Solving Statistical Mechanics using Variational Autoregressive Networks

Dian Wu,[1] Lei Wang,[2, *] and Pan Zhang[3, †]

[1]*School of Physics, Peking University, Beijing 100871, China*
[2]*Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China*
[3]*Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, China*

We propose a general framework for solving statistical mechanics of systems with a 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. The network 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 classical systems, including 2-d Ising models, Hopfield model, Sherrington–Kirkpatrick spin glasses, 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.

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

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

where $\beta = 1/T$ is the inverse temperature and Z is the partition function. *Statistical mechanics* problems concern about how to estimate the free energy $F = -\frac{1}{\beta} \ln Z$ of the given 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 $p(\mathbf{s})$. Usually the closeness between two probabilities is measured by Kullback–Leibler (KL) divergence [1]

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

with

$$\beta F_q = \sum_{\mathbf{s}} q_\theta(\mathbf{s}) \left[\beta E(\mathbf{s}) + \ln q_\theta(\mathbf{s}) \right] \qquad (3)$$

is the variational free energy given by 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, which is an upperbound to the true free energy F, with respect to the parameters θ.

One of the most popular variational approach, namely the variational mean-field method originates from statistical physics, 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 single spin i. 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 nonlinear equations which are known as the *naïve mean-field* (NMF) equations in statistical physics. Despite its simplicity, NMF approximations have been used in various applications in statistical physics, statistical inference and machine learning [2, 3]. Although NMF gives an upper bound to the true free energy F, typically it is not accurate since it completely ignores the correlation between variables. Many approaches, which essentially adopt different variational formula (or ansatzs) for the joint distributions, have been developed to give better estimate (although not always an upper bound) of the free energy. These form a family of mean-field approximations [2], including Bethe approximation [4, 5], Thouless–Anderson–Palmer equations [6], and Kikuchi loop expansions [7].

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 paper, 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 models 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 (3) is efficiently computable. The models we adopted here is named *autoregressive networks*, where the joint probability of all variables is expressed as product of *conditional probabilities* [8–11]

$$q_\theta(\mathbf{s}) = \prod_{i=1}^{N} q_\theta(s_i \mid s_1, \ldots, s_{i-1}), \qquad (4)$$
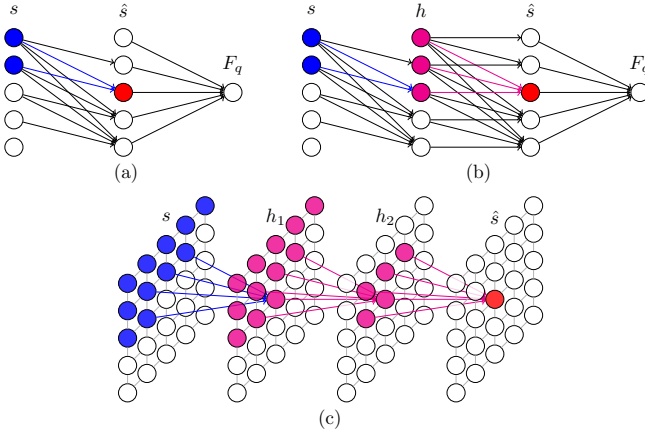
Figure 1. Autoregressive networks with different architectures for variational free energy calculation. The spin configuration $\mathbf{s}$ is the input to the network, $\hat{\mathbf{s}}$ is the output of the network, and h denotes hidden layer. The loss function F_q is given by Eq. (3) and Eq. (4). 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 2-d lattice. Only connections in a convolution kernel are shown for clarity.

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

The simplest autoregressive network is depicted in Fig. 1(a), which known as the *fully visible sigmoid belief net* [9]. The input of the network is a configuration $\mathbf{s} \in \{\pm1\}^N$ with a predetermined order, and the output $\hat{s}_i = \sigma\left(\sum_{j<i} W_{ij}s_j\right)$ 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, $\hat{s}_i = q(s_i = +1 \mid \mathbf{s}_{<i})$, 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 $q(s_i \mid \mathbf{s}_{<i}) = \hat{s}_i\delta_{s_i,+1} + (1 - \hat{s}_i)\delta_{s_i,-1}$.

