space I: Replica matrices and functionals*, Physica A: Statistical Mechanics and its Applications, **359**, 222-240 (2006), https://doi.org/10.1016/j.physa.2005.05.077.

[55] A.Y. Khrennikov, M. Nilson, *p-Adic Deterministic and Random Dynamics. Mathematics and Its Applications*, vol 574, Springer, Dordrecht (2004), https://doi.org/10.1007/978-1-4020-2660-7_8.

[56] Andrei Khrennikov, Brunello Tirozzi, *Learning of p-adic neural networks.* Stochastic processes, physics and geometry: new interplays, II (Leipzig, 1999), 395–401, CMS Conf. Proc., 29, Amer. Math. Soc., Providence, RI, (2000).

[57] Neal Koblitz, *p-Adic Numbers, p-adic Analysis, and Zeta-Functions. Graduate Texts in Mathematics No. 58,* New York, Springer-Verlag (1984), ISBN 978-1-4612-1112-9.

[58] Anatoly N. Kochubei, *Pseudo-differential equations and stochastics over non-Archimedean fields*, New York, Marcel Dekker, Inc. (2001), ISBN 0824706552.

[59] Anatoly N. Kochubei, Mustafa R. Sait-Ametov, *Interaction measures on the space of distributions over the field of p-adic numbers*, Infin. Dimens. Anal. Quantum Probab. Relat. Top. **6**(3), 389–411 (2003), https://doi.org/10.1142/S0219025703001353.

[60] Honglak Lee, Roger Grosse, Rajesh Ranganath, Andrew Y. Ng, *Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations.*ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning 2009, Pages 609–616, https://doi.org/10.1145/1553374.1553453.

[61] Edwin León-Cardenal, W. A. Zúñiga-Galindo, *An introduction to the theory of local zeta functions from scratch*, Rev. Integr. Temas Mat. 3**7** no. 1, 45–76 (2019), https://doi.org/10.18273/revint.v37n12019004.

[62] E. Y. Lerner, M. D. Misarov, *Scalar models in* $p-$*adic quantum field theory and hierarchical models*, Theor. Math. Phys. **78**, 177–184 (1989), https://doi.org/10.1007/BF01018683.

[63] Nicolas Le Roux, Yoshua Bengio, *Representational power of restricted Boltzmann machines and deep belief networks*, Neural Comput. **20** , no. 6, 1631–1649 (2008), doi: https://doi.org/10.1162/neco.2008.04-07-510.

[64] Nicolas Le Roux, Yoshua Bengio, *Deep belief networks are compact universal approximators*, Neural Comput. **22**, no. 8, 2192–2207 (2010), https://doi.org/10.1162/neco.2010.08-09-1081.
[65] Soon Hoe Lim, *Understanding recurrent neural networks using nonequilibrium response theory*, J. Mach. Learn. Res. **22**, Paper No. 47, 48 pp. (2021), https://www.jmlr.org/papers/volume22/20-620/20-620.pdf.
[66] Soon Hoe Lim, N. Benjamin Erichson, Liam Hodgkinson, Michael W. Mahoney, *Noisy Recurrent Neural Networks in Advances* in Neural Information Processing Systems 34 (NeurIPS 2021), Edited by: M. Ranzato and A. Beygelzimer and Y. Dauphin and P.S. Liang and J. Wortman Vaughan. Pages 5124–5137. Curran Associates, Inc., (2021).
[67] A. Maiti, K. Stoner and J. Halverson, *Symmetry-via-duality: Invariant neural network densities from parameter-space correlators*, https://arxiv.org/abs/2106.00694.
[68] M. L. Mendoza-Martínez, J. A. Vallejo, W. A. Zúñiga-Galindo, *Acausal quantum theory for non-Archimedean scalar fields*, Rev. Math. Phys. **31**(4), 1950011, 46 pp. (2019), https://doi.org/10.1142/S0129055X19500119.
[69] P. Mehta, M. Bukov, C-H. Wang, A. G. R. Day, C. Richardson et al., *A high-bias, low-variance introduction to Machine Learning for physicists*, Phys. Rep. **810**, 1–124 (2019), doi: 10.1016/j.physrep.2019.03.001.
[70] M. Mézard, A. Montanari, *Information, Physics, and Computation*, New York: Oxford Univ. Press, (2009), doi 10.1093/acprof:oso/9780198570837.001.0001.
[71] M. D. Missarov, *Random fields on the adele ring and Wilson's renormalization group*, Annales de l'institut Henri Poincaré (A): Physique Theorique **50**(3) , 357– 367 (1989), http://www.numdam.org/item/AIHPA_1989__50_3_357_0/.
[72] M. D. Missarov, *$p-$Adic φ^4-theory as a functional equation problem*, Lett. Math. Phys. **39**, 253-260 (1997), https://doi.org/10.1023/A:1007303725761.
[73] M. D. Missarov, *$p-$Adic renormalization group solutions and the Euclidean renormalization group conjectures,* p-Adic Numbers Ultrametric Anal. Appl. **4**, 109-114 (2012), https://doi.org/10.1134/S2070046612020033.
[74] M. D. Missarov, *The continuum limit in the fermionic hierarchical model*, Theoret. and Math. Phys. **118**, 32–40 (1999), https://doi.org/10.1007/BF02557193.
[75] M. Mézard, G. Parisi, M. A. Virasoro, *Spin glass theory and beyond*, World Scientific, (1987), https://doi.org/10.1142/0271.
[76] Moritz Helias, DavidDahmen, *Statistical field theory for neural networks. Lecture Notes in Physics, 970*, Springer, Cham (2020), https://doi.org/10.1007/978-3-030-46444-8.
[77] F. Mukhamedov and H. Akın, "Phase transitions for p-adic Potts model on the Cayley tree of order three," *J. Stat. Mech. Theory Exp.* **7**, P07014, (2013).
[78] F. Mukhamedov, M. Saburov, O. Khakimov, *On p-adic Ising-Vannimenus model on an arbitrary order Cayley tree*, J. Stat. Mech. Theory Exp. **5**, P05032, (2015), https://doi.org/10.1088/1742-5468/2015/05/P05032.
[79] G. Parisi and N. Sourlas, *p-Adic numbers and replica symmetry breaking*, Eur. Phys. J. B **14**, 535–542 (2000), https://doi.org/10.1007/s100510051063.
[80] R. Rammal, G. Toulouse, M. A. Virasoro, *Ultrametricity for physicists*, Rev. Modern Phys. **58**, 765–788 (1986), https://link.aps.org/doi/10.1103/RevModPhys.58.765.
[81] D. A. Roberts, S. Yaida and B. Hanin, *The principles of deep learning theory*, https://arxiv.org/abs/2106.10165.
[82] Barry Simon, *The $P(\phi)_2$ Euclidean (quantum) field theory*, Princeton University Press (1974), ISBN 9780691645490.
[83] Ya. G. Sinaǐ , *Theory of phase transitions: rigorous results*,Pergamon Press (1982), ISBN 1483126072.
[84] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, Surya Ganguli, *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*, Proceedings of the 32nd International Conference on Machine Learning, PMLR **37**, 2256-2265, (2015), http://proceedings.mlr.press/v37/sohl-dickstein15.pdf.
[85] Oded Shor, Amir Glik, Amit Yaniv-Rosenfeld, Avi Valevski, Abraham Weizman, Andrei Khrennikov, Felix Benninger, *EEG p-adic quantum potential accurately identifies depression, schizophrenia and cognitive decline*, PLoS ONE 16(8): e0255529 (2021), https://doi.org/10.1371/journal.pone.0255529.
[86] M. H. Taibleson, *Fourier analysis on local fields*, Princeton University Press (1975), ISBN 9780691645162.

