



## Neural Monte Carlo renormalization group

Jui-Hui Chung and Ying-Jer Kao 

*Center for Theoretical Physics and Department of Physics, National Taiwan University, Taipei 10607, Taiwan*

 (Received 6 November 2020; revised 7 December 2020; accepted 19 May 2021; published 22 June 2021)

The key idea behind the renormalization group (RG) transformation is that properties of physical systems with very different microscopic makeups can be characterized by a few universal parameters. However, finding a systematic way to construct RG transformation for particular systems remains difficult due to the many possible choices of the weight factors in the RG procedure. Here we show, by identifying the conditional distribution in the restricted Boltzmann machine and the weight factor distribution in the RG procedure, that a valid real-space RG transformation can be learned without prior knowledge of the physical system. This neural Monte Carlo RG algorithm allows for direct computation of the RG flow and critical exponents. Our results establish a solid connection between the RG transformation in physics and the deep architecture in machine learning, paving the way for further interdisciplinary research.

DOI: [10.1103/PhysRevResearch.3.023230](https://doi.org/10.1103/PhysRevResearch.3.023230)

### I. INTRODUCTION

The renormalization group (RG) [1] formalism provides a systematic method for quantitative analysis of critical phenomena. Among all the RG schemes, the real-space renormalization group (RSRG), first proposed by Kadanoff [2], is the most intuitive and natural way to perform RG transformations on lattice models [3]. These methods allow for a straightforward construction of the critical surface and calculation of the critical exponents using numerical methods such as the Monte Carlo renormalization group (MCRG) [4–6]. However, the RSRG transformation typically generates long-range couplings not present in the original Hamiltonian and truncation is necessary to make the method manageable. From the physical point of view, we expect the range of the renormalized interactions of a physical lattice system near the fixed point to be short. Finding a systematic way to construct a coarse-graining scheme for particular Hamiltonians is crucial for the success of any RSRG scheme. The fundamental difficulty lies in the enormous degrees of freedom in choosing the weight factors for the RG transformation. Several attempts in the past have been made to find the optimal transformation. Swendsen proposed an optimal MCRG scheme by introducing variational parameters into the RG procedure [7]. Blöte *et al.* proposed the modification of the Hamiltonian and the weight factors such that the corrections to scaling would be small [8]. Ron *et al.* proposed a choice of parameters such that the critical exponent of interest was nearly constant during the MCRG iterations [9]. However, it remains unclear how to determine the weight factors without prior knowledge of the system.

A general guideline in searching for a RG transformation is to identify important degrees of freedom in the RG flow. However, it is difficult *a priori* to determine which degrees of freedom should be retained. This resembles the question in machine learning (ML) on how to extract relevant features from raw data. Deep learning [10] using deep neural networks (DNNs) has significantly improved the machine's ability in many areas such as speech recognition [11], object recognition [12], and Go and video game playing [13–15], as well as aided discoveries in various fields of physics [16–20]. Multiple layers of representation are used to learn distinct features directly from the training data. The similarity between the structure of the DNN and the course-graining schemes in statistical physics has inspired many efforts to establish connection between variational RG [21] and unsupervised learning of DNNs [22–30]. Here we address a different question: How can we train a DNN to obtain a good RSRG transformation? This issue is partially addressed from an information-theoretic perspective [25,26], where an optimal RG transformation is obtained by maximizing the real-space mutual information (RSMI). However, the proposed RSMI algorithm requires a mutual information proxy to probe the effective temperature (coupling) of the system along the RG flow, rendering it less practical. A more direct and transparent method that enables direct computation of the corresponding RG flow and critical exponents is thus highly coveted.

Here we present a scheme called neural Monte Carlo RG that parametrizes the RG transformation in terms of a restricted Boltzmann machine (RBM) [31]. The RG transformation can be learned by minimizing the Kullback-Leibler (KL) divergence between the system distribution and the marginal weight factor distribution [defined in Eq. (5)]. This provides an explicit link between the RG transformation and the RBM, allowing us to use the modern ML techniques to find an RG transformation. In addition, the scheme is readily integrated with the MCRG techniques to directly determine the effective couplings along the RG flow and critical

---

*Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.*

exponents. We demonstrate the accuracy of this approach for the two- and three-dimensional classical Ising models.

## II. PARAMETRIZATION OF THE REAL-SPACE RENORMALIZATION GROUP

Consider a generic lattice Hamiltonian

$$H(\sigma) = \sum_{\alpha} K_{\alpha} S_{\alpha}(\sigma), \quad (1)$$

where the interactions  $S_{\alpha}$  are combinations of the original spins  $\sigma$  and the  $K_{\alpha}$  are the corresponding coupling constants where the usual Boltzmann factor  $-\beta = -1/k_B T$  is conventionally absorbed [3]. A general RG transformation [3,26] can be written as

$$e^{H'(\mu)} = \sum_{\sigma} P(\mu|\sigma) e^{H(\sigma)}, \quad (2)$$

with weight factor  $P(\mu|\sigma)$ , where  $\mu = \pm 1$  correspond to the renormalized spins in the renormalized Hamiltonian  $H'(\mu) = \sum_{\alpha} K'_{\alpha} S_{\alpha}(\mu)$  with renormalized couplings  $K'_{\alpha}$ . We parametrize the weight factors as

$$P(\mu|\sigma) = \frac{1}{\sum_{\mu} \exp(\sum_{ij} W_{ij} \sigma_i \mu_j)} \exp\left(\sum_{ij} W_{ij} \sigma_i \mu_j\right), \quad (3)$$

where  $W_{ij}$  are variational parameters to be optimized [see Fig. 1(a)]. In particular, if  $W_{ij}$  are infinite in a local block of spins and zero everywhere else, then we recover the majority-rule transformation [4]. Importantly, this parametrization satisfies the so-called trace condition

$$\sum_{\mu} P(\mu|\sigma) = 1, \quad (4)$$

which is required to correctly reproduce thermodynamics [3,24,26]. To make a connection to the RBM in the following discussion, we define the weight factor distribution as

$$P(\sigma, \mu) = \frac{1}{Z} \exp\left(\sum_{ij} W_{ij} \sigma_i \mu_j\right), \quad (5)$$

where  $Z = \sum_{\sigma, \mu} \exp(\sum_{ij} W_{ij} \sigma_i \mu_j)$ . The weight factor (3) is then simply the condition distribution of the weight factor distribution, that is, we have  $P(\mu|\sigma) = P(\sigma, \mu) / \sum_{\mu} P(\sigma, \mu)$ .

An RBM is a generative model that is a staple deep learning tool to solve tasks that involve unsupervised learning [32,33]. Hidden layers of an RBM can extract meaningful features from the data [34]. In this regard, an RBM with fewer hidden variables than the visible variables resembles coarse graining in the RG, first pointed out by Mehta and Schwab [22]. However, their proposed mapping from the variational RG procedure to unsupervised training of a DNN does not satisfy the trace condition (4) and thus does not constitute a proper RG (see Appendix B for a detailed comparison). Here we propose a direct mapping between the RBM and the weight factors such that Eq. (4) is naturally satisfied.

An RBM can be written in terms of weights  $W_{ij}$ , hidden variables  $h_j$ , and visible variables  $v_i$  as

$$Q(v, h) = \frac{1}{Z_{\text{RBM}}} \exp\left(\sum_{ij} W_{ij} v_i h_j\right), \quad (6)$$

where  $Z_{\text{RBM}} = \sum_{v, h} \exp(\sum_{ij} W_{ij} v_i h_j)$ . The empirical feature distribution  $\hat{p}'(h)$  can be extracted from the empirical distribution  $\hat{p}(v)$  through

$$\hat{p}'(h) = \sum_v Q(h|v) \hat{p}(v), \quad (7)$$

where  $Q(h|v) = Q(v, h) / \sum_h Q(v, h)$  is the conditional distribution of the hidden variables, given the values of the visible variables [32]. The parameters for the RBM are chosen by minimizing the KL divergence between the empirical distribution  $\hat{p}(v)$  and the marginal distribution  $\sum_h Q(v, h)$ ,

$$D_{\text{KL}}\left(\hat{p}(v) \left\| \sum_h Q(v, h)\right.\right), \quad (8)$$

where  $D(p||q) = \sum_{\sigma} p(\sigma) \log[p(\sigma)/q(\sigma)]$  for two discrete distributions  $p(\sigma)$  and  $q(\sigma)$ .

