Editors' Suggestion

Solving Statistical Mechanics Using Variational Autoregressive Networks

Dian Wu,[1] Lei Wang,[2,3,4,*] and Pan Zhang[5,†]
[1]School of Physics, Peking University, Beijing 100871, China
[2]Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China
[3]CAS Center for Excellence in Topological Quantum Computation, University of Chinese Academy of Sciences, Beijing 100190, China
[4]Songshan Lake Materials Laboratory, Dongguan, Guangdong 523808, China
[5]Key Laboratory of Theoretical Physics, Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, China

(Received 8 November 2018; published 28 February 2019)

We propose a general framework for solving statistical mechanics of systems with finite size. The approach extends the celebrated variational mean-field approaches using autoregressive neural networks, which support direct sampling and exact calculation of normalized probability of configurations. It computes variational free energy, estimates physical quantities such as entropy, magnetizations and correlations, and generates uncorrelated samples all at once. Training of the network employs the policy gradient approach in reinforcement learning, which unbiasedly estimates the gradient of variational parameters. We apply our approach to several classic systems, including 2D Ising models, the Hopfield model, the Sherrington-Kirkpatrick model, and the inverse Ising model, for demonstrating its advantages over existing variational mean-field methods. Our approach sheds light on solving statistical physics problems using modern deep generative neural networks.

DOI: 10.1103/PhysRevLett.122.080602

Key words:
variational mean-field approaches(变分平均场近似),
autoregressive neural networks(自回归神经网络),
policy gradient approach(策略梯度算法：一种深度增强学习算法)

Consider a statistical physics model such as the celebrated Ising model, the joint probability of spins $\mathbf{s} \in \{\pm 1\}^N$ follows the Boltzmann distribution

$$p(\mathbf{s}) = \frac{e^{-\beta E(\mathbf{s})}}{Z}, \tag{1}$$

where $\beta = 1/T$ is the inverse temperature and Z is the partition function. Given a problem instance, *statistical mechanics* problems concern about how to estimate the free energy $F = -(1/\beta)\ln Z$ of the instance, how to compute macroscopic properties of the system such as magnetizations and correlations, and how to sample from the Boltzmann distribution efficiently. Solving these problems are not only relevant to physics, but also find broad applications in fields like Bayesian inference where the Boltzmann distribution naturally acts as posterior distribution, and in combinatorial optimizations where the task is equivalent to study zero temperature phase of a spin-glass model.

When the system has finite size, computing exactly the free energy belongs to the class of #P-hard problems, hence is in general intractable. Therefore, usually one employs approximate algorithms such as variational approaches. The variational approach adopts an *Ansatz* for the joint distribution $q_\theta(\mathbf{s})$ parametrized by variational parameters θ, and adjusts them so that $q_\theta(\mathbf{s})$ is as close as possible to the Boltzmann distribution $p(\mathbf{s})$. The closeness between two distributions is measured by Kullback-Leibler (KL) divergence [1]

KL散度/相对熵

$$D_{\mathrm{KL}}(q_\theta \| p) = \sum_{\mathbf{s}} q_\theta(\mathbf{s}) \ln\left(\frac{q_\theta(\mathbf{s})}{p(\mathbf{s})}\right) = \beta(F_q - F), \tag{2}$$

where

Variational inference/approximation

$$F_q = \frac{1}{\beta}\sum_{\mathbf{s}} q_\theta(\mathbf{s})[\beta E(\mathbf{s}) + \ln q_\theta(\mathbf{s})] \tag{3}$$

变分自由能

is the variational free energy corresponding to distribution $q_\theta(\mathbf{s})$. Since the KL divergence is non-negative, minimizing the KL divergence is equivalent to minimizing the variational free energy F_q, an upper bound to the true free energy F.

One of the most popular variational approaches, namely the variational mean-field method, assumes a factorized variational distribution $q_\theta(\mathbf{s}) = \prod_i q_i(s_i)$, where $q_i(s_i)$ is the marginal probability of the ith spin. In such parametrization, the variational free energy F_q can be expressed as an analytical function of parameters $q_i(s_i)$, as well as its derivative with respect to $q_i(s_i)$. By setting the derivatives to zero, one obtains a set of iterative equations, known as the *naïve mean-field* (NMF) equations. Despite its simplicity, NMF has been used in various applications in statistical physics, statistical inference, and machine learning [2,3]. Although NMF gives an upper bound to the physical free energy F, typically it is not accurate, since it completely ignores the correlation between variables. Other approaches, which essentially adopt different variational *Ansätze* for $q_\theta(\mathbf{s})$, have been developed to give better

estimate (although not always an upper bound) of the free energy. These *Ansätze*, including Bethe approximation [4,5], Thouless-Anderson-Palmer equations [6], and Kikuchi loop expansions [7], form a family of mean-field approximations [2].

