PHYSICAL REVIEW E **96**, 051301(R) (2017)

Exploring cluster Monte Carlo updates with Boltzmann machines

Lei Wang[*]

Beijing National Lab for Condensed Matter Physics and Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China

(Received 8 March 2017; revised manuscript received 30 October 2017; published 16 November 2017)

Boltzmann machines are physics informed generative models with broad applications in machine learning. They model the probability distribution of an input data set with latent variables and generate new samples accordingly. Applying the Boltzmann machines back to physics, they are ideal recommender systems to accelerate the Monte Carlo simulation of physical systems due to their flexibility and effectiveness. More intriguingly, we show that the generative sampling of the Boltzmann machines can even give different cluster Monte Carlo algorithms. The latent representation of the Boltzmann machines can be designed to mediate complex interactions and identify clusters of the physical system. We demonstrate these findings with concrete examples of the classical Ising model with and without four-spin plaquette interactions. In the future, automatic searches in the algorithm space parametrized by Boltzmann machines may discover more innovative Monte Carlo updates.

DOI: 10.1103/PhysRevE.96.051301

Introduction. There have been endless efforts made toward inventing new Monte Carlo algorithms for the efficient simulation of challenging physical problems ever since its invention [1]. Innovative Monte Carlo algorithms such as Refs. [2–6] represent landmark achievements in computational physics. In certain cases, they even outperform the hardware accelerations from Moore's law, i.e., running these modern algorithms on decades-old computers would be faster than running traditional algorithms on the fastest supercomputers of today [7]. Orders of magnitude acceleration not only concerns efficiency and energy consumption but also allows us to discover qualitative new physical phenomena [8].

Recently, there have been heated efforts to systematically improve the Monte Carlo sampling efficiency for physical problems using ideas and techniques from machine learning [9–13]. The basic idea is to construct surrogate models based on past samples, then use them to guide future sampling. Although similar ideas were discussed repeatedly in statistics literature [14–17], there are two notable features of recent attempts in the physics contexts [9–13]: using simple surrogate models with a clear physical meaning [10,12] and using Boltzmann machines (BMs) [9]. The BM is a historic model in machine learning [18,19] and has played a crucial role in the recent resurgence of deep learning [20]. Using the BM to model physical distributions [21] and accelerate Monte Carlo sampling [9] opens new possibilities for algorithmic innovations because they can suggest novel Monte Carlo update strategies instead of merely acting as cheaper surrogate models.

As illustrated in Fig. 1, the BM consists of stochastic visible ($\mathbf{s}$) and hidden ($\mathbf{h}$) variables. An energy function $E(\mathbf{s},\mathbf{h})$ specifies the connectivity and interaction between these units. Their joint probability distribution follows the Boltzmann distribution $p(\mathbf{s},\mathbf{h}) = e^{-E(\mathbf{s},\mathbf{h})}$. The hidden units of the BM act as internal representations and mediate interactions between the visible units. After tracing out the hidden units, the marginal probability $p(\mathbf{s}) = \sum_{\mathbf{h}} p(\mathbf{s},\mathbf{h})$ can approximate arbitrarily complex probability distributions over the visible variables since BM is a universal probability approximator [22–24]. By tuning the parameters in the energy function one can therefore use $p(\mathbf{s})$ to model certain target probability distributions $\pi(\mathbf{s})$ of a data set. The expressive powers of BMs were investigated recently both from machine learning [25,26] and physics perspectives [27–30]. See Refs. [21,31–33] for other recent applications of the BM to quantum and statistical physics problems.

A successfully trained BM can capture the salient features of the input data. For example, the BM learns about some building blocks from an image data set of handwritten digits [34]. By simulating a trained BM as a statistical physics system, one can generate new samples from the learned distribution. Reference [9] uses samples generated from a BM as Monte Carlo proposals. To keep the physical simulation unbiased, the BM recommended update of the visible units $\mathbf{s} \to \mathbf{s}'$ is accepted according to the Metropolis-Hastings rule [1,35] (see Appendix A),

$$A(\mathbf{s} \to \mathbf{s}') = \min\left[1, \frac{p(\mathbf{s})}{p(\mathbf{s}')}\frac{\pi(\mathbf{s}')}{\pi(\mathbf{s})}\right]. \tag{1}$$

Equation (1) shows that the BM guides the Monte Carlo sampling by exploiting the learned probability distribution. In particular, one can even achieve a rejection-free Monte Carlo simulation scheme if the BM perfectly captures the target probability distribution $p(\mathbf{s}) \sim \pi(\mathbf{s})$. Using Eq. (1) is advantageous as long as the simulation of the BM is cheaper than the original model. For example, Ref. [9] employs a restricted architecture of BM where the connections are limited to being between the visible and hidden units [36]. Such a restricted BM can be sampled efficiently by blocked Gibbs sampling alternating between the hidden and visible units.

Moreover, Ref. [9] finds that the generative sampling of the restricted BM appears to exploit the collective density correlations learned from the Monte Carlo data. This suggests that besides being a cheap surrogate and a general recommender engine for Monte Carlo simulations, BM may help us find out conceptually new efficient Monte Carlo updates with its feature discovery ability.