Motivated by the similarity between Eqs. (2) and (7), we identify the conditional distribution  $Q(h|v)$  in the RBM with our parametrized weight factor  $P(\mu|\sigma)$  and associate the hidden and visible variables in the RBM with the renormalized and original spins, respectively. In analogy to the optimization scheme of an RBM, we propose an optimal choice of the parameters in the weight factors by minimizing the KL divergence between the system distribution and the marginal

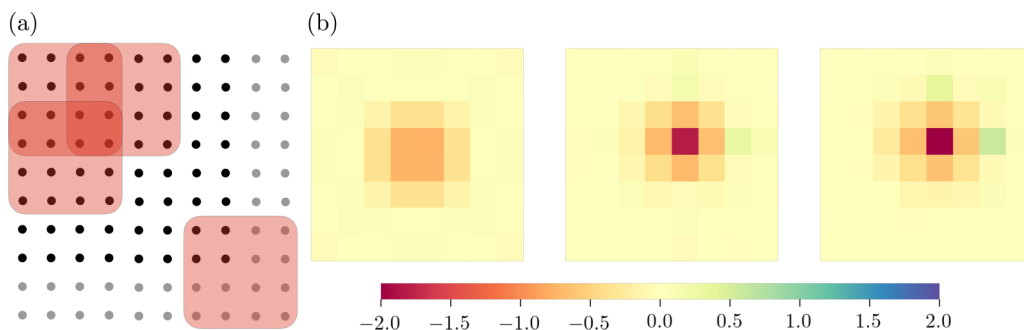


FIG. 1. RG transformation. (a) RG transformation of original spins (black dots) using overlapping parametrized weight factors (red square) as in Eq. (3). The opaque black dots are the periodic copies of the original spins. In this illustration,  $8^2$  spins are renormalized to  $4^2$  spins, leading to a scale factor equal to 2. (b) The  $8^2$  filters are learned on a  $32^2$  Ising model at critical NN coupling  $K_1 \simeq 0.4407$ . From left to right, we show the development of the filters at the 10th, 30th, and 50th epochs corresponding to Fig. 2(a).

weight factor distribution

$$D_{\text{KL}}\left(\frac{1}{Z}e^{H(\sigma)}\left\|\sum_{\mu}P(\sigma,\mu)\right.\right), \quad (9)$$

which can be carried out using standard ML techniques.

### III. STOCHASTIC OPTIMIZATION FOR THE OPTIMAL CRITERION

The optimization problem is solved by the stochastic gradient descent, where the parameters are updated through decrementing them in the direction of the gradient of the KL divergence. We replace the system distribution  $e^{H(\sigma)}/Z$  by its empirical distribution  $\hat{p}(\sigma)$  over Monte Carlo samples drawn from the Wolff algorithm [35] and write the KL divergence (9) as an expectation value over the empirical distribution

$$D_{\text{KL}}\left(\hat{p}(\sigma)\left\|\sum_{\mu}P(\sigma,\mu)\right.\right). \quad (10)$$

The gradient  $G_{ij}$  of the KL divergence (10) with respect to  $W_{ij}$  can be derived as

$$G_{ij} = \sum_{\sigma} \hat{p}(\sigma) \partial_{W_{ij}} F(\sigma) - \sum_{\sigma} P(\sigma) \partial_{W_{ij}} F(\sigma), \quad (11)$$

where  $F(\sigma)$  is the free energy defined as  $F(\sigma) = -\log \sum_{\mu} \exp(\sum_{ij} W_{ij} \sigma_i \mu_j)$ . The first term in Eq. (11) is simply a sample average of the derivative of the free energy and can be readily computed. The second term is approximated using the contrastive divergence algorithm [36] ( $\text{CD}_k$ ), where the expectation value is calculated from samples drawn from a Markov chain initialized with data distribution and implemented by Gibbs sampling with  $k$  Markov steps.

We update the weights in the direction of negative gradients

$$W_{ij}^{(k+1)} = W_{ij}^{(k)} - G_{ij}^{(k)}, \quad (12)$$

where the superscript of the weight  $W^{(k)}$  indicates the number of training epochs the weight has descended. We initialize  $W^{(0)}$  randomly around zero. Along the gradient descent we obtain a sequence of weight factors, which can be used to compute critical exponents and renormalized couplings, to see what feature distribution [ $\hat{p}(h)$  in Eq. (7)] the RBM is trying to learn. For translationally invariant systems, translationally

invariant parametrization of the weight factor distribution (5) can be achieved via convolution [37].

### IV. TWO-DIMENSIONAL ISING MODEL

To validate our scheme, we first consider the two-dimensional (2D) ferromagnetic Ising model

$$H(\sigma) = K_1 S_{\text{NN}} = K_1 \sum_{\langle ij \rangle} \sigma_i \sigma_j, \quad K_1 > 0, \quad (13)$$

where  $\sigma_i = \pm 1$ ,  $K_1$  is the nearest-neighbor coupling, and  $S_{\text{NN}}$  denotes the collection of nearest-neighbor interspin interactions. In the following, we consider a 2D lattice of size  $32^2$  with the periodic boundary condition. We analyze the learned weight factors' ability to remove long-range interactions by directly calculating the renormalized couplings and extract critical exponents [38]. The number of epochs required for the parameters to reach convergence is on the order of  $10^1$  for a training sample size of  $10^4$  and batch size of  $10^1$ , which typically takes from seconds to several minutes on a workstation with a single GPU.

Figure 1 shows the weight factors along the optimization process [at 10th, 30th, and 50th epochs corresponding to Fig. 2(a)] learned with a translationally invariant filter of size  $8^2$ . The filters are initialized uniformly around zero. Localized features emerge after a few epochs of training and progressively aggregate toward the center, in agreement with the conventional wisdom that renormalized and original spins close to one another should couple more strongly than those farther apart [39]. On the other hand, the RBM also picks up nonlocal correlations between the renormalized and original spins, where the interaction strength falls off exponentially with distance.

We proceed to investigate the effect of the criterion of minimizing KL divergence to see what the machine is trying to learn. In Fig. 2(a) we show the thermal critical exponents calculated from weight factors  $W^{(k)}$  along the optimization flow. At the beginning of the training, the partially optimized weight gives a poor estimate of the thermal critical exponent at the first iteration of RG transformation. After the 30th epoch, the value grows rapidly and converges to the exact value. In Figs. 2(b) and 2(c) we use the weights obtained at each training epoch to calculate the renormalized coupling

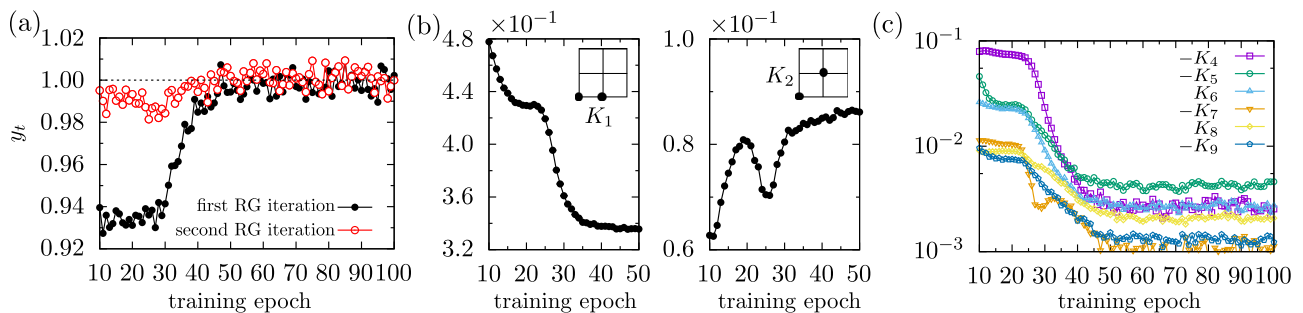


FIG. 2. Evolution of the critical exponents and coupling parameters during training. (a) Thermal critical exponent calculated from the weights obtained along the learning process. (b) Short-range renormalized coupling parameters, the nearest neighbor  $K_1$  and next-nearest neighbor  $K_2$ , as a function of the training epoch. Insets indicate the corresponding couplings in real space. (c) Longer-range renormalized coupling parameters (see Appendix A for details).