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 model as building blocks, we can design more complex and expressive models, while preserving the autoregressive property. For example, we can add more layers of hidden variables to the network, as shown in Fig. 1(b).

When the system has structures, e.g. lying on a 2-d lattice, a classic network architecture designed specifically for it is

the convolutional networks [8], which respect 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 model is known as the 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.

The autoregressive models are one of the leading generative models which 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 learns the autoregressive model from training data using maximum likelihood estimation, which minimizes the KL divergence between empirical training data distribution $p_{\text{data}}(\mathbf{s})$ and the model, $D_{\text{KL}}(p_{\text{data}} \| q_\theta)$. Whereas in our variational free energy calculation, the goal is to reduce the *reversed* KL divergence $D_{\text{KL}}(q_\theta \| p)$. Therefore, we train the model using data produced by its own. 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 Eq. (4). A nice feature of autoregressive models is that one can draw independent samples from the joint distribution efficiently by sampling each variable in the predetermined order. Moreover, one have 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 model distribution $q_\theta(\mathbf{s})$, 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.

Moreover, the gradient of the variational free energy with respect to model parameters reads [18]

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

In practice, we evaluate Eq. (5) by drawing a batch of independent samples directly from the variational distribution, and estimate the quantity in the square bracket. Furthermore, we employ the control variates method of [19] to reduce the variance in the gradient estimator [18]. In the context of reinforcement learning [20], $q_\theta(\mathbf{s})$ is a stochastic policy which produce instances of $\mathbf{s}$, and the term in the square bracket of (3) is the reward signal. Thus, learning according to Eq. (5) amounts to the policy gradient algorithm. We note that the variational Monte Carlo studies of quantum states [21] employ a similar gradient estimator for variational parameters. However, the variational autoregressive networks enjoy unbiased estimate

of the gradient using efficient direct sampling instead relying on the correlated Markov chains. We perform the stochastic gradient descent optimization of the parameters θ using the gradient information.

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, and tensor networks [22, 23]. When compared with existing frameworks, the features of our method are: giving an upper bound to the true free energy; efficiently generating independent samples without needing Markov chain, and is ideal for parallelization (on GPUs); computing physical observables, such as the energy or correlations, using a sufficiently large amount of samples without any auto-correlations.

Numerical experiments To demonstrate ability of the VAN approach in terms of accuracy of variational free energy and quality of sampling, we perform experiments on the classic prototype for statistical mechanics problem, the Ising model, where the energy function of configuration **s** is given by $E(\mathbf{s}) = -\sum_{(ij)} J_{ij} s_i s_j$, with (ij) denoting pair of connections. With different choices of J, we cover systems on different topologies: 2-d square and triangular lattices, fully connected systems; as well as systems with different behaviors: ferromagnetic, anti-ferromagnetic, glassy, and as associative memory.

We first apply our approach to the ferromagnetic Ising model on 2-d square lattice with periodic boundary condition, which admits exact solution [24]. We have tested two types of network architectures, the 2-d convolution (Conv) and densely connected (Dense) respectively, to verify our assumption that taking into account the lattice structure is beneficial. More details on the implementation are described in appendices.

The relative error of the free energy given by the autoregressive networks, NMF, and Bethe approximation are shown in Fig. 2(a). The figure shows that deep autoregressive networks significantly improve the accuracy of the variational calculation. The maximum relative error is around the critical point, where the system develops long range correlations. We have observed that 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.

