PTEP

Prog. Theor. Exp. Phys. **2023** 063A01(17 pages)
DOI: 10.1093/ptep/ptad061

p-Adic statistical field theory and convolutional deep Boltzmann machines

W. A. Zúñiga-Galindo [1], C. He [2], and B. A. Zambrano-Luna [1]

[1] *University of Texas Rio Grande Valley, School of Mathematical and Statistical Sciences, One West University Blvd., Brownsville, TX 78520, USA*
[2] *Oklahoma State University, Department of Mathematics, MSCS 425, Stillwater, OK, USA*
*E-mail: wilson.zunigagalindo@utrgv.edu
The author was partially supported by the Lokenath Debnath Endowed Professorship.

Received February 15, 2023; Accepted May 9, 2023; Published May 10, 2023

Understanding how deep learning architectures work is a central scientific problem. Recently, a correspondence between neural networks (NNs) and Euclidean quantum field theories has been proposed. This work investigates this correspondence in the framework of p-adic statistical field theories (SFTs) and neural networks. In this case, the fields are real-valued functions defined on an infinite regular rooted tree with valence p, a fixed prime number. This infinite tree provides the topology for a continuous deep Boltzmann machine (DBM), which is identified with a statistical field theory on this infinite tree. In the p-adic framework, there is a natural method to discretize SFTs. Each discrete SFT corresponds to a Boltzmann machine with a tree-like topology. This method allows us to recover the standard DBMs and gives new convolutional DBMs. The new networks use $O(N)$ parameters while the classical ones use $O(N^2)$ parameters.

Subject Index A12, A44, A45, B37, B38

1. Introduction

Deep neural networks have been successfully applied to various tasks, including self-driving cars, natural language processing, and visual recognition, among many others [1,2]. There is a consensus about the need to develop a theoretical framework to understand how deep learning architectures work. Recently, physicists have proposed the existence of a correspondence between neural networks (NNs) and quantum field theories (QFTs), more precisely statistical field theory (SFT); see Refs. [3–12] and references therein. This correspondence takes different forms depending on the architecture of the networks involved.

In Ref. [12], a study of the above-mentioned correspondence was initiated in the framework of the non-Archimedean statistical field theory (SFT). In this case, the background space (the set of real numbers) is replaced by a set of p-adic numbers, where p is a fixed prime number. The p-adics are organized in a tree-like structure; this feature facilitates the description of hierarchical architectures. In Ref. [12], a p-adic version of the convolutional deep Boltzmann machines (DBMs) is introduced where only binary data are considered and with no implementation. By

Funded by SCOAP³. SCOAP3 and OUP support the goals of the International Year of Basic Sciences for Sustainable Development

adapting the mathematical techniques introduced by Le Roux and Benigio in Ref. [13], the author shows that these machines are universal approximators for binary data tasks.

In this article, we continue discussing the correspondence between SFTs and NNs in the p-adic framework. Compared with Ref. [12], here we consider more general architectures and data types. We note that dealing with general data is challenging both in theory and implementation. We argue that p-adic analysis still provides the right framework to understand the dynamics of NNs with large tree-like hierarchical architectures. In our approach, a NN corresponds to the discretization of a p-adic SFT. The discretization process is carried out in a rigorous and general way. Moreover, such discretization allows us to obtain many recently developed DBMs. For instance, the NNs constructed in Ref. [5] are a particular case of those introduced here. We also discuss the implementation of a class of p-adic convolutional networks and obtain desired results on a feature detection task based on hand-writing images of decimal digits.

The main novelty of our p-adic convolutional DBMs is that they use significantly fewer parameters than the conventional ones. A detailed discussion is given in Sect. 6. We note that the connections between p-adic numbers and neural networks have been considered before. Neural networks whose states are p-adic numbers were studied in Refs. [14,15]. These models are completely different from those considered here. These ideas have been used to develop non-Archimedean models of brain activity and mental processes [16]. In Refs. [17,18], p-adic versions of cellular neural networks were studied. These models involved abstract evolution equations.

2. p-adic numbers

In this section, we introduce basic concepts for the p-adic numbers. For more detailed information, refer to Appendix A.

From now on, p denotes a fixed prime number. Any non-zero p-adic number x has a unique expansion of the form

$$x = x_{-k}p^{-k} + x_{-k+1}p^{-k+1} + \cdots + x_0 + x_1 p + \cdots ,$$

with $x_{-k} \neq 0$, where k is an integer, and the x_j are numbers from the set $\{0, 1, \ldots, p-1\}$. The set of all possible numbers of such a form constitutes the field of p-adic numbers $\mathbb{Q}_p$. There are natural field operations, sum and multiplication, on p-adic numbers; see, e.g., Ref. [19]. There is also a natural norm in $\mathbb{Q}_p$ defined as $|x|_p = p^k$ where k depends on x, for a non-zero p-adic number x.

The field of p-adic numbers with the distance induced by $|\cdot|_p$ is a complete ultrametric space. The ultrametric property refers to the fact that $|x-y|_p \leq \max\{|x-z|_p, |z-y|_p\}$ for any x, y, z in $\mathbb{Q}_p$. In this article, we work with p-adic integers, which are p-adic numbers satisfying $-k \geq 0$. All such p-adic integers constitute the unit ball $\mathbb{Z}_p$. The unit ball is closed under addition and multiplication, so it is a commutative ring. Throughout this article, we work mainly with locally constant functions supported in the unit ball, i.e., with functions of type $\varphi : \mathbb{Z}_p \to \mathbb{R}$, such that $\varphi(a+x) = \varphi(a)$ for all x in $\mathbb{Z}_p$. The simplest example of such a function is the chararcterisstic function $1_{\mathbb{Z}_p}(x)$ of the unit ball $\mathbb{Z}_p$: $1_{\mathbb{Z}_p}(x) = 1$ if $|x|_p \leq 1$, otherwise $1_{\mathbb{Z}_p}(x) = 0$. To check that $1_{\mathbb{Z}_p}(a+x) = 1_{\mathbb{Z}_p}(a)$, we use that $\mathbb{Z}_p$ is closed under addition. If $|x|_p \leq p^{-l}$, where the integer l is fixed and independent of a. We denote by $\mathcal{D}(\mathbb{Z}_p)$ the real vector space of test functions supported in the unit ball. There is a natural integration theory so that $\int_{\mathbb{Z}_p} \varphi(x)\, dx$ gives a

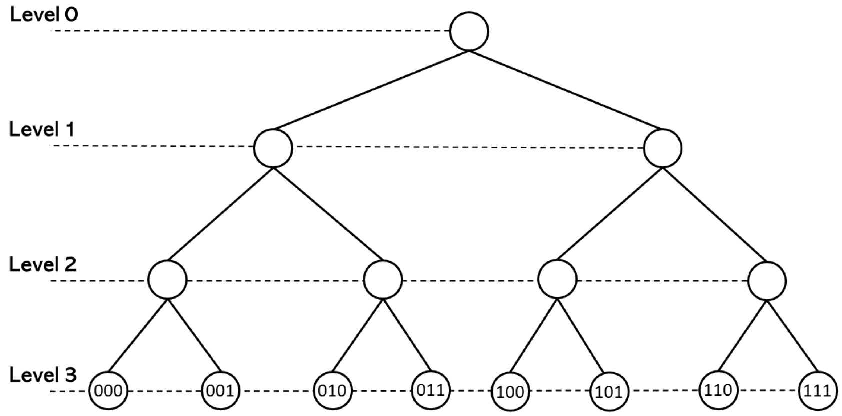

Fig. 1. The rooted tree associated with the group $\mathbb{Z}_2/2^3\mathbb{Z}_2$. The elements of $\mathbb{Z}_2/2^3\mathbb{Z}_2$ have the form $i = i_0 + i_1 2 + i_2 2^2$, $i_0, i_1, i_2 \in \{0, 1\}$. The distance satisfies $-\log_2|i - j|_2$ =level of the first common ancestor of i, j.