In this Rapid Communication, we demonstrate the BM's power by exact constructions of cluster updates and present

[*] wanglei@iphy.ac.cn

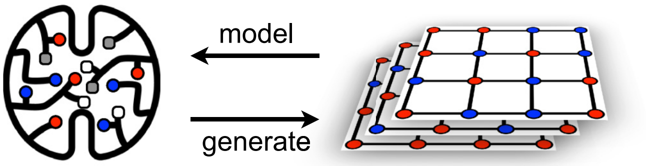

FIG. 1. A schematic plot of the Boltzmann machine and its typical use in machine learning. The Boltzmann machine consists of stochastic variables (red and blue dots) and hidden units (white and gray squares) connected into a network. Different colors denote active or inactive states of various units. The joint probability distribution of these variables follows a Boltzmann distribution. By adjusting the structure and parameters of the BM, it models the target probability distribution of input data as the marginal probability distribution of visible variables. The sampling of the Boltzmann machine generates new samples according to the learned distribution. Here, we show that a BM with an appropriately designed architecture can suggest efficient cluster Monte Carlo algorithms in its generative sampling.

a general framework to fully exploit its potential. The crucial insight is that the hidden units of the BM can mediate complex interactions between the visible units and identify clusters of the visible units. The generative sampling of the BM then automatically proposes efficient cluster updates. To encourage these desired features, it is crucial to design the BM in a suitable architecture and allow its parameters to adapt to the physical distribution via learning.

Example: Ising model. To make the discussions concrete, we start with the classical Ising model and show that the generative sampling of the BM encompasses a wide range of celebrated cluster algorithms [3,37,38]. The Boltzmann weight of the Ising model reads

$$\pi(\mathbf{s}) = \exp\left(\beta J \sum_{\ell} \prod_{i\in\ell} s_i\right), \tag{2}$$

where $\beta = 1/T$ is the inverse temperature and J is the coupling constant. We consider ferromagnetic coupling $J > 0$ in the following for clarity. The considerations are nevertheless general and valid for the antiferromagnetic case as well. Equation (2) consists of a summation over links ℓ of a lattice and a product over Ising spins $s_i \in \{-1,1\}$ residing on the vertices connected by the link.

To devise a BM inspired cluster update of the Ising model, we consider the architecture illustrated in Fig. 2(a). We view the Ising spins as visible variables and introduce binary hidden variables $h_\ell \in \{0,1\}$ on the links of the lattice. Coupling of these units gives rise to the following energy function,

$$E(\mathbf{s},\mathbf{h}) = -\sum_{\ell}\left(W\prod_{i\in\ell} s_i + b\right)h_\ell. \tag{3}$$

Equation (3) is a high-order BM [39] because the interaction consists of three-spin interactions (one hidden unit and two visible units). Similar architectures were discussed in the machine learning literature under the name three-way Boltzmann machines [40–42]. In light of the translational invariance of the Ising model (2) we use the same connection weight W and bias b for all the links. Therefore the BM energy function (3) only contains two free parameters.

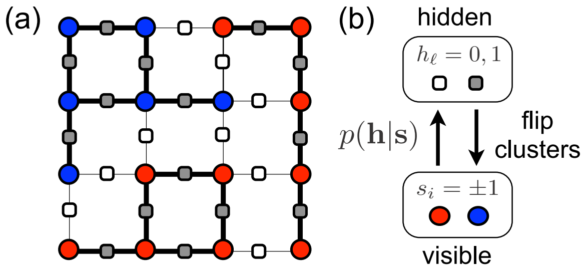

FIG. 2. (a) The Boltzmann machine (3) reproduces cluster Monte Carlo algorithms of the Ising model (2). Solid dots residing on the vertices are the visible units representing the Ising spins. Red and blue colors denote Ising spin up and down. The squares residing on the links are the binary hidden units, where the white and gray colors indicate the inactive ($h_\ell = 0$) or active ($h_\ell = 1$) status of the hidden unit. The effective interaction between the visible units can either be W (thick links) or 0 (thin links). (b) The sampling of the BM. Given the visible units, we sample the hidden units according to Eq. (4). The inactive hidden units (white squares) divide the visible units into disconnected components which can be flipped collectively at random.

To perform a generative sampling of the BM (3) we proceed in two steps by exploiting its particular architecture shown in Fig. 2. First, given a set of visible Ising spins, we can readily perform a direct sampling of the hidden units. This is because the conditional probability factorizes into products over each link, $p(\mathbf{h}|\mathbf{s}) = p(\mathbf{s},\mathbf{h})/p(\mathbf{s}) = \prod_\ell p(h_\ell|\mathbf{s})$, where

$$p(h_\ell = 1|\mathbf{s}) = \sigma\left(W\prod_{i\in\ell} s_i + b\right), \tag{4}$$

and $\sigma(z) = 1/(1+e^{-z})$ is the sigmoid activation function. As shown in Fig. 2(a), the inactive hidden units (white squares) divide the lattice into disconnected components since $h_\ell = 0$ in Eq. (3) decouples the visible Ising spins residing on the link ℓ. Next, one can identify connected components using the union-find algorithm [43,44] and flip all the visible Ising spins within each component collectively at random. This cluster move with respect to the statistical weight of the BM (3) due to the Z_2 symmetry of the visible Ising spins in the energy function.

