Quantum field-theoretic machine learning

Dimitrios Bachtis,[1,*] Gert Aarts,[2,3,†] and Biagio Lucini[1,4,‡]

[1] *Department of Mathematics, Swansea University, Bay Campus, SA1 8EN, Swansea, Wales, UK*
[2] *Department of Physics, Swansea University, Singleton Campus, SA2 8PP, Swansea, Wales, UK*
[3] *European Centre for Theoretical Studies in Nuclear Physics and Related Areas (ECT*) & Fondazione Bruno Kessler Strada delle Tabarelle 286, 38123 Villazzano (TN), Italy*
[4] *Swansea Academy of Advanced Computing, Swansea University, Bay Campus, SA1 8EN, Swansea, Wales, UK*

(Dated: February 18, 2021)

We derive machine learning algorithms from discretized Euclidean field theories, making inference and learning possible within dynamics described by quantum field theory. Specifically, we demonstrate that the ϕ^4 scalar field theory satisfies the Hammersley-Clifford theorem, therefore recasting it as a machine learning algorithm within the mathematically rigorous framework of Markov random fields. We illustrate the concepts by minimizing an asymmetric distance between the probability distribution of the ϕ^4 theory and that of target distributions, by quantifying the overlap of statistical ensembles between probability distributions and through reweighting to complex-valued actions with longer-range interactions. Neural networks architectures are additionally derived from the ϕ^4 theory which can be viewed as generalizations of conventional neural networks and applications are presented. We conclude by discussing how the proposal opens up a new research avenue, that of developing a mathematical and computational framework of machine learning within quantum field theory.

I. INTRODUCTION

Relativistic quantum fields [1] are formulated on Minkowski space where intricate mathematical problems related to the hyperbolic geometry emerge. By recasting Minkowski space as Euclidean significant simplifications can be obtained for certain cases: The hyperbolic problems are transformed to be elliptic, the Poincaré group becomes the Euclidean group where a positive definite scalar product emerges, noncommuting operators are expressed as random variables and causality is formulated as a Markov property.

Of high importance is the reverse direction: that of arriving at a quantum field in Minkowski space by constructing it from one in Euclidean space. To make such prospects attainable a rigorous mathematical framework for quantum fields had to be established, and a series of relevant contributions led to advances known as constructive quantum field theory [2–4]. A connection between probability theory and quantum field theory was then established when quantum fields were constructed from Euclidean fields that satisfy Markov properties [5, 6].

Recently, applications of deep learning [7], a class of machine learning algorithms which are able to hierarchically extract abstract features in data, have emerged in the physical sciences [8], including in lattice field theories [9–16] and in the study of phase transitions [17–23]. Insights on machine learning algorithms have been obtained from the perspective of statistical physics [24–32], particularly within the theory of spin glasses [33], or in relation to Gaussian processes [34–38].

A notable case of these algorithms is the framework of Markov random fields [39], which introduces Markov properties on a graph-based representation to encode probability distributions over high-dimensional spaces. As quantum field theory and probability theory are evidently connected analytically [6], and computational investigations of quantum fields are feasible through the framework of lattice field theory [40], a new challenge is anticipated to emerge: namely that of investigating machine learning from the perspective of quantum fields.

In this manuscript, we derive machine learning algorithms from discretized Euclidean field theories, making inference and learning possible within dynamics described by quantum field theory. From the mathematical point of view, we explore if the ϕ^4 scalar field theory on a square lattice satisfies the Hammersley-Clifford theorem, therefore recasting it as a Markov random field which can complete machine learning tasks. From the equivalent perspective of physics, we treat the ϕ^4 scalar field theory as a system with inhomogeneous coupling constants and we search based on its dynamics, which comprise local interactions, for the optimal values of the coupling constants that are able to complete a machine learning task. Specifically we consider the minimization of an asymmetric distance between the probability distribution of the ϕ^4 theory and that of target distributions. We also quantify the overlap of statistical ensembles between probability distributions and investigate if reweighting to the parameter space of complex-valued actions with longer-range interactions is possible by utilizing instead the probability distribution of the approximating local inhomogeneous action.

We then proceed to derive neural network architectures from the ϕ^4 scalar field theory which can progressively extract features of increased abstraction in data. We explore the implications of including a local symmetry-

* dimitrios.bachtis@swansea.ac.uk
† g.aarts@swansea.ac.uk
‡ b.lucini@swansea.ac.uk

breaking term in the ϕ^4 Markov random field, and rearrange the lattice topology to derive a ϕ^4 neural network which can be viewed as a generalization of conventional neural network architectures. Based on the equivalence between the ϕ^4 scalar field theory and the Ising model under a certain limit, we discuss how the ϕ^4 neural network can provide novel physical insights to the interpretability of a notable class of machine learning algorithms. Finally, we conclude by discussing how the introduction of ϕ^4 machine learning algorithms opens up a new research avenue, that of developing, computationally and analytically, a framework of machine learning within quantum field theory.

II. THE ϕ^4 SCALAR FIELD THEORY AS A MARKOV RANDOM FIELD

Let Λ be a finite set whose points represent the sites of a physical model, and let Λ have an additional structure, for instance consider that the spacing between the sites might be known and that the sites are connected. We now consider that the points of Λ lie on the vertices of a finite graph $\mathcal{G} = (\Lambda, e)$, where e is the set of edges on $\mathcal{G}$. If $i, j \in \Lambda$ and there exists an edge between i and j then i and j are called neighbours and the set of all neighbours of a considered point i will be denoted by $\mathcal{N}_i$. A clique is a subset of Λ where the points are pairwise connected, and a clique is called maximal if no additional point can be included such that the resulting set is still a clique. We will denote a maximal clique as c and the set of all maximal cliques as C. For an illustration of the concepts see Fig. 1 and for rigorous results see Refs. [39, 41].

In addition we associate to each point $i \in \Lambda$ a random variable $\phi_{i,i\in\Lambda}$ and we will call $\phi = \{\phi_i\}$ a state or configuration of the system. Given a graph $\mathcal{G} = (\Lambda, e)$, the set of random variables define a Markov random field if the associated probability distribution p fulfills the local Markov property with respect to $\mathcal{G}$. The local Markov property denotes that a variable ϕ_i is conditionally independent of all other variables given its neighbors $\mathcal{N}_i$, i.e:

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

A probability distribution is then related with the events generated by a Markov random field through the Hammersley-Clifford theorem [39]:

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 nonnegative 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.

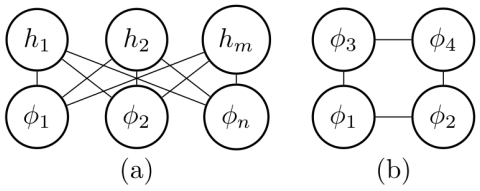

FIG. 1. (a) A bipartite graph. The maximal cliques correspond to the sites associated with the random variables $\{\phi_1, h_1\}$, $\{\phi_1, h_2\}$, $\{\phi_1, h_m\}$, $\{\phi_2, h_1\}$, $\{\phi_2, h_2\}$, $\{\phi_2, h_m\}$, $\{\phi_n, h_1\}$, $\{\phi_n, h_2\}$, $\{\phi_n, h_m\}$. (b) A square lattice. The maximal cliques correspond to the sites associated with the random variables $\{\phi_1, \phi_2\}$, $\{\phi_1, \phi_3\}$, $\{\phi_3, \phi_4\}$ and $\{\phi_2, \phi_4\}$.

We will demonstrate that the ϕ^4 scalar field theory satisfies the Hammersley-Clifford theorem and is therefore a Markov random field. The two-dimensional ϕ^4 theory is described by the Euclidean Langrangian:

$$\mathcal{L}_E = \frac{\kappa}{2}(\nabla\phi)^2 + \frac{\mu_0^2}{2}\phi^2 + \frac{\lambda}{4}\phi^4, \tag{3}$$

where the action that regularizes the continuum theory on a square lattice is:

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

The quantities $\kappa_L, \mu_L^2, \lambda_L$ are dimensionless parameters, one of which is deprecated and can be absorbed by rescaling the fields [42]. Nevertheless, consider the set of variables $w = \kappa_L$, $a = (\mu_L^2 + 4\kappa_L)/2$, $b = \lambda_L/4$ as inhomogeneous and the resulting action as:

$$S(\phi;\theta) = -\sum_{\langle ij\rangle} w_{ij}\phi_i\phi_j + \sum_i a_i\phi_i^2 + \sum_i b_i\phi_i^4, \tag{5}$$

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

$$p(\phi;\theta) = \frac{\exp\big[-S(\phi;\theta)\big]}{\int_{\boldsymbol{\phi}}\exp[-S(\boldsymbol{\phi},\theta)]d\boldsymbol{\phi}}. \tag{6}$$

The ϕ^4 scalar field theory is formulated on a graph $\mathcal{G} = (\Lambda, e)$ where Λ is the set of lattice sites and e the set of edges or pairwise interactions. For a square lattice only nearest neighbors define a maximal clique (see Fig. 1). Since we search for arbitrary, strictly positive potential functions ψ_c per maximal clique $c \in C$, we can multiply ψ_c with nonnegative functions of subsets of c [43], i.e. with functions of one-site cliques. We then arrive, after considering the imposed boundary conditions, at a nonunique choice of potential function:

$$\psi_c = \exp\left[-w_{ij}\phi_i\phi_j + \frac{1}{4}(a_i\phi_i^2 + a_j\phi_j^2 + b_i\phi_i^4 + b_j\phi_j^4)\right], \tag{7}$$

where i, j are nearest neighbors. As the potential functions ψ_c are nonnegative the quantity $\ln \psi_c$ can be defined, and the probability distribution $p(\phi;\theta)$ can be factorized as:

$$p(\phi;\theta) = \frac{\exp\left[\sum_{c\in C} \ln \psi_c(\phi)\right]}{\int_{\boldsymbol{\phi}} \exp\left[\sum_{c\in C} \ln \psi_c(\boldsymbol{\phi})\right] d\boldsymbol{\phi}} = \frac{1}{Z} \prod_{c\in C} \psi_c(\phi). \tag{8}$$

To summarize, the discretized ϕ^4 scalar field theory satisfies the Hammersley-Clifford theorem and the local Markov property and is therefore a Markov random field. To understand intuitively the meaning of the local Markov property, consider the more familiar case satisfied by a Markov chain $P(\phi^{k+1}|\phi^k, \ldots, \phi^0) = P(\phi^{k+1}|\phi^k)$. This property declares that given a certain state ϕ^k a future state ϕ^{k+1} depends only on the current state ϕ^k, and not on states that preceded it, such as ϕ^{k-1}. The local Markov property of Eq. 1 extends this concept to higher dimensions by giving it a spatial representation via a Markov random field. For the case of the ϕ^4 scalar field theory the variational parameters θ are the coupling constants $\theta = \{w_{ij}, a_i, b_i\}$. By considering that the probability $p(\phi;\theta)$ of the Markov random field depends on the parameters θ a variety of machine learning tasks can then be completed.