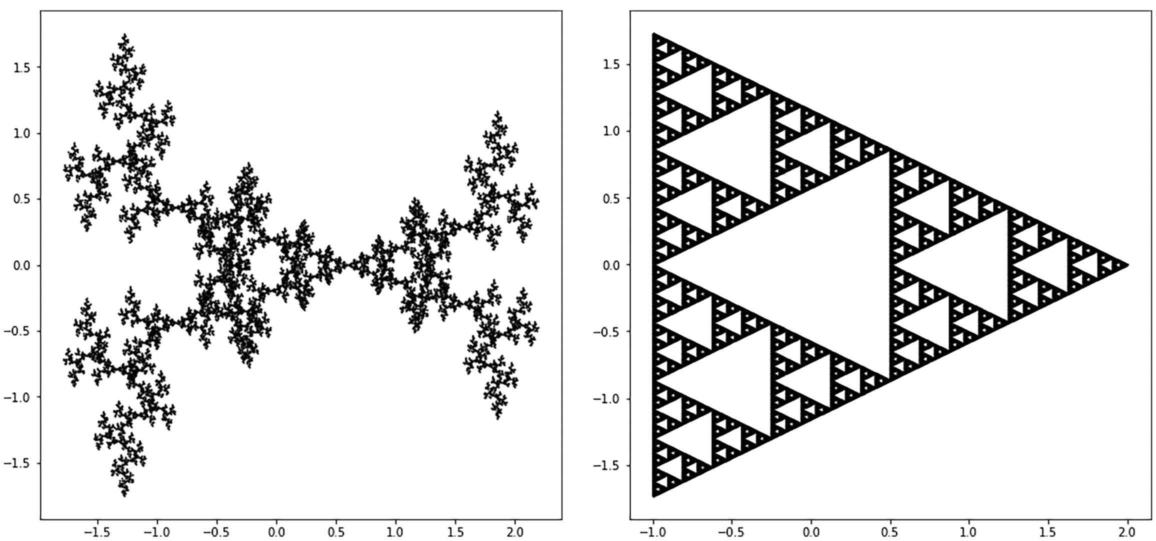

Fig. 2. Based upon Ref. [20], we construct an embedding $\mathfrak{f} : \mathbb{Z}_p \to \mathbb{R}^2$. The figure shows the images of $\mathfrak{f}(\mathbb{Z}_2)$ and $\mathfrak{f}(\mathbb{Z}_3)$. This computation requires a truncation of the p-adic integers. We use $\mathbb{Z}_2/2^{14}\mathbb{Z}_2$ and $\mathbb{Z}_3/3^{10}\mathbb{Z}_3$, respectively.

well-defined real number. The measure dx is the so-called Haar measure of $\mathbb{Q}_p$. Further details are given in Appendix A.3.

Since the p-adic numbers are infinite series, any computational implementation involving these numbers requires a truncation process: $x \longmapsto x_0 + x_1 p + \ldots + x_{l-1}p^{l-1}$, $l \geq 1$. The set of all truncated integers mod p^l is denoted as $G_l = \mathbb{Z}_p/p^l\mathbb{Z}_p$. This set can be represented as a rooted tree with l levels; see Fig. 1.

The unit ball $\mathbb{Z}_p$ is an infinite rooted tree with fractal structure; see Fig. 2. Appendix A provides a review of the basic aspects of the p-adic analysis required here. We note that the word "field" will be used here in two different contexts throughout the article. In a mathematical context, we refer to algebraic fields; in a physical context, we refer to Euclidean quantum fields.

3. Non-Archimedean $\{\boldsymbol{v}, \boldsymbol{h}\}^4$-statistical field theories

We fix $a(x)$, $b(x)$, $c(x)$, $d(x) \in \mathcal{D}(\mathbb{Z}_p)$, $e \in \mathbb{R}$, and an integrable function $w(x) : \mathbb{Z}_p \to \mathbb{R}$. A *p-adic continuous Boltzmann machine (or a p-adic continuous BM)* is a statistical field theory involving two scalars fields $\boldsymbol{v}$, $\boldsymbol{h}$. 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*. We assume that the field $\{\boldsymbol{v}, \boldsymbol{h}\}$ performs thermal fluctuations and that the expectation value of the field is zero.

The size of the fluctuations is controlled by an energy functional of the form

$$E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) := E(\boldsymbol{v}, \boldsymbol{h}) = E_0(\boldsymbol{v}, \boldsymbol{h}) + E_{\text{int}}(\boldsymbol{v}, \boldsymbol{h}),$$

where $\boldsymbol{\theta} = (w, a, b, c, d, e)$. The first term

$$E_0(\boldsymbol{v}, \boldsymbol{h}) = -\int_{\mathbb{Z}_p} a(x)\boldsymbol{v}(x)\,dx - \int_{\mathbb{Z}_p} b(x)\boldsymbol{h}(x)\,dx + \frac{e}{2}\int_{\mathbb{Z}_p} \boldsymbol{v}^2(x)\,dx + \frac{e}{2}\int_{\mathbb{Z}_p} \boldsymbol{h}^2(x)\,dx$$

is an analogue of the *free-field energy*. The second term

$$E_{\text{int}}(\boldsymbol{v}, \boldsymbol{h}) = -\iint_{\mathbb{Z}_p\times\mathbb{Z}_p} \boldsymbol{h}(y)\,w(x-y)\,\boldsymbol{v}(x)\,dxdy + \int_{\mathbb{Z}_p} c(x)\boldsymbol{v}^4(x)\,dx + \int_{\mathbb{Z}_p} d(x)\boldsymbol{h}^4(x)\,dx$$

is an analogue of the *interaction energy*. The results presented in this section are valid for more general functionals in which the first term in $E_{\text{int}}(\boldsymbol{v}, \boldsymbol{h})$ is replaced by

$$\iint_{\mathbb{Z}_p\times\mathbb{Z}_p} \boldsymbol{h}(y)\,w(x, y)\,\boldsymbol{v}(x)\,dxdy.$$

The motivation behind the definition of the energy functionals $E(\boldsymbol{v}, \boldsymbol{h})$ is that the discretizations of these functionals give the energy functionals considered in Refs. [5,13,21,22].

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

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

where K_B is the Boltzmann constant and T is the temperature constant. We normalize in such a way that $K_B T = 1$. The measure $d\boldsymbol{v}d\boldsymbol{h}$ is ill defined. However, it is expected that such a measure can be defined rigorously by a limit process. The statistical field theory corresponding to the energy functional $E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})$ is defined as the probability measure

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

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

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 $\mathbb{I}\coprod\mathbb{K} = \{1, 2, \ldots, n\}$, where $\coprod$ is the disjoint union, we set

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

These functions are also called the n-point Green functions.

To study 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 $\{\boldsymbol{v}, \boldsymbol{h}\}$:

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

now 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

$$Z(J_0, J_1) = \frac{1}{Z_0^{\text{phys}}} \int d\boldsymbol{v} d\boldsymbol{h}\, e^{-E(\boldsymbol{v}, \boldsymbol{h}, J_0, J_1)},$$

where

$$Z_0^{\text{phys}} = \int d\boldsymbol{v} d\boldsymbol{h}\, e^{-E_0(\boldsymbol{v}, \boldsymbol{h})}.$$

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:

$$G^{(n)}_{\mathbb{I},\mathbb{K}}(x_1, \ldots, x_n) = \frac{1}{Z} \left[\prod_{i \in \mathbb{I}} \frac{\delta}{\delta J_0(x_i)} \prod_{j \in \mathbb{K}} \frac{\delta}{\delta J_1(x_j)} Z(J_0, J_1) \right]_{\substack{J_0 = 0 \\ J_1 = 0}},$$

where $Z = \frac{Z^{\text{phys}}}{Z_0^{\text{phys}}}$. The functional $Z(J_0, J_1)$ is called the *generating functional* of the theory.

The description of the $\{\boldsymbol{v}, \boldsymbol{h}\}^4$-SFTs presented above is based in the classical version of these theories [23,24]. In Ref. [25], a mathematically rigorous formulation of ϕ^4-SFTs is presented; the fields are functions from $\mathbb{Q}_p^N$ into $\mathbb{R}$, with N arbitrary. We expect that this theory can be extended to the $\{\boldsymbol{v}, \boldsymbol{h}\}^4$-SFTs presented here.