Combining the two steps in Fig. 2(b) forms an update of the visible units of the BM. Recommending the update to the Ising model Monte Carlo simulation, it is accepted with the probability (1) (see Appendix A). In the trivial case of $W = 0$, the marginal probability $p(\mathbf{s}) = \prod_\ell (1 + e^{W\prod_{i\in\ell} s_i + b})$ of the BM (3) is independent of the visible spins and Eq. (1) reduces to the ordinary Metropolis algorithm where b controls how many spins we attempt to flip together. While matching $p(\mathbf{s})$ and the Ising model Boltzmann weight (2) we obtain a rejection-free Monte Carlo scheme. The resulting condition

$$\frac{1+e^{b+W}}{1+e^{b-W}} = e^{2\beta J} \tag{5}$$

can always be satisfied with appropriately chosen W and b. It is instructive to examine the BM recommended updates in two limiting cases.

In the limit of $b \to -\infty$, the solution of Eq. (5) reads $W + b = \ln(e^{2\beta J} - 1)$. Thus, the conditional sampling of the hidden units (4) will set $h_\ell = 1$ with probability $\sigma(W + b) = 1 - e^{-2\beta J}$ if the link connects to two parallel spins $\prod_{i\in\ell} s_i = 1$, while it will always set the hidden unit to inactive $h_\ell = 0$ if the link connects to antiparallel spins. Combined with the random cluster flip of visible units, this BM recommended update shown in Fig. 2(b) exactly reproduces the Swendsen-Wang cluster algorithm [3] of the Ising model.

While in the opposite limit $b \to \infty$, the solution of Eq. (5) approaches to $W = \beta J$. In this limit, all the hidden units are frozen to $h_\ell = 1$ because the activation function in Eq. (4) saturates regardless of whether or not the visible Ising spins are aligned. The BM (3) then trivially reproduces the Ising model statistics by copying its coupling constant βJ to the connection weight W. In this limit the BM recommended update shown in Fig. 2(b) is a trivial global flip of the visible Ising spins.

In between the above two limiting cases, the BM still recommends valid rejection-free Monte Carlo updates for the Ising model. These updates correspond to Niedermayer's cluster algorithm [37] where the sites are randomly connected into clusters according to Eq. (4) and the clusters may contain misaligned visible spins. The bias parameter b in Eq. (4) controls the activation threshold of the hidden units and thus affects the average cluster size. In essence, the sampling of BM (3) is a form of the dual Monte Carlo algorithm [45,46], which encompasses the Kandel-Domany cluster Monte Carlo framework [38]. The framework is based on the Fortuin-Kasteleyn transformation [47,48], where the Monte Carlo sampling alternates between the physical degrees of freedom and auxiliary graphical variables. BM represents these auxiliary variables with the hidden units.

Example: Ising model with plaquette interactions. The potential of BM goes beyond reproducing existing algorithmic frameworks [3,37,38]. By further exploiting its power from latent representations one can make nontrivial algorithmic discoveries. We illustrate this using the plaquette Ising model [10] as an example. The Boltzmann weight reads

$$\pi(\mathbf{s}) = \exp\left(\beta J \sum_{\ell} \prod_{i\in\ell} s_i + \beta K \sum_{\wp} \prod_{i\in\wp} s_i\right), \tag{6}$$

where the second term contains four-spin interactions on each square plaquette denoted by $\wp$. We consider $K > 0$ for concreteness. Since no simple and efficient cluster algorithm is known, Ref. [10] fits the Boltzmann weight (6) to an ordinary Ising model with two-spin interactions and proposes Monte Carlo updates by simulating the latter model with cluster algorithms [4,49]. However, the acceptance rates decrease for large systems due to a mismatch between the surrogate model and the original physical model. The approach ends up showing similar scaling behavior as the single spin-flip update algorithm.

Here, we construct a BM which suggests an efficient, unbiased, and rejection-free cluster Monte Carlo algorithm for Eq. (6). First, we decompose the four-spin plaquette interaction using the Hubbard-Stratonovich (HS) transformation [50,51]

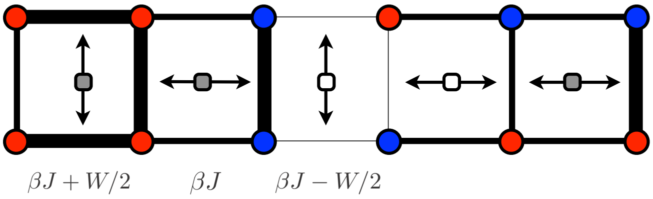

FIG. 3. The Boltzmann machine (8) suggests a new cluster update for the plaquette Ising model (6). Red/blue dots on the vertices denote the visible Ising spins, and white/gray squares in the plaquette center denote the hidden units. The double arrows point to the two parallel links $\ell_\wp, \bar{\ell}_\wp$ composing the plaquette $\wp$. The hidden units are sampled directly according to Eq. (9) where the breakup of the plaquette into parallel links is chosen at random. Once the hidden units are given, Eq. (8) reduces to an inhomogeneous Ising model where the visible spins interact with modified coupling strengths, indicated by the thicknesses of the links.