III. MACHINE LEARNING WITH THE ϕ^4 SCALAR FIELD THEORY

A. Learning without predefined data

Consider a target probability distribution $q(\phi)$ of an arbitrary statistical system. An asymmetric measure of the distance between the two probability distributions $p(\phi;\theta)$ and $q(\phi)$ can be defined, which is called the Kullback-Leibler divergence [39]:

$$KL(p||q) = \int_{-\infty}^{\infty} p(\boldsymbol{\phi};\theta) \ln \frac{p(\boldsymbol{\phi};\theta)}{q(\boldsymbol{\phi})} d\boldsymbol{\phi} \geq 0. \tag{9}$$

The Kullback-Leibler divergence is nonnegative and equal to zero when the two probability distributions exactly match one another. We emphasize that the Kullback-Leibler divergence does not satisfy the triangle inequality and it therefore cannot be classified as a proper distance as it isn't symmetric. It is the quantity $KL(p||q) + KL(q||p)$ which is a true metric. The Kullback-Leibler divergence will be called an asymmetric distance to retain the intuitive picture that it establishes a measure of the difference between two probability distributions.

By searching for an optimal set of coupling constants $\theta = \{w_{ij}, a_i, b_i\}$ we can minimize the Kullback-Leibler divergence so that the probability distribution of the ϕ^4 scalar field theory $p(\phi;\theta)$ will converge to the target probability distribution $q(\phi)$. Once minimization is conducted a Markov chain Monte Carlo simulation can be initiated for $p(\phi;\theta)$ to draw samples that would be representative of the target distribution $q(\phi)$. Let us consider the case where the target probability distribution $q(\phi)$ is that of an arbitrary statistical system with partition function $Z_{\mathcal{A}}$ and it has a Boltzmann form $q(\phi) = \exp[-\mathcal{A}]/Z_{\mathcal{A}}$. Any additional parameter, such as the inverse temperature, is absorbed within the Hamiltonian or lattice action $\mathcal{A}$. By substituting $q(\phi)$ and $p(\phi;\theta)$ in Eq. 9 we arrive at:

$$-\ln Z_{\mathcal{A}} \leq \langle \mathcal{A} - S \rangle_{p(\phi;\theta)} - \ln Z. \tag{10}$$

By considering that the terms $F_{\mathcal{A}} = -\ln Z_{\mathcal{A}}$ and $F = -\ln Z$ are equal to the free energy, the above equation can be equivalently expressed as:

$$F_{\mathcal{A}} \leq \langle \mathcal{A} - S \rangle_{p(\phi;\theta)} + F \equiv \mathcal{F}, \tag{11}$$

where $\mathcal{F}$ is the variational free energy. As a result Eq. 11 sets a rigorous upper bound to the calculation of the free energy $F_{\mathcal{A}}$ of the target system and this bound $\mathcal{F}$ is dependent on calculations conducted entirely on the distribution $p(\phi;\theta)$ of the ϕ^4 Markov random field. This indicates that one can map an arbitrary system to a ϕ^4 scalar field theory by minimizing an asymmetric distance between the probability distributions of the two systems.

A gradient-based approach can then be implemented to minimize the variational free energy $\mathcal{F}$ via its derivatives in terms of the parameters θ:

$$\frac{\partial \mathcal{F}}{\partial \theta_i} = \langle \mathcal{A} \rangle \Big\langle \frac{\partial S}{\partial \theta_i} \Big\rangle - \Big\langle \mathcal{A} \frac{\partial S}{\partial \theta_i} \Big\rangle + \Big\langle S \frac{\partial S}{\partial \theta_i} \Big\rangle - \langle S \rangle \Big\langle \frac{\partial S}{\partial \theta_i} \Big\rangle, \tag{12}$$

where all expectation values are calculated under the probability distribution $p(\phi;\theta)$ of the ϕ^4 scalar field theory. Derivations can be found in Appendix A. The variational parameters are then updated at each epoch t of the minimization process through:

$$\theta^{(t+1)} = \theta^{(t)} - \eta * \mathcal{L}, \tag{13}$$

where η is the learning rate and $\mathcal{L} = \partial \mathcal{F}/\partial \theta^{(t)}$. After the minimization process we anticipate that $\mathcal{F} \approx F_A$ and as a result $p(\phi;\theta) \approx q(\phi)$.

To illustrate the approach we consider as a target system a ϕ^4 lattice action $\mathcal{A}$ with longer-range interactions and complex-valued coupling constants, defined as:

$$\mathcal{A} = \sum_{k=1}^{5} g_k \mathcal{A}^{(k)} = g_1 \sum_{\langle ij \rangle_{nn}} \phi_i \phi_j + g_2 \sum_i \phi_i^2 \tag{14}$$

$$+ g_3 \sum_i \phi_i^4 + g_4 \sum_{\langle ij \rangle_{nnn}} \phi_i \phi_j + i g_5 \sum_i \phi_i^2. \tag{15}$$

The notation nn and nnn denotes nearest neighbor and next-nearest neighbor interactions and the lattice action is complex due to the $g_5 \mathcal{A}^{(5)}$ term. The combination of the g_2 and g_5 parameters introduces a complex coupling constant in the mass term. The coupling constants have values $g1 = g4 = -1$, $g_2 = 1.52425$, $g_3 = 0.175$ and

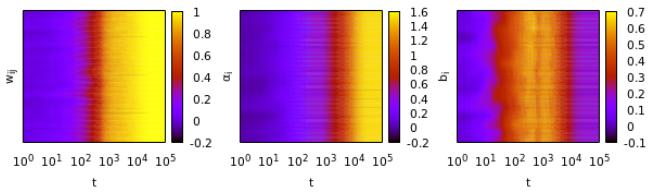

FIG. 2. Variational parameters $\theta = \{w_{ij}, a_i, b_i\}$ versus epochs t on logarithmic scale. The figures depict the evolution of the parameters θ towards the expected values of the coupling constants in the target homogeneous action.

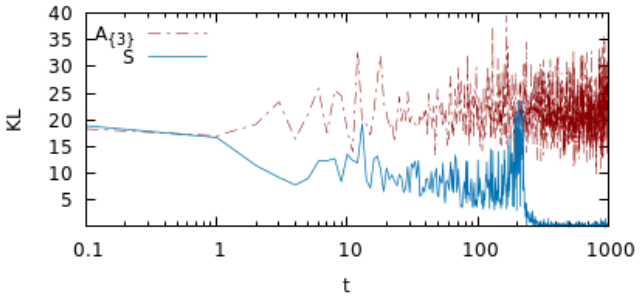

FIG. 3. Estimated Kullback-Leibler divergence versus epoch t on logarithmic scale. The probability distributions of actions $\mathcal{A}_{\{3\}}$ and S are compared with the one of $\mathcal{A}_{\{4\}}$. Only the action S is updated at each epoch based on a finite sample of fixed size. For action $\mathcal{A}_{\{3\}}$ 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.

$g_5 = 0.15$. The values for g_1, g_2 and g_3 have been chosen near the critical point of the second-order phase transition for the system with a local homogeneous action for which $g_4 = g_5 = 0$. We will present three applications for lattices of size $L = 4$ at each dimension: firstly, a proof-of-principle demonstration will be conducted to verify that the inhomogeneous action S (see Eq. 5) can learn the local lattice action $\mathcal{A}_{\{3\}} = \sum_{k=1}^{3} g_k \mathcal{A}^{(k)}$. Secondly, we will discuss that by considering the local lattice action $\mathcal{A}_{\{3\}}$ it is impossible to reweight to the full action $\mathcal{A}$ due to insufficient overlap of statistical ensembles, but there exists an inhomogeneous representation of $\mathcal{A}_{\{3\}}$ equal to S for which this is possible. Finally we will demonstrate that S can approximate $\mathcal{A}$ sufficiently to simultaneously extrapolate observables in the parameter space of the complex action $\mathcal{A}$ along the trajectory of a considered coupling constant and we will discuss how to successfully define the allowed reweighting range.