4. Discrete SFTs and *p*-adic discrete Boltzmann machines

A central difference between the p-adic SFTs and the classical ones is that, in the p-adic case, the discretization process can be carried out in an easy rigorous way. More specifically, the discretization of a p-adic SFT is constructed by restricting the energy functional $E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})$ to a finite-dimensional vector subspace $\mathcal{D}^l(\mathbb{Z}_p)$ of the space of test functions $\mathcal{D}(\mathbb{Z}_p)$. The test functions in $\mathcal{D}^l(\mathbb{Z}_p)$ have the form

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

where $i = i_0 + i_1 p + \cdots + i_{l-1} p^{l-1} \in G_l = \mathbb{Z}_p / p^l \mathbb{Z}_p$, $l \geq 1$, and $\Omega(p^l |x - i|_p)$ is the characteristic function of the ball $B_{-l}(i)$. Here, it is important to notice that G_l is a finite, Abelian, additive group. In the p-adic world, the discrete functions are a particular case of the p-adic continuous functions, more precisely, $\mathcal{D}(\mathbb{Z}_p) = \cup_{l \in \mathbb{N}} \mathcal{D}^l(\mathbb{Z}_p)$ and $\mathcal{D}^l(\mathbb{Z}_p) \subset \mathcal{D}^{l+1}(\mathbb{Z}_p)$. There is no Archimedean counterpart of this result.

By taking $\boldsymbol{v}, \boldsymbol{h} \in \mathcal{D}^l(\mathbb{Z}_p)$ and l sufficiently large, the restriction $E_l(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})$ of the energy functional $E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})$ to $\mathcal{D}^l(\mathbb{Z}_p)$ has the form

$$\begin{aligned} E_l(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) = &-\sum_{j \in G_l} \sum_{k \in G_l} w_k v_{j+k} h_j - \sum_{j \in G_l} a_j v_j - \sum_{j \in G_l} b_j h_j + \frac{e}{2} \sum_{j \in G_l^N} v_j^2 + \frac{e}{2} \sum_{j \in G_l} h_j^2 \\ &+ \sum_{j \in G_l} c_j v_j^4 + \sum_{j \in G_l} d_j h_j^4. \end{aligned} \tag{1}$$

We refer the reader to Appendix B for further details of this calculation. From now on, we refer to Eq. (1) as the energy functional for a discrete $\{\boldsymbol{v}, \boldsymbol{h}\}^4$-SFT.

In the general case, $E_l(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})$ is a p-adic analogue of the $\{\boldsymbol{v}, \boldsymbol{h}\}^4$ neural networks introduced in Ref. [5]. In this article, each hidden state h_j and visible state v_i interact through a weight w_{ij}. Therefore, this requires the number of G_l^2 for weights w, whereas our counterpart requires only G_l. See Sect. 6 for further discussion.

We now attach to the discrete energy functional a p-adic discrete BM. For any visible and hidden states, $\boldsymbol{v} = [v_i]_{i \in G_l}$ and $\boldsymbol{h} = [h_i]_{i \in G_l}$, the Boltzmann probability distribution attached to $E_l(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})$ is given by

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

When there is no risk of confusion, we will omit $\boldsymbol{\theta}$ in the notation. When $c_j = d_j = 0$ for all $j \in G_l$, the energy functional $E_l(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})$ corresponds to a p-adic analogue of the convolutional deep belief networks studied in Ref. [21].

We note that the energy functional E_l has translational symmetry; i.e., E_l is invariant under the transformations $j \to j + j_0$, $k \to k + k_0$ for any $j_0, k_0 \in G_l$. This transformation is well defined since G_l is an additive group. In the case of applications to image processing, the group property also implies that the convolution operation does not alter image dimensions. The convolutional p-adic discrete BMs introduced here are a specific type of DBM (also called deep belief networks, DBNs).

5. Experimental results

We implement a p-adic discrete Boltzmann machine for processing binary images, then $\boldsymbol{v}, \boldsymbol{h} : \mathbb{Z}_p \to \{0, 1\} \subset \mathbb{R}$; in this case, we use the following energy functional:

$$E_l(\boldsymbol{v}, \boldsymbol{h}; \theta) = -\sum_{j \in G_l} \sum_{\substack{k \in G_l \\ |k|_p \le p^{-N}}} w_k v_{j+k} h_j - \sum_{j \in G_l} a_j v_j - \sum_{j \in G_l} b_j h_j$$

for some natural number $0 \le N \le l$. Note that, compared to Eq. (1), the quadratic and biquadratic terms are omitted since they do not play any role in the case in which $\boldsymbol{v}, \boldsymbol{h}$ are binary variables. The condition $N \le l$ implies that the convolution operation is restricted to a small neighborhood of radius p^{-N} for each pixel. The condition $N = 0$ means that the convolution involves all the points in the image.

Our numerical experiment is based on the MNIST dataset, where each image is considered as a sample of the visible state. Our first task is to train the network to maximize the log-likelihood of the visible states. We choose $p = 3$ and $l = 6$ since the image dimension is $3^3 \times 3^3$. In general, p and l depend on the size of the images to be processed. Typically p is chosen as a small prime number, e.g., 2 or 3.

To tune the parameters $a_j, b_j, w_j, j \in G_6$ of the network, we first transform each image I into a test function Test$(I) \in \mathcal{D}^6(\mathbb{Z}_3)$. The test function Test$(I)$ is defined in terms of the tree structure of G_6 in the following way: we define I as the root of the tree. Later we divide I into three even horizontal slices (sub-images). These sub-images are the vertices at level 1, and they are the children of I. Each sub-image I_j at level 1 is then divided vertically into three sub-images; these are the children of I_j. All the 3^2 new sub-images correspond to the vertices at level 2. We repeat this process until we reach level 6. At level 6, each vertex corresponds to a pixel, and we denote it by a value I_i for $i \in G_6$. Then Test(I) is defined as $\sum_{i \in G_6} I_i \Omega(3^6 |x - i|_3)$. See Fig. 3 for the construction of the tree corresponding to a $3^2 \times 3^2$ image. Figure 4 shows a graph of a test

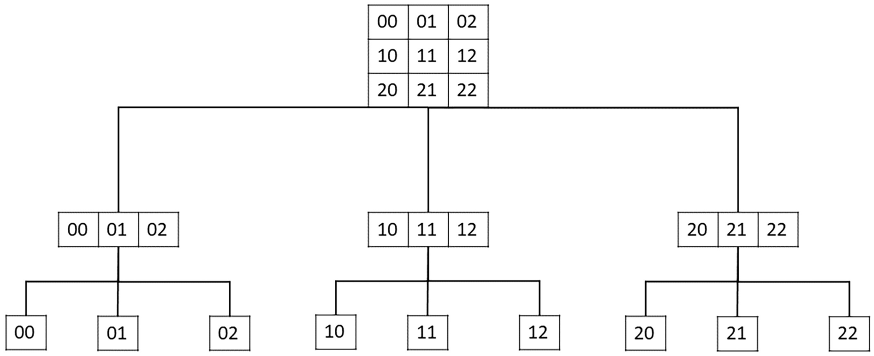

Fig. 3. The construction of the tree corresponding to a 3×3 image.