However, on systems with strong interactions and on a factor graph with loops of different lengths (such as lattices), mean-field approximations usually give very limited performance. The major difficulty for the mean-field methods in this case is to give a powerful, yet tractable variation form of joint distribution $q_\theta(\mathbf{s})$. In this Letter, we generalize the existing variational mean-field methods to a much more powerful and general framework using autoregressive neural networks.

Variational autoregressive networks.—The recently developed neural networks give us ideal methods for parameterizing variational distribution $q_\theta(\mathbf{s})$ with a strong representational power. The key ingredient of employing them to solve statistical mechanics problem is to design neural networks such that the variational free energy [Eq. (3)] is efficiently computable. The method we adopted here is named *autoregressive networks*, where the joint probability of all variables is expressed as product of conditional probabilities [8–11]

why just use j < i spins?

$$q_\theta(\mathbf{s}) = \prod_{i=1}^{N} q_\theta(s_i|s_1, \ldots, s_{i-1}), \tag{4}$$

and the factors are parametrized as neural networks. We denote using Eq. (4) as an *Ansatz* for the variational calculation and Eq. (3) as a *variational autoregressive networks* (VAN) approach for statistical mechanics.

The simplest autoregressive network is depicted in Fig. 1(a), which is known as the *fully visible sigmoid belief network* [9]. The input of the network is a configuration $\mathbf{s} \in \{\pm 1\}^N$ with a predetermined order, and the output $\hat{s}_i = \sigma(\sum_{j<i} W_{ij} s_j)$ has the same dimension as the input. We see that the network is parametrized by a triangular matrix W, which ensures that $\hat{s}_i$ is independent with s_j when $j \geq i$. This is named as *autoregressive property* in machine learning literatures. The sigmoid activation function $\sigma(\cdot)$ ranges in (0,1), so we can expect that $\hat{s}_i$ represents a probability with proper normalization. Namely, $\hat{s}_i = q(s_i = +1|\mathbf{s}_{<i})$, which means the conditional probability of s_i being $+1$, given the configuration of spins in front of it, $\mathbf{s}_{<i}$, in the predetermined order of variables. Thus, given a configuration $\mathbf{s}$ as the input to the network, the joint distribution of the input variables can be expressed as the product of conditional probabilities, and each factor is a Bernoulli distribution $q(s_i|\mathbf{s}_{<i}) = \hat{s}_i \delta_{s_i,+1} + (1 - \hat{s}_i)\delta_{s_i,-1}$.

There have been many discussions in the machine learning community on how to make the autoregressive network deeper and more expressive, and how to increase the generalization power by sharing weights [10–14]. Using the simplest one-layer network as building blocks, we can design more complex and expressive networks, while preserving the autoregressive property. For example, we can add more layers of hidden variables to the network, as shown in Fig. 1(b).

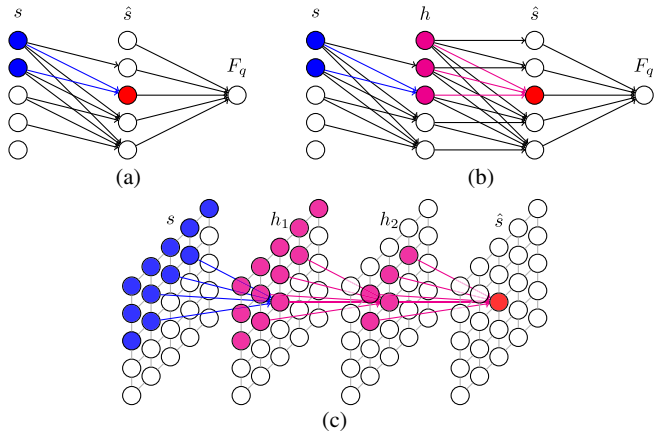

FIG. 1. Autoregressive networks with different architectures for variational free energy calculation. The spin configuration $\mathbf{s}$ is the input to the network, $\hat{\mathbf{s}}$ is the output of the network, and h denotes hidden layer. The loss function F_q is given by Eq. (3) and Eq. (4). The colored sites denote the receptive field of a site in $\hat{\mathbf{s}}$. (a) The network has only one layer, which is densely connected, while the autoregressive property hold. (b) The network has a hidden layer. (c) The network has masked convolution layers on 2D lattice. Only connections in a convolution kernel are shown for clarity.