We now initialize the ϕ^4 Markov random field with inhomogeneous coupling constants θ which are randomly drawn from a Gaussian distribution and consider as a target system in Eq. 10 the local lattice action $\mathcal{A}_{\{3\}}$. We anticipate that the optimal solution is the one where the inhomogeneous coupling constants θ of the ϕ^4 Markov random field will converge to the homogeneous constants g_1, g_2 and g_3 of the target ϕ^4 scalar field theory. Details about the simulations can be found in Appendix B. The time evolution for the parameters θ is depicted in Fig 2 and details of the training process can be found in Appendix B. After training is conducted the parameters θ have converged to the homogeneous constants of the target system with precision of order of magnitude of 10^{-8} for all cases. It then becomes clear that given sufficient training time the two systems become identical.

The overlap of statistical ensembles can be quantified through the Kullback-Leibler divergence. We consider the probability distribution $p(\boldsymbol{\phi}; \theta)$, described by the local inhomogeneous action S, and we minimize the Kullback-Leibler divergence to approximate the target distribution of action $\mathcal{A}_{\{4\}}$ which is denoted as $q(\phi)$. In addition, we simultaneously estimate the Kullback-Leibler divergence between the distributions of $\mathcal{A}_{\{3\}}$ and $\mathcal{A}_{\{4\}}$ to quantify their overlap of statistical ensembles. The results are depicted in Fig. 3 where it is evident that the local inhomogeneous action S produces a probability distribution which approximates $\mathcal{A}_{\{4\}}$ exceedingly better than the probability distribution of $\mathcal{A}_{\{3\}}$. This tentatively indicates that while S and $\mathcal{A}_{\{3\}}$ have the same form of lattice action, the inhomogeneity present in the former allows for the construction of richer representations of probability distributions. As a result, histogram reweighting [44] from local inhomogeneous actions to regions of parameter space that are inaccessible to the local homogeneous action might be possible.

We proceed to discuss the precise implications of the equivalence between the approximating distribution $p(\phi; \theta)$ of action S and the target distribution $q(\phi)$ of action $\mathcal{A}_{\{4\}}$. The definition of the expectation value $\langle O \rangle_P$ of an arbitrary observable O in a system that has some equilibrium occupation probabilities P is:

$$\langle O \rangle_P = \sum_{\boldsymbol{\phi}} O_{\boldsymbol{\phi}} P(\boldsymbol{\phi}), \tag{16}$$

where the sum is over all possible states $\boldsymbol{\phi}$ of the system. After the Kullback-Leibler divergence between the distributions $p(\phi; \theta)$ and $q(\phi)$ is minimized we anticipate that:

$$p(\phi; \theta) \approx q(\phi), \tag{17}$$

which instantly implies, based on Eq.16, that:

$$\langle O \rangle_{p(\phi;\theta)} \approx \langle O \rangle_{q(\phi)}. \tag{18}$$

To clarify further, observables, such as the lattice action $\mathcal{A}_{\{4\}}$ should yield approximately equal values when calculated from samples drawn from either distribution $p(\phi; \theta)$ or $q(\phi)$ even though the two distributions have different actions S and $\mathcal{A}_{\{4\}}$, respectively. To express these ideas in a more formal manner, we now consider the expectation value of an arbitrary observable as obtained during a Monte Carlo simulation (e.g. see Refs [20, 21])

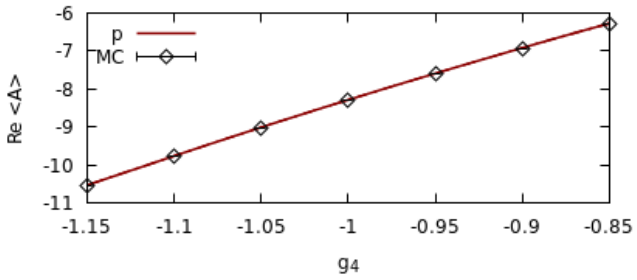

FIG. 4. Real part of the complex lattice action $\mathcal{A}$ versus coupling constant g_4. The results are obtained by reweighting from the Markov random field distribution p to the distribution of the complex action $\mathcal{A}$. The statistical errors are comparable with the width of the line. The results are compared with Monte Carlo (MC) and reweighting from the distribution of the real action $\mathcal{A}_{\{4\}}$ to $\mathcal{A}$.

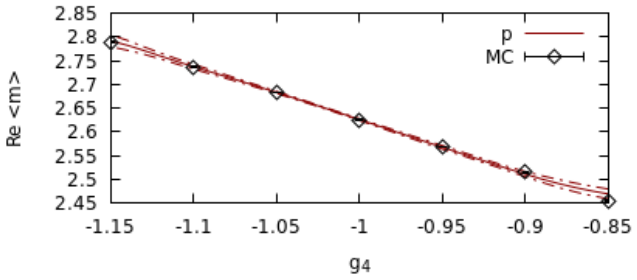

FIG. 5. Real part of the magnetization m versus coupling constant g_4. The results are obtained by reweighting from the Markov random field distribution p to the target distribution of the complex action $\mathcal{A}$. The associated statistical errors are depicted by the dashed lines. The results are compared with Monte Carlo (MC) and reweighting from the distribution of the real action $\mathcal{A}_{\{4\}}$ to $\mathcal{A}$.

in the target system with action $\mathcal{A}_{\{4\}}$:

$$\langle O\rangle_{q(\phi)} = \frac{\sum_{l=1}^{N} \tilde{p}_l^{-1} O_l \exp[-\sum_{k=1}^{4} g_k \mathcal{A}_l^{(k)}]}{\sum_{l=1}^{N} \tilde{p}_l^{-1} \exp[-\sum_{k=1}^{4} g_k \mathcal{A}_l^{(k)}]}, \tag{19}$$

where $\tilde{p}$ are the probabilities used to sample from the equilibrium distribution and N the number of samples that we have obtained during the Monte Carlo simulation. There are two fundamentally different ways to proceed in calculating the expectation value of the above equation by relying instead on the approximating probability distribution $p(\phi;\theta)$.

The first is to draw a subset of samples from $p(\phi;\theta)$ and then conjecture, based on Eq. 17, that these N samples have been produced instead by the distribution $q(\phi)$. This would have been equivalent to considering $\tilde{p} = q(\phi)$ in Eq. 19 but a systematic error would be introduced based on the accuracy in which the probability distribution $p(\phi;\theta)$ approximates $q(\phi)$. The second approach again relies on drawing a subset of samples from the distribution $p(\phi;\theta)$, but this time we will consider that $p(\phi;\theta) \neq q(\phi)$ and that the samples have been produced directly from $p(\phi;\theta)$ of Eq. 6 with action S. This is equivalent to conducting a reweighting step to arrive from distribution $p(\phi;\theta)$ to $q(\phi)$ under a sufficient overlap of ensembles. We anticipate that this reweighting step is possible to achieve due to the minimization of the Kullback-Leibler divergence between the two distributions $p(\phi;\theta)$ to $q(\phi)$ and their approximate equivalence.

We will follow the second approach and implement a reweighting technique, details of which can be found in Appendix C, to simultaneously extrapolate observables in the parameter space of the full action $\mathcal{A}$ which includes complex couplings and longer-range interactions:

$$\langle O\rangle = \frac{\sum_{l=1}^{N} O_l \exp[S_l - g'_j \mathcal{A}_l^{(j)} - \sum_{k=1,k\neq j}^{5} g_k \mathcal{A}_l^{(k)}]}{\sum_{l=1}^{N} \exp[S_l - g'_j \mathcal{A}_l^{(j)} - \sum_{k=1,k\neq j}^{5} g_k \mathcal{A}_l^{(k)}]}. \tag{20}$$

The equation above can be interpreted as two distinct simultaneous reweighting steps. Firstly the probability distribution $p(\phi;\theta)$ of the ϕ^4 Markov random field with action S is reweighted to the distribution $q(\phi)$ with action $\mathcal{A}_{\{4\}}$ but with a shifted coupling constant g'_j. This acts as a correction step to ensure that the proper distribution is reached from $p(\phi;\theta)$ and it additionally allows an extrapolation along the direction of the parameter space described by coupling g'_j. Secondly there is a reweighting step to reach the distribution described by the complex lattice action $\mathcal{A}$, which includes the imaginary part $g_5\mathcal{A}^{(5)}$. Any arbitrary observable can be reweighted in parameter space, such as machine learning derived observables [21], and Hamiltonian-agnostic reweighting [20] could additionally be explored.

We consider that $j = 4$ and we extrapolate observables along the trajectory of the g'_4 coupling constant for a continuous range of values $g'_4 \in [-0.85, -1.15]$. We recall that the ϕ^4 Markov random field was trained to approximate the action $\mathcal{A}_{\{4\}}$ where $g_4 = -1$. Results for the magnetization and the internal energy, obtained with reweighting from the probability distribution $p(\phi;\theta)$ to the full action $\mathcal{A}$ are depicted in Figs. 4 and 5. The results are compared with Monte Carlo simulations conducted on action $\mathcal{A}_{\{4\}}$ which are combined with reweighting to the full complex distribution to allow for a comparison with the ones from $p(\phi;\theta)$. It is evident that the results depicted agree within statistical errors with the Monte Carlo extrapolations. Details about the statistical error analysis can be found in Appendix D.