$$\exp\left(\beta K \prod_{i\in\wp} s_i\right) = \frac{e^{-\beta K}}{2} \sum_{h_\wp\in\{0,1\}} \exp\left[W\left(h_\wp - \frac{1}{2}\right)\mathcal{F}_\wp(\mathbf{s})\right], \tag{7}$$

where $W = \operatorname{acosh}(e^{2\beta K})$ is the coupling strength between the binary HS field $h_\wp$ and the sum of two-spin products $\mathcal{F}_\wp(\mathbf{s}) = \prod_{i\in\ell_\wp} s_i + \prod_{i\in\bar{\ell}_\wp} s_i$ defined for the plaquette. The two parallel links $\ell_\wp$ and $\bar{\ell}_\wp$ constitute the plaquette $\wp$ (see Fig. 3). Equation (7) is equivalent to the discrete HS transformation widely adopted for the Hubbard models [52]. Regarding the HS field $h_\wp$ as a hidden unit, the following BM,

$$E(\mathbf{s},\mathbf{h}) = -\sum_{\ell}\left[\beta J + W \sum_{\wp}\left(h_\wp - \frac{1}{2}\right) \times (\delta_{\ell\ell_\wp} + \delta_{\ell\bar{\ell}_\wp})\right]\prod_{i\in\ell} s_i, \tag{8}$$

exactly reproduces Eq. (6) after marginalization. Since Eq. (7) holds for arbitrary partition of the plaquette into two links $\ell_\wp \cup \bar{\ell}_\wp = \wp$ and $\ell_\wp \cap \bar{\ell}_\wp = \varnothing$, we choose a vertical or horizontal breakup at random for each plaquette.

Simulation of the BM (8) suggests an efficient cluster update for the original plaquette Ising model (6). First of all, sampling the hidden variables given the visible Ising spins is straightforward since the conditional probability factorizes over plaquettes $p(\mathbf{h}|\mathbf{s}) = \prod_\wp p(h_\wp|\mathbf{s})$, where

$$p(h_\wp = 1|\mathbf{s}) = \sigma(W\mathcal{F}_\wp(\mathbf{s})). \tag{9}$$

Therefore, the hidden unit of each plaquette activates independently given the local features $\mathcal{F}_\wp(\mathbf{s})$. Next, once the hidden variables are given, the BM (8) corresponds to an Ising model with two-spin interactions only, shown in Fig. 3. One can sample it efficiently using the cluster updates [3] by taking into account the randomly modified coupling strengths. As discussed above, this amounts to introducing another set of hidden units which plays the role of auxiliary graphical variables. Finally, according to Eq. (1), the updates of the visible

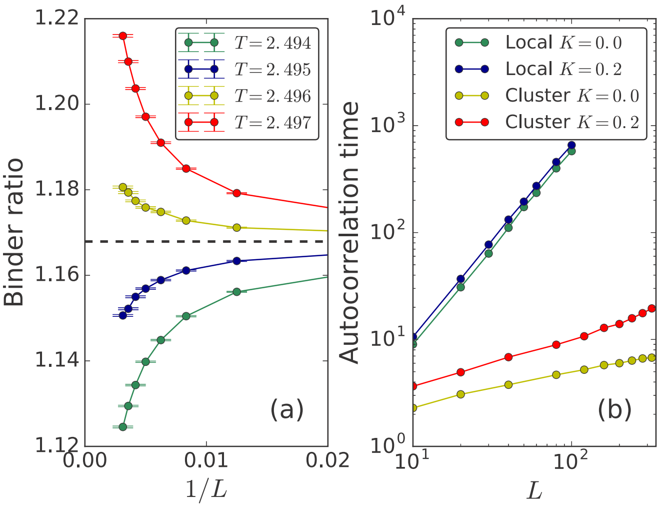

FIG. 4. Results for the Ising model with four-spin plaquette interactions (6) on square lattices with linear length L. (a) Binder ratio obtained using the cluster update suggested by BM (8) at $K/J = 0.2$. The dashed line indicates the universal value for the two-dimensional Ising universality class. (b) The cluster update improves the energy autocorrelation time by orders of magnitude at the critical point compared to the local update.

Ising spins are always accepted because the BM (8) exactly reproduces the statistics of the plaquette Ising model (6).

To demonstrate the efficiency of the BM inspired cluster update, we simulate the plaquette Ising model (6) in the vicinity of the critical point and compare its performance to the simple local update algorithm. Figure 4(a) shows the Binder ratio $\langle(\sum_i s_i)^4\rangle / \langle(\sum_i s_i)^2\rangle^2$ for various system sizes at $K/J = 0.2$, which indicates a critical temperature $T/J = 2.4955(5)$. The black dashed line indicates the universal critical value of the Binder ratio 1.1679 corresponding to the two-dimensional Ising universality class [53]. Figure 4(b) shows the energy autocorrelation times [7] of the local updates and the cluster updates at the critical point, both measured in units of Monte Carlo sweeps of the visible spins [54]. The local updates exhibit the same scaling for the Ising model ($K = 0$) and the plaquette Ising model ($K = 0.2$), while the cluster updates are orders of magnitude more efficient than the local updates. The dynamic exponent of the cluster algorithm is also significantly reduced compared to the local update.