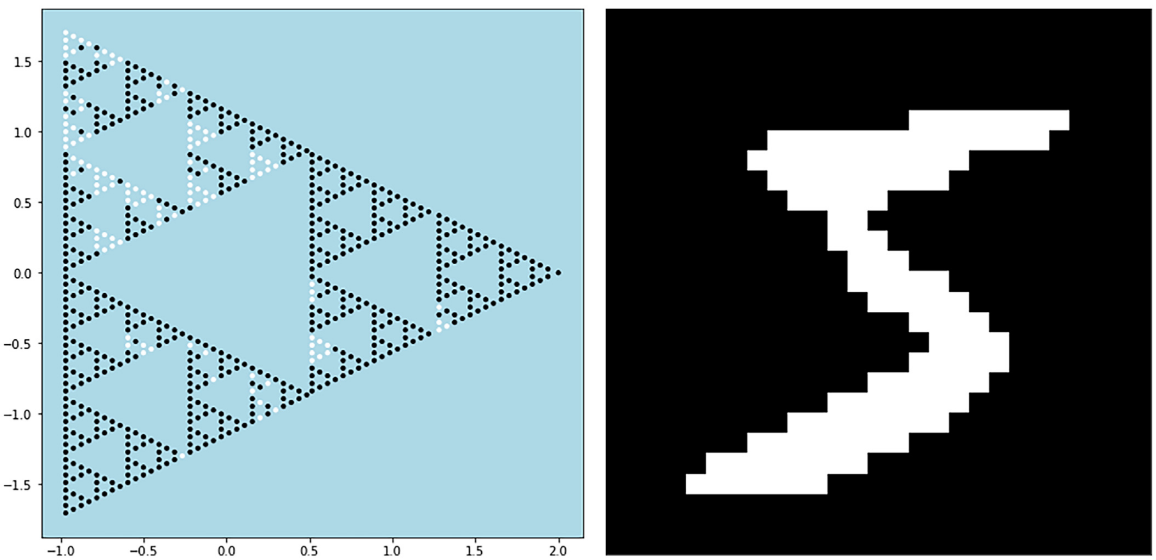

Fig. 4. The processing of date using p-adic networks requires the transformation of the actual data to its p-adic form. The left image shows a visualization of the test function corresponding to the right image. The visualization uses the embedding $\mathfrak{f} : \mathbb{Z}_p \to \mathbb{R}^2$ shown in Fig. 2.

function corresponding to an image from the MNIST data set. For further details, the reader may consult the appendix to Ref. [18].

We note that, in a p-adic discrete deep BM, the visible and hidden states are functions on a finite tree. Only the vertices at the top level, which are marked as orange and blue balls, are allowed to have states. See Fig. 5 for the case $p = 2$, $l = 2$. The remaining trees' vertices (see the black dots in Fig. 5) only codify the hierarchical relation between states. The visible and hidden vertices connected by the same type of lines share the same weight w.

We adapted the contrastive divergence learning method to the p-adic framework [26]. The technical details are presented in Appendix C.2. We implement two different types of networks. In the first type, the function w_k is supported in the entire tree G_6; in the second type, the function w_k is supported in a proper subset of the tree G_6. We use the full MNIST handwritten digits, without considering labels, to train a six-layer three-adic feature detector. The results are shown in Fig. 6.

After the processing by the network is completed, it is necessary to transform the test function into an image.

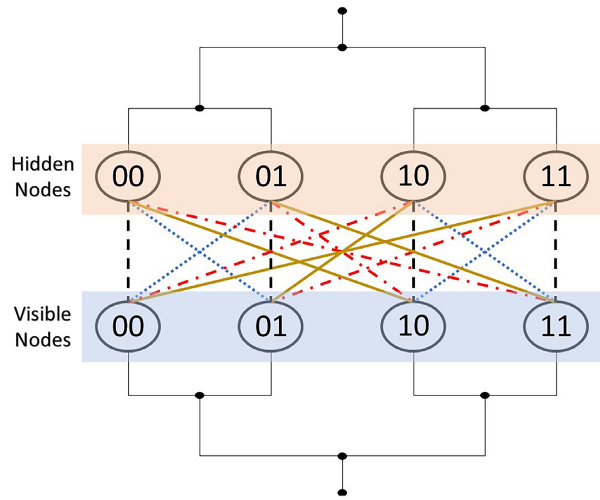

Fig. 5. A p-adic BM for $p = 2$, $l = 2$.

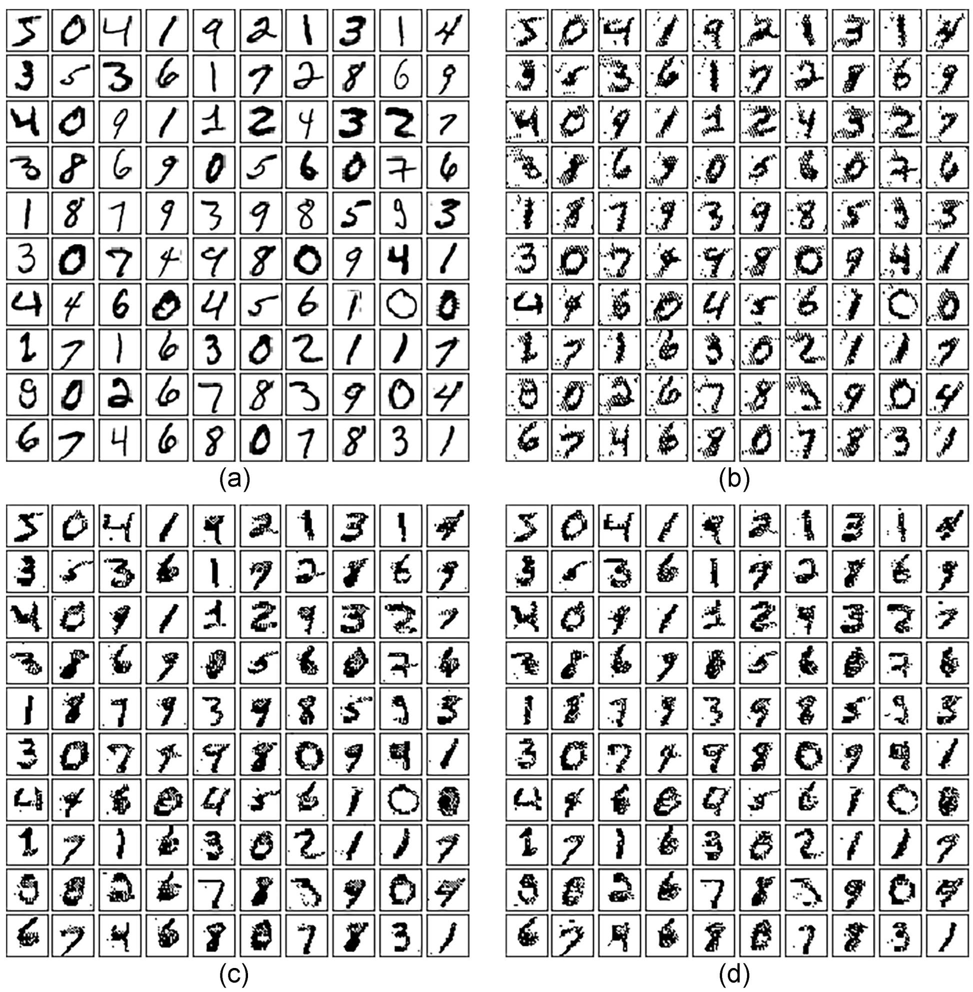

Fig. 6. (a) is the original input image. (b) is the reconstructed image using Gibbs sampling with w having $\mathbb{Z}_3$ as support. (c) is the reconstructed image with w supported in $3^3\mathbb{Z}_3$. (d) is the reconstructed image with w supported in $3^4\mathbb{Z}_3$.

6. Conclusions

The standard BMs [26] and the ϕ^4-neural networks introduced in Ref. [5] are particular classes of p-adic discrete BMs. Indeed, if $w(x, y)$ is a test function and the interaction between the visible and hidden field has the form $-\iint_{\mathbb{Z}_p\times\mathbb{Z}_p}\boldsymbol{h}(y)\, w(x, y)\, \boldsymbol{v}(x)\, dxdy$, then in the corresponding discrete energy functional, after a suitable rescaling of the weights $w_{k,j}$, the interaction of the visible and hidden states takes the form $-\sum_{j\in G_l}\sum_{k\in G_l} w_{k,j} v_k h_j$. Here $[w_{k,j}]$ is an ordinary matrix, which means that its entries $w_{k,j}$ depend neither on the algebraic structure of G_l nor on its topology.

In the case in which the interaction between the visible and hidden fields has the form $-\iint_{\mathbb{Z}_p\times\mathbb{Z}_p}\boldsymbol{h}(y)\, w(x-y)\, \boldsymbol{v}(x)\, dxdy$, the corresponding discrete energy functional depends on the group structure of G_l, and the corresponding neural network is a particular case of a DBM; see Fig. 5.