When reweighting is implemented to extrapolate to the probability distribution of a complex action or as a correction step in the case of an approximating distribution the question of how to strictly define the reweighting range emerges. This can be achieved, formally, through the calculation of weight functions which are dependent on the underlying histograms. Specifically, we consider as an example in Eq. 20 the expectation value of the action S. In addition, instead of expressing Eq. 20 as a sum over

each action S_l calculated on a configuration ϕ we instead reformulate it in terms of each uniquely sampled action S in the Monte Carlo dataset after the construction of histograms. The expectation value is then:

$$\langle S \rangle = \sum_{S} S\mathcal{W}(S), \tag{21}$$

where the sum is over uniquely sampled actions S and $\mathcal{W}(S)$ is a weight function which is equal to:

$$\mathcal{W}(S) = \frac{\sum_{\Re[\mathcal{A}'],\Im[\mathcal{A}']} h(S,\Re[\mathcal{A}'],\Im[\mathcal{A}'])\exp[S-\Re[\mathcal{A}']-i\Im[\mathcal{A}']]}{\sum_{S,\Re[\mathcal{A}'],\Im[\mathcal{A}']} h(S,\Re[\mathcal{A}'],\Im[\mathcal{A}'])\exp[S-\Re[\mathcal{A}']-i\Im[\mathcal{A}']]}, \tag{22}$$

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 in which we are interested to extrapolate towards during reweighting. Reweighting can be achieved either by including novel terms in the action or by shifting its corresponding coupling constant if the term already exists. Of particular interest is also the quantity $\mathcal{W}'(S)$ where the exponentials are chosen equal to one and which is proportional to the actual histograms of the action in the corresponding Monte Carlo dataset. This quantity can additionally serve as an indication of the reweighting range.

We proceed to calculate the weight functions $\mathcal{W}(S)$ for each uniquely sampled action S in a considered extrapolation range. The results are depicted in Fig. 6 where an overlap between distinct weight functions that are adjacent in parameter space to the coupling constant $g_4 = -1$ is observed. We recall that reweighting extrapolations are accurate only when the method successfully predicts the form of histograms at the extrapolated point in parameter space based on the histograms present at the initial dataset. When the coupling constant is $g'_4 = -0.8$ major inconsistencies can be noticed. This indicates that reweighting extrapolations to $g'_4 = -0.8$ would be inaccurate as the form of the weight functions cannot be successfully predicted.

We emphasize that reweighting from the local homogeneous action $\mathcal{A}_{\{3\}}$ to the full action $\mathcal{A}$ is not possible. The inclusion of an imaginary term and a longer range interaction does not produce a sufficient overlap of ensembles. Results are depicted in Fig. 7. We recall that the local homogeneous action $\mathcal{A}_{\{3\}}$ has coupling constant $g_4 = 0$ and the target distribution of action $\mathcal{A}_{\{4\}}$ includes a term with coupling constant $g'_4 = -1.0$. It is clear that the values of the lattice action lie at an entirely different scale and inconsistencies begin to emerge when $g'_4 = -0.2$. Reweighting to the full action is then impossible from the probability distribution of action $\mathcal{A}_{\{3\}}$. However, the local inhomogeneous action S is able to achieve reweighting to the full distribution of the action $\mathcal{A}$. Consequently the opportunity to map improved lattice actions, which include longer-range interactions, to local inhomogeneous actions is a prospect that is open to explore. This can be achieved by minimizing the asymmetric distance between their associated probability distributions.

B. Learning with predefined data

The preceding results do not require any predefined data to be used as input within the training process since configurations were obtained during the gradient-based

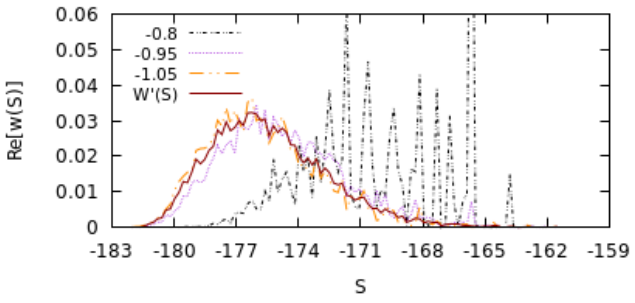

FIG. 6. Real part of the weight function $\mathcal{W}(S)$ versus lattice action S for considered coupling constants $g'_4 \in [-1.05, -0.8]$. The results are obtained by reweighting from the local inhomogeneous action S to the complex action $\mathcal{A}$ which includes longer-range interactions.

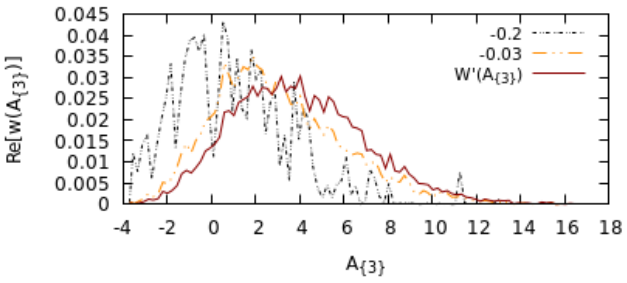

FIG. 7. Real part of the weight function $\mathcal{W}(\mathcal{A}_{\{3\}})$ versus lattice action $\mathcal{A}_{\{3\}}$ for considered coupling constants g'_4. The results are obtained by reweighting from action $\mathcal{A}_{\{3\}}$ to the complex action $\mathcal{A}$ which includes longer-range interactions.

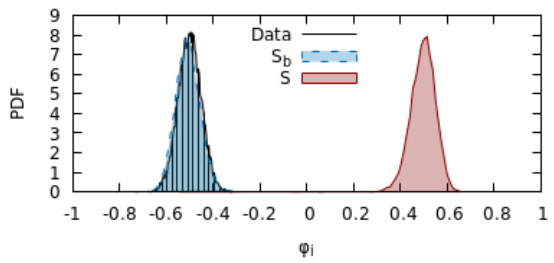

FIG. 8. Probability density function versus lattice value ϕ_i for a Euclidean action S that is Z_2 invariant and S_b which includes a local symmetry-breaking term.

approach. However, there exist cases where one has already obtained a set of available data, which could comprise configurations of a system, experimental data, or a set of images, and whose probability distribution is of unknown form. The obtained dataset then explicitly encodes an empirical probability distribution $q(\phi)$ that is a representation of the complete probability distribution of the system. The empirical distribution $q(\phi)$ can still be learned by minimizing instead the opposite divergence:

$$KL(q||p) = \int_{-\infty}^{\infty} q(\boldsymbol{\phi}) \ln \frac{q(\boldsymbol{\phi})}{p(\boldsymbol{\phi};\theta)} d\boldsymbol{\phi} \geq 0. \tag{23}$$

By expanding the above equation we arrive at:

$$KL(q||p) = \langle \ln q(\phi) \rangle_{q(\phi)} - \langle \ln p(\phi;\theta) \rangle_{q(\phi)}. \tag{24}$$

The first right hand term is constant and the minimization of $KL(q||p)$ is therefore equivalent to the maximization of the second right hand term under the training data:

$$\frac{\partial \ln p(\phi;\theta)}{\partial \theta} = \left\langle \frac{\partial S}{\partial \theta} \right\rangle_{p(\phi;\theta)} - \frac{\partial S}{\partial \theta}. \tag{25}$$

The variational parameters are now updated according to Eq. 13 where $\mathcal{L} = -\partial \ln p(\phi;\theta^{(t)})/\partial \theta^{(t)}$.

To illustrate the concepts we now create a dataset from a Gaussian distribution with $\mu = -0.5$ and $\sigma = 0.05$ which encodes an empirical distribution $q(\phi)$. The information about the form of $q(\phi)$ will not be introduced in Eq. 23 because the training will instead be conducted on the obtained data. To clarify further, the same approach can be established for any obtained dataset, without the need to even infer the underlying form of the distribution. After successful training, Markov chain Monte Carlo simulations can be implemented based on the distribution $p(\phi;\theta)$ of the ϕ^4 Markov random field to draw samples that would be representative of the unknown target distribution $q(\phi)$. Additional details can be found in Appendix A.

We anticipate, due to the invariance under the Z_2 symmetry in the lattice action S, that the symmetric distribution with $\mu = 0.5$ might be additionally reproduced.

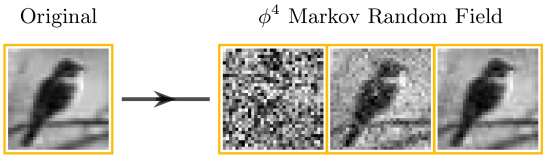

FIG. 9. Original image and equilibration of the Markov random field after 1, 10 and 50 steps.

If this feature is not desirable then a local symmetry-breaking term of the form $\sum_i r_i \phi_i$ can be included in the action S to favor configurations that will explicitly reproduce $q(\phi)$. The Hammersley-Clifford theorem is still satisfied and results for the symmetric action S and the action S_b which includes a symmetry-breaking term are depicted in Fig. 8. We observe for the symmetric case that while the algorithm has been trained on one of the probable solutions it is able to produce additional solutions that are invariant under the inherent symmetry, whereas this feature has been eliminated for the broken-symmetry case where the probability distribution $q(\phi)$ is explicitly reproduced.

Markov random fields are widely applied to problems in computer vision, image segmentation and compression, as well as image analysis [45]. Every problem that is formulated as an energy or lattice action minimization problem can be solved by implementing Markov random fields. Since the ϕ^4 scalar field theory satisfies the Hammersley-Clifford theorem and is therefore a Markov random field it can be implemented to complete such tasks. We therefore consider as $q(\phi)$ in Eq. 24 the configuration of an image from the CIFAR-10 dataset [46], which we will map to the action of the inhomogeneous ϕ^4 theory of Eq. 5. In essence, we search for the optimal values of the coupling constants, which describe the local interactions in the ϕ^4 scalar field theory, that can reproduce the considered image as a configuration in the equilibrium distribution of the system. In Fig. 9, results are depicted after training the ϕ^4 theory. We observe that by initializing a Markov chain the configurations of the equilibrium distribution converge to an accurate representation of the original image.