Next, we apply the variational approach to the frustrated antiferromagnetic Ising model on 2-d triangular lattice with periodic boundary condition. The Ising spins does not order even at the ground state due to frustration. Fig. 2(b) shows the entropy per site versus inverse temperature β for various lat-
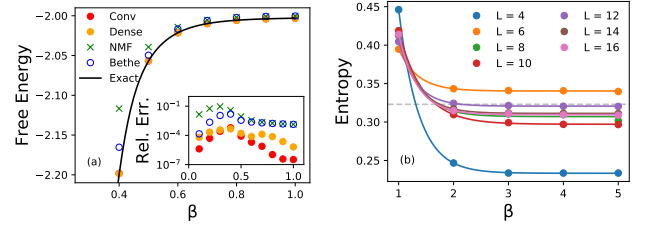

Figure 2. (a) Free energy per site and its relative error of ferromagnetic Ising model on 16×16 square lattice with periodic boundary condition. (b) Entropy per site of anti-ferromagnetic Ising model on triangular lattices of various sizes L with periodic boundary condition. The exact result (dashed line) at $T = 0$ and $L \to \infty$ is $S/N = 0.323066$ [25, 26].

tice 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 the autoregressive network variational ansatz correctly captures the exponentially large number of ground states. In comparison, describing such feature has been challenging to conventional MCMC and mean-field approaches.

Next, to demonstrate the ability of the model on capturing multiple modes at low temperature, we consider the Hopfield model [27], which is a generalized Ising model with couplings composed of P random patterns, $J_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu$, 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 [28, 29]. 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 4 modes, corresponding to 2 stored patterns and their mirrors (due to $\mathbb{Z}_2$ symmetry). We start training our model at $\beta = 0.3$ and slowly anneal the temperature to $\beta = 1.5$. At each step, we collect configurations sampled from the trained autoregressive network and show their log probability $\ln(p)$ as a surface plot in Fig. 3. The sampled configurations are projected into a two-dimensional space spanned by two stored patterns, thus in the figures X and Y-axes are overlap (inner product, normalized to the range $[-1, 1]$) between the configuration and two patterns respectively. Z-axis is the log probability.

The figure shows that at an high temperature with $\beta = 0.3$, samplings are not correlated with the two stored patterns, the system is in the paramagnetic state. The log probability landscape is quite flat, as the Gibbs measure is dominated by entropy. While when the inverse temperature β is increased to 1.5, we can see clearly from the right panel of the Fig. 3 that four peaks of probabilities completely emerge, having dominating probability with respect to 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
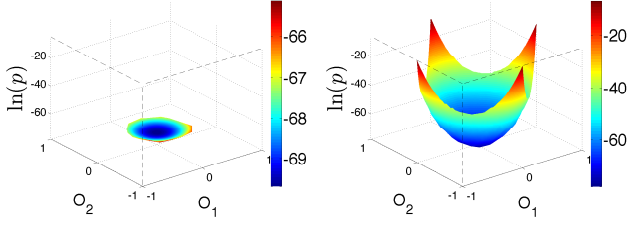
Figure 3. Log probability of sampled configurations from VAN learned for an Hopfield model with $N = 100$ spins, and $P = 2$ orthogonal patterns, on the two-dimensional space spanned by the two patterns. X-axis (O_1) and Y-axis (O_2) represent inner product (overlap) between each sampled configuration and the first and the second stored patterns respectively. Our network uses single layer and only $N(N-1)/2$ parameters. In the left panel, $\beta = 0.3$, system is in the paramagnetic phase. And in the right panel, $\beta = 1.5$ and system is in the retrieval phase. Note the different scales in the colorbars.
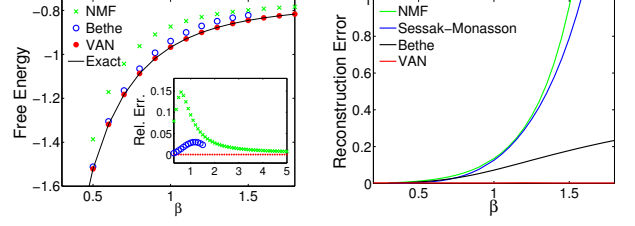