CNN

When the system has structures, e.g., lying on a 2D lattice, a classic network architecture designed specifically for it is the convolutional network [8], which respects the locality and the translational symmetry of the system. To ensure the autoregressive property, one can put a mask on the convolution kernel, so that the weights are not zero only for half of the kernel, and $\hat{s}_i$ is independent of s_j with $j < i$ in the predetermined order. The receptive field of the masked convolution through multiple layers is shown in Fig. 1(c). This kind of structured autoregressive networks is known as PixelCNN [15], which has achieved state-of-the-art results in modeling and generating natural images. In additional, by using the dilated convolutions the autoregressive WaveNet [16] can capture long-range correlations in audio signals, and has achieved remarkable performance in real-world speech synthesis.

mask ? window

autoregressive property

PixelCNN

The autoregressive networks are one of the leading generative models that find wide applications under the general purpose of density estimations [15–17]. A key difference between our work and those machine learning applications is that for density estimation one trains the network from the training data using maximum likelihood estimation, i.e., minimizing the KL divergence between empirical training data distribution $p_{\text{data}}(\mathbf{s})$ and the network, $D_{\text{KL}}(p_{\text{data}}\|q_\theta)$. Whereas in our variational free

encoder-decoder

energy calculation, the goal is to reduce the *reversed* KL divergence $D_{\mathrm{KL}}(q_\theta \| p)$. Therefore, we train the network using data produced by itself. The only input of our calculation is the energy function of the statistical mechanics problem, and no training data from the target Boltzmann distribution is assumed.

The variational free energy in Eq. (3) can be regarded as a scalar loss function over the parameters θ of the autoregressive network of Eq. (4). A nice feature of autoregressive networks is that one can draw independent samples efficiently by sampling each variable in the predetermined order. Moreover, one has direct access to the normalized probability $q_\theta(\mathbf{s})$ of any given sample. Exploiting these properties, one can replace the summation over all possible configurations weighted by $q_\theta(\mathbf{s})$ by samplings from the network, and evaluate the entropy and energy terms respectively in Eq. (3). Thanks to the direct-sampling ability, the estimated variational free energy provides an exact upper bound to the true free energy of the model. 直接采样?

The gradient of the variational free energy with respect to network parameters reads [18]

$$\beta \nabla_\theta F_q = \mathbb{E}_{\mathbf{s}\sim q_\theta(\mathbf{s})}\{[\beta E(\mathbf{s}) + \ln q_\theta(\mathbf{s})]\nabla_\theta \ln q_\theta(\mathbf{s})\}. \quad (5)$$

We perform the stochastic gradient descent optimization on the parameters θ. Furthermore, we employ the control variates method of Ref. [24] to reduce the variance in the gradient estimator [18]. In the context of reinforcement learning [25], $q_\theta(\mathbf{s})$ is a stochastic policy which produces instances of $\mathbf{s}$, and the term in the square bracket of Eq. (3) is the reward signal. Thus, learning according to Eq. (5) amounts to the policy gradient algorithm. We note that the variational studies of quantum states [26] employ a similar gradient estimator. However, the variational autoregressive networks enjoy unbiased estimate of the gradient using efficient direct sampling instead of relying on the correlated Markov chains.

To the best of our knowledge, the variational framework using deep autoregressive networks for statistical mechanics has not been explored before. Our method can be seen as an extension to the variational mean-field methods with a more expressive variational *Ansatz*. Its representational power comes from recently developed deep neural networks with guarantee of *universal expressive power* [8]. Rather than a specific model, we consider our approach as a general framework, analogous to existing frameworks such as Markov chain Monte Carlo (MCMC), mean-field methods, and tensor networks [27,28]. When compared with existing frameworks, the features of VAN are these: giving an upper bound to the true free energy; efficiently generating independent samples without needing Markov chains, which is ideal for parallelization (on GPUs); and computing physical observables, such as the energy and correlations, using a sufficiently large amount of samples without any autocorrelations.