IV. ϕ^4 NEURAL NETWORKS

When the aim of the machine learning task is to study intricate probability distributions, deep learning algorithms that include multiple layers in the neural network architecture can be implemented. These layers progressively transform data to arrive at increasingly abstract representations, allowing for increased expressivity and representational capacity in the model. Such cases of deep learning algorithms can be constructed from the dynamics of the ϕ^4 scalar field theory.

We consider that part of the random variables ϕ_i on the lattice sites are visible and correspond to a set of

observations and the remaining are hidden variables h_j, which capture dependencies on a set of training data, given as input to ϕ_i. In addition, to make the connection with the computer science literature we consider a bipartite graph which imposes the restriction that interactions are exclusively between the ϕ and the h variables (see Fig. 1). We therefore recast the ϕ^4 neural network as a variant of a restricted Boltzmann machine (RBM) [47–50], which is able to model continuous data. Alternative parametrizations of the graph structure are open to explore. A joint probability distribution $p(\phi, h; \theta)$ is then defined, based on a lattice action $S(\phi, h; \theta)$:

$$S(\phi, h; \theta) = -\sum_{i,j} w_{ij}\phi_i h_j + \sum_i r_i \phi_i + \sum_i a_i \phi_i^2 \quad (26)$$

$$+ \sum_i b_i \phi_i^4 + \sum_j s_j h_j + \sum_j m_j h_j^2 + \sum_j n_j h_j^4, \quad (27)$$

which also gives rise to a new expression, based on Eq. 23, for the derivative of the log-likelihood $\ln p(\phi, \theta)$:

$$\frac{\partial \ln p(\phi;\theta)}{\partial \theta} = \left\langle \frac{\partial S}{\partial \theta} \right\rangle_{p(\phi,h;\theta)} - \left\langle \frac{\partial S}{\partial \theta} \right\rangle_{p(h|\phi;\theta)}, \quad (28)$$

where the set of variational parameters is now $\theta = \{w_{ij}, r_i, a_i, b_i, s_j, m_j, n_j\}$. The conditional distributions of the visible and the hidden variables are $p(\phi|h;\theta) = \prod_i p(\phi_i|h)$ and $p(h|\phi;\theta) = \prod_j p(h_j|\phi)$. Derivations can be found in Appendix A.

By considering certain values of parameters in the ϕ^4 neural network of Eq. 26 one can arrive at other neural network architectures, all of which are special cases of a ϕ^4 Markov random field. For instance by choosing $b_i = n_j = 0$ one obtains a Gaussian-Gaussian RBM [47, 49]. If $b_i = n_j = m_j = 0$ and $h_j \in \{-1, 1\}$ then the architecture is a Gaussian-Bernoulli RBM[47, 49]. Of particular interest could be the choice of $m_j = n_j = 0$ and $h_j \in \{-1, 1\}$ which would reduce to a ϕ^4-Bernoulli RBM, a case with a non-linear sigmoid function that, to our knowledge, has not been studied before. We emphasize that the ϕ^4 Bernoulli RBM is anticipated to have substantial representational capacity due to the presence of the non-linear sigmoid function in the hidden layer [51].

It is a well-known fact that the ϕ^4 scalar field theory of Eq. 4, a model with continuous degrees of freedom, reduces to an Ising model under the limit κ_L fixed, $\lambda_L \to \infty$ and $\mu_L^2 \to -\infty$ [42]. The ϕ^4-Bernoulli RBM can then be interpreted as a ϕ^4 neural network where certain lattice sites have reached the Ising limit, allowing for novel physical insights. It is important to recall that, with the inclusion of two hidden layer layers, deep variants of restricted Boltzmann machines are universal approximators of probability distributions [52].

To demonstrate the applicability of the ϕ^4 neural network of Eq. 26, we train it on the first forty examples of the Olivetti faces dataset [53] using 4096 visible units and 32 hidden unit to observe if meaningful features are learned. A subset of the learned features, i.e. the coupling constants w_{ij} for a fixed j, are depicted in Fig. 10. We observe that the neural network has learned hidden features which comprise abstract face shapes and characteristics. The hidden units can then serve as input to a new ϕ^4 neural network to progressively extract abstract features in data [54].

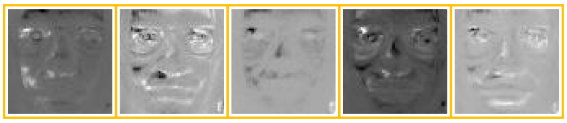

FIG. 10. Example features learned in the hidden layer of the ϕ^4 neural network.

V. CONCLUSIONS

In this manuscript we derived machine learning algorithms from discretized Euclidean field theories. Specifically we demonstrated that the ϕ^4 scalar field theory on a square lattice satisfies the Hammersley-Clifford theorem and is therefore a Markov random field that can be used for inference and learning. By recasting the ϕ^4 theory within a mathematically rigorous framework a variety of theorems, as well as training algorithms, are available and an overview can be found in Ref. [39]. As the resulting algorithm has inhomogeneous coupling constants it can additionally be investigated from the perspective of spin glasses [33], and enhanced sampling can be obtained based on computational techniques from statistical mechanics [55, 56], or model-specific algorithms [57, 58].

The Kullback-Leibler divergence can be utilized to quantify the overlap of statistical ensembles between probability distributions. Specifically, we demonstrated that the ϕ^4 scalar field theory with inhomogeneous coupling constants is able to absorb longer range interactions and observables can be reweighted to the parameter space of complex actions using the approximating probability distribution. The prospect of constructing improved lattice actions [59, 60] based on local inhomogeneous representations is open to explore.

In principle any arbitrary system can be mapped to a ϕ^4 scalar field theory with inhomogeneous coupling constants by minimizing an asymmetric distance of their probability distributions based on Eq. 10. The concepts are therefore anticipated to be generally applicable to systems within condensed matter physics, lattice field theories and statistical mechanics. To enhance the accuracy a variant of a neural network architecture can be implemented which is proven to be a universal approximator of a probability distribution [52]. In the manuscript such variants have been presented as special cases of a ϕ^4 neural network.

The resulting ϕ^4 machine learning algorithm of Sections II and III retains the topology of the lattice structure and the boundary conditions, but differs from the conventional ϕ^4 scalar field theory due to the inhomogeneous coupling constants. To employ the tools of quantum field theory a framework involving the replica method is required, but the theories can still be formulated in terms of the functional integral with an additional averaging over the space of couplings [61]. It is noted that in our formulation the couplings are inhomogeneous but not random as they are determined during the minimization process.

We emphasize that prior arguments considering the Hammersley-Clifford theorem hold for arbitrary dimensions and one could therefore construct a d-dimensional Markov random field to initiate analytical or computational investigations. The factorization of a lattice action in terms of products of potential functions, a step that is required to recast a system as a Markov random field, depends on the topology of the graph structure and different topologies yield different maximal cliques. An equivalence between local, pairwise and global Markov properties of a graph structure can also be rigorously proven [39]. Through the construction of quantum fields in Minkowski space from Markov fields in Euclidean space [6], a new research avenue is envisaged, namely that of developing a computational and mathematical framework of machine learning within quantum field theory.

VI. ACKNOWLEDGEMENTS

The authors received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 813942. The work of GA and BL has been supported in part by the UKRI Science and Technology Facilities Council (STFC) Consolidated Grant ST/P00055X/1. The work of BL is further supported in part by the Royal Society Wolfson Research Merit Award WM170010 and by the Leverhulme Foundation Research Fellowship RF-2020-461\9. Numerical simulations have been performed on the Swansea SUNBIRD system. This system is part of the Supercomputing Wales project, which is part-funded by the European Regional Development Fund (ERDF) via Welsh Government. We thank COST Action CA15213 THOR for support.

Appendix A: Derivations

1. ϕ^4 Markov Random Field

The Kullback-Leibler divergence, which is repeated here for convenience, defines an asymmetric measure of the distance between the distribution of the machine learning algorithm $p(\phi;\theta)$ and an unknown target distribution $q(\phi)$:

$$KL(p||q) = \int_{-\infty}^{\infty} p(\boldsymbol{\phi};\theta) \ln \frac{p(\boldsymbol{\phi};\theta)}{q(\boldsymbol{\phi})} d\boldsymbol{\phi} \geq 0. \tag{A1}$$

By expanding the above equation we arrive at:

$$\langle \ln p(\phi;\theta) \rangle_{p(\phi;\theta)} - \langle \ln q(\phi) \rangle_{p(\phi;\theta)} \geq 0, \tag{A2}$$

where $\langle\rangle_{p(\phi;\theta)}$ denotes the expectation value under the probability distribution $p(\phi;\theta)$. If the two probability distributions are substituted to be of Boltzmann form, $p(\phi;\theta) = \exp[-S]/Z$, $q(\phi) = \exp[-\mathcal{A}]/Z_{\mathcal{A}}$, we arrive at:

$$-\langle \ln Z_{\mathcal{A}} \rangle_{p(\phi;\theta)} \leq \langle \mathcal{A} - S \rangle_{p(\phi;\theta)} - \langle \ln Z \rangle_{p(\phi;\theta)}. \tag{A3}$$

The terms $\langle \ln Z \rangle_{p(\phi;\theta)}$ are constant in terms of expectation values and we therefore obtain:

$$-\ln Z_{\mathcal{A}} \leq \langle \mathcal{A} - S \rangle_{p(\phi;\theta)} - \ln Z. \tag{A4}$$