Figure 4. (*Left*:) Comparison of the free energy obtained using different methods, on an SK model with $N = 20$ spins. VAN uses single layer network with $N(N-1)/2$ parameters. The inset shows relative errors to exact values in a larger β regime. (*Right*:) The reconstruction error in the inverse Ising problem given by different methods. The underlying model is an SK model with $N = 20$ spins. Data correlation used for reconstruction are computed exactly by enumerating all 2^N configurations. VAN uses a two-layer autoregressive network with 2050 parameters.

the features of the whole landscape despite that those modes are separated by high barriers due to the first-order transition in the Hopfield model.

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 [30], because they have infinite number of pure states, in the picture of replica symmetry breaking [31]. Here we apply the developed method to the classic Sherrington–Kirkpatrick [32] model 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 [33]). 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. Parisi formula [31]. 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 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.

As a proof of concept, we do experiments on a SK model with a small system size $N = 20$, because in this way we can enumerate all 2^N configurations, compute the exact values of free energy, then evaluate the performance of our approach. As opposed to models defined on lattices, there is no topology structure to apply convolution. We choose to use the simplest autoregressive network with only one layer and total number of parameters equals $N(N-1)/2$, which is even smaller than that used in the belief propagation, $N(N-1)$.

In Fig. 4(a) we show the obtained free energy compared

against naïve mean-field (NMF), and Bethe approximation via iterating the belief propagation equations (BP) [5]. In our experiments, BP stopped converging at the spin glass phase when $\beta > 1.0$, then we used a large damping factor to force it to converge. However, in this instance damping only works up to $\beta = 1.5$, so we report Bethe free energy with $\beta \le 1.5$. From the figure we can see that the free energy of our method is much better than mean-field and Bethe free energy, and are even indistinguishable to the exact free energies. This is quite remarkable considering that our model adopts only $N(N-1)/2$ parameters. We also checked that our approach not only gives a good estimate on free energy, energy and entropy, it also obtains accurate 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 Ising problem* which asks to reconstruct couplings and external field of an Ising (spin glass) model given magnetizations and correlations of the underlying model [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 correlations of the model which are given as data. The problem has been studied for a long time especially in the field of statistical mechanics [34], mainly using mean-field based methods.

The adaptation of our method for the inverse problem is straight forward by repeating the following two-step procedure until the correlations of VAN are close enough to the (given) correlations of the underlying Ising model: (1) Learn a VAN according the Ising model with an existing J_{ij} by minimizing the variational free energy; (2) Compute pairwise correlations of VAN via direct sampling, then update J_{ij} according to the difference between two sets of correlations. Our results are shown in the Fig. 4 right, where we can see that our method works much better than the popular mean-field methods of naïve mean-field [35, 36], Sessak–Monasson small-correlation expansions [37], and that based on Bethe approximation [38, 39], especially in the glassy phase with $\beta > 1$.

Outlooks In the present paper we have focused on binary spins. However, it is straightforward to generalize the approach to Potts models and models with continuous variables. For systems defined on a 2-d lattice we have shown how to adopt convolutions for respecting 2-d structure of the underlying factor graph [15]. This strategy can be extended straightforwardly to systems on 3-d lattices using 3-d convolutions, and to graphical models on an arbitrary factor graph using e.g. graph convolution networks [40] with proper filters.

We anticipate that our application 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 [41]. Another example of application would be the combinatorial optimizations and constraint satisfaction problems, in which finding the optimal configurations and solutions corresponds 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, a fascinating direction for future work would be even more deeply incorporating successful physics and machine learning concepts (such as renormalization group and dilated convolution) into the network architecture design, e.g. the WaveNet [16]. This would allow to scale to much larger problem size, or even to the thermodynamic limit.