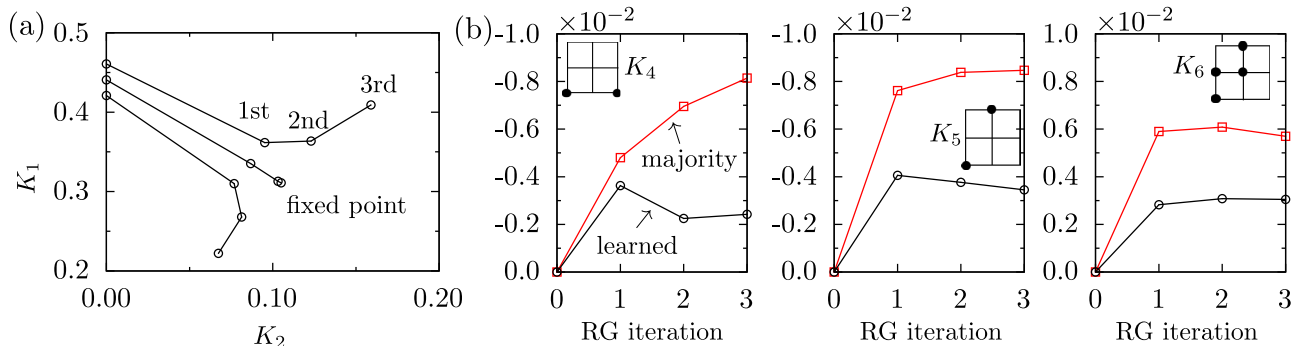


FIG. 3. Renormalization group flow. (a) Flow of nearest-  $K_1$  and next-nearest-neighbor  $K_2$  coupling parameters calculated from the converged weights along the renormalized group flow. The trajectory starts at critical couplings ( $K_1 \simeq 0.4407$  and  $K_2 = 0$ ) and flows to the renormalized couplings at the first, second, and third RG iterations. (b) Flow of long-range coupling parameters along the renormalized group trajectory for majority-rule transformation and learned transformation. Insets indicate the corresponding couplings in real space.

parameters along the training trajectory. The renormalized couplings, in machine-learning terms, completely describe the energy model underlying the empirical feature distribution [see Eq. (7)] extracted by the machine for the Ising empirical distribution. In Fig. 2(b) we see that the interactions are dominated by nearest- ( $K_1$ ) and next-nearest- ( $K_2$ ) neighbor couplings. The values for the longer-range interactions flow progressively towards zero as shown in Fig. 2(c). The trend shows that our criterion aims to remove longer-range coupling parameters in the renormalized Hamiltonian.

Figure 3(a) shows the RG flow diagram projected on the short-range coupling parameters’ subspace for the learned weight factors. The RG trajectory starting from the nearest-neighbor critical point flows rapidly to a fixed point. Slightly away from the critical point, the coupling parameters flow away to the infinite- (zero-) temperature trivial fixed points. Figures 3(b) and 3(c) show the renormalized coupling parameters along the RG flow. The coupling parameters coarse grained with the learned weight factors reach  $K_1 = 0.3109(3)$ ,  $K_2 = 0.1051(2)$ , and  $K_3 = -0.0184(2)$  at the third RG iteration. The values for longer-range interactions are greatly suppressed compared to those obtained by the majority-rule transformation.

Table I shows the critical exponents of the 2D Ising model computed using both the RBM and majority-rule transformations. Surprisingly, although the weights are learned without any prior knowledge of the model, the exponent is very close to the exact value at the first iteration of the renormalization transformation giving  $y_t = 1.000(2)$  for filter sizes  $8^2$  and  $16^2$ , consistent with the exact value within the statistical

error. Equally surprising is that the RBM trained on such a small amount of training data with only  $10^4$  samples can generalize well. In contrast, the majority-rule transformation gives  $y_t = 0.975(3)$  at the first RG iteration. Table I also shows that  $2^2$  and  $4^2$  filters are inadequate to produce exact critical exponents in the first iteration of the renormalization transformation. Even though the convergence for the thermal critical exponents looks extremely good, the scheme overestimates the magnetic critical exponents in the first RG iteration. The discrepancy in the magnetic exponents was also noted previously [7,40], indicating a separate filter is needed.

The weight factors considered in the literature are mostly short range [41] (decimation and majority transformation), i.e., they only couple one renormalized spin to a few original spins in the immediate vicinity. However, despite the seeming locality, these weight factors generally lead to an infinite proliferation of interactions upon renormalizing. With our proposed criterion, the learned weight factors contain nonlocal terms that work as counter terms, making the renormalization transformation more local; therefore, only a few short-range interactions are produced during the RG transformation. We note that the strategy along this line of transferring the complexity in renormalized Hamiltonian to the weight factors has yielded the first exactly soluble RG transformation [39].

Next we consider the antiferromagnetic Ising model on a square lattice with nearest-neighbor interactions

$$H(\sigma) = K_1 \sum_{\langle i,j \rangle} \sigma_i \sigma_j, \quad K_1 < 0. \quad (14)$$

TABLE I. Thermal and magnetic critical exponents of the 2D ferromagnetic Ising model. Results are obtained on a  $32^2$  lattice using the learned weight factors and the majority-rule transformation. Here  $N_r$  is the number of RG iterations. Seven (four) coupling terms are used for even (odd) interactions. The exact values are  $y_t = 1$  and  $y_h = 1.875$ .

Critical exponent	$N_r$	Majority	Filter size			
			$2^2$	$4^2$	$8^2$	$16^2$
$y_t$	1	0.975(3)	0.974(1)	0.975(2)	1.000(2)	1.000(2)
$y_t$	2	1.000(3)	1.000(1)	1.000(3)	1.000(1)	1.000(2)
$y_h$	1	1.8804(2)	1.8845(1)	1.8887(2)	1.8941(5)	1.8917(1)
$y_h$	2	1.8758(3)	1.8771(1)	1.8801(1)	1.8827(3)	1.8810(2)



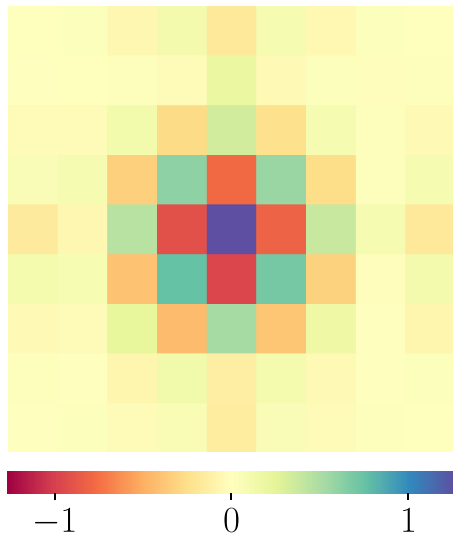


FIG. 4. Learned filter for the antiferromagnetic model. The values of the learned filters are staggered on sublattices. This pattern is learned without any prior knowledge of the system.

This model can be mapped to the ferromagnetic Ising model through a sublattice rotation, so the even (thermal) critical exponent is expected to be the same as for the ferromagnetic Ising model. Due to the sublattice structure, the conventional majority-rule transformation with scale factor  $b = 2$  or  $3$  fails to generate a proper renormalization group flow for this model and hence cannot be used to extract critical exponents. To circumvent this difficulty, Swendsen adopted an RG transformation introduced by van Leeuwen [42] with a scale factor  $b = \sqrt{5}$  such that all sites in a block are on the same sublattice [43]. Here we train a  $9^2$  filter on a  $54^2$  lattice without any restrictions. Figure 4 shows that the filter learns the proper staggering structure on different sublattices, which leads to an improved thermal critical exponent at the first RG iteration (Table II).

