Machine learning with quantum field theories

Dimitrios Bachtis,[a,*] **Gert Aarts**[b,c] **and Biagio Lucini**[a,d]

[a] *Swansea University,*
 Department of Mathematics, Bay Campus, SA1 8EN, Swansea, Wales, United Kingdom

[b] *Swansea University,*
 Department of Physics, Singleton Campus, SA2 8PP, Swansea, Wales, United Kingdom

[c] *European Centre for Theoretical Studies in Nuclear Physics and Related Areas (ECT*)*
 & Fondazione Bruno Kessler Strada delle Tabarelle 286, 38123 Villazzano (TN), Italy

[d] *Swansea University,*
 Swansea Academy of Advanced Computing, Bay Campus, SA1 8EN, Swansea, Wales, United Kingdom
 E-mail: dimitrios.bachtis@swansea.ac.uk, g.aarts@swansea.ac.uk,
 b.lucini@swansea.ac.uk

The precise equivalence between discretized Euclidean field theories and a certain class of probabilistic graphical models, namely the mathematical framework of Markov random fields, opens up the opportunity to investigate machine learning from the perspective of quantum field theory. In this contribution we will demonstrate, through the Hammersley-Clifford theorem, that the ϕ^4 scalar field theory on a square lattice satisfies the local Markov property and can therefore be recast as a Markov random field. We will then derive from the ϕ^4 theory machine learning algorithms and neural networks which can be viewed as generalizations of conventional neural network architectures. Finally, we will conclude by presenting applications based on the minimization of an asymmetric distance between the probability distribution of the ϕ^4 machine learning algorithms and target probability distributions.

---

*Speaker

1. Introduction

The transition to Euclidean space and the discretization of quantum field theories on a graph representation are two fundamental steps undertaken to formulate quantum fields within the framework of lattice field theory. Lattice field theory, which shares inherent connections with statistical physics and probability theory, provides a rigorous approach to define mathematically and to simulate numerically quantum field theories. It is through the probabilistic perspective that a connection between lattice field theories and a certain class of machine learning algorithms, namely the mathematical framework of Markov random fields, can be rigorously established [1].

A Markov random field can be defined as a set of random variables on a graph structure that satisfy Markov properties. Markov properties are significant conditions of locality that emerge across different research fields. Within the constructive aspect of quantum field theory, Markov properties appear as required conditions to rigorously construct quantum fields in Minkowski space based on Markov fields in Euclidean space [2]. In machine learning [3], algorithms that satisfy Markov properties have been used extensively in problems ranging from computer vision to computational biology, where the imposed condition of locality is exploited to uncover local structures in data.

In this contribution, we derive machine learning algorithms and neural networks from lattice field theories, opening up the opportunity to investigate machine learning directly within quantum field theory. Specifically, we demonstrate, through the Hammersley-Clifford theorem, that the ϕ^4 scalar field theory on a square lattice satisfies the local Markov property with respect to the graph structure, therefore recasting it as a Markov random field. We then derive neural network architectures from the ϕ^4 theory which can be viewed as generalizations of conventional architectures. Finally, we present applications pertinent to the minimization of an asymmetric distance between the probability distribution of the ϕ^4 machine learning algorithms and target probability distributions.

2. The ϕ^4 theory as a Markov random field

We consider a physical model which is represented by a finite set Λ whose points lie on the vertices of a finite graph $\mathcal{G} = (\Lambda, e)$, where e is the set of edges in $\mathcal{G}$. Two points $i, j \in \Lambda$ are called neighbours if there exists an edge between i and j. A clique is a subset of Λ whose points are neighbours, and a clique is called maximal if no additional point can be included such that the resulting set is still a clique, see Fig. 1. We now associate to each point $i \in \Lambda$ a random variable ϕ_i.
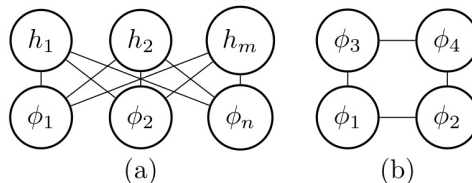

Figure 1: A bipartite graph (a) and a square lattice (b). Examples of maximal cliques are $\{\phi_1, h_1\}$ and $\{\phi_3, \phi_4\}$, respectively.