VAN: variational autoregressive network

Numerical experiments.—To demonstrate the ability of VAN in terms of accuracy of the variational free energy and estimated physical quantities, we perform experiments on Ising models. The energy of the configuration $\mathbf{s}$ is given by $E(\mathbf{s}) = -\sum_{\langle ij\rangle} J_{ij} s_i s_j$, with $\langle ij\rangle$ denoting pair of connections. With different choices of the coupling matrix J, we cover systems on different topologies: 2D square and triangular lattices, and fully connected systems. We also cover systems with different behaviors: ferromagnetic, antiferromagnetic, glassy, and as associative memory.

We first apply our approach to the ferromagnetic Ising model on 2D square lattice with periodic boundary condition, which admits an exact solution [29]. We have tested two types of network architectures, the 2D convolution (conv) and densely connected (dense), respectively, to verify that taking into account the lattice structure is beneficial. More details on the implementation are discussed in the Supplemental Material [18].

The free energy given by VAN, compared with NMF and Bethe approximations, is shown in Fig. 2(a). The figure shows that VAN significantly outperforms the two traditional methods. The maximum relative error is around the critical point, where the system develops long range correlations. Also, the network architecture with convolution layers performs significantly better than dense connection, since it respects the two-dimensional nature of the lattice, which is particularly beneficial when the correlation is short ranged. However, around criticality, they exhibits similar performance.

Then, we apply our approach to the frustrated antiferromagnetic Ising model on 2D triangular lattice with a periodic boundary condition. Figure 2(b) shows the entropy per site versus inverse temperature β for various lattice sizes. Reaching a finite entropy density indicates that the system processes an exponentially large number of degenerate ground states. Extrapolation of $\beta \to \infty$ shows that VAN correctly captures the exponentially large number of ground states. In comparison, describing such feature has

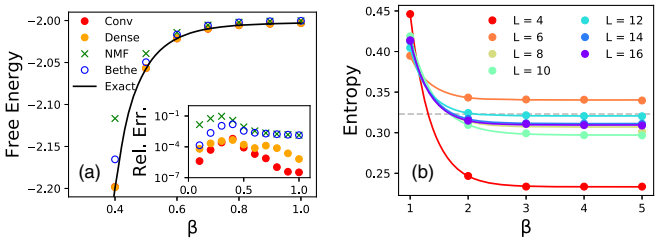

FIG. 2. (a) Free energy per site and its relative error of ferromagnetic Ising model on 16×16 square lattice with periodic boundary condition. (b) Entropy per site of antiferromagnetic Ising model on triangular lattices of various sizes L with periodic boundary condition. The exact result (dashed line) at $T = 0$ and $L \to \infty$ is $S/N = 0.323\,066$ [30,31]. The curves for $L = 8$, 14, 16 are almost overlapped.

been challenging to conventional MCMC and mean-field approaches.

Next, to demonstrate the ability of capturing multiple states at low temperature, we consider the Hopfield model [32], where N spins are connected to each other. The couplings composed of P random patterns, $J_{ij} = (1/N)\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu$, with $\{\xi^\mu\} \in \{\pm 1\}^N$ denoting a random pattern. At a low temperature with P small, the system has a retrieval phase where all P patterns are remembered by the system; hence there are P pure states in the system [33,34]. The experiments are carried out on a Hopfield network with $N = 100$ spins and $P = 2$ orthogonal random patterns. At low temperature the energy (probability) landscape contains four modes, corresponding to two stored patterns and their mirrors (due to $\mathbb{Z}_2$ symmetry). As opposed to models defined on lattices, there is no topology structure to apply convolution, so we use a simplest VAN with only one layer and $N(N-1)/2$ parameters. We start training our network at $\beta = 0.3$ and slowly anneal the temperature to $\beta = 1.5$. At each temperature, we sample configurations from the trained VAN, and show their log probability in Fig. 3.

The figure shows that at high temperature with $\beta = 0.3$, samplings are not correlated with the two stored patterns, and the system is in the paramagnetic state. The log probability landscape is quite flat, as the Gibbs measure is dominated by entropy. When β is increased to 1.5, four peaks of probability emerge and dominate over other configurations. These four peaks touch coordinates $[1, 0]$, $[0, 1]$, $[-1, 0]$, and $[0, -1]$ in the X-Y plane, which correspond exactly to the two patterns and their mirrors. This is an evidence that our approach avoids collapsing into a single mode, and gives samplings capturing the features of the whole landscape, despite that those modes are separated by high barriers.