### V. THREE-DIMENSIONAL ISING MODEL

The scheme can be easily generalized to higher dimensions as long as we can train an RBM to represent an RG transformation. Table III shows the thermal critical exponents computed using learned filters starting at a system size of  $64^3$ . The filters at the first and second RG iterations are learned. The filters for the subsequent RG iterations use the same filter obtained in the second iteration. We compare the results with the values obtained from the majority rule [44]. Only the first

TABLE II. Thermal critical exponents for the antiferromagnetic model. The learned filter size is  $9^2$  and the scale factor for the coarse graining is  $b = 3$ . The van Leeuwen coarse-graining rule [42] has a scale factor  $b = \sqrt{5}$ . The result using van Leeuwen's coarse-graining rule is taken from [4].

$N_r$	van Leeuwen's	Learned
1	0.883	0.953(1)
2	0.997	0.997(2)

TABLE III. Thermal and magnetic critical exponents of the 3D Ising model. Results are obtained on a  $64^3$  lattice using the learned weight factors and the majority-rule transformation. Here  $N_r$  is the number of RG iterations. The first 20 coupling terms from [44] are used for even and odd interactions. The accepted values are  $y_t \simeq 1.587$  and  $y_h \simeq 2.482$  [45].

Critical exponent	$N_r$	Majority [44]	Filter size		
			$2^3$	$4^3$	$8^3$
$y_t$	1	1.425(3)	1.531(6)	1.300(4)	1.318(2)
$y_t$	2	1.509(2)	1.568(2)	1.521(2)	1.548(1)
$y_t$	3	1.547(2)	1.579(2)	1.556(4)	1.566(1)
$y_t$	4	1.563(9)	1.587(3)	1.558(6)	1.555(2)
$y_h$	1	2.4578(5)	2.515(1)	2.377(1)	2.3819(5)
$y_h$	2	2.4603(2)	2.4940(2)	2.4670(2)	2.4916(1)
$y_h$	3	2.4721(4)	2.4875(3)	2.4770(2)	2.4854(3)
$y_h$	4	2.476(1)	2.4850(8)	2.4815(1)	2.4845(8)

20 couplings out of the total 53 couplings in Ref. [44] are used. The  $2^3$  learned filter gives the exponent closest to the best estimate from the finite-size scaling result  $y_t = 1.587$  [45]. We find that the eight values of the  $2^3$  learned filter are homogeneous, with an average value of  $0.5254(2)$  for the first-iteration filter, which is very close to the tuned optimal choice of  $0.4314$  in Ref. [9]. The weight value at the second iteration is  $0.5057(9)$ .

Table IV shows the thermal critical exponents  $y_t$  starting at various system sizes using the  $8^2$  learned filter. Comparing the results for lattice sizes  $64^3$  and  $32^3$ , we find that up to the second RG iteration all results for  $y_t$  agree within the statistical errors. We conclude that the finite-size effect is significant when coarse graining  $8^3$  lattices down to  $4^3$ . We perform the finite-size correction outlined in Refs. [8,44] and the estimate of  $y_t$  becomes  $1.5681(34)$  at the third RG iteration and  $1.5771(50)$  at the fourth RG iteration. Linear extrapolation of the finite-size corrected results to the infinite coarse-graining level yields  $y_t = 1.5860(109)$ .

### VI. REAL-SPACE MUTUAL INFORMATION

The RSMI measures the information that the knowledge of the environmental degrees of freedom  $\mathcal{E}$  gives about the relevant degrees of freedom  $\mathcal{H}$  and is defined as

$$I(\mathcal{H}; \mathcal{E}) = \sum_{\mathcal{H}, \mathcal{E}} P(\mathcal{H}, \mathcal{E}) \log \left( \frac{P(\mathcal{H}, \mathcal{E})}{P(\mathcal{H})P(\mathcal{E})} \right). \quad (15)$$

If  $\mathcal{E}$  completely determines  $\mathcal{H}$ , then the information gained is maximized and  $I(\mathcal{H}; \mathcal{E})$  reduces to the self-information (the entropy) of the relevant degrees of freedom  $\mathcal{H}$ , which itself is upper bounded by the logarithm of all possible configurations of  $\mathcal{H}$ . The RSMI scheme argues that an optimal RG transformation can be obtained by maximizing the RSMI [25,26].

Adopting the definition in Refs. [25,26], we consider a system described by a quadripartite distribution  $P(\mathcal{V}, \mathcal{E}, \mathcal{H}, \mathcal{O})$  [Fig. 5(a)]. We define the RSMI of the system as  $I(\mathcal{H}; \mathcal{E})$ , i.e., the mutual information between hidden and environment

TABLE IV. Finite-size corrections for the 3D Ising model. The filter size is  $8^3$ . The second to fifth columns show the thermal critical exponents for different lattice sizes. In the sixth and seventh columns we give estimates of finite-size corrections when coarse graining the  $16^3$  lattice to  $8^3$  and the  $8^3$  lattice to  $4^3$ , respectively [8,44]. In the last column we give estimates of finite-size corrected exponents and we linear extrapolated these two numbers to infinite RG iterations.

$N_r$	Lattice size				Finite-size corrections		
	$64^3$	$32^3$	$16^3$	$8^3$	$16^3 \rightarrow 8^3$	$8^3 \rightarrow 4^3$	$64^3$ (corrected)
1	1.3182(15)	1.3216(17)	1.3212(14)	1.2953(26)	-0.0030(21)	0.0230(30)	
2	1.5475(9)	1.5476(13)	1.5244(17)		-0.0010(15)	0.0230(20)	
3	1.5661(12)	1.5433(22)			0.0021(32)	0.0227(25)	1.5681(34)
4	1.5546(21)					0.0225(46)	1.5771(50)
$\infty$							1.5860(109)

random variables. The relevant distributions needed to compute  $I(\mathcal{H}; \mathcal{E})$  are appropriate marginals of  $P(\mathcal{V}, \mathcal{E}, \mathcal{H}, \mathcal{O})$ .

Here we consider a  $4^2$  Ising model with the periodic boundary condition where RSMI can be computed exactly. We train a  $3^2$  filter on the system to obtain a learned weight factor distribution. Figure 5(b) shows the partition of the lattice into visible (orange), environmental (green), hidden (top left red square), and other (top right, bottom left, and bottom right red squares) random variables. Figure 5(c) shows the evolution of RSMI during training. Random initialization of the filters gives zero RSMI, and as the training progresses, the RSMI saturates to the upper bound  $\ln 2 \simeq 0.693$ .

Similar behavior also appears in the two-dimensional antiferromagnetic ( $3^2$  lattice with  $3^2$  filter size) and three-dimensional Ising model ( $2^3$  lattice with  $2^3$  filter size),

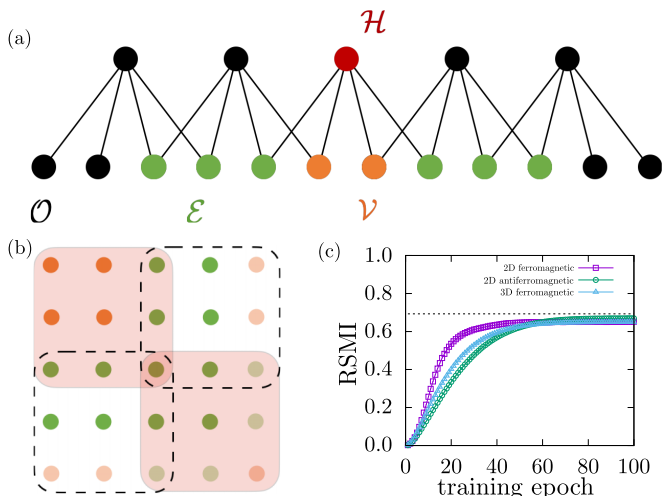


FIG. 5. Real-space mutual information for the 2D ferromagnetic Ising model. (a) Schematic decomposition of a system described by a quadripartite distribution  $P(\mathcal{V}, \mathcal{E}, \mathcal{H}, \mathcal{O})$  over visible, environmental, hidden, and other random variables. (b) Decomposition for 2D ferromagnetic Ising model. The  $3^2$  squares represent the hidden variables which connect to the overlapping visible variables. The opaque dots are the periodic copies of the visible variables. The system is partitioned into visible (orange), environmental (green), hidden (top left red square), and other (top right, bottom left, and bottom right red squares) random variables. (c) RSMI as a function of training epochs for 2D ferromagnetic, 2D antiferromagnetic, and 3D ferromagnetic Ising models. As the training progresses, the RSMI saturates to the upper bound  $\ln 2 \simeq 0.693$ .

where the RSMI increases and plateaus at the maximum possible value. Although the RG transformation and the RSMI values may be subject to strong finite-size effects due to the small sizes we study, the results suggest a possible connection between the RSMI scheme and neural MCRG; more theoretical and numerical studies are necessary to establish the connection.