Given a graph $\mathcal{G} = (\Lambda, e)$ the random variables define a Markov random field if the associated probability distribution $p(\phi)$ satisfies the local Markov property with respect to the graph structure $\mathcal{G}$:

$$p(\phi_i | (\phi_j)_{j \in \Lambda}) = p(\phi_i | (\phi_j)_{j \in \mathcal{N}_i}), \tag{1}$$

where $\mathcal{N}_i$ is the set of neighbours of a given point i. We will prove the local Markov property for a set of random variables using the Hammersley-Clifford theorem.

Theorem 1 (Hammersley-Clifford). *A strictly positive distribution p satisfies the local Markov property of an undirected graph $\mathcal{G}$, if and only if p can be represented as a product of strictly positive potential functions ψ_c over $\mathcal{G}$, one per maximal clique $c \in C$, i.e.,*

$$p(\phi) = \frac{1}{Z} \prod_{c \in C} \psi_c(\phi), \tag{2}$$

where $Z = \int_{\boldsymbol{\phi}} \prod_{c \in C} \psi_c(\boldsymbol{\phi}) d\boldsymbol{\phi}$ is the partition function and $\boldsymbol{\phi}$ are all possible states of the system.

We now consider the two-dimensional ϕ^4 theory on a square lattice with the Euclidean action

$$S_E = -\kappa_L \sum_{\langle ij \rangle} \phi_i \phi_j + \frac{(\mu_L^2 + 4\kappa_L)}{2} \sum_i \phi_i^2 + \frac{\lambda_L}{4} \sum_i \phi_i^4. \tag{3}$$

The dimensionless parameters $\kappa_L, \mu_L^2, \lambda_L$ are now considered as inhomogeneous and are redefined as $w = \kappa_L, a = (\mu_L^2 + 4\kappa_L)/2, b = \lambda_L/4$, giving rise to the action

$$S(\phi; \theta) = -\sum_{\langle ij \rangle} w_{ij} \phi_i \phi_j + \sum_i a_i \phi_i^2 + \sum_i b_i \phi_i^4, \tag{4}$$

where the set of coupling constants is $\theta = \{w_{ij}, a_i, b_i\}$, and the associated Boltzmann probability distribution is:

$$p(\phi; \theta) = \frac{\exp\left[-S(\phi; \theta)\right]}{\int_{\boldsymbol{\phi}} \exp[-S(\boldsymbol{\phi}, \theta)] d\boldsymbol{\phi}}. \tag{5}$$

The ϕ^4 lattice field theory is, by definition, formulated on a graph $\mathcal{G} = (\Lambda, e)$. A nonunique choice of a potential function ψ_c which factorizes the probability distribution in terms of maximal cliques $c \in C$ is

$$\psi_c = \exp\left[-w_{ij} \phi_i \phi_j + \frac{1}{4}(a_i \phi_i^2 + a_j \phi_j^2 + b_i \phi_i^4 + b_j \phi_j^4)\right], \tag{6}$$

where i, j are nearest neighbours.

By demonstrating that the strictly positive distribution $p(\phi; \theta)$ of the ϕ^4 scalar field theory on a square lattice can be represented as a product of strictly positive potential functions ψ_c over the maximal cliques of the graph $\mathcal{G}$, we have verified that the ϕ^4 theory is a Markov random field.

3. Machine learning with ϕ^4 Markov random fields

3.1 Learning without predefined data

Within the probabilistic perspective a wide class of machine learning applications can be expressed as the minimization of a distance metric between two probability distributions. Specifically, we are searching for the optimal values of the variational parameters θ, which are the coupling constants of the ϕ^4 theory, that are able to minimize the considered distance metric.