Compared with the landscape of Hopfield model in the retrieval phase which exhibits several local minima in the energy and probability landscape, models in the spin glass phase are considerably more complex [35], because they have an infinite number of pure states, in the picture of replica symmetry breaking [36]. Here we apply our method to the classic Sherrington-Kirkpatrick (SK) model [37], where N spins are connected to each other by couplings J_{ij} drawn from Gaussian distribution with variance $1/N$. So far the tensor network approaches do not apply to this model because of long range interactions and the disorder, which causes negative Z issue [38]. On the thermodynamic limit with $N \to \infty$ where the free energy concentrates to its mean value averaged over disorder, using for example replica method and cavity method, and replica symmetry breaking, i.e., the Parisi formula [36]. On a single instance of SK model, the algorithm version of the cavity method, belief propagation, or Thouless-Anderson-Paler [6] equations apply as message passing algorithms. On large systems in the replica symmetry phase, the message passing algorithms converge and the obtained Bethe free energy is a good approximation, but in the replica symmetry breaking phase they fail to converge. Also notice that even in the replica symmetry phase, Bethe free energy is not an upper bound to the true free energy.

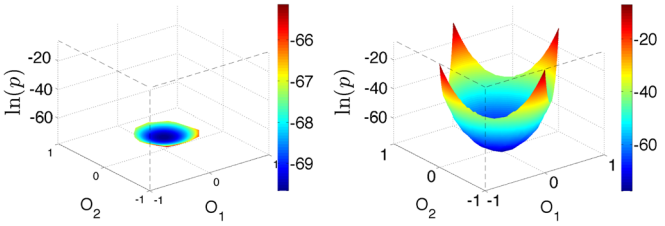

FIG. 3. Log probability of sampled configurations from VAN trained for a Hopfield model with $N = 100$ spins, and $P = 2$ orthogonal patterns. The sampled configurations are projected onto the two-dimensional space spanned by the two patterns. X axis (O_1) and Y axis (O_2) are the overlap (inner product, normalized to $[-1, 1]$) between each sampled configuration and the two patterns, respectively. (a) $\beta = 0.3$, and the system is in the paramagnetic phase. (b) $\beta = 1.5$, and the system is in the retrieval phase. Note the different scales in the color bars.

As a proof of concept, we use a small system size $N = 20$, so we can enumerate all 2^N configurations, compute the exact value of free energy, then evaluate the performance of our approach. Again, we use a simple VAN with only one layer.

In Fig. 4(a) we show the free energy obtained from VAN, compared with NMF and Bethe approximations. The free energy from VAN is much better than NMF and Bethe, and even indistinguishable to the exact value. This is quite remarkable considering that VAN adopts only $N(N-1)/2$ parameters, which is even smaller than that used in the belief propagation, $N(N-1)$. We also checked that our approach not only gives a good estimate on free energy, it also obtains accurate energy, entropy, magnetizations, and correlations.

The ability of solving *ordinary* statistical mechanics problems also gives us the ability to solve *inverse* statistical mechanics problems. A prototype problem is the *inverse*

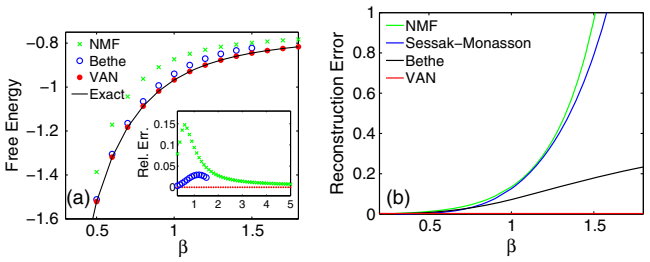

FIG. 4. (a) Free energy of SK model with $N = 20$ spins. The inset shows relative errors to exact values in a larger β regime. Bethe converges only when $\beta \leq 1.5$. (b) The reconstruction error in the inverse Ising problem. The underlying model is an SK model with $N = 20$ spins. VAN uses a network with two layers (a hidden layer and an output layer).

Ising problem, which asks us to reconstruct the couplings of an Ising (spin glass) model, given the correlations [18]. It is well known that the Ising model is the maximum entropy model given the first and the second moments, so the couplings are uniquely determined by the correlations. The problem has been studied for a long time especially in the field of statistical mechanics [39], mainly using mean-field based methods.