## VII. CONCLUSION

We have demonstrated a scheme based on an RBM that is capable of learning an RG transformation from Monte Carlo samples. The similarity between the standard RBM and the weight factors means that we can take advantage of the progress in the ML architectures and techniques to parametrize and train the filters for the RG. This algorithm is flexible and can be directly applied to disordered systems [46]. Although we focus on the RBM with binary variables, for models with continuous variables such as the Heisenberg model, one can use Gaussian-Bernoulli RBMs to better model the RG transformation [47]. Generalization of the present scheme to quantum systems should be straightforward by a quantum-to-classical mapping of the  $d$ -dimensional quantum system to the  $(d + 1)$ -dimensional classical system [48]. On the other hand, how to extend the present scheme to study models with emergent degrees of freedom such as dimer and ice-type models remains an open question which requires further study.

The code that generates data used in this paper is available from [49].

## ACKNOWLEDGMENTS

The authors thank Yantao Wu for fruitful discussions. This work was supported by Ministry of Science and Technology of Taiwan under Grants No. 108-2112-M-002-020-MY3 and No. 107-2112-M-002-016-MY3 and partly supported by National Center of Theoretical Science of Taiwan. We are grateful to the National Center for High-Performance Computing for computer time and facilities.

## APPENDIX A: MONTE CARLO RENORMALIZATION GROUP

Here we summarize the MCRG method used to calculate the critical exponents and renormalized coupling parameters

from Monte Carlo samples for a given filter [38]. To determine the critical exponents, we need to calculate the derivatives of the transformation

$$T_{\alpha\beta}^{(n+1)} \equiv \frac{\partial K_{\alpha}^{(n+1)}}{\partial K_{\beta}^{(n)}}, \quad (\text{A1})$$

which is given by the solution of the linear equation [4]

$$\frac{\partial \langle S_{\gamma}^{(n+1)} \rangle}{\partial K_{\beta}^{(n)}} = \sum_{\alpha} \frac{\partial \langle S_{\gamma}^{(n+1)} \rangle}{\partial K_{\alpha}^{(n+1)}} \frac{\partial K_{\alpha}^{(n+1)}}{\partial K_{\beta}^{(n)}}. \quad (\text{A2})$$

Here  $\langle S_{\gamma}^{(n)} \rangle$  is the expectation of the spin combinations at the  $n$ th RG iterations. The derivatives of these expectation values of the spin combinations are obtained from the correlation functions

$$\frac{\partial \langle S_{\gamma}^{(n+1)} \rangle}{\partial K_{\beta}^{(n)}} = \langle S_{\gamma}^{(n+1)} S_{\beta}^{(n)} \rangle - \langle S_{\gamma}^{(n+1)} \rangle \langle S_{\beta}^{(n)} \rangle, \quad (\text{A3})$$

$$\frac{\partial \langle S_{\gamma}^{(n+1)} \rangle}{\partial K_{\alpha}^{(n+1)}} = \langle S_{\gamma}^{(n+1)} S_{\alpha}^{(n+1)} \rangle - \langle S_{\gamma}^{(n+1)} \rangle \langle S_{\alpha}^{(n+1)} \rangle. \quad (\text{A4})$$

Given a set of spin configurations sampled from some Hamiltonian  $H = \sum_{\alpha} K_{\alpha} S_{\alpha}$ , we would like to infer the coupling parameters of  $H$ . We define a specific spin-dependent expectation

$$\langle S_{\alpha,l} \rangle_l \equiv \frac{1}{z_l} \sum_{\sigma_l} S_{\alpha,l} e^{\gamma_l}, \quad (\text{A5})$$

where  $z_l = \sum_{\sigma_l} e^{H_l}$ ,  $H_l = \sum_{\alpha} K_{\alpha} S_{\alpha,l}$ , and  $S_{\alpha,l}$  are combinations of spins in  $S_{\alpha}$  that include only  $\sigma_l$ . Here  $z_l$  and  $H_l$ , and hence  $\langle S_{\alpha,l} \rangle_l$ , depend on spins neighboring  $\sigma_l$ . The summation of  $\sigma_l$  can be carried out analytically and we obtain the formula

$$\langle S_{\alpha,l} \rangle_l = \hat{S}_{\alpha,l} \tanh \left[ \sum_{\beta} K_{\beta} \hat{S}_{\beta,l} \right], \quad (\text{A6})$$

where  $S_{\alpha,l} \equiv \sigma_l \hat{S}_{\alpha,l}$ .

The correlation functions can then be written in another form as

$$\frac{1}{Z} \sum_{\sigma} S_{\alpha} e^H = \frac{1}{Z} \sum_{\sigma} \left[ \frac{1}{m_{\alpha}} \sum_l \langle S_{\alpha,l} \rangle_l \right] e^{H(\sigma)}, \quad (\text{A7})$$

where  $m_{\alpha}$  is the number of spins in the combination  $S_{\alpha}$ . Introducing a second set of coupling parameters  $\{\tilde{K}_{\alpha}\}$ , we define

$$\langle \tilde{S}_{\alpha} \rangle = \frac{1}{Z} \sum_{\sigma} \left\{ \frac{1}{m_{\alpha}} \sum_l \hat{S}_{\alpha,l} \tanh \left[ \sum_{\beta} \tilde{K}_{\beta} \hat{S}_{\beta,l} \right] \right\} e^{H(\sigma)}. \quad (\text{A8})$$

It can be shown that  $\{\langle S_{\alpha} \rangle\} = \{\langle \tilde{S}_{\alpha} \rangle\}$  if and only if  $\{K_{\alpha}\} = \{\tilde{K}_{\alpha}\}$ .

Figure 6 shows the couplings used for the calculation of the renormalized coupling parameters for the two-dimensional Ising model. The first seven even couplings in Fig. 6(a) are used to compute the thermal critical exponent. The odd couplings in Fig. 6(b) are used to compute the magnetic critical exponent.

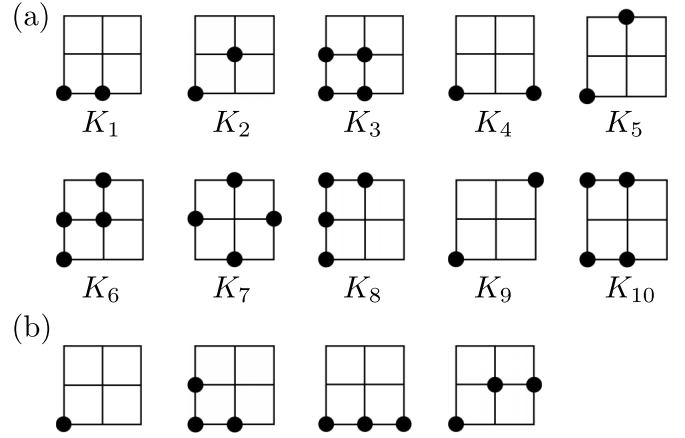


FIG. 6. 2D couplings. (a) Couplings used for the calculation of renormalized coupling parameters. The first seven are used for the calculation of the thermal critical exponent. (b) The four couplings used to compute the magnetic critical exponent.