As a metric we consider the Kullback-Leibler divergence, which can be viewed as an asymmetric distance between the probability distribution $p(\phi; \theta)$ of the ϕ^4 machine learning algorithms and a target probability distribution $q(\phi)$:

$$KL(p||q) = \int_{-\infty}^{\infty} p(\boldsymbol{\phi}; \theta) \ln \frac{p(\boldsymbol{\phi}; \theta)}{q(\boldsymbol{\phi})} d\boldsymbol{\phi} \geq 0. \tag{7}$$

The Kullback-Leibler divergence is a nonnegative quantity which becomes zero if and only if the two probability distributions $p(\phi; \theta)$ and $q(\phi)$ are equal. Our aim is to minimize this quantity through a variational approach, specifically a gradient-based algorithm. After a successful minimization, the two probability distributions will be equal and we can use the probability distribution $p(\phi; \theta)$ to draw samples from the target distribution $q(\phi)$.

We consider a target Boltzmann probability distribution $q(\phi) = \exp[-\mathcal{A}]/Z_{\mathcal{A}}$ which describes an arbitrary statistical system or an arbitrary lattice field theory. By substituting the probability distributions $p(\phi)$ and $q(\phi)$ in the Kullback-Leibler divergence we obtain:

$$F_{\mathcal{A}} \leq \langle \mathcal{A} - S \rangle_{p(\phi; \theta)} + F \equiv \mathcal{F}, \tag{8}$$

where $\mathcal{F}$ is the variational free energy and $\langle O \rangle_{p(\phi; \theta)}$ denotes the expectation value of an observable O under the probability distribution $p(\phi; \theta)$. There are two important observations for Eq. 8: first, it sets a rigorous upper bound to the calculation of the free energy $F_{\mathcal{A}}$ of the target system with action $\mathcal{A}$ and, second, the minimization of this quantity is entirely dependent on calculations conducted under the probability distribution $p(\phi; \theta)$ of the ϕ^4 theory. The implication of these observations is that the system with target action $\mathcal{A}$ can be approximated and, subsequently, sampled using exclusively configurations drawn from the probability distribution $p(\phi; \theta)$ of the ϕ^4 theory.

To minimize the variational free energy $\mathcal{F}$ we implement a gradient-based approach

$$\frac{\partial \mathcal{F}}{\partial \theta_i} = \langle \mathcal{A} \rangle \left\langle \frac{\partial S}{\partial \theta_i} \right\rangle - \left\langle \mathcal{A} \frac{\partial S}{\partial \theta_i} \right\rangle + \left\langle S \frac{\partial S}{\partial \theta_i} \right\rangle - \langle S \rangle \left\langle \frac{\partial S}{\partial \theta_i} \right\rangle, \tag{9}$$

and the variational parameters θ are then updated at each epoch t based on the relation:

$$\theta^{(t+1)} = \theta^{(t)} - \eta * \mathcal{L}, \tag{10}$$

where η is the learning rate and $\mathcal{L} = \partial \mathcal{F} / \partial \theta^{(t)}$.

We now consider as a target system a ϕ^4 lattice action $\mathcal{A}$ defined as

$$\mathcal{A} = \sum_{k=1}^{5} g_k \mathcal{A}^{(k)} = g_1 \sum_{\langle ij \rangle_{nn}} \phi_i \phi_j + g_2 \sum_i \phi_i^2 + g_3 \sum_i \phi_i^4 + g_4 \sum_{\langle ij \rangle_{nnn}} \phi_i \phi_j + i g_5 \sum_i \phi_i^2, \tag{11}$$

where nn and nnn denote nearest-neighbor and next-nearest neighbor interactions, the $g_4 \mathcal{A}^{(4)}$ term is a longer-range interaction, and the action is complex due to the $g_5 \mathcal{A}^{(5)}$ term. The coupling constants can have arbitrary values but for this example we consider $g_1 = g_4 = -1$, $g_2 = 1.52425$, $g_3 = 0.175$ and $g_5 = 0.15$, see Ref. [1].