The adaptation of our method for the inverse problem is straightforward by repeating the following two steps, until the correlations given by VAN are close enough to the given correlations of the underlying model: (1) train a VAN according to the Ising model with an existing J_{ij} by minimizing the variational free energy; (2) compute correlations via direct sampling from the VAN, then update J_{ij} to minimize the difference between the two sets of correlations. We use our approach to reconstruct an SK model with $N = 20$ spins, and the given correlations are computed exactly by enumerating all 2^N configurations. The VAN uses two layers with 2000 parameters. The results are shown in Fig. 4(b). Our method works much better than the popular mean-field methods of naïve mean-field [40,41], Sessak-Monasson small-correlation expansions [42], and those based on a Bethe approximation [43,44], especially in the glassy phase with $\beta > 1$.

Outlooks.—In the present Letter, we have focused on binary spins. However, it is straightforward to generalize the approach to Potts models and models with continuous variables. We also notice that, for continuous variables and with a regular structure, a flow-based model together with a renormalization group has been proposed for the variational free energy minimization problem [45]. For systems defined on a 2D lattice, we have shown how to adopt convolutions for respecting the 2D structure of the underlying factor graph [15]. This strategy can be extended straightforwardly to systems on 3D lattices using 3D convolutions, and to graphical models on an arbitrary factor graph using, e.g., graph convolution networks [46] with proper filters.

We anticipate that our method will find immediate applications in a broad range of disciplines. For example, it can be applied directly to statistical inference problems, where the Boltzmann distribution in statistical mechanics becomes the posterior distribution of Bayesian inference [47]. Another example of application would be the combinatorial optimizations and constraint satisfaction problems, in which finding the optimal configurations and solutions correspond to finding ground states of spin glasses, and counting the number of solutions corresponds to computing entropy at zero temperature.

So far our approach is rather a proof of concept of a promising variational framework on statistical physics problems. Building on the current work, an interesting direction for future work would be even more deeply incorporating successful physics and machine learning concepts (such as a renormalization group and dilated convolution) into the network architecture design, e.g., the WaveNet [16]. This would allow us to scale to much larger problem size, or even to the thermodynamic limit.

The main limitation of our method is that the variational free energy calculation relies on sampling of the model; hence it is slower than canonical variational mean-field message passing algorithms, which compute variational free energy directly using model parameters. We also notice that the sampling process can be sped up by caching intermediate activations in the sampling procedure as explored in Refs. [48,49]. Or, one may use alternative model such as inverse autoregressive flow [50], which supports parallel sampling.

A pytorch implementation of our model and algorithms is available at Ref. [51].

We thank Zhiyuan Xie and Haijun Zhou for discussions. L. W. is supported by the Ministry of Science and Technology of China under the Grant No. 2016YFA0300603, the National Natural Science Foundation of China under the Grant No. 11774398, and the Strategic Priority Research Program of Chinese Academy of Sciences Grant No. XDB28000000. P. Z. is supported by Key Research Program of Frontier Sciences, CAS, Grant No. QYZDB-SSW-SYS032 and Project 11747601 of National Natural Science Foundation of China.

*wanglei@iphy.ac.cn
†panzhang@itp.ac.cn

[1] D. J. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, England, 2003).
[2] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, Mach. Learn. **37**, 183 (1999).
[3] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag, New York, 2006).
[4] H. A. Bethe, Proc. R. Soc. A **150**, 552 (1935).
[5] J. S. Yedidia, W. T. Freeman, and Y. Weiss, Understanding belief propagation and its generalizations, in *Exploring Artificial Intelligence in the New Millennium*, edited by G. Lakemeyer and B. Nebel (Morgan Kaufmann Publishers Inc., San Francisco, 2003), pp. 239–369.
[6] D. J. Thouless, P. W. Anderson, and R. G. Palmer, Philos. Mag. **35**, 593 (1977).
[7] R. Kikuchi, Phys. Rev. **81**, 988 (1951).
[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016).
[9] B. J. Frey, *Graphical Models for Machine Learning and Digital Communication* (MIT Press, Cambridge, MA, 1998).
[10] B. Uria, M.-A. Côté, K. Gregor, I. Murray, and H. Larochelle, J. Mach. Learn. Res. **17**, 7184 (2016).
[11] M. Germain, K. Gregor, I. Murray, and H. Larochelle, in *International Conference on Machine Learning, Lille Paris* (2015), pp. 881–889, http://proceedings.mlr.press/.