## APPENDIX B: COMPARISON WITH OTHER RBM-BASED SCHEMES

### 1. RG transformation and normalizing condition

Consider again a general RG transformation

$$e^{H'(\mu)} = \sum_{\sigma} P(\mu|\sigma) e^{H(\sigma)}, \quad (\text{B1})$$

where  $P(\mu|\sigma)$  is the weight factor. The weight factor is required to satisfy the trace condition

$$\sum_{\mu} P(\mu|\sigma) = 1. \quad (\text{B2})$$

We argue that the trace condition is indispensable, since the condition leads to the invariance of free energy under renormalization and the fundamental relation

$$f(K) = b^{-d} f(K'), \quad (\text{B3})$$

where  $f(K)$  is the free energy density of the system in the thermodynamic limit. For  $K$  consisting of nearest-neighbor coupling and a magnetic field, under suitable transformation, we could arrive at  $f(t, h) = b^{-d} f(b^{y_t} t, b^{y_h} h)$ , where  $y_t$  and  $y_h$  are the often-sought-after critical thermal and magnetic exponents, respectively.

In the following, we review the schemes proposed in Refs. [22,25] and point out the shortcomings in each scheme.

### 2. Variational RG and Mehta and Schwab's mapping

In Ref. [22] the weight factor is defined as

$$P_W(\mu|\sigma) = \exp \left( \sum_{ij} W_{ij} \sigma_i \mu_j - H(\sigma) \right). \quad (\text{B4})$$

Here  $H(\sigma)$  is the original Hamiltonian, e.g.,  $H(\sigma) = K \sum_{\langle ij \rangle} \sigma_i \sigma_j$ . The  $W_{ij}$  are the variational parameters. The form of the weight factor does not satisfy the trace condition and in general it is not possible to choose the parameters  $W_{ij}$  to satisfy the trace condition (B2). The fundamental relation (B3) is only approximated.

We note that in the original procedure of the variational renormalization group [21], the form of the weight factor is chosen with variational parameters such that for all values of variational parameters the weight factor must satisfy the trace condition. The variational parameters are used instead to optimize the lower bound of the approximated free energy density.

We define a distribution of the weight factor with variational parameters  $W_{ij}$ ,

$$P_W(\sigma) = \frac{\sum_{\mu} \exp(\sum_{ij} W_{ij} \sigma_i \mu_j)}{\sum_{\sigma} \sum_{\mu} \exp(\sum_{ij} W_{ij} \sigma_i \mu_j)}. \quad (\text{B5})$$

In Ref. [22] the variational parameters are chosen to make

$$D_{\text{KL}}\left(\frac{e^{H(\sigma)}}{Z} \parallel P_W(\sigma)\right) \quad (\text{B6})$$

as small as possible. This completely fixes the variational parameters, leaving no room for optimizing the lower bound free energy approximation. That is to say, the variational approximation in machine learning (B6) and the variational approximation of the variational renormalization theory work at completely different levels.

The rationale of the criterion (B6) for choosing the variational parameters is that it is a necessary but not sufficient condition for the trace condition to be satisfied:

$$\sum_{\mu} \exp\left(\sum_{ij} W_{ij} \sigma_i \mu_j - H(\sigma)\right) = 1$$

implies

$$e^{H(\sigma)} = \sum_{\mu} \exp\left(\sum_{ij} W_{ij} \sigma_i \mu_j\right).$$

The normalization factor  $\sum_{\sigma} \sum_{\mu} \exp(\sum_{ij} W_{ij} \sigma_i \mu_j)$  is equal to the partition function for the original Hamiltonian, denoted by  $Z$ . Therefore, the divergence (B6) is exactly zero. The criterion is not sufficient since when

$$e^{f(\sigma)} / \sum e^{f(\sigma)} = e^{g(\sigma)} / \sum e^{g(\sigma)},$$

we have

$$e^{f(\sigma)-g(\sigma)} = \sum e^{f(\sigma)} / \sum e^{g(\sigma)},$$

where the trace condition fails up to some unknown constant not necessarily equal to one.

On the other hand, with the parametrized form of the weight factor as in (B4), the renormalized Hamiltonian would then describe the marginal distribution  $P_W(\mu)$  of the RBM. We define  $P_W(\mu)$  to be

$$P_W(\mu) = \frac{\sum_{\sigma} \exp(\sum_{ij} W_{ij} \sigma_i \mu_j)}{\sum_{\sigma} \sum_{\mu} \exp(\sum_{ij} W_{ij} \sigma_i \mu_j)}. \quad (\text{B7})$$

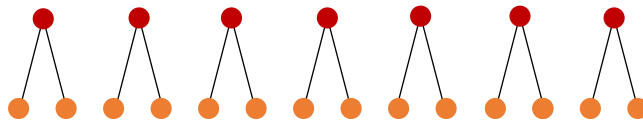


FIG. 7. Weight factor factorized as identical copies of local weight factors.

Carrying out the RG transformation (B1) for the weight factors (B4) gives

$$\begin{aligned} e^{H'(\mu)} &= \sum_{\sigma} \exp\left(\sum_{ij} W_{ij} \sigma_i \mu_j - H(\sigma)\right) e^{H(\sigma)} \\ &= \sum_{\sigma} \exp\left(\sum_{ij} W_{ij} \sigma_i \mu_j\right). \end{aligned} \quad (\text{B8})$$

The normalization factor  $\sum_{\sigma} \sum_{\mu} \exp(\sum_{ij} W_{ij} \sigma_i \mu_j)$  is thus equal to the partition function  $Z'$  for the renormalized Hamiltonian irrespective of the choice of the variational parameters  $W_{ij}$ . Therefore,

$$P_W(\mu) = \frac{e^{H'(\mu)}}{Z'}. \quad (\text{B9})$$

In this respect, we can say that the hidden variables of the machine are described by the renormalized Hamiltonian.

### 3. Real-space mutual information algorithm

In Ref. [25] the weight factor factorizes as

$$P_{\Lambda}(\mu|\sigma) = \prod_j P_{\Lambda}(\mathcal{H}_j|\mathcal{V}_j), \quad (\text{B10})$$

where  $\mathcal{H}_j = \{\mu_j\}$  consists of a single renormalized spin and  $\mathcal{V}_j = \{\sigma_j^1, \sigma_j^2\}$  consists of two original spins in the case of a one-dimensional system (and  $2^2$  in the case of a two-dimensional system) (see Fig. 7). The local weight factor is parametrized as

$$P_{\Lambda}(\mathcal{H}_j|\mathcal{V}_j) = \frac{\exp(\sum_i \Lambda_i \mu_j \sigma_j^i)}{\sum_{\mu} \exp(\sum_i \Lambda_i \mu_j \sigma_j^i)}. \quad (\text{B11})$$

The variational parameter  $\Lambda$  is obtained through only a single copy of the local weight factor and hence we omit the subscript  $j$  in the following. Consider a single copy of the local weight factor where the local visible spins  $\mathcal{V}$  are embedded among the buffer  $\mathcal{B}$ , environmental  $\mathcal{E}$ , and other  $\mathcal{O}$  spins which collectively form the original system spins  $\mathcal{X}$  (see Fig. 8). Construct two proxies  $P_{\Theta_1}(\mathcal{V})$  and  $P_{\Theta_2}(\mathcal{V}, \mathcal{E})$  in the form of RBMs trained on the restriction of  $\mathcal{X} = (\mathcal{V}, \mathcal{B}, \mathcal{E}, \mathcal{O})$  Monte Carlo (MC) samples (from the Boltzmann equilibrium distribution of the Hamiltonian of concerned). Define  $P_{\Lambda}(\mathcal{E}, \mathcal{H}) = \sum_{\mathcal{V}} P_{\Theta_2}(\mathcal{V}, \mathcal{E}) P_{\Lambda}(\mathcal{H}|\mathcal{V})$ ,  $P_{\Lambda}(\mathcal{H}) = \sum_{\mathcal{V}} P_{\Theta_1}(\mathcal{V}) P_{\Lambda}(\mathcal{H}|\mathcal{V})$ , and  $P(\mathcal{E}) = \sum_{\mathcal{V}} P_{\Theta_2}(\mathcal{V}, \mathcal{E})$ . The variational parameters  $\Lambda$  are chosen to make

$$I_{\Lambda}(\mathcal{H}; \mathcal{E}) = \sum_{\mathcal{H}, \mathcal{E}} P_{\Lambda}(\mathcal{E}, \mathcal{H}) \log\left(\frac{P_{\Lambda}(\mathcal{E}, \mathcal{H})}{P_{\Lambda}(\mathcal{H})P(\mathcal{E})}\right) \quad (\text{B12})$$