A first proof-of-principle demonstration is to train a ϕ^4 Markov random field with lattice size $L = 4$ in each dimension and randomly initialized coupling constants in the action S to approximate the lattice action $\mathcal{A}_{\{3\}}$. After training the ϕ^4 Markov random field for 10^5 epochs, the coupling
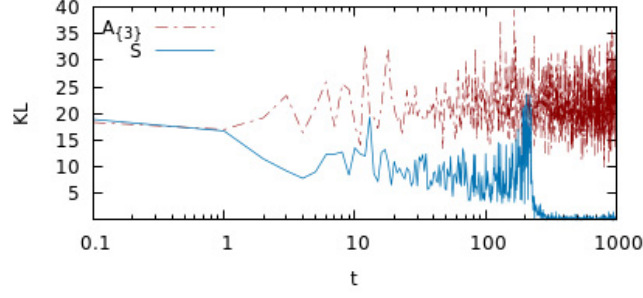
Figure 2: Estimated Kullback-Leibler divergence versus epoch t on logarithmic scale. The probability distributions of actions $\mathcal{A}_{\{3\}}$ and S are compared with the one of $\mathcal{A}_{\{4\}}$. Only the action S is updated at each epoch based on a finite sample of fixed size. For action $\mathcal{A}_{\{3\}}$ results are depicted based on a finite sample of equal size to allow for a direct comparison of the two quantities at each epoch t.

constants have converged to the anticipated values g_1, g_2 and g_3 with precision of order or magnitude 10^{-8} (see Ref. [1]) and the two systems have therefore become identical.

To compare the representational capacity between the ϕ^4 Markov random field with inhomogeneous action S versus a standard ϕ^4 action $\mathcal{A}_{\{3\}}$ we can estimate the Kullback-Leibler divergence during the training process. Specifically, we estimate the asymmetric distances of the probability distributions with actions S and $\mathcal{A}_{\{3\}}$ against the probability distribution of the action $\mathcal{A}_{\{4\}}$, which includes longer-range interactions. The results are depicted in Fig. 2, where the inhomogeneous action S approximates the target system exceedingly better than the standard action, indicating that it has increased representational capacity.

We will now investigate how accurately a ϕ^4 Markov random field has approximated a target probability distribution of action $\mathcal{A}_{\{4\}}$ which includes longer-range interactions. Explicitly, we will investigate if reweighting can be implemented based on the approximating action S to extrapolate in the parameter space of the full action $\mathcal{A}$. We consider the following reweighting procedure:

$$\langle O \rangle = \frac{\sum_{l=1}^{N} O_l \exp[S_l - g_j' \mathcal{A}_l^{(j)} - \sum_{k=1, k \neq j}^{5} g_k \mathcal{A}_l^{(k)}]}{\sum_{l=1}^{N} \exp[S_l - g_j' \mathcal{A}_l^{(j)} - \sum_{k=1, k \neq j}^{5} g_k \mathcal{A}_l^{(k)}]}. \tag{12}$$

This form of reweighting acts as a correction step between the probability distributions of the actions S and $\mathcal{A}_{\{4\}}$ and simultaneously extrapolates along the trajectory of the coupling constant g_j' and the
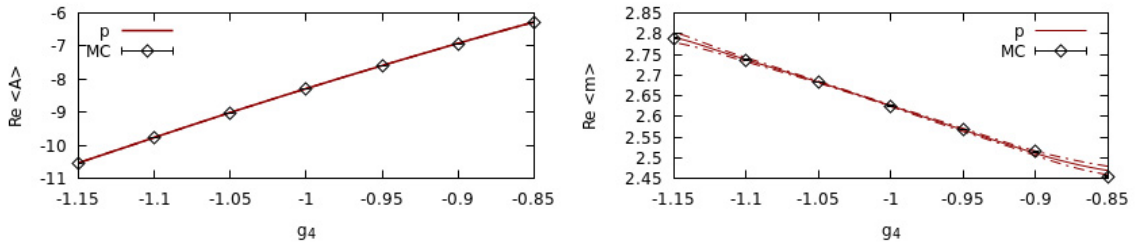