The condition $l \geq l_0$, for some constant l_0, means that a p-adic discrete BM admits arbitrarily large copies; this is a consequence of the fact that p-adic numbers have a tree-like structure. We expect that for l sufficiently large the statistical properties of the network can be studied using a p-adic continuous SFT.

Our numerical experiments show that p-adic discrete convolutional DBMs alone can be used to process real data; this opens the possibility of using these networks as layers in specialized NNs.

In Ref. [12] the first author conjectured that the limit

$$\frac{e^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z^{\text{phys}}} d\boldsymbol{v} d\boldsymbol{h} = \lim_{l\to\infty} \boldsymbol{P}_l(\boldsymbol{v}, \boldsymbol{h})\, d^{\#G_l}\boldsymbol{v}\, d^{\#G_l}\boldsymbol{h} \tag{2}$$

exists in some sense; here $d^{\#G_l}\boldsymbol{x}$ denotes the Lebesgue measure of $\mathbb{R}^{\#G_l}$. Then the correlation between the network activity in different regions of the underlying tree G_l can be understood by computing the correlation functions of the corresponding continuous SFT.

For practical applications the NNs should be discrete entities. This type of NN naturally corresponds with discrete SFTs. To use Euclidean QFT to study NNs, it is convenient to have continuous versions of these networks. Thus, a clear way of passing between discrete SFTs to continuous ones is required. The existence of the limit (2) is a very difficult problem in classical QFT. In Ref. [25] the existence of this limit was established for p-adic ϕ^4-theories involving one scalar field; we expect that these techniques can be extended to the QFT case considered here.

It is widely accepted in the artificial intelligence community that the probability distributions $\boldsymbol{P}_l(\boldsymbol{v}, \boldsymbol{h})$ should approximate any finite probability distribution very well, which means that the corresponding NNs are universal approximators. We argue that this property is connected with the topology and the structure of the NNs, and that the problem of designing good architectures for NNs is out of the scope of QFT. We expect that QFT techniques will be useful to understand the qualitative behavior of large NNs, which can be well approximated as 'continuous' NNs.

The study of the correspondence between p-adic Euclidean QFTs and NNs is just starting. We envision that the next step will be to develop perturbative calculations of the correlation functions via Feynman diagrams, to study connections with Ginzburg–Landau theory, and to develop practical applications of the p-adic convolutional BMs.

Funding

Open Access funding: SCOAP[3].

A. Basic facts of *p*-adic analysis

In this section we review some of the basic results of p-adic analysis required in this article. For a detailed exposition on p-adic analysis the reader may consult Refs. [27,28–30]. For a quick review of p-adic analysis the reader may consult Ref. [31].

A.1. The field of p-adic numbers

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

$$|x|_p = \begin{cases} 0 & \text{if } x = 0 \\ p^{-\gamma} & \text{if } x = p^{\gamma}\dfrac{a}{b}, \end{cases} \tag{A1}$$

where a and b are integers coprime with p. The integer $\gamma = \mathrm{ord}_p(x) := \mathrm{ord}(x)$, with $\mathrm{ord}(0) := +\infty$, 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 $|x+y|_p \leq \max\{|x|_p, |y|_p\}$. As a topological space $\mathbb{Q}_p$ is homeomorphic to a Cantor-like subset of the real line; see, e.g., Refs. [27,20,28].

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

$$x = p^{\mathrm{ord}(x)} \sum_{j=0}^{\infty} x_j p^j,$$

where $x_j \in \{0, 1, 2, \ldots, p-1\}$ and $x_0 \neq 0$. In addition, for any $x \in \mathbb{Q}_p \setminus \{0\}$ we have

$$|x|_p = p^{-\mathrm{ord}(x)}.$$

A.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 its 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 $\mathbb{Z}_p$*. The balls are both open and closed subsets in $\mathbb{Q}_p$. We use $\Omega(p^{-r}|x-a|_p)$ to denote the characteristic function of the ball $B_r(a)$. 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., Sect. 1.3 in Ref. [27] or Sect. 1.8 in Ref. [28]. The balls and spheres are compact subsets. Thus $\left(\mathbb{Q}_p, |\cdot|_p\right)$ is a locally compact topological space.

A.2.1. Tree-like structures Any p-adic integer i admits an expansion of the form $i = i_k p^k + i_{k+1}p^{k+1} + \cdots$ for some $k \geq 0$, $i_k \neq 0$. The set of p-adic truncated integers modulo p^l, $l \geq 1$, consists of all the integers of the form $i = i_0 + i_1 p + \cdots + i_{l-1}p^{l-1}$. These numbers form a complete set of representatives for the elements of the additive group $G_l = \mathbb{Z}_p/p^l\mathbb{Z}_p$, which is isomorphic to the set of integers $\mathbb{Z}/p^l\mathbb{Z}$ (written in base p) modulo p^l. By restricting $|\cdot|_p$ to G_l, it becomes a normed space, and $|G_l|_p = \{0, p^{-(l-1)}, \ldots, p^{-1}, 1\}$. With the metric induced by $|\cdot|_p$, 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 to the possible values of the digit i_0 in the p-adic expansion of i. Each of these vertices is connected to the root by a non-directed edge. At level k, with $2 \leq k \leq l+1$, 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}. See Fig. 1. The balls $B_{-r}(a) = a + p^r\mathbb{Z}_p$ are infinite rooted trees.

A.3. The Haar measure

Since $(\mathbb{Z}_p, +)$ is a locally compact topological group, there exists a Haar measure dx, which is invariant under translations; i.e., $d(x+a) = dx$ [32]. If we normalize this measure by the condition $\int_{\mathbb{Z}_p} dx = 1$, then dx is unique. It follows immediately that

$$\int_{B_r(a)} dx = \int_{a+p^{-r}\mathbb{Z}_p} dx = p^r \int_{\mathbb{Z}_p} dy = p^r, \ r \in \mathbb{Z}.$$

On a few occasions we use the 2D Haar measure $dxdy$ of the additive group $(\mathbb{Z}_p \times \mathbb{Z}_p, +)$ and normalize this measure 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 Ref. [31] and references therein.

A.4. The Bruhat–Schwartz space in the unit ball

A real-valued function φ defined on $\mathbb{Z}_p$ is called a *Bruhat–Schwartz function (* or a *test function)* if for any $x \in \mathbb{Z}_p$ there exists an integer $l(x) \in \mathbb{Z}$ such that

$$\varphi(x + x') = \varphi(x) \text{ for any } x' \in B_{l(x)}. \tag{A2}$$

The $\mathbb{R}$-vector space of Bruhat–Schwartz functions supported in the unit ball is denoted by $\mathcal{D}(\mathbb{Z}_p)$. For $\varphi \in \mathcal{D}(\mathbb{Z}_p)$, the largest number $l = l(\varphi)$ satisfying Eq. (A2) is called the *exponent of local constancy (* or the *parameter of constancy) of* φ. A function φ in $\mathcal{D}(\mathbb{Z}_p)$ can be written as

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

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

B. Discretization of the energy functional

The elements of $G_l = \mathbb{Z}_p/p^l\mathbb{Z}_p$, $l \geq 1$, have the form $\boldsymbol{i} = i_0 + i_1 p + \cdots + i_{l-1}p^{l-1}$, where the i_k are p-adic digits. We denote by $\mathcal{D}^l(\mathbb{Z}_p)$ the $\mathbb{R}$-vector space of all test functions of the form

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

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

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

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

where $\hookrightarrow$ denotes a continuous embedding.