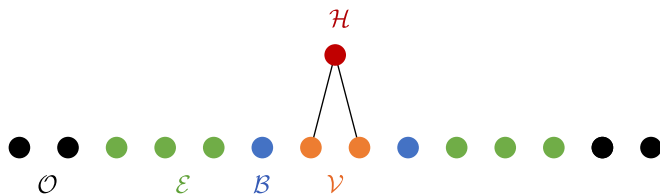


FIG. 8. Schematic decomposition of the system spins into the visible ( $\mathcal{V}$ ), buffer ( $\mathcal{B}$ ), environmental ( $\mathcal{E}$ ) and other ( $\mathcal{O}$ ) spins, respectively. We refer to a local block of hidden spins  $\mathcal{H}$ .

as large as possible, where the distributions needed on the right-hand side are defined as above. Since  $P(\mathcal{E})$  is independent of  $\Lambda$ , we instead maximize a proxy  $A_\Lambda = \sum_{\mathcal{H}, \mathcal{E}} P_\Lambda(\mathcal{E}, \mathcal{H}) \log[P_\Lambda(\mathcal{E}, \mathcal{H})/P_\Lambda(\mathcal{H})]$  of mutual information. However, to evaluate the proxy  $A_\Lambda$ , further approximations have to be made.

In order to perform a quantitative analysis, Koch-Janusz and Ringel constructed a “thermometer” function  $T(A_\Lambda)$  which maps the proxy  $A_\Lambda$  to the temperature. The thermometer works to extract the effective temperature of the renormalized system. Constructing such a thermometer requires the generation of sets of MC samples at different temperatures. For each set of samples, one can compute the proxy  $A_\Lambda$  and hence know the mapping from  $A_\Lambda$  to the temperature  $T$  for this set of samples. For a given type of system (e.g., Ising), we can write  $T(A_\Lambda)$  as  $T(T_0, L, b, l)$ , where  $T_0$  is the temperature of the initially prepared system,  $L$  is the initial system size,  $b$  is the scale factor, and  $l$  is the scaling length ( $l = 0$  means the original system,  $l = 1$  means one-step renormalization, and so on). We can then fit a function to these sets of samples and construct the thermometer. Koch-Janusz and Ringel [25] postulated a scaling function of the form  $f((L/b^l)^{1/\nu})$  related to the effective renormalized temperature  $T(T_0, L, b, l)$  as

$$\frac{T(T_0, L, b, l) - T_c}{T_0 - T_c} = f((L/b^l)^{1/\nu}), \quad (\text{B13})$$

where  $T_c$  is the critical temperature of the original system. Finally, one could collapse the plot of  $(T - T_c)/(T_0 - T_c)$  as a function of  $(L/b^l)^{1/\nu}$  to estimate the values of  $\nu$  and  $T_c$ .

#### 4. Neural Monte Carlo renormalization group

In the present work we define the weight factor as

$$P_W(\mu|\sigma) = \frac{\exp(\sum_{ij} W_{ij} \sigma_i \mu_j)}{\sum_\mu \exp(\sum_{ij} W_{ij} \sigma_i \mu_j)}. \quad (\text{B14})$$

where, for the translationally invariant system, the variational parameters are shift invariant; that is, for different  $j$  and  $j'$  we have

$$W_{ij} = W_{[(i+j'-j) \bmod N]j'} \quad (\text{B15})$$

in the case of the one-dimensional system. The weight factor satisfies the trace condition for all values of  $W_{ij}$ . Let us define a joint distribution out of this weight factor

$$P_W(\mu, \sigma) = \frac{\exp(\sum_{ij} W_{ij} \sigma_i \mu_j)}{\sum_\sigma \sum_\mu \exp(\sum_{ij} W_{ij} \sigma_i \mu_j)}. \quad (\text{B16})$$

Here  $P_W(\mu, \sigma)$  has exactly the same form of a RBM and the weight factor can be viewed as the conditional distribution  $P_W(\mu|\sigma) = P_W(\mu, \sigma) / \sum_\mu P_W(\mu, \sigma)$ .

Consider one of the breakthroughs in the realm of deep learning where Hinton introduced a greedy layerwise unsupervised learning algorithm (see Sec. 2.3 of [32]). Denote by  $P_W(\mu|\sigma)$  the posterior over  $\mu$  associated with the trained RBM (we recall that  $\sigma$  is the observed input). This gives rise to a (feature) empirical distribution  $p'(\mu)$  over the hidden variables  $\mu$  when  $\sigma$  is sampled from the data’s empirical distribution  $p(\sigma)$ :

$$p'(\mu) = \sum_\sigma P_W(\mu|\sigma) p(\sigma). \quad (\text{B17})$$

The samples of  $\mu$  with empirical distribution  $p'(\mu)$  become the input for another layer of the RBM. We can view the RBM to work as extracting features  $\mu$  from inputs  $\sigma$ .

Note the similarity between the RG transformation (B1) and the feature extraction process (B17). We could postulate that the input distribution  $p(\sigma)$  is determined by some Hamiltonian  $H(\sigma)$  where  $p(\sigma) = e^{H(\sigma)}/Z$ . We postulate that the posterior distribution  $P_W(\mu|\sigma)$  of an RBM works as a weight factor to do the RG transformation:  $e^{H'(\mu)} = \sum_\sigma P_W(\mu|\sigma) e^{H(\sigma)}$ . Hence the feature extraction process (B17) becomes a necessary condition for the system to perform the RG transformation. In other words, the feature distribution extracted by the machine is described by the renormalized Hamiltonian.

Now the variational parameters in the weight factor  $P_W(\mu|\sigma)$  are free to change. All choices of parameters should derive a well-defined RG transformation. The criterion for choosing the parameters is entirely arbitrary from the perspective of doing the RG: We do not know *a priori* what weights  $W_{ij}$  could give a “nicer” RG flow. A nice RG flow, however, should bring the original Hamiltonian closer to the fixed point fast. Also, it should remove long-range coupling parameters for practical purposes of performing the RG and, loosely speaking, for killing the irrelevant scaling fields. Critical exponents and the coupling parameters can be easily computed using the MCRG techniques described in Appendix A.

In the realm of machine learning, the weights of an RBM are chosen to make the divergence (B6) as small as possible. We note that the criterion is entirely machine-learning theoretical. In contrast, in Ref. [22] the criterion also serves as a necessary condition for the weight factor to satisfy the trace condition, a notion which is RG theoretical.

#### APPENDIX C: CRITICAL MAGNETIC EXPONENT

In Fig. 9(a) we show the critical magnetic exponents calculated from the learned weight factors. As the training epoch

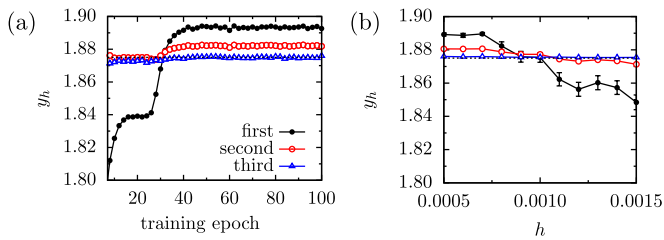


FIG. 9. (a) Critical magnetic exponents for weight factors along the training epoch. (b) Critical magnetic exponents for weight factors learned on samples with an external magnetic field. The dotted line shows the exact value of  $\gamma_h = 1.875$ .

increases, the evolution of the magnetic exponent shows behavior similar to that of the thermal exponent, although the converged value overshoots the exact value. It is interesting to see that at some point along the training epochs, the value of the magnetic exponent at the first RG iteration meets that of the second and third.