[87] V. S. Vladimirov, I. V. Volovich, E. I. Zelenov, *p-Adic analysis and mathematical physics,* Singapore, World Scientific (1994), http://dx.doi.org/10.1142/1581.
[88] I. V. Volovich, *Number theory as the ultimate physical theory*, p-Adic Numbers Ultrametric Anal. Appl. **2**, 77–87 (2010), https://doi.org/10.1134/S2070046610010061.
[89] S. Yaida, *Non-Gaussian processes and neural networks at finite widths*, https://arxiv.org/abs/1910.00019.
[90] A. V. Zabrodin, *Non-Archimedean strings and Bruhat-Tits trees*, Comm. Math. Phys. **123**, 463–483 (1989), https://doi.org/10.1007/BF01238811.
[91] B.A. Zambrano-Luna, W.A. Zúñiga-Galindo, *p-adic Cellular Neural Networks*, J Nonlinear Math Phys (2022), https://doi.org/10.1007/s44198-022-00071-8.
[92] W. A. Zúñiga-Galindo, *Non-Archimedean statistical field theory*, Rev. Math. Phys, https://doi.org/10.1142/S0129055X22500222.
[93] W. A. Zúñiga-Galindo, *Eigen's paradox and the quasispecies model in a non-Archimedean framework*, Physica A: Statistical Mechanics and its Applications, **602**, 127648 (2022), https://doi.org/10.1016/j.physa.2022.127648.
[94] W. A. Zúñiga-Galindo , *Ultrametric diffusion, rugged energy landscapes and transition networks*, Physica A: Statistical Mechanics and its Applications **597**, 127221 (2022), https://doi.org/10.1016/j.physa.2022.127221.
[95] W. A. Zúñiga-Galindo, B. A. Zambrano-Luna, E. León-Cardenal, *Graphs, local zeta functions, log-Coulomb gases, and phase transitions at finite temperature*, J. Math. Phys. **63**, 013506 (2022), https://doi.org/10.1063/5.0070683.
[96] W. A. Zúñiga-Galindo, Sergii M., Torba, *Non-Archimedean Coulomb gases*, J. Math. Phys. **61**, 013504 (2020) https://doi.org/10.1063/1.5127191.
[97] W. A. Zúñiga-Galindo, *Non-Archimedean white noise, pseudodifferential stochastic equations, and massive Euclidean fields,* J. Fourier Anal. Appl. **23**, 288–323 (2017), https://doi.org/10.1007/s00041-016-9470-1.
[98] W. A. Zúñiga-Galindo, *Pseudodifferential equations over non-Archimedean spaces. Lectures Notes in Mathematics 2174,* Switzerland, Springer, (2016), https://doi.org/10.1007/978-3-319-46738-2.
[99] W. A. Zúñiga-Galindo, *The non-Archimedean stochastic heat equation driven by Gaussian noise*, J. Fourier Anal. Appl. **21**, 600–627 (2015), https://doi.org/10.1007/s00041-014-9383-9.

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`