Figure 3: Real part of the complex lattice action $\mathcal{A}$ (left) and real part of the magnetization m (right) versus coupling constant g_4. The statistical errors are comparable with the width of the line.
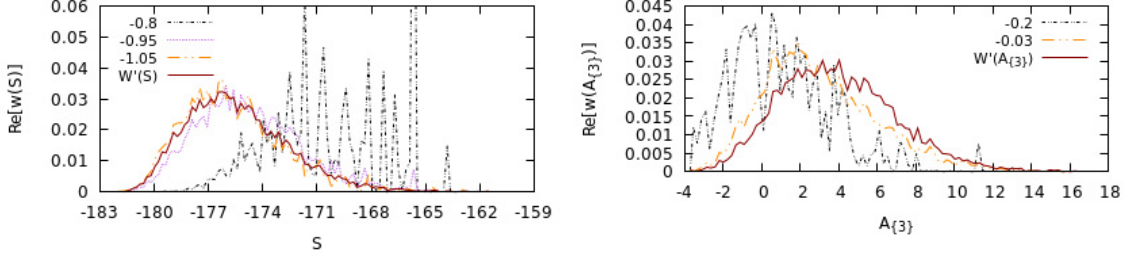
Figure 4: Real part of the weight function $\mathcal{W}(S)$ versus lattice action S (left). Real part of the weight function $\mathcal{W}(\mathcal{A}_{\{3\}})$ versus lattice action $\mathcal{A}_{\{3\}}$ (right). The results are for considered coupling constants g'_4.

complex term $g_5\mathcal{A}^{(5)}$. Details can be found in Refs. [1, 4–6].

We consider $j = 4$ and extrapolate the observables of the action $\mathcal{A}$ and the magnetization m for values of the coupling constant $g'_4 \in [-1.15, -0.85]$. The results are depicted in Fig. 3 where it is evident that they agree within statistical errors with independent calculations on Monte Carlo datasets supplemented with reweighting to the complex action. The results certify that reweighting to the parameter space of the full action $\mathcal{A}$ is possible based on the approximating action S.

The remaining question is how to strictly define the permitted reweighting range during an extrapolation in parameter space based on an approximating action S. This can be achieved through the construction of weight functions

$$\mathcal{W}(S) = \frac{\sum_{\Re[\mathcal{A}'],\Im[\mathcal{A}']} h(S, \Re[\mathcal{A}'], \Im[\mathcal{A}']) \exp[S - \Re[\mathcal{A}'] - i\Im[\mathcal{A}']]}{\sum_{S,\Re[\mathcal{A}'],\Im[\mathcal{A}']} h(S, \Re[\mathcal{A}'], \Im[\mathcal{A}']) \exp[S - \Re[\mathcal{A}'] - i\Im[\mathcal{A}']]}, \tag{13}$$

where $\mathcal{A}' = g'_j \mathcal{A}^{(j)} + \sum_{k=1,k\neq j}^{5} g_k \mathcal{A}^{(k)}$. The quantity $h(S, \Re[\mathcal{A}'], \Im[\mathcal{A}'])$ is a multi-dimensional histogram of the inhomogeneous action S as well as each action term.

The calculations of the weight functions $\mathcal{W}(S)$ are depicted in Fig. 4 (left). We observe that weight functions which are constructed for coupling constants adjacent to $g_4 = -1$ overlap. In addition, inconsistencies can be observed for $g'_4 = -0.8$, indicating that reweighting extrapolations beyond the range of this specific value are inacurrate. We remark that histogram reweighting from the action $\mathcal{A}_3$ to the full action is impossible due to an insufficient overlap of statistical ensembles. The result is depicted in Fig. 4 (right). Inconsistencies in reweighting begin to emerge for the value $g'_4 = -0.2$, therefore verifying that the parameter space which corresponds to the coupling constant $g'_4 = -1.0$ in the action $\mathcal{A}$ cannot be reached from the parameter space of the action $\mathcal{A}_3$.