General framework. To sum up, we outline a general framework of exploring cluster updates using the following BM,

$$E(\mathbf{s},\mathbf{h}) = E(\mathbf{s}) - \sum_{\alpha} [W_\alpha \mathcal{F}_\alpha(\mathbf{s}) + b_\alpha] h_\alpha, \tag{10}$$

where $\mathcal{F}_\alpha(\mathbf{s})$ is a feature of the visible units and $h_\alpha \in \{0,1\}$ is the corresponding hidden variable. W_α and b_α are the connection weight and bias, and α is the index for various features. For example, Eqs. (3) and (8) used the features defined on the links ($\alpha = \ell$) and on plaquettes ($\alpha = \wp$), respectively. In general, one is free to design features consisting of long-range interactions or even multispin interactions [55]. There are several crucial points in the general structure of Eq. (10). First, one can easily sample the hidden units conditioned on these features since there is no interaction between the hidden variables, i.e., Eq. (10) is a semirestricted BM. The activation probability of each hidden unit is $p(h_\alpha = 1|\mathbf{s}) = \sigma(W_\alpha \mathcal{F}_\alpha(\mathbf{s}) + b_\alpha)$ [cf. Eqs. (4) and (9)]. Second, once the hidden units are given, Eq. (10) reduces to an effective model for the visible spins, which should be easier to sample compared to the original problem. For example, one can randomly flip each disconnected component separated by the inactive hidden units if $E(\mathbf{s}) = 0$ and $\mathcal{F}_\alpha(\mathbf{s}) = \mathcal{F}_\alpha(-\mathbf{s})$. Alternatively, one can build another BM to simplify the sampling of Eq. (10) given the hidden units. One can even apply this idea iteratively and build a hierarchy of BMs.

Next, it is important to choose appropriate features $\mathcal{F}_\alpha(\mathbf{s})$ in Eq. (10) such that the BM correctly reproduces the target physical distribution. A good feature design is likely to exploit the knowledge of the original physical problem. Alternatively, one can start with a general BM architecture with many common features, such as links or plaquettes, and adjust their corresponding weight and bias parameters to maximize the efficiency of the proposed Monte Carlo updates. In this way, one translates the structure learning of BM into a more tractable parameter learning problem. In general, it is not possible for one to find out the optimal parameters of the BM (10) analytically as we did in this Rapid Communication. However, one can readily adopt fully fledged machine learning algorithms to carry out the BM parameter learning from data. For example, one can perform either unsupervised learning [21,56] or supervised learning of the physical distribution $\pi(\mathbf{s})$ based on the Monte Carlo data [9]. Ultimately, we anticipate a reinforcement learning [57] approach which directly searches for an optimal update policy in the algorithmic space parametrized by the BM (10).

In closing, we note many cluster quantum Monte Carlo algorithms [58–60] share the framework of Refs. [38,45,46]. The generalization of the hidden units to higher integers or even continuous variables is likely to increase the capacity of the BM. One can include higher-order self-interactions of the hidden variables in Eq. (10) in this case. To this end, these BMs provide concrete parametrizations of valid Monte Carlo update policies which can be optimized through learning. This approach opens a promise of discovering practically useful Monte Carlo algorithms for a broad range of problems, such as frustrated magnets or correlated fermions where known efficient cluster updates are rare (see Appendix B). Learning BM parameters from data are particularly useful for fermionic problems [9,61–63] because of their Monte Carlo weights involving nonlocal fermion determinants which cannot be handled analytically.

Acknowledgments. L.W. thanks L. Huang and Y.-f. Yang for collaborations on Refs. [9,12] and acknowledges J. Chen, Y. Deng, J. Liu, Y. Qi, Z.-Y. Meng, and T. Xiang for useful discussions. L.W. thanks Li Huang for comments on the manuscript and an anonymous referee for comments on the proof of Eq. (1). This work is supported by the Ministry of Science and Technology of China under the Grant No. 2016YFA0302400 and National Natural Science Foundation of China under Grant No. 11774398. The simulation is performed at Tianhe-2 supercomputer in National Supercomputer Center in Guangzhou. We use the ALPS library [64] for the Monte Carlo data analysis.

APPENDIX A: DETAILED BALANCE CONDITION OF EQ. (1)

The acceptance probability of the recommender update from the restricted Boltzmann machine is derived in Ref. [9]. We repeat the derivation for the BM considered in the main text for the convenience of the readers.