The restriction of E to the subspace $\mathcal{D}^l(\mathbb{Z}_p)$ gives a discretization of E denoted as E_l. Indeed, by assuming that

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

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

we have

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

$$= \iint\limits_{p^l\mathbb{Z}_p\times p^l\mathbb{Z}_p} w((i-j)+(x-y))dxdy,$$

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

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

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

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

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

and consequently

$$\begin{aligned} E_l(\boldsymbol{v},\boldsymbol{h}) = &-\sum_{\substack{i,\ j\in G_l\\ i\neq j}} \boldsymbol{v}(i)\,\boldsymbol{h}(j)\left(\iint\limits_{p^l\mathbb{Z}_p\times p^l\mathbb{Z}_p} w(i-j+x-y)dxdy\right)\\ &- p^{-l}\sum_{i\in G_l} a(i)\,\boldsymbol{v}(i) - p^{-l}\sum_{i\in G_l} b(i)\,\boldsymbol{h}(i) + \frac{ep^{-l}}{2}\sum_{i\in G_l}\boldsymbol{v}^2(i) + \frac{ep^{-l}}{2}\sum_{i\in G_l}\boldsymbol{h}^2(i)\\ &+ \frac{p^{-l}}{2}\sum_{i\in G_l} c(i)\,\boldsymbol{v}^4(i) + \frac{p^{-l}}{2}\sum_{i\in G_l} d(i)\,\boldsymbol{h}^4(i). \end{aligned}$$

We take $v_i = \boldsymbol{v}(i)$, $h_i = \boldsymbol{h}(i)$,

$$w_{i-j} = \iint\limits_{p^l\mathbb{Z}_p\times p^l\mathbb{Z}_p} w((i-j)+(x-y))dxdy,$$

$a_i = p^{-l}a(i)$, $b_i = p^{-l}b(i)$, $c_i = p^{-l}c(i)$, $d_i = p^{-l}d(i)$, for $i, j \in G_l$, and $\boldsymbol{\theta} = \left\{w_{ij}, a_i, b_i, c_i, d_i\right\}$. We also rescale e to ep^l, then

$$\begin{aligned} E_l(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta}) = &-\sum_{i,\ j\in G_l} w_{i-j}v_ih_j - \sum_{i\in G_l} a_iv_i - \sum_{i\in G_l} b_ih_i\\ &+ \frac{e}{2}\sum_{i\in G_l} v_i^2 + \frac{e}{2}\sum_{i\in G_l} h_i^2 + \sum_{i\in G_l} c_iv_i^4 + \sum_{i\in G_l} d_ih_i^4. \end{aligned}$$

We now recall that G_l is an additive group, then

$$\sum_{i,\ j\in G_l} w_{i-j}v_ih_j = \sum_{j\in G_l}\sum_{k\in G_l} w_k v_{j+k}h_j,$$

and consequently

$$E_l\left(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta}\right)=-\sum_{j\in G_l}\sum_{k\in G_l}w_k v_{j+k}h_j-\sum_{j\in G_l}a_j v_j-\sum_{j\in G_l}b_j h_j$$
$$+\frac{e}{2}\sum_{j\in G_l^N}v_j^2+\frac{e}{2}\sum_{j\in G_l}h_j^2+\sum_{j\in G_l}c_j v_j^4+\sum_{j\in G_l}d_j h_j^4.$$

C. Some probability distributions

From now on, we assume that visible and hidden fields are binary variables. However, most of our mathematical formulation is valid under the assumption that visible and hidden fields are discrete variables. We set

$$\boldsymbol{V}=\{V_1,V_2,\ldots,V_N\},\quad \boldsymbol{H}=\{H_1,H_2,\ldots,H_N\},$$

$N=p^l$, to be the respective visible and hidden random variable sets. The random variables $(\boldsymbol{V},\boldsymbol{H})$ take values $(\boldsymbol{v},\boldsymbol{h})\in\{0,1\}^{2N}$. For the sake of simplicity, we will identify the random vector $\boldsymbol{V}$ with $\boldsymbol{v}$, and the random vector $\boldsymbol{H}$ with $\boldsymbol{h}$. We 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, h_i. With this notation, $\boldsymbol{v}=[v_i]_{i\in G_l}$ is a realization of the visible field, and $\boldsymbol{h}=[h_i]_{i\in G_l}$ is a realization of the hidden field. The joint distribution of the random vectors $(\boldsymbol{v},\boldsymbol{h})$ is given by the following Boltzmann probability distribution:

$$\boldsymbol{P}_l(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})=\frac{\exp\left(-E_l\left((\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})\right)\right)}{Z_l},\tag{C1}$$

where

$$Z_l=\sum_{\boldsymbol{v},\boldsymbol{h}}\exp\left(-E_l\left(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta}\right)\right),$$

and E_l is defined in Eq. (1).

By using the joint distribution of the visible field and the hidden field (C1), we compute the marginal probability distributions as follows:

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

Since we are assuming that $\boldsymbol{v}$, $\boldsymbol{h}$ are binary, the energy functional takes the form

$$E_l(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})=-\sum_{k\in G_l}\sum_{j\in G_l}w_k v_{j+k}h_j-\sum_{j\in G_l}a_j v_j-\sum_{j\in G_l}b_j h_j.$$

The classical BM has the advantage of independence between the visible units as well as the hidden units. The p-adic BM shares the same advantage; i.e., by fixing the hidden field $\boldsymbol{h}$, the random variables v_i, $i\in G_l$, become independent. An analogous assertion is valid if we fix the visible field $\boldsymbol{v}$. More precisely, the conditional probability distributions satisfy

$$\boldsymbol{P}_l(\boldsymbol{v}\mid\boldsymbol{h};\boldsymbol{\theta})=\prod_{j\in G_l}\boldsymbol{P}_l\left(v_j\mid\boldsymbol{h};\boldsymbol{\theta}\right),\quad\text{and}\quad \boldsymbol{P}_l(\boldsymbol{h}\mid\boldsymbol{v};\boldsymbol{\theta})=\prod_{j\in G_l}\boldsymbol{P}_l\left(h_j\mid\boldsymbol{v};\boldsymbol{\theta}\right).$$

Indeed, by direct computation, we have

$$\boldsymbol{P}_l(\boldsymbol{v} \mid \boldsymbol{h}; \boldsymbol{\theta}) = \frac{\boldsymbol{P}_l(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}{\boldsymbol{P}_l(\boldsymbol{h}; \boldsymbol{\theta})}$$

$$= \frac{\exp\left(-E_l\left(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}\right)\right)}{\sum_{\boldsymbol{v}} \exp\left(-E_l\left(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}\right)\right)} = \frac{\prod_{j\in G_l} \exp\left(\sum_{k\in G_l} w_k v_j h_{j-k} + a_j v_j\right)}{\sum_{\boldsymbol{v}} \prod_{j\in G_l} \exp\left(\sum_{k\in G_l} w_k v_j h_{j-k} + a_j v_j\right)}$$

$$= \prod_{j\in G_l} \frac{\exp\left(v_j \sum_{k\in G_l} w_k h_{j-k} + a_j v_j\right)}{\sum_{v_j} \exp\left(v_j \sum_{k\in G_l} w_k h_{j-k} + a_j v_j\right)} = \prod_{j\in G_l} \frac{\boldsymbol{P}_l\left(v_j, \boldsymbol{h}; \boldsymbol{\theta}\right)}{\sum_{v_j} \boldsymbol{P}_l\left(v_j, \boldsymbol{h}; \boldsymbol{\theta}\right)} = \prod_{j\in G_l} \boldsymbol{P}_l\left(v_j \mid \boldsymbol{h}; \boldsymbol{\theta}\right).$$

Similarly, we can prove that

$$\boldsymbol{P}_l(\boldsymbol{h} \mid \boldsymbol{v}; \boldsymbol{\theta}) = \prod_{j\in G_l} \boldsymbol{P}_l\left(h_j \mid \boldsymbol{v}; \boldsymbol{\theta}\right). \tag{C2}$$

C.1. Gradient of the log-likelihood

The log-likelihood giving a single visible state $\boldsymbol{v}$ is given by

$$\ln \boldsymbol{P}_l(\boldsymbol{v}|\boldsymbol{\theta}) = \ln \frac{1}{Z} \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h})} = \ln \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h})} - \ln \sum_{\boldsymbol{h},\boldsymbol{v}} e^{-E(\boldsymbol{v},\boldsymbol{h})}.$$