By denoting the right hand part as $\mathcal{F}$, the derivative in terms of a variational parameter θ_i is equal to:

$$\frac{\partial \mathcal{F}}{\partial \theta_i} = \frac{\partial \langle \mathcal{A} \rangle_{p(\phi;\theta)}}{\partial \theta_i} - \frac{\partial \langle S \rangle_{p(\phi;\theta)}}{\partial \theta_i} - \frac{\partial(-\ln Z)}{\partial \theta_i}, \tag{A5}$$

where each term is calculated as:

$$\frac{\partial \langle \mathcal{A} \rangle_{p(\phi;\theta)}}{\partial \theta_i} = \frac{\partial}{\partial \theta_i} \left[\frac{\int_{\boldsymbol{\phi}} \mathcal{A}(\boldsymbol{\phi}) \exp[-S(\boldsymbol{\phi};\theta)] d\boldsymbol{\phi}}{\int_{\boldsymbol{\phi}} \exp[-S(\boldsymbol{\phi};\theta)] d\boldsymbol{\phi}} \right] = -\Big\langle \mathcal{A} \frac{\partial S}{\partial \theta_i} \Big\rangle_{p(\phi;\theta)} + \langle \mathcal{A} \rangle_{p(\phi;\theta)} \Big\langle \frac{\partial S}{\partial \theta_i} \Big\rangle_{p(\phi;\theta)}, \tag{A6}$$

$$\frac{\partial \langle S \rangle_{p(\phi;\theta)}}{\partial \theta_i} = \frac{\partial}{\partial \theta_i} \left[\frac{\int_{\boldsymbol{\phi}} S(\boldsymbol{\phi};\theta) \exp[-S(\boldsymbol{\phi};\theta)] d\boldsymbol{\phi}}{\int_{\boldsymbol{\phi}} \exp[-S(\boldsymbol{\phi};\theta)] d\boldsymbol{\phi}} \right] = \Big\langle \frac{\partial S}{\partial \theta_i} \Big\rangle_{p(\phi;\theta)} - \Big\langle S \frac{\partial S}{\partial \theta_i} \Big\rangle_{p(\phi;\theta)} + \langle S \rangle_{p(\phi;\theta)} \Big\langle \frac{\partial S}{\partial \theta_i} \Big\rangle_{p(\phi;\theta)}, \tag{A7}$$

$$\frac{\partial(-\ln Z)}{\partial \theta_i} = -\frac{\int_{\boldsymbol{\phi}} \frac{\partial}{\partial \theta_i}(-S(\boldsymbol{\phi};\theta)) \exp[-S(\boldsymbol{\phi};\theta)] d\boldsymbol{\phi}}{\int_{\boldsymbol{\phi}} \exp[-S(\boldsymbol{\phi};\theta)] d\boldsymbol{\phi}} = \Big\langle \frac{\partial S}{\partial \theta_i} \Big\rangle_{p(\phi;\theta)}, \tag{A8}$$

By substituting we arrive at:

$$\frac{\partial \mathcal{F}}{\partial \theta_i} = -\Big\langle \mathcal{A}\frac{\partial S}{\partial \theta_i}\Big\rangle + \langle \mathcal{A}\rangle\Big\langle \frac{\partial S}{\partial \theta_i}\Big\rangle - \cancel{\Big\langle \frac{\partial S}{\partial \theta_i}\Big\rangle} + \Big\langle S\frac{\partial S}{\partial \theta_i}\Big\rangle - \langle S\rangle\Big\langle \frac{\partial S}{\partial \theta_i}\Big\rangle + \cancel{\Big\langle \frac{\partial S}{\partial \theta_i}\Big\rangle}. \tag{A9}$$

A gradient based approach can be implemented based on the above equation to learn a target known probability distribution.

On the opposite direction if a set of data is available for which the probability distribution is unknown the alternative Kullback-Leibler divergence can be considered:

$$KL(q||p) = \int_{-\infty}^{\infty} q(\boldsymbol{\phi}) \ln \frac{q(\boldsymbol{\phi})}{p(\boldsymbol{\phi};\theta)} d\boldsymbol{\phi}. \tag{A10}$$

By expanding the right-hand side we arrive at the expression:

$$KL(q||p) = \langle \ln q(\phi)\rangle_{q(\phi)} - \langle \ln p(\phi;\theta)\rangle_{q(\phi)}. \tag{A11}$$

Minimizing the Kullback-Leibler divergence is equivalent to the maximization of the term $\langle \ln p(\phi;\theta)\rangle_{q(\phi)}$, which is:

$$\langle \ln p(\phi;\theta)\rangle_{q(\phi)} = \frac{1}{N}\sum_x \ln p(\phi^{(x)};\theta), \tag{A12}$$

where x is a training example and N the number of training data. For the case of the Markov random field the derivative of the log-likelihood is:

$$\begin{aligned}
\frac{\partial \ln p(\phi;\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta}\left[\ln \frac{\exp[-S(\phi;\theta)]}{\int_{\boldsymbol{\phi}} \exp[-S(\boldsymbol{\phi};\theta)]d\boldsymbol{\phi}}\right] = \frac{\partial}{\partial \theta}\left[\ln \exp[-S(\phi;\theta)] - \ln \int_{\boldsymbol{\phi}} \exp[-S(\boldsymbol{\phi};\theta)]d\boldsymbol{\phi}\right] \\
&= \frac{\partial}{\partial \theta}(-S(\phi;\theta)) - \frac{\int_{\boldsymbol{\phi}} \frac{\partial}{\partial \theta}(-S(\boldsymbol{\phi};\theta))\exp[-S(\boldsymbol{\phi};\theta)]d\boldsymbol{\phi}}{\int_{\boldsymbol{\phi}} \exp[-S(\boldsymbol{\phi};\theta)]d\boldsymbol{\phi}} = \frac{\partial}{\partial \theta}(-S(\phi;\theta)) - \int_{\boldsymbol{\phi}} p(\phi;\theta)\frac{\partial(-S(\phi;\theta))}{\partial \theta}d\boldsymbol{\phi} \\
&= \frac{\partial}{\partial \theta}(-S(\phi;\theta)) - \Big\langle \frac{\partial}{\partial \theta}(-S(\phi;\theta))\Big\rangle_{p(\phi;\theta)},
\end{aligned}$$

where the term outside the expectation values is calculated on the training examples.

2. ϕ^4 Neural Network

The case of the quantum field-theoretic neural network is more complicated due to the joint probability distribution of the visible units ϕ and the hidden units h:

$$p(\phi, h;\theta) = \frac{\exp[-S(\phi,h;\theta)]}{\int_{\boldsymbol{\phi},\boldsymbol{h}} \exp[-S(\boldsymbol{\phi},\boldsymbol{h};\theta)]d\boldsymbol{\phi}d\boldsymbol{h}}. \tag{A13}$$

From the joint probability distribution we can define marginal probability distributions via:

$$p(\phi;\theta) = \int_{\boldsymbol{h}} p(\phi,\boldsymbol{h};\theta)d\boldsymbol{h} = \frac{\int_{\boldsymbol{h}} \exp[-S(\phi,\boldsymbol{h};\theta)]d\boldsymbol{h}}{\int_{\boldsymbol{\phi},\boldsymbol{h}} \exp[-S(\boldsymbol{\phi},\boldsymbol{h};\theta)]d\boldsymbol{\phi}d\boldsymbol{h}},$$

$$p(h;\theta) = \int_{\boldsymbol{\phi}} p(\boldsymbol{\phi},h;\theta)d\boldsymbol{\phi} = \frac{\int_{\boldsymbol{\phi}} \exp[-S(\boldsymbol{\phi},h;\theta)]d\boldsymbol{\phi}}{\int_{\boldsymbol{\phi},\boldsymbol{h}} \exp[-S(\boldsymbol{\phi},\boldsymbol{h};\theta)]d\boldsymbol{\phi}d\boldsymbol{h}},$$

as well as conditional probability distributions through:

$$p(\phi|h;\theta) = \frac{p(\phi,h;\theta)}{p(h;\theta)} = \frac{\exp[-S(\phi,h;\theta)]dh}{\int_{\boldsymbol{\phi}}\exp[-S(\boldsymbol{\phi},h;\theta)]d\boldsymbol{\phi}} \tag{A14}$$

$$= \frac{\exp[\sum_{i,j} w_{ij}\phi_i h_j - \sum_i r_i\phi_i - \sum_i a_i\phi_i^2 - \sum_i b_i\phi_i^4 - \cancel{\sum_j s_j h_j} - \cancel{\sum_j m_j h_j^2} - \cancel{\sum_j n_j h_j^4}]}{\int_{\boldsymbol{\phi}}\exp[\sum_{i,j} w_{ij}\phi_i h_j - \sum_i r_i\phi_i - \sum_i a_i\phi_i^2 - \sum_i b_i\phi_i^4 - \cancel{\sum_j s_j h_j} - \cancel{\sum_j m_j h_j^2} - \cancel{\sum_j n_j h_j^4}]d\boldsymbol{\phi}} \tag{A15}$$

$$= \frac{\prod_i \exp[\phi_i \sum_j w_{ij}h_j - r_i\phi_i - a_i\phi_i^2 - b_i\phi_i^4]}{\int_{\boldsymbol{\phi}}\prod_i \exp[\phi_i \sum_j w_{ij}h_j - r_i\phi_i - a_i\phi_i^2 - b_i\phi_i^4]d\boldsymbol{\phi}} \tag{A16}$$

$$= \prod_i \frac{\exp[\phi_i \sum_j w_{ij}h_j - r_i\phi_i - a_i\phi_i^2 - b_i\phi_i^4]}{\int_{\boldsymbol{\phi_i}}\exp[\phi_i \sum_j w_{ij}h_j - r_i\phi_i - a_i\phi_i^2 - b_i\phi_i^4]d\boldsymbol{\phi_i}} \tag{A17}$$

$$= \prod_i p(\phi_i|h), \tag{A18}$$

Similarly:

$$p(h|\phi;\theta) = \frac{p(\phi,h;\theta)}{p(\phi;\theta)} = \frac{\exp[-S(\phi,h;\theta)]}{\int_{\boldsymbol{h}}\exp[-S(\phi,\boldsymbol{h};\theta)]d\boldsymbol{h}} = \prod_j p(h_j|\phi). \tag{A19}$$