First of all, the Metropolis-Hastings [1,35] acceptance rate of the physical model satisfies

$$A(\mathbf{s}\to\mathbf{s}') = \min\left[1, \frac{T(\mathbf{s}'\to\mathbf{s})}{T(\mathbf{s}\to\mathbf{s}')}\frac{\pi(\mathbf{s}')}{\pi(\mathbf{s})}\right]. \tag{A1}$$

The transition probability is determined by the simulation of the BM, where we sample alternatingly between the hidden and visible units, i.e., $T(\mathbf{s}\to\mathbf{s}') = \sum_{\mathbf{h}} F_{\mathbf{h}}(\mathbf{s}\to\mathbf{s}')p(\mathbf{h}|\mathbf{s})$. Here, $F_{\mathbf{h}}$ denotes the update of the visible units given the hidden variables $\mathbf{h}$. The condition on $F_{\mathbf{h}}$ is that it respects the joint probability distribution of the BM for the given hidden units $\mathbf{h}$,

$$F_{\mathbf{h}}(\mathbf{s}\to\mathbf{s}')p(\mathbf{s},\mathbf{h}) = F_{\mathbf{h}}(\mathbf{s}'\to\mathbf{s})p(\mathbf{s}',\mathbf{h}). \tag{A2}$$

In Ref. [9] we have used $F_{\mathbf{h}}(\mathbf{s}\to\mathbf{s}') = p(\mathbf{s}'|\mathbf{h})$, which is a factorized probability distribution in the restricted BM, while for the general BMs consider in this Rapid Communication, we adopted more sophisticated updates such as a cluster flip of the visible units. These cluster updates also satisfy Eq. (A2) because they keep the energy $E(\mathbf{s},\mathbf{h})$ unchanged.

Using Eq. (A2), the ratio of the transition probability satisfies

$$\begin{aligned}\frac{T(\mathbf{s}\to\mathbf{s}')}{T(\mathbf{s}'\to\mathbf{s})} &= \frac{\sum_{\mathbf{h}} F_{\mathbf{h}}(\mathbf{s}\to\mathbf{s}')p(\mathbf{s},\mathbf{h})}{\sum_{\mathbf{h}} F_{\mathbf{h}}(\mathbf{s}'\to\mathbf{s})p(\mathbf{s}',\mathbf{h})}\frac{p(\mathbf{s}')}{p(\mathbf{s})}\\ &= \frac{p(\mathbf{s}')}{p(\mathbf{s})}.\end{aligned} \tag{A3}$$

Substituting Eq. (A3) into the Metropolis-Hastings acceptance probability (A1), we obtain Eq. (1) in the main text.

APPENDIX B: BOLTZMANN MACHINES FOR FRUSTRATED SPIN MODELS

We discuss applications of the general framework (10) outlined in the main text to more challenging problems of frustrated spins. As a concrete example, we consider the fully frustrated Ising model (FFIM) [65] and note that other frustrated systems such as the antiferromagnetic Ising model on the triangular lattice can be treated similarly. FFIM is defined on a square lattice where all plaquettes are frustrated. When dividing the lattice into corner-sharing plaquettes, each plaquette consists of three ferromagnetic and one antiferromagnetic coupling as illustrated in Fig. 5. The Boltzmann weight reads

$$\pi(\mathbf{s}) = \exp\left(\beta J\sum_{\wp}\mathcal{F}_{\wp}(\mathbf{s})\right), \tag{B1}$$

where $\mathcal{F}_{\wp}(\mathbf{s}) = s_1s_2 + s_1s_3 + s_2s_4 - s_3s_4$ is a feature defined for each corner-sharing plaquette, as shown in Fig. 5. Assuming $J>0$, the interaction between s_3 and s_4 is thus antiferromagnetic while the other three interactions are ferromagnetic. The FFIM model is challenging for the conventional Swendsen-Wang cluster algorithm at low temperature [3] because the cluster extends to the whole lattice.

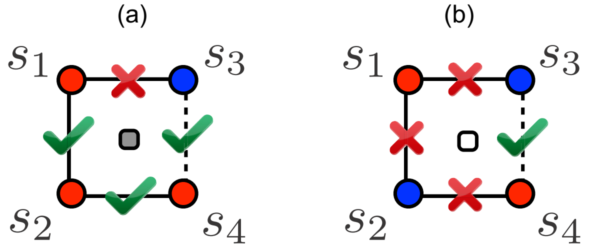

FIG. 5. A plaquette of the fully frustrated Ising model. The solid and dashed lines indicate ferromagnetic and antiferromagnetic coupling, respectively. The red and blue dots denote up and down Ising spins. (a) A plaquette configuration with three satisfied bonds and one unsatisfied bond, and $\mathcal{F}_{\wp}(\mathbf{s}) = 2$. (b) A plaquette configuration with three unsatisfied bonds and one satisfied bond, and $\mathcal{F}_{\wp}(\mathbf{s}) = -2$. The gray/white squares in the plaquette center denote the hidden units in the active/inactive status.