Taking the derivative with respect to the parameters gives the following mean-like representation:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \ln \boldsymbol{P}_l(\boldsymbol{v}|\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}} \ln \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h})} - \frac{\partial}{\partial \boldsymbol{\theta}} \ln \sum_{\boldsymbol{h},\boldsymbol{v}} e^{-E(\boldsymbol{v},\boldsymbol{h})} \\
&= -\frac{1}{\sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h})}} \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h})} \frac{\partial}{\partial \boldsymbol{\theta}} E(\boldsymbol{v}, \boldsymbol{h}) + \frac{1}{\sum_{\boldsymbol{v},\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h})}} \sum_{\boldsymbol{v},\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h})} \frac{\partial}{\partial \boldsymbol{\theta}} E(\boldsymbol{v}, \boldsymbol{h}) \\
&= -\sum_{\boldsymbol{h}} \boldsymbol{P}_l(\boldsymbol{h}|\boldsymbol{v}) \frac{\partial}{\partial \boldsymbol{\theta}} E(\boldsymbol{v}, \boldsymbol{h}) + \sum_{\boldsymbol{v},\boldsymbol{h}} \boldsymbol{P}_l(\boldsymbol{v}, \boldsymbol{h}) \frac{\partial}{\partial \boldsymbol{\theta}} E(\boldsymbol{v}, \boldsymbol{h}).
\end{aligned} \tag{C3}$$

In the case of multiple visible states $S = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_s\}$, the log-likelihood is defined in the average sense; i.e., $\frac{1}{s} \sum_{\boldsymbol{v}\in S} \ln \boldsymbol{P}_l(\boldsymbol{v}|\boldsymbol{\theta})$.

Taking the derivative with respect to w_k gives

$$\begin{aligned}
\frac{\partial}{\partial w_k} \ln \boldsymbol{P}_l(\boldsymbol{v}|\boldsymbol{\theta}) &= \sum_{\boldsymbol{h}} \left(\boldsymbol{P}_l(\boldsymbol{h}|\boldsymbol{v}) \sum_{j\in G_l} v_{j+k} h_j\right) - \sum_{\boldsymbol{v},\boldsymbol{h}} \left(\boldsymbol{P}_l(\boldsymbol{h}, \boldsymbol{v}) \sum_{j\in G_l} v_{j+k} h_j\right) \\
&= \sum_{\boldsymbol{h}} \left(\boldsymbol{P}_l(\boldsymbol{h}|\boldsymbol{v}) \sum_{j\in G_l} v_{j+k} h_j\right) - \sum_{\boldsymbol{v}} \boldsymbol{P}_l(\boldsymbol{v}) \left(\sum_{\boldsymbol{h}} \left(\boldsymbol{P}_l(\boldsymbol{h}|\boldsymbol{v}) \sum_{j\in G_l} v_{j+k} h_j\right)\right).
\end{aligned} \tag{C4}$$

Now, let $\square$ be the ordered indexes of all the positive states in $\boldsymbol{v}$:

$$
\begin{aligned}
\sum_{\boldsymbol{h}}\left(\boldsymbol{P}_l(\boldsymbol{h}|\boldsymbol{v})\sum_{i\in G_l} v_{i+k}h_i\right) &= \sum_{\boldsymbol{h}}\boldsymbol{P}_l(\boldsymbol{h}|\boldsymbol{v})\left(\sum_{i\in\square} h_{i-k}\right)\\
&= \sum_{\boldsymbol{h}_{\square}}\sum_{\boldsymbol{h}_{G_l\setminus\square}}\prod_{i\in\square}\boldsymbol{P}_l(h_{i-k}|\boldsymbol{v})\prod_{i\in G_l\setminus\square}\boldsymbol{P}_l(h_{i-k}|\boldsymbol{v})\left(\sum_{i\in\square} h_{i-k}\right)\\
&= \left(\sum_{\boldsymbol{h}_{\square}}\prod_{i\in\square}\boldsymbol{P}_l(h_{i-k}|\boldsymbol{v})\left(\sum_{i\in\square} h_{i-k}\right)\right)\underbrace{\left(\sum_{\boldsymbol{h}_{G_l\setminus\square}}\prod_{i\in G_l\setminus\square}\boldsymbol{P}_l(h_{i-k}|\boldsymbol{v})\right)}_{=1}\\
&= \sum_{\boldsymbol{h}_{\square}}\left(\prod_{i\in\square}\boldsymbol{P}_l(h_{i-k}|\boldsymbol{v})\sum_{i\in\square} h_{i-k}\right)\\
&= \sum_{i\in\square}\boldsymbol{P}_l(h_{i-k}=1|\boldsymbol{v}) = \sum_{i\in G_l}\boldsymbol{P}_l(h_{i-k}=1|\boldsymbol{v})v_i.
\end{aligned}
\tag{C5}
$$

Combining Eqs. (C4) and (C5) gives

$$
\frac{\partial \log \boldsymbol{P}_l(\boldsymbol{v}|\boldsymbol{\theta})}{w_k} = \sum_{i\in G_l}\boldsymbol{P}_l(h_{i-k}=1|\boldsymbol{v})v_i - \sum_{\boldsymbol{v}}\boldsymbol{P}_l(\boldsymbol{v})\left(\sum_{i\in G_l}\boldsymbol{P}_l(h_{i-k}=1|\boldsymbol{v})v_i\right). \tag{C6}
$$

Note that the above formula is different from the classical BM (see Eq. (29) in Ref. [26]). We derive the derivatives with respect to a_j and b_j similarly as in the classical BM:

$$
\begin{aligned}
\frac{\partial}{\partial a_j}\ln \boldsymbol{P}_l(\boldsymbol{v}|\boldsymbol{\theta}) &= -\sum_{\boldsymbol{h}}\boldsymbol{P}_l(\boldsymbol{h}|\boldsymbol{v})\frac{\partial}{\partial a_j}E_l(\boldsymbol{v},\boldsymbol{h}) + \sum_{\boldsymbol{v},\boldsymbol{h}}\boldsymbol{P}_l(\boldsymbol{h},\boldsymbol{v})\frac{\partial}{\partial a_j}E_l(\boldsymbol{v},\boldsymbol{h})\\
&= \sum_{\boldsymbol{h}}\boldsymbol{P}_l(\boldsymbol{h}|\boldsymbol{v})v_j - \sum_{\boldsymbol{v},\boldsymbol{h}}\boldsymbol{P}_l(\boldsymbol{h},\boldsymbol{v})v_j = v_j - \sum_{\boldsymbol{v}}\boldsymbol{P}_l(\boldsymbol{v})v_j,
\end{aligned}
\tag{C7}
$$

and

$$
\begin{aligned}
\frac{\partial}{\partial b_j}\ln \boldsymbol{P}_l(\boldsymbol{v}|\boldsymbol{\theta}) &= -\sum_{\boldsymbol{h}}\boldsymbol{P}_l(\boldsymbol{h}|\boldsymbol{v})\frac{\partial}{\partial b_j}E_l(\boldsymbol{v},\boldsymbol{h}) + \sum_{\boldsymbol{v},\boldsymbol{h}}\boldsymbol{P}_l(\boldsymbol{h},\boldsymbol{v})\frac{\partial}{\partial b_j}E_l(\boldsymbol{v},\boldsymbol{h})\\
&= \sum_{\boldsymbol{h}}\boldsymbol{P}_l(\boldsymbol{h}|\boldsymbol{v})h_j - \sum_{\boldsymbol{v},\boldsymbol{h}}\boldsymbol{P}_l(\boldsymbol{h},\boldsymbol{v})h_j = \boldsymbol{P}_l(h_j=1|\boldsymbol{v}) - \sum_{\boldsymbol{v}}\boldsymbol{P}_l(h_j=1|\boldsymbol{v}).
\end{aligned}
\tag{C8}
$$

C.2. Contrastive divergence learning

As in the classical case, the exact computation of the gradient of the log-likelihood involves an exponential number of terms; see Eq. (C6). We adopt the contrastive divergence (CD) method, introduced by Hinton [33], to the p-adic case to approximate the minimization of the gradient of the log-likelihood.