The gradient of the log-likelihood for the case of the quantum field-theoretic neural network is:

$$\frac{\partial \ln p(\phi;\theta)}{\partial\theta} = \frac{\partial}{\partial\theta}\left[\ln\frac{\int_{\boldsymbol{h}}\exp[-S(\phi,\boldsymbol{h};\theta)]d\boldsymbol{h}}{\int_{\boldsymbol{\phi},\boldsymbol{h}}\exp[-S(\boldsymbol{\phi},\boldsymbol{h};\theta)]d\boldsymbol{\phi}d\boldsymbol{h}}\right] \tag{A20}$$

$$= \frac{\partial}{\partial\theta}\left[\ln\int_{\boldsymbol{h}}\exp[-S(\phi,\boldsymbol{h};\theta)]d\boldsymbol{h} - \ln\int_{\boldsymbol{\phi},\boldsymbol{h}}\exp[-S(\boldsymbol{\phi},\boldsymbol{h};\theta)]d\boldsymbol{\phi}d\boldsymbol{h}\right] \tag{A21}$$

$$= \frac{\int_{\boldsymbol{h}}\frac{\partial}{\partial\theta}(-S(\phi,\boldsymbol{h};\theta))\exp[-S(\phi,\boldsymbol{h};\theta)]d\boldsymbol{h}}{\int_{\boldsymbol{h}}\exp[-S(\phi,\boldsymbol{h};\theta)]d\boldsymbol{h}} - \frac{\int_{\boldsymbol{\phi},\boldsymbol{h}}\frac{\partial}{\partial\theta}(-S(\boldsymbol{\phi},\boldsymbol{h};\theta))\exp[-S(\boldsymbol{\phi},\boldsymbol{h};\theta)]d\boldsymbol{\phi}d\boldsymbol{h}}{\int_{\boldsymbol{\phi},\boldsymbol{h}}\exp[-S(\boldsymbol{\phi},\boldsymbol{h};\theta)]d\boldsymbol{\phi}d\boldsymbol{h}} \tag{A22}$$

$$= \int_{\boldsymbol{h}} p(\boldsymbol{h}|\phi;\theta)\frac{\partial}{\partial\theta}(-S(\phi,\boldsymbol{h};\theta))d\boldsymbol{h} - \int_{\boldsymbol{\phi},\boldsymbol{h}} p(\boldsymbol{\phi},\boldsymbol{h};\theta)\frac{\partial}{\partial\theta}(-S(\boldsymbol{\phi},\boldsymbol{h};\theta))d\boldsymbol{\phi}d\boldsymbol{h} \tag{A23}$$

$$= \left\langle\frac{\partial}{\partial\theta}(-S(\phi,h;\theta))\right\rangle_{p(h|\phi;\theta)} - \left\langle\frac{\partial}{\partial\theta}(-S(\phi,h;\theta))\right\rangle_{p(\phi,h;\theta)}. \tag{A24}$$

We approximate the last expression in the above equation for each parameter θ using contrastive divergence. Specifically, the visible units ϕ are set equal to a specific training example $\phi^{(x)}$ and then based on the conditional distribution $p(h|\phi^{(x)})$ a set of hidden units $h^{(x)}$ is sampled. The hidden units $h^{(x)}$ are then utilized to sample a new set of visible units $\phi^{(x+1)}$ and the approach is repeated for k steps:

$$CD_k = \left\langle\frac{\partial}{\partial\theta}(-S(\phi^{(0)},h;\theta))\right\rangle_{p(h|\phi^{(0)};\theta)} - \left\langle\frac{\partial}{\partial\theta}(-S(\phi^{(k)},h;\theta))\right\rangle_{p(h|\phi^{(k)},;\theta)}, \tag{A25}$$

where for the considered cases we use $k = 1$.

Appendix B: Simulation details and Hyper-Parameters

The ϕ^4 scalar field theory is a system with continuous degrees of freedom $-\infty < \phi < +\infty$. To sample the system we implement Markov chain Monte Carlo sampling with the Metropolis algorithm, where we consider one step as equivalent to updating a number of lattice sites equal to the volume of the system. The question of how to properly choose a new state additionally arises. When the training data have values which lie at a specific interval, the aim of the machine learning algorithm is to learn a probability distribution which reproduces them. The new state can then be chosen by sampling uniformly between the minimum and maximum value, therefore guaranteeing that every state is reachable under an arbitrary large number of Monte Carlo steps. For the case of the hidden units in the ϕ^4 neural network we impose the same restriction, even though the hidden units could, in principle, remain unconstrained, i.e. $-\infty < h < +\infty$. We also emphasize that during the gradient process of the Markov random field we retain one Markov chain to

conduct the necessary calculations.

The learning rate that produced Figs. 2 and 3 is 10^{-3} and 10^{-2}, respectively. The sample size is chosen equal to 50 before updating the variational parameters θ. The image in Fig. 9 has size $32 * 32$ and its continuous values lie between $[-1, 1]$. The Markov random field was trained with learning rate 0.1 and 4×10^4 epochs. The parameters that produced Fig. 8 are a learning rate of 0.1, 400 epochs and a batch size of 4. For the results depicted in Fig. 10 the ϕ^4 neural network has 4096 visible units, 32 hidden units, learning rate 0.1, batch size of 5 and was trained for 10^4 epochs on the first 40 examples of the Olivetti faces dataset.

Appendix C: Histogram Reweighting

We consider the numerical estimator for an arbitrary observable $\langle O\rangle$ in the full complex action $\mathcal{A}$ which we aim to sample:

$$\langle O\rangle = \frac{\sum_{l=1}^{N} O_l \tilde{p_l}^{-1} \exp[-g'_j \mathcal{A}_l^{(j)} - \sum_{k=1,k\neq j}^{5} g_k \mathcal{A}_l^{(k)}]}{\sum_{l=1}^{N} \tilde{p_l}^{-1} \exp[-g'_j \mathcal{A}_l^{(j)} - \sum_{k=1,k\neq j}^{5} g_k \mathcal{A}_l^{(k)}]}, \tag{C1}$$

where N is the subset of Monte Carlo samples and $\tilde{p}$ are the probabilities used to sample from the equilibrium distribution. We have expressed the numerical estimator in a form that simultaneously allows extrapolation along the trajectory of a coupling constant g'_j. We will now substitute the probabilities $\tilde{p}$ for the probabilities of the inhomogeneous ϕ^4 Markov random field:

$$\tilde{p_l} = \frac{\exp[-S_l]}{\int_{\boldsymbol{\phi}} \exp[-S_{\boldsymbol{\phi}}] d\boldsymbol{\phi}}, \tag{C2}$$

where the sum is over all possible states ϕ of the system and we arrive at the reweighting equation:

$$\langle O\rangle = \frac{\sum_{l=1}^{N} O_l \exp[S_l - g'_j \mathcal{A}_l^{(j)} - \sum_{k=1,k\neq j}^{5} g_k \mathcal{A}_l^{(k)}]}{\sum_{l=1}^{N} \exp[S_l - g'_j \mathcal{A}_l^{(j)} - \sum_{k=1,k\neq j}^{5} g_k \mathcal{A}_l^{(k)}]}. \tag{C3}$$

Given a subset of samples drawn from the equilibrium distribution of the ϕ^4 Markov random field, which is described by the action S, one can extrapolate observables to the full distribution of the action $\mathcal{A}$ which includes longer range interactions and complex-valued terms along the trajectory of a coupling constant g'_j.

To compare the reweighting extrapolations from the ϕ^4 Markov random field to the full action, we additionally implement reweighting from the simulated action $A_{\{4\}}$. In this form of reweighting we consider again Eq. C1 and we substitute $\tilde{p}$ for:

$$\tilde{p_l} = \frac{\exp[-\sum_{k=1}^{4} g_k \mathcal{A}_l^{(k)}]}{\int_{\boldsymbol{\phi}} \exp[-\sum_{k=1}^{4} g_k \mathcal{A}_{\boldsymbol{\phi}}^{(k)}] d\boldsymbol{\phi}}, \tag{C4}$$

where we consider for this specific case that $g'_j = g_j$, arriving at equation:

$$\langle O\rangle = \frac{\sum_{l=1}^{N} O_l \exp[-\Im A_l]}{\sum_{l=1}^{N} \exp[-\Im A_l]}. \tag{C5}$$

One observable of interest is the magnetization which is defined as:

$$m = \frac{1}{V}\left|\sum_i \phi_i\right|, \tag{C6}$$

where $V = L * L$ is the volume of the system.

Appendix D: Binning Analysis

Statistical errors are calculated with the binning method on the obtained Monte Carlo datasets. Each dataset with 10^4 minimally correlated configurations is split into $n = 10$ datasets where calculations of observables O are conducted. The standard deviation for an observable O is then obtained through:

$$\sigma_O = \sqrt{\frac{1}{n-1}(\overline{O^2} - \overline{O}^2)}. \tag{D1}$$

[1] J. Zinn-Justin, *Quantum Field Theory and Critical Phenomena* (Oxford University Press, Oxford, 2002).

[2] J. Glimm and A. Jaffe, *Quantum Physics: A Functional Integral Point of View* (Springer, New York, NY, 1987).

[3] G. Velo and A. Wightman, *Constructive Quantum Field Theory: The 1973 "Ettore Majorana" International School of Mathematical Physics*, Lecture Notes in Physics (Springer Berlin Heidelberg, 1973).