[12] Y. Bengio and S. Bengio, in *Advances in Neural Information Processing Systems, Denver, CO, USA*, edited by edited by T.K. Leen, T. G. Dietterich, and V. Tresp (2000), pp. 400–406.
[13] H. Larochelle and I. Murray, in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA* (2011), pp. 29–37, http://proceedings.mlr.press/.
[14] K. Gregor, I. Danihelka, A. Mnih, C. Blundell, and D. Wierstra, arXiv:1310.8499.
[15] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, in *International Conference on Machine Learning, New York City, NY, USA* (2016), pp. 1747–1756, http://proceedings.mlr.press/.
[16] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, in Speech Synthesis Workshops, Sunnyvale, CA, USA (2016), p. 125, http://ssw9.talp.cat/.
[17] G. Papamakarios, T. Pavlakou, and I. Murray, arXiv:1705.07057.
[18] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.122.080602 for (1) an introduction to autoregressive networks; (2) the derivation of the gradient estimator and the variance reduction trick; (3) discussions on the zero variance condition; (4) backgrounds of the inverse Ising problem; (5) more results on the Sherrington-Kirkpatrick model, the Hopfield model, and the inverse SK model; (6) heat capacity and critical temperature for the Ising model; and (7) details on the network structure and the training process. The Supplemental Material includes Refs. [19–23].
[19] R. J. Williams, Mach. Learn. **8**, 229 (1992).
[20] M. Mezard and T. Mora, J. Physiol. Paris **103**, 107 (2009).
[21] K. He, X. Zhang, S. Ren, and J. Sun, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778.
[22] D. A. Moore, in *NIPS Workshop on Advances in Approximate Bayesian Inferece, Barcelona, Spain* (2016).
[23] D. P. Kingma and J. Ba, arXiv:1412.6980.
[24] A. Mnih and K. Gregor, arXiv:1402.0030.
[25] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 1998).
[26] G. Carleo and M. Troyer, Science **355**, 602 (2017).
[27] M. Levin and C. P. Nave, Phys. Rev. Lett. **99**, 120601 (2007).
[28] Z. Y. Xie, H. C. Jiang, Q. N. Chen, Z. Y. Weng, and T. Xiang, Phys. Rev. Lett. **103**, 160601 (2009).
[29] L. Onsager, Phys. Rev. **65**, 117 (1944).
[30] G. H. Wannier, Phys. Rev. **79**, 357 (1950).
[31] G. H. Wannier, Phys. Rev. B **7**, 5017 (1973).
[32] J. J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).
[33] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **32**, 1007 (1985).
[34] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. Lett. **55**, 1530 (1985).
[35] M. Mézard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*, Vol. 9 (World Scientific, Singapore, 1987).
[36] G. Parisi, J. Phys. A **13**, 1101 (1980).
[37] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).
[38] C. Wang, S.-M. Qin, and H.-J. Zhou, Phys. Rev. B **90**, 174201 (2014).
[39] H. C. Nguyen, R. Zecchina, and J. Berg, Adv. Phys. **66**, 197 (2017).
[40] H. J. Kappen and F. de Borja Rodríguez Ortiz, Neural Comput. **10**, 1137 (1998).
[41] Y. Roudi, J. Tyrcha, and J. Hertz, Phys. Rev. E **79**, 051915 (2009).
[42] V. Sessak and R. Monasson, J. Phys. A **42**, 055001 (2009).
[43] H. C. Nguyen and J. Berg, J. Stat. Mech. (2012) P03004.
[44] F. Ricci-Tersenghi, J. Stat. Mech. (2012) P08015.
[45] S.-H. Li and L. Wang, Phys. Rev. Lett. **121**, 260601 (2018).
[46] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, IEEE Signal Process. Mag. **34**, 18 (2017).
[47] L. Zdeborová and F. Krzakala, Adv. Phys. **65**, 453 (2016).
[48] T. L. Paine, P. Khorrami, S. Chang, Y. Zhang, P. Ramachandran, M. A. Hasegawa-Johnson, and T. S. Huang, arXiv:1611.09482.
[49] P. Ramachandran, T. L. Paine, P. Khorrami, M. Babaeizadeh, S. Chang, Y. Zhang, M. A. Hasegawa-Johnson, R. H. Campbell, and T. S. Huang, arXiv:1704.06001.
[50] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, arXiv:1606.04934.
[51] https://github.com/wdphy16/stat-mech-van.