3.2 Learning with predefined data

In the previous examples, established based on the minimization of the Kullback-Leibler divergence of Eq. 7, we considered that the probability distribution $q(\phi)$ and the corresponding action $\mathcal{A}$ were known. However, there exist cases where one might have already obtained a dataset, for instance through a Monte Carlo simulation, as experimental data or based on a set of images. These datasets then encode an empirical probability distribution $q(\phi)$ whose form is unknown. One can then ask if we can learn the appropriate coupling constants in the action S of the ϕ^4 theory that are able to approximate this target probability distribution $q(\phi)$ which is explicitly encoded in the
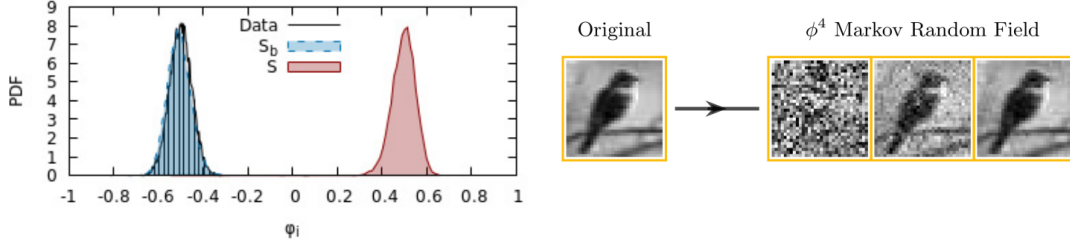
Figure 5: Probability density function versus lattice value ϕ_i for a Euclidean action S that is Z_2 invariant and S_b which includes a local symmetry-breaking term (left). Original image and equilibration of the Markov random field after 1, 10 and 50 steps (right).

available dataset. The problem can be solved using an alternative expression of the Kullback-Leibler divergence:

$$KL(q||p) = \int_{-\infty}^{\infty} q(\boldsymbol{\phi}) \ln \frac{q(\boldsymbol{\phi})}{p(\boldsymbol{\phi}; \theta)} d\boldsymbol{\phi} \geq 0. \tag{14}$$

By substituting and taking the derivative in terms of the variational parameters θ we obtain:

$$\frac{\partial \ln p(\phi; \theta)}{\partial \theta} = \left\langle \frac{\partial S}{\partial \theta} \right\rangle_{p(\phi; \theta)} - \frac{\partial S}{\partial \theta}, \tag{15}$$

and the update rule of the parameters θ at each epoch t is given according to Eq. 10, where $\mathcal{L} = -\partial \ln p(\phi; \theta^{(t)})/\partial \theta^{(t)}$.

We reiterate that the method is applicable to an arbitrary probability distribution $q(\phi)$. For instance, one can consider as $q(\phi)$ the empirical distribution encoded in a dataset sampled from a Gaussian distribution with $\mu = -0.5$ and $\sigma = 0.05$. We anticipate, due to the Z_2 invariance in the lattice action, that the symmetric probability distribution is equiprobable in being reproduced. To explicitly reproduce the desired distribution a term of the form $\sum_i r_i \phi_i$ can be included in the action: this term breaks the symmetry of the system explicitly. The results are depicted in Fig. 5 (left) where we observe the anticipated behaviour.

We additionally demonstrate the applicability of the method in an image from the CIFAR-10 dataset. Specifically, we search for the optimal values of the coupling constants in the ϕ^4 action that are able to reproduce the image as a configuration in the equilibrium distribution. The results are verified in Fig. 5 (right). We remark the method can be potentially extended to applications pertinent to the renormalization group [7], where it can be used to calculate the appropriate coupling constants which describe renormalized systems.

3.3 Machine learning with ϕ^4 neural networks

To derive a neural network architecture from the ϕ^4 theory we will introduce within the graph structure a new set of variables h_j which are called hidden. In addition, we will consider that the topology of the graph structure is that of a bipartite graph, and that the interactions are exclusively between the visible ϕ and the hidden h variables. The lattice action which describes this system is

$$S(\phi, h; \theta) = -\sum_{i,j} w_{ij} \phi_i h_j + \sum_i r_i \phi_i + \sum_i a_i \phi_i^2 + \sum_i b_i \phi_i^4 + \sum_j s_j h_j + \sum_j m_j h_j^2 + \sum_j n_j h_j^4. \tag{16}$$

Figure 6: Example features learned in the hidden layer of the ϕ^4 neural network.