The approximation of Eqs. (C6)–(C9) using the contrastive convergence method can be represented as follows:

$$\frac{\partial \log \boldsymbol{P}_l(\boldsymbol{v}|\boldsymbol{\theta})}{w_k} \approx \sum_{i\in G_l} \boldsymbol{P}_l(h_{i-k}=1|\boldsymbol{v}^{(0)})v_i - \sum_{i\in G_l} \boldsymbol{P}_l(h_{i-k}=1|\boldsymbol{v}^{(m)})v_i^{(m)},$$

$$\frac{\partial \log \boldsymbol{P}_l(\boldsymbol{v}|\boldsymbol{\theta})}{a_j} \approx v_j^{(0)} - v_j^{(m)},$$

$$\frac{\partial \log \boldsymbol{P}_l(\boldsymbol{v}|\boldsymbol{\theta})}{b_j} \approx \boldsymbol{P}_l(h_j=1|\boldsymbol{v}^{(0)}) - \boldsymbol{P}_l(h_j=1|\boldsymbol{v}^{(m)}), \tag{C9}$$

where m is a predetermined positive integer, $\boldsymbol{v}^{(0)}$ is a training example, and $\boldsymbol{v}^{(m)}$ is a sample of the Gibbs chain after m steps. More precisely, we implement the Gibbs sampling in the following way. First, we obtain a sample of $\boldsymbol{h}^{(0)}$ using the conditional distribution $P_l(\boldsymbol{h}|\boldsymbol{v}^{(\mathbf{0})})$. Then, we obtain a sample of $\boldsymbol{v}^{(1)}$ using $P_l(\boldsymbol{v}|\boldsymbol{h}^{(0)})$. We repeat this process until we get $\boldsymbol{v}^{(m)}$.

The following formulas are utilized in the calculation. Let $\boldsymbol{h}_{-i}$ denote the state of all hidden units expect for the ith one:

$$\begin{aligned}
\boldsymbol{P}_l(h_i=1|\boldsymbol{v}) = \boldsymbol{P}_l(h_i=1|\boldsymbol{h}_{-i},\boldsymbol{v}) &= \frac{\boldsymbol{P}_l(h_i=1,\boldsymbol{h}_{-i},\boldsymbol{v})}{\boldsymbol{P}_l(\boldsymbol{h}_{-i},\boldsymbol{v})} \\
&= \frac{\boldsymbol{P}_l(h_i=1,\boldsymbol{h}_{-i},\boldsymbol{v})}{\boldsymbol{P}_l(h_i=1,\boldsymbol{h}_{-i},\boldsymbol{v})+\boldsymbol{P}_l(h_i=0,\boldsymbol{h}_{-i},\boldsymbol{v})} = \frac{1}{1+\dfrac{\boldsymbol{P}_l(h_i=0,\boldsymbol{h}_{-i},\boldsymbol{v})}{\boldsymbol{P}_l(h_i=1,\boldsymbol{h}_{-i},\boldsymbol{v})}} \\
&= \frac{1}{1+\dfrac{1}{\exp\left(\sum_{k\in G_l} w_k v_{i+k} + b_i\right)}} =: \sigma\left(\sum_{k\in G_l} w_k v_{i+k} + b_i)\right).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\boldsymbol{P}_l(h_i=0|\boldsymbol{v}) &= \sigma\left(-\sum_{k\in G_l} w_k v_{i+k} - b_i\right), \\
\boldsymbol{P}_l(v_i=1|\boldsymbol{h}) &= \sigma\left(\sum_{k\in G_l} w_{i-k} h_k + a_i\right), \\
\boldsymbol{P}_l(v_i=0|\boldsymbol{h}) &= \sigma\left(-\sum_{k\in G_l} w_{i-k} h_k - a_i\right).
\end{aligned} \tag{C10}$$

References

1 Y. LeCun, Y. Bengio, and G. E. Hinton, Nature **521**, 436 (2015).
2 Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli, Ann. Rev. Condens. Matter Phys. **11**, 501 (2020).
3 A. Buice Michael and C. Chow Carson, J. Stat. Mech. Theory Exp. **3**, P03003 (2013).
4 A. Buice Michael and D. Cowan Jack, Phys. Rev. E **75**, 051919 (2007).
5 D. Bachtis, G. Aarts, and B. Lucini, Phys. Rev. D **103**, 074510 (2021).
6 E. Dyer and G. Gur-Ari, arXiv:1909.11304 [cs.LG] [Search inSPIRE].
7 H. Erbin, V. Lahoche, and D. Ousmane Samary, Mach. Learn. Sci. Technol. **3**, 015027 (2022).
8 J. Halverson, A. Maiti, and K. Stoner, Mach. Learn. Sci. Technol. **2**, 035002 (2021).
9 A. Maiti, K. Stoner, and J. Halverson, arXiv:2106.00694 [cs.LG] [Search inSPIRE].

10 M. Helias and D. Dahmen, Statistical Field Theory for Neural Networks (Springer, Cham, 2020), Lecture Notes in Physics, Vol. 970.
11 S. Yaida, Proc. Mathematical and Scientific Machine Learning, p. 165 (2020).
12 W. A. Zúñiga-Galindo, Physica A **612**, 128492 (2023).
13 N. Le Roux and Y. Bengio, Neural Comput. **20**, 1631 (2008).
14 S. Albeverio, A. Khrennikov, and B. Tirozzi, Math. Models Methods Appl. Sci. **9**, 1417 (1999).
15 A. Khrennikov and B. Tirozzi, CMS Conf. Proc. 29, p. 395 (2000).
16 A. Khrennikov, Information Dynamics in Cognitive, Psychological, Social and Anomalous Phenomena (Springer, Berlin, 2004).
17 B. A. Zambrano-Luna and W. A. Zúñiga-Galindo, J. Nonlinear Math. Phys. **30**, 34 (2023).
18 B. A. Zambrano-Luna and W. A. Zúñiga-Galindo, Physica D **446**, 133668 (2023).
19 N. Koblitz, p-Adic Numbers, p-Adic Analysis, and Zeta-Functions (Springer, New York, 1984), Graduate Texts in Mathematics, Vol. 58.
20 D. V. Chistyakov, Theor. Math. Phys. **109**, 1495 (1997).
21 L. Honglak, G. Roger, R. Rajesh, and Y. Ng. Andrew, Commun. ACM, **54**, 95 (2011).
22 N. Le Roux and Y. Bengio, Neural Comput. **22**, 2192 (2010).
23 H. Kleinert and V. Schulte-Frohlinde, Critical Properties of ϕ^4-Theories (World Scientific, Singapore, 2001).
24 G. Mussardo, Statistical Field Theory: An Introduction to Exactly Solved Models of Statistical Physics (Oxford University Press, Oxford, UK, 2010).
25 W. A. Zúñiga-Galindo, Rev. Math. Phys. **34**, 2250022 (2022).
26 A. Fischer and C. Igel, An introduction to restricted Boltzmann machines, in Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: CIARP 2012, eds. L. Alvarez, M. Mejail, L. Gomez, and J. Jacobo (Springer, Berlin, 2012), Lecture Notes in Computer Science, Vol. 7441.
27 V. S. Vladimirov, I. V. Volovich, and E. I. Zelenov, p-Adic Analysis and Mathematical Physics (World Scientific, Singapore, 1994).
28 S. Albeverio, A. Yu. Khrennikov, and V. M. Shelkovich, Theory of p-Adic Distributions: Linear and Nonlinear Models (Cambridge University Press, Cambridge, UK, 2010).
29 A. N. Kochubei, Pseudo-differential Equations and Stochastics over Non-Archimedean Fields (Marcel Dekker, New York, 2001).
30 M. H. Taibleson, Fourier Analysis on Local Fields (Princeton University Press, Princeton, NJ, 1975).
31 M. Bocardo-Gaspar, H. García-Compeán, and W. A. Zúñiga-Galindo, Lett. Math. Phys. **109**, 1167 (2019).
32 P. Halmos, Measure Theory (Van Nostrand, New York, 1950).
33 G. E. Hinton, Neural Comput. **14**, 1771 (2002).