Specifying the general BM in Eq. (10) to a BM with hidden units coupled to the plaquette features, we have

$$E(\mathbf{s},\mathbf{h}) = -\sum_{\wp}[W\mathcal{F}_{\wp}(\mathbf{s}) + b]h_{\wp}. \tag{B2}$$

One can sample the BM (B2) efficiently following the general recipe outlined in the main text. First, one samples the hidden units for each plaquette according to the conditional probability $p(h_{\wp}=1|\mathbf{s}) = \sigma(W\mathcal{F}_{\wp}(\mathbf{s})+b)$. Next, for a given set of hidden units, one is free to update the visible Ising spins under the condition (A2). A valid approach is to keep the energy of the BM (B2) unchanged. Since an inactive hidden unit $h_{\wp}=0$ vanishes the contribution of the corresponding plaquette in the BM energy function, one only needs to pay attention to those plaquettes with active hidden units $h_{\wp}=1$. For each of these active plaquettes one can freeze two satisfied bonds if $F_{\wp}(\mathbf{s}) = 2$, while one freezes two unsatisfied bonds if $F_{\wp}(\mathbf{s}) = -2$. This keeps the energies of these active plaquettes unchanged. Finally, randomly flipping each disconnected component formed by these frozen bonds conserves the energy of the BM (B2).

The above sampling strategy of the BM suggests a nonlocal update of the FFIM with acceptance rate (1), where $p(\mathbf{s}) = \prod_{\wp}(1+e^{W\mathcal{F}_{\wp}(\mathbf{s})+b})$. Similar to the examples discussed in the main text, the BM exactly captures the FFIM model probability distribution (B1) when the condition $(1+e^{b+2W})/(1+e^{b-2W}) = e^{4\beta J}$ holds. One therefore obtains a family of rejection-free cluster Monte Carlo updates. In the limit of $b\to-\infty$, the BM update reproduces the algorithm of Refs. [66,67], which is known to be efficient for FFIM at low temperature.

Without using prior knowledge to specify the BM structure of Eq. (B2), one can start with a general BM with hidden units coupled to various typical features such as links and plaquettes. One can set different adjustable weights and biases for each connection. Optimizing these parameters with respect to a cost function representing the efficiency of the recommended Monte Carlo updates will search in the parametrized algorithm space. In such a way, one translates the BM structure design into a parameter learning task.

[1] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

[2] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, Phys. Lett. B **195**, 216 (1987).

[3] R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **58**, 86 (1987).

[4] U. Wolff, Phys. Rev. Lett. **62**, 361 (1989).

[5] H. G. Evertz, G. Lana, and M. Marcu, Phys. Rev. Lett. **70**, 875 (1993).

[6] N. Prokof'ev and B. Svistunov, Phys. Rev. Lett. **87**, 160601 (2001).

[7] V. Ambegaokar and M. Troyer, Am. J. Phys. **78**, 150 (2010).

[8] E. P. Bernard and W. Krauth, Phys. Rev. Lett. **107**, 155704 (2011).

[9] L. Huang and L. Wang, Phys. Rev. B **95**, 035105 (2017).

[10] J. Liu, Y. Qi, Z. Y. Meng, and L. Fu, Phys. Rev. B **95**, 041101(R) (2017).

[11] J. Liu, H. Shen, Y. Qi, Z. Y. Meng, and L. Fu, Phys. Rev. B **95**, 241104(R) (2017).

[12] L. Huang, Y.-f. Yang, and L. Wang, Phys. Rev. E **95**, 031301(R) (2017).

[13] X. Y. Xu, Y. Qi, J. Liu, L. Fu, and Z. Y. Meng, Phys. Rev. B **96**, 041119(R) (2017).

[14] R. M. Neal, *Bayesian Learning for Neural Networks* (Springer, Berlin, 1996).

[15] C. E. Rasmussen, J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, in *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, edited by J. M. Bernardo *et al.* (Clarendon, Oxford, UK, 2003), pp. 651–659.

[16] J. S. Liu, *Monte Carlo Strategies in Scientific Computing* (Springer, Berlin, 2008).

[17] F. Liang, C. Liu, and R. Carroll, *Advanced Markov Chain Monte Carlo methods: Learning from Past Samples*, Wiley Series in Computational Statistics Vol. 714 (Wiley, Hoboken, NJ, 2011).

[18] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, Cognit. Sci. **9**, 147 (1985).

[19] G. E. Hinton and T. J. Sejnowski, in *Parallel Distributed Processing*, edited by D. E. Rumelhart, J. L. McClelland, and PDP Research Group (MIT Press, Cambridge, MA, 1986), Vol. 1, pp. 282–317.

[20] G. E. Hinton and R. R. Salakhutdinov, Science **313**, 504 (2006).

[21] G. Torlai and R. G. Melko, Phys. Rev. B **94**, 165134 (2016).

[22] Y. Freund and D. Haussler, in *Proceeding NIPS'91: Proceedings of the 4th International Conference on Neural Information Processing Systems* (Morgan Kaufmann Publishers, San Francisco, CA, 1994), pp. 912–919.

[23] N. Le Roux and Y. Bengio, Neural Comput. **20**, 1631 (2008).

[24] G. Montufar and N. Ay, Neural Comput. **23**, 1306 (2011).

[25] G. F. Montúfar, J. Rauh, and N. Ay, in *Advances in Neural Information Processing Systems*, edited by M. I. Jordan (MIT Press, Cambridge, MA, 2011), pp. 415–423.