[4] E. Seiler, *Gauge Theories as a Problem of Constructive Quantum Field Theory and Statistical Mechanics* (Springer, Berlin, Heidelberg, 1982).

[5] E. Nelson, Probability theory and euclidean field theory, in *Constructive Quantum Field Theory*, edited by G. Velo and A. Wightman (Springer Berlin Heidelberg, Berlin, Heidelberg, 1973) pp. 94–124.

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

[7] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, USA, 2016) `http://www.deeplearningbook.org`.

[8] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine

learning and the physical sciences, Reviews of Modern Physics **91**, 10.1103/revmodphys.91.045002 (2019).
[9] G. Kanwar, M. S. Albergo, D. Boyda, K. Cranmer, D. C. Hackett, S. Racanière, D. J. Rezende, and P. E. Shanahan, Equivariant flow-based sampling for lattice gauge theory, Phys. Rev. Lett. **125**, 121601 (2020).
[10] P. E. Shanahan, D. Trewartha, and W. Detmold, Machine learning action parameters in lattice quantum chromodynamics, Phys. Rev. D **97**, 094506 (2018).
[11] K. Zhou, G. Endrődi, L.-G. Pang, and H. Stöcker, Regressive and generative neural networks for scalar field theory, Phys. Rev. D **100**, 011501 (2019).
[12] D. Bachtis, G. Aarts, and B. Lucini, Mapping distinct phase transitions to a neural network, Phys. Rev. E **102**, 053306 (2020).
[13] M. N. Chernodub, H. Erbin, V. A. Goy, and A. V. Molochkov, Topological defects and confinement with machine learning: The case of monopoles in compact electrodynamics, Phys. Rev. D **102**, 054501 (2020).
[14] S. Blücher, L. Kades, J. M. Pawlowski, N. Strodthoff, and J. M. Urban, Towards novel insights in lattice field theory with explainable machine learning, Phys. Rev. D **101**, 094507 (2020).
[15] M. Favoni, A. Ipp, D. I. Müller, and D. Schuh, Lattice gauge equivariant convolutional neural networks (2020), arXiv:2012.12901 [hep-lat].
[16] K. A. Nicoli, C. J. Anders, L. Funcke, T. Hartung, K. Jansen, P. Kessel, S. Nakajima, and P. Stornati, Estimation of thermodynamic observables in lattice field theories with deep generative models (2021), arXiv:2007.07115 [hep-lat].
[17] E. L. van Nieuwenburg, Y.-H. Liu, and S. Huber, Learning phase transitions by confusion, Nature Physics **13**, 435 (2017).
[18] J. Carrasquilla and R. G. Melko, Machine learning phases of matter, Nature Physics **13**, 431 (2017).
[19] C. Giannetti, B. Lucini, and D. Vadacchino, Machine learning as a universal tool for quantitative investigations of phase transitions, Nuclear Physics B **944**, 114639 (2019).
[20] D. Bachtis, G. Aarts, and B. Lucini, Adding machine learning within hamiltonians: Renormalization group transformations, symmetry breaking and restoration, Phys. Rev. Research **3**, 013134 (2021).
[21] D. Bachtis, G. Aarts, and B. Lucini, Extending machine learning classification capabilities with histogram reweighting, Phys. Rev. E **102**, 033303 (2020).
[22] L. Wang, Discovering phase transitions with unsupervised learning, Phys. Rev. B **94**, 195105 (2016).
[23] A. Tanaka and A. Tomiya, Detection of phase transition via convolutional neural networks, Journal of the Physical Society of Japan **86**, 063001 (2017), https://doi.org/10.7566/JPSJ.86.063001.
[24] E. Agliari, A. Barra, P. Sollich, and L. Zdeborová, Machine learning and statistical physics: preface, Journal of Physics A: Mathematical and Theoretical **53**, 500401 (2020).
[25] L. Zdeborova and F. Krzakala, Statistical physics of inference: thresholds and algorithms, Advances in Physics **65**, 453 (2016), https://doi.org/10.1080/00018732.2016.1211393.
[26] S. Goldt, M. S. Advani, A. M. Saxe, F. Krzakala, and L. Zdeborová, Dynamics of stochastic gradient descent for two-layer neural networks in the teacher–student setup, Journal of Statistical Mechanics: Theory and Experiment **2020**, 124010 (2020).
[27] D. Alberici, A. Barra, P. Contucci, and E. Mingione, Annealing and replica-symmetry in deep boltzmann machines, Journal of Statistical Physics **180**, 665 (2020).
[28] E. Agliari, D. Migliozzi, and D. Tantari, Non-convex multi-species hopfield models, J. Stat. Phys. **172** (2018).
[29] A. Barra, G. Genovese, P. Sollich, and D. Tantari, Phase transitions in restricted boltzmann machines with generic priors, Phys. Rev. E **96** (2017).
[30] A. Barra, G. Genovese, P. Sollich, and D. Tantari, Phase diagram of restricted boltzmann machines and generalized hopfield networks with arbitrary priors, Phys. Rev. E **97** (2018).
[31] M. Mézard, Mean-field message-passing equations in the hopfield model and its generalizations, Phys. Rev. E **95** (2017).
[32] A. Barra, A. Bernacchia, E. Santucci, and P. Contucci, On the equivalence of hopfield networks and boltzmann machines, Neural Netw. **34** (2012).
[33] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications* (World Scientific, Singapore, 1987).
[34] J. Halverson, A. Maiti, and K. Stoner, Neural Networks and Quantum Field Theory, (2020), arXiv:2008.08601 [cs.LG].
[35] J. Lee, Y. Bahri, R. Novak, S. Schoenholz, J. Pennington, and J. Sohl-dickstein, Deep neural networks as gaussian processes (6th International Conference on Learning Representations, Vancouver, BC, Canada, 2018).
[36] A. G. de G. Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani, Gaussian process behaviour in wide deep neural networks, in *International Conference on Learning Representations* (2018).
[37] R. Novak, L. Xiao, Y. Bahri, J. Lee, G. Yang, D. A. Abolafia, J. Pennington, and J. Sohl-dickstein, Bayesian deep convolutional networks with many channels are gaussian processes, in *International Conference on Learning Representations* (2019).
[38] A. Garriga-Alonso, C. E. Rasmussen, and L. Aitchison, Deep convolutional networks as shallow gaussian processes, in *International Conference on Learning Representations* (2019).
[39] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques* (The MIT Press, 2009).
[40] K. G. Wilson, Confinement of quarks, Phys. Rev. D **10**, 2445 (1974).
[41] C. J. Preston, *Gibbs States on Countable Sets*, Cambridge Tracts in Mathematics (Cambridge University Press, 1974).
[42] A. Milchev, D. W. Heermann, and K. Binder, Finite-size scaling analysis of the $\phi 4$ field theory on the square lattice, Journal of Statistical Physics **44**, 749 (1986).
[43] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag, Berlin, Heidelberg, 2006).
[44] A. M. Ferrenberg and R. H. Swendsen, New monte carlo technique for studying phase transitions, Phys. Rev. Lett. **61**, 2635 (1988).
[45] A. Blake, P. Kohli, and C. Rother, *Markov Random Fields for Vision and Image Processing* (The MIT Press, 2011).

[46] A. Krizhevsky, Learning multiple layers of features from tiny images (2009).

[47] P. Smolensky, Information processing in dynamical systems: Foundations of harmony theory, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations* (MIT Press, Cambridge, MA, USA, 1986) p. 194–281.

[48] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, A learning algorithm for boltzmann machines, Cognitive Science **9**, 147 (1985).

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

[50] G. E. Hinton, A practical guide to training restricted boltzmann machines, in *Neural Networks: Tricks of the Trade: Second Edition*, edited by G. Montavon, G. B. Orr, and K.-R. Müller (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012) pp. 599–619.

[51] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, Greedy layer-wise training of deep networks, in *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06 (MIT Press, Cambridge, MA, USA, 2006) p. 153–160.

[52] O. Krause, A. Fischer, T. Glasmachers, and C. Igel, Approximation properties of DBNs with binary hidden units and real-valued visible units, in *Proceedings of the 30th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 28, edited by S. Dasgupta and D. McAllester (PMLR, Atlanta, Georgia, USA, 2013) pp. 419–426.

[53] This dataset contains a set of face images taken between April 1992 and April 1994 at AT&T Laboratories Cambridge.

[54] G. E. Hinton and R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science **313**, 504 (2006).

[55] E. Marinari and G. Parisi, Simulated tempering: A new monte carlo scheme, Europhysics Letters (EPL) **19**, 451 (1992).

[56] J. Lee, New monte carlo algorithm: Entropic sampling, Phys. Rev. Lett. **71**, 211 (1993).

[57] R. C. Brower and P. Tamayo, Embedded dynamics for φ^4 theory, Phys. Rev. Lett. **62**, 1087 (1989).

[58] W. Loinaz and R. S. Willey, Monte carlo simulation calculation of the critical coupling constant for two-dimensional continuum φ^4 theory, Phys. Rev. D **58**, 076003 (1998).

[59] W. Bietenholz, R. Brower, S. Chandrasekharan, and U.-J. Wiese, Progress on perfect lattice actions for qcd, Nuclear Physics B - Proceedings Supplements **53**, 921 (1997), lattice 96.

[60] W. Bietenholz and U.-J. Wiese, Perfect actions with chemical potential, Physics Letters B **426**, 114 (1998).

[61] M. Jain and V. Vanchurin, Generating functionals for quantum field theories with random potentials, Journal of High Energy Physics **2016**, 107 (2016).