We next use training samples generated with an external magnetic field to train our weight factors and use them to calculate the critical magnetic exponents. Figure 9(b) shows a crossing of the magnetic critical exponents when varying the magnetic field. The learned weight factor is able to predict well the magnetic exponent at the first RG iteration at a field of  $h \simeq 0.001$ .

- [1] K. G. Wilson, Renormalization group and critical phenomena. I. Renormalization group and the Kadanoff scaling picture, *Phys. Rev. B* **4**, 3174 (1971).
- [2] L. P. Kadanoff, Scaling laws for Ising models near  $T_c$ , *Phys. Phys. Fiz.* **2**, 263 (1966).
- [3] T. Niemeijer and J. van Leeuwen, in *Phase Transitions and Critical Phenomena*, edited by C. Domb and M. S. Green (Academic, New York, 1976), Vol. 6, Chap. 7, pp. 425–505.
- [4] R. H. Swendsen, Monte Carlo Renormalization Group, *Phys. Rev. Lett.* **42**, 859 (1979).
- [5] Y. Wu and R. Car, Variational Approach to Monte Carlo Renormalization Group, *Phys. Rev. Lett.* **119**, 220602 (2017).
- [6] Y. Wu and R. Car, Determination of the critical manifold tangent space and curvature with Monte Carlo renormalization group, *Phys. Rev. E* **100**, 022138 (2019).
- [7] R. H. Swendsen, Optimization of Real-Space Renormalization-Group Transformations, *Phys. Rev. Lett.* **52**, 2321 (1984).
- [8] H. W. J. Blöte, J. R. Heringa, A. Hoogland, E. W. Meyer, and T. S. Smit, Monte Carlo Renormalization of the 3D Ising Model: Analyticity and Convergence, *Phys. Rev. Lett.* **76**, 2613 (1996).
- [9] D. Ron, A. Brandt, and R. H. Swendsen, Surprising convergence of the Monte Carlo renormalization group for the three-dimensional Ising model, *Phys. Rev. E* **95**, 053305 (2017).
- [10] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, *Nature (London)* **521**, 436 (2015).
- [11] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *IEEE Signal Process. Mag.* **29**, 82 (2012).
- [12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, in *Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran, Red Hook, 2012), pp. 1097–1105.
- [13] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, Mastering the game of Go with deep neural networks and tree search, *Nature (London)* **529**, 484 (2016).
- [14] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vechnyevets, *et al.*, Grandmaster level in StarCraft II using multi-agent reinforcement learning, *Nature (London)* **575**, 350 (2019).
- [15] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, Human-level control through deep reinforcement learning, *Nature (London)* **518**, 529 (2015).
- [16] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, *Science* **355**, 602 (2017).
- [17] E. P. L. van Nieuwenburg, Y.-H. Liu, and S. D. Huber, Learning phase transitions by confusion, *Nat. Phys.* **13**, 435 (2017).
- [18] J. Carrasquilla and R. G. Melko, Machine learning phases of matter, *Nat. Phys.* **13**, 431 (2017).
- [19] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, *Rev. Mod. Phys.* **91**, 045002 (2019).
- [20] J. Carrasquilla, Machine learning for quantum matter, *Adv. Phys.: X* **5**, 1797528 (2020).
- [21] L. P. Kadanoff, A. Houghton, and M. C. Yalabik, Variational approximations for renormalization group transformations, *J. Stat. Phys.* **14**, 171 (1976).
- [22] P. Mehta and D. J. Schwab, An exact mapping between the variational renormalization group and deep learning, [arXiv:1410.3831](https://arxiv.org/abs/1410.3831).
- [23] H. W. Lin, M. Tegmark, and D. Rolnick, Why Does Deep and Cheap Learning Work So Well?, *J. Stat. Phys.* **168**, 1223 (2017).
- [24] D. J. Schwab and P. Mehta, Comment on “Why does deep and cheap learning work so well?” [H. W. Lin, M. Tegmark, and D. Rolnick, *J. Stat. Phys.* **168**, 1223 (2017)], [arXiv:1609.03541](https://arxiv.org/abs/1609.03541).
- [25] M. Koch-Janusz and Z. Ringel, Mutual information, neural networks and the renormalization group, *Nat. Phys.* **14**, 578 (2018).
- [26] P. M. Lenggenhager, D. E. Gökmen, Z. Ringel, S. D. Huber, and M. Koch-Janusz, Optimal Renormalization Group

- Transformation from Information Theory, *Phys. Rev. X* **10**, 011037 (2020).
- [27] S. Iso, S. Shiba, and S. Yokoo, Scale-invariant feature extraction of neural network and renormalization group flow, *Phys. Rev. E* **97**, 053304 (2018).
- [28] S. S. Funai and D. Giataganas, Thermodynamics and feature extraction by machine learning, *Phys. Rev. Research* **2**, 033415 (2020).
- [29] S. Efthymiou, M. J. S. Beach, and R. G. Melko, Super-resolving the Ising model with convolutional neural networks, *Phys. Rev. B* **99**, 075113 (2019).
- [30] J.-H. Chung and Y.-J. Kao, Optimal real-space renormalization-group transformations with artificial neural networks, [arXiv:1912.09005](https://arxiv.org/abs/1912.09005).
- [31] P. Smolensky, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, edited by D. E. Rumelhart, J. L. McClelland, and PDP Research Group (MIT Press, Cambridge, 1986), pp. 194–281.
- [32] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, in *Advances in Neural Information Processing Systems 20*, edited by J. Platt, D. Koller, Y. Singer, and S. Roweis (Curran, Red Hook, 2007), pp. 153–160.
- [33] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, 2016).
- [34] A. Krizhevsky, Learning multiple layers of features from tiny images, University of Toronto report, 2009 (unpublished).
- [35] U. Wolff, Collective Monte Carlo Updating for Spin Systems, *Phys. Rev. Lett.* **62**, 361 (1989).
- [36] G. E. Hinton, Training products of experts by minimizing contrastive divergence, *Neural Comput.* **14**, 1771 (2002).
- [37] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, *Proceedings of the 26th Annual International Conference on Machine Learning* (ACM, New York, 2009), pp. 609–616.
- [38] R. H. Swendsen, Monte Carlo Calculation of Renormalized Coupling Parameters, *Phys. Rev. Lett.* **52**, 1165 (1984).
- [39] H. J. Hilhorst, M. Schick, and J. M. J. van Leeuwen, Differential Form of Real-Space Renormalization: Exact Results for Two-Dimensional Ising Models, *Phys. Rev. Lett.* **40**, 1605 (1978).
- [40] R. Gupta, Open problems in Monte Carlo renormalization group: Application to critical phenomena, *J. Appl. Phys.* **61**, 3605 (1987).
- [41] L. P. Kadanoff and A. Houghton, Numerical evaluations of the critical properties of the two-dimensional Ising model, *Phys. Rev. B* **11**, 377 (1975).
- [42] J. Van Leeuwen, Singularities in the Critical Surface and Universality for Ising-Like Spin Systems, *Phys. Rev. Lett.* **34**, 1056 (1975).
- [43] R. H. Swendsen, Monte Carlo renormalization-group studies of the  $d = 2$  Ising model, *Phys. Rev. B* **20**, 2080 (1979).
- [44] C. F. Baillie, R. Gupta, K. A. Hawick, and G. S. Pawley, Monte Carlo renormalization-group study of the three-dimensional Ising model, *Phys. Rev. B* **45**, 10438 (1992).
- [45] M. Hasenbusch, Finite size scaling study of lattice models in the three-dimensional Ising universality class, *Phys. Rev. B* **82**, 174433 (2010).
- [46] J.-S. Wang and R. H. Swendsen, Monte Carlo renormalization-group study of Ising spin glasses, *Phys. Rev. B* **37**, 7745 (1988).
- [47] G. E. Hinton and R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, *Science* **313**, 504 (2006).
- [48] M. A. Novotny and D. P. Landau, Monte Carlo renormalization group for quantum systems, *Phys. Rev. B* **31**, 1449 (1985).
- [49] <https://github.com/unixtomato/nmcrg>.