By setting certain coupling constants to zero or by considering that the visible or the hidden variables are binary one can arrive at different neural network architectures. These architectures have been extensively studied in computer science, for instance, see Ref. [3] and references therein. Specifically, if $b_i = n_j = 0$ one obtains a Gaussian-Gaussian restricted Boltzmann machine and if $b_i = n_j = m_j = 0$ and $h_j \in \{-1, 1\}$ then the architecture is a Gaussian-Bernoulli restricted Boltzmann machine. We emphasize that the ϕ^4 theory reduces to an Ising model under the limit κ_L positive and fixed, $\lambda_L \to \infty$ and $\mu_L^2 \to -\infty$. This equivalence could potentially shed novel insights into this class of machine learning algorithms which were initially inspired by Ising models.

To demonstrate that the ϕ^4 neural network of Eq. 16 is capable of extracting meaningful features during the learning of a probability distribution with predefined data, we train it on the first forty examples of the Olivetti faces dataset. A representative example of the learned features, i.e. the coupling constants w_{ij} for a fixed j, are depicted in Fig. 6. We observe that the neural network has learned features which correspond to abstract face shapes.

4. Conclusions

In this contribution we derived machine learning algorithms and neural networks from lattice field theories, opening up the opportunity to investigate machine learning directly within quantum field theory [1]. Through the Hammersley-Clifford theorem, we established a precise equivalence between the ϕ^4 theory on a square lattice and the mathematical framework of Markov random fields. We additionally derived neural networks from the ϕ^4 theory which can be viewed as generalizations of a notable class of machine learning algorithms, namely that of restricted Boltzmann machines. Finally, applications pertinent to physics and to computer science were presented.

Numerous research directions can be envisaged. These include the derivation of machine learning algorithms from lattice gauge theories, the investigation of the inhomogeneity in the coupling constants of these systems and the exploration for the emergence of novel phase transitions within quantum field-theoretic machine learning. Finally, through the construction of quantum fields in Minkowski space from Markov fields in Euclidean space, quantum field-theoretic machine learning can be additionally explored based on a mathematical perspective, one that was firstly established within constructive quantum field theory and probability theory.

In summary, the derivation of quantum field-theoretic machine learning algorithms opens up a new research avenue, that of developing a mathematical and computational framework of machine learning within quantum field theory. As a result, the work paves the way for nonperturbative investigations of the dynamics which describe neural networks and machine learning algorithms,

a prospect which can be achieved from the perspective of physics and within the highly successful framework of lattice field theory.

References

[1] D. Bachtis, G. Aarts, B. Lucini, *Quantum field-theoretic machine learning*, Phys. Rev. D **103**, 074510 (2021).

[2] E. Nelson, *Construction of quantum fields from Markoff fields*, Journal of Functional Analysis **12**, 97 (1973).

[3] A. Fischer, C. Igel, *Training restricted Boltzmann machines: An introduction*, Pattern Recognition **47**, 25 (2014).

[4] D. Bachtis, G. Aarts, B. Lucini, *Adding machine learning within Hamiltonians: Renormalization group transformations, symmetry breaking and restoration*, Phys. Rev. Research **3**, 013134 (2021).

[5] D. Bachtis, G. Aarts, B. Lucini, *Mapping distinct phase transitions to a neural network*, Phys. Rev. E **102**, 053306 (2020).

[6] D. Bachtis, G. Aarts, B. Lucini, *Extending machine learning classification capabilities with histogram reweighting*, Phys. Rev. E **102**, 033303 (2020).

[7] D. Bachtis, G. Aarts, F. Di Renzo, B. Lucini, *Inverse renormalization group in quantum field theory*, arXiv:2107.00466 [hep-lat] (2021).