[26] J. Martens, A. Chattopadhya, T. Pitassi, and R. Zemel, in *Advances in Neural Information Processing Systems 26*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, New York, 2013), pp. 2877–2885.

[27] J. Chen, S. Cheng, H. Xie, L. Wang, and T. Xiang, arXiv:1701.04831.

[28] D.-L. Deng, X. Li, and S. Das Sarma, Phys. Rev. X **7**, 021021 (2017).

[29] X. Gao and L.-M. Duan, Nat. Commun. **8**, 662 (2017).

[30] Y. Huang and J. E. Moore, arXiv:1701.06246.

[31] G. Carleo and M. Troyer, Science **355**, 602 (2017).

[32] D.-L. Deng, X. Li, and S. Das Sarma, arXiv:1609.09060.

[33] G. Torlai and R. G. Melko, Phys. Rev. Lett. **119**, 030501 (2017).

[34] http://deeplearning.net/tutorial/rbm.html.

[35] W. K. Hastings, Biometrika **57**, 97 (1970).

[36] P. Smolensky, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, edited by D. E. Rumelhart, J. L. McClelland, and P. R. Group (MIT Press, Cambridge, MA, 1986), Vol. 1, pp. 194–281.

[37] F. Niedermayer, Phys. Rev. Lett. **61**, 2026 (1988).

[38] D. Kandel and E. Domany, Phys. Rev. B **43**, 8539 (1991).

[39] T. J. Sejnowski, in *Neural Networks for Computing*, edited by J. S. Denker, AIP Conf. Proc. Vol. 151 (AIP, New York, 1986), pp. 398–403.

[40] R. Memisevic and G. Hinton, in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2007), pp. 1–8.

[41] R. Memisevic and G. E. Hinton, Neural Comput. **22**, 1473 (2010).

[42] A. Krizhevsky and G. E. Hinton, in *2010 International Conference on Artificial Intelligence and Statistics* (IEEE, Piscataway, NJ, 2010), pp. 621–628.

[43] R. Sedgewick and K. D. Wayne, *Algorithms* (Addison-Wesley, Boston, 2011).

[44] J. Gubernatis, N. Kawashima, and P. Werner, *Quantum Monte Carlo Methods: Algorithms for Lattice Models* (Cambridge University Press, Cambridge, UK, 2016).

[45] N. Kawashima and J. E. Gubernatis, Phys. Rev. E **51**, 1547 (1995).

[46] D. M. Higdon, J. Am. Stat. Assoc. **93**, 585 (1998).

[47] P. W. Kasteleyn and C. M. Fortuin, J. Phys. Soc. Jpn. Suppl. **26**, 11 (1969).

[48] C. M. Fortuin and P. W. Kasteleyn, Physica **57**, 536 (1972).

[49] G. T. Barkema and J. F. Marko, Phys. Rev. Lett. **71**, 2070 (1993).

[50] R. Stratonovich, Sov. Phys. Dokl. **2**, 416 (1957).

[51] J. Hubbard, Phys. Rev. Lett. **3**, 77 (1959).

[52] J. E. Hirsch, Phys. Rev. B **28**, 4059(R) (1983).

[53] J. Salas and A. D. Sokal, J. Stat. Phys. **98**, 551 (2000).

[54] M. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics* (Oxford University Press, Oxford, UK, 1999).

[55] Introducing two set of hidden units coupled respectively to the features on links $\mathcal{F}_\ell(\mathbf{s}) = \Pi_{i\in\ell} s_i$ and on plaquette $\mathcal{F}_\wp(\mathbf{s}) = \Pi_{i\in\wp} s_i$ gives an alternative cluster algorithms for the plaquette Ising model Eq. (6). The resulting algorithm is a simple generalization of the Swendsen-Wang algorithm [3] to the case of the plaquette unit. However, this algorithm is less efficient than the one presented in the main text because it has larger average cluster sizes.

[56] G. E. Hinton, Neural Comput. **14**, 1771 (2002).

[57] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 1998).

[58] N. Kawashima and J. E. Gubernatis, J. Stat. Phys. **80**, 169 (1995).

[59] H. G. Evertz, Adv. Phys. **52**, 1 (2003).

[60] N. Kawashima and K. Harada, J. Phys. Soc. Jpn. **73**, 1379 (2004).

[61] H. Yokoyama and H. Shiba, J. Phys. Soc. Jpn. **56**, 1490 (1987).

[62] R. Blankenbecler, D. J. Scalapino, and R. L. Sugar, Phys. Rev. D **24**, 2278 (1981).

[63] S. Chandrasekharan, Phys. Rev. D **82**, 025007 (2010).

[64] B. Bauer *et al.*, J. Stat. Mech.: Theor. Exp. (2011) P05001.

[65] J. Villain, J. Phys. C **10**, 1717 (1977).

[66] D. Kandel, R. Ben-Av, and E. Domany, Phys. Rev. Lett. **65**, 941 (1990).

[67] D. Kandel, R. Ben-Av, and E. Domany, Phys. Rev. B **45**, 4700 (1992).