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

[1] D. J. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, 2003).
[2] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, Machine Learning **37**, 183 (1999).
[3] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag New York, Inc., 2006).
[4] H. A. Bethe, Proc. R. Soc. Lond. A **150**, 552 (1935).
[5] J. S. Yedidia, W. T. Freeman, and Y. Weiss, in *International Joint Conference on Artificial Intelligence* (2001).
[6] D. J. Thouless, P. W. Anderson, and R. G. Palmer, Philosophical Magazine **35**, 593 (1977).
[7] R. Kikuchi, Phys. Rev. **81**, 988 (1951).
[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016).
[9] B. J. Frey, *Graphical Models for Machine Learning and Digital Communication* (MIT Press, 1998).
[10] B. Uria, M.-A. Côté, K. Gregor, I. Murray, and H. Larochelle, The Journal of Machine Learning Research **17**, 7184 (2016).
[11] M. Germain, K. Gregor, I. Murray, and H. Larochelle, in *International Conference on Machine Learning* (2015) pp. 881–889.
[12] Y. Bengio and S. Bengio, in *Advances in Neural Information Processing Systems* (2000) pp. 400–406.
[13] H. Larochelle and I. Murray, in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (2011) pp. 29–37.
[14] K. Gregor, I. Danihelka, A. Mnih, C. Blundell, and D. Wierstra, in *International Conference on Machine Learning* (2014) pp. 1242–1250.
[15] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, in *International Conference on Machine Learning* (2016) pp. 1747–1756.
[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* (2016) p. 125.
[17] G. Papamakarios, T. Pavlakou, and I. Murray, (2017), arXiv:1705.07057.
[18] See supplementary materials for derivation of the gradient estimator and variance reduction trick, details of the network and training, and more discussions on the inverse Ising problem.
[19] A. Mnih and K. Gregor, (2014), arXiv:1402.0030.
[20] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, 1998).
[21] G. Carleo and M. Troyer, Science **355**, 602 (2017).
[22] M. Levin and C. P. Nave, Phys. Rev. Lett. **99**, 120601 (2007).
[23] Z. Y. Xie, H. C. Jiang, Q. N. Chen, Z. Y. Weng, and T. Xiang, Phys. Rev. Lett. **103**, 160601 (2009).
[24] L. Onsager, Phys. Rev. **65**, 117 (1944).
[25] G. H. Wannier, Phys. Rev. **79**, 357 (1950).
[26] G. H. Wannier, Phys. Rev. B **7**, 5017 (1973).
[27] J. J. Hopfield, Proc. Natl. Acad. Sci. USA **79**, 2554 (1982).
[28] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **32**, 1007 (1985).
[29] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. Lett. **55**, 1530 (1985).
[30] 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 Publishing Company, 1987).
[31] G. Parisi, J. Phys. A **13**, 1101 (1980).
[32] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).
[33] C. Wang, S.-M. Qin, and H.-J. Zhou, Phys. Rev. B **90**, 174201 (2014).
[34] H. C. Nguyen, R. Zecchina, and J. Berg, Advances in Physics **66**, 197 (2017).
[35] H. J. Kappen and F. de Borja Rodríguez Ortiz, Neural Computation **10**, 1137 (1998).
[36] Y. Roudi, J. Tyrcha, and J. Hertz, Phys. Rev. E **79**, 051915 (2009).
[37] V. Sessak and R. Monasson, J. Phys. A **42**, 055001 (2009).
[38] F. Ricci-Tersenghi, J. Stat. Mech. **2012**, P08015 (2012).
[39] H. C. Nguyen and J. Berg, J. Stat. Mech. **2012**, P03004 (2012).
[40] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, IEEE Signal Processing Magazine **34**, 18 (2017).
[41] L. Zdeborová and F. Krzakala, Advances in Physics **65**, 453 (2016).
[42] R. J. Williams, Machine Learning **8**, 229 (1992).
[43] H. J. Kappen and F. de Borja Rodríguez Ortiz, in *Advances in Neural Information Processing Systems* (1998) pp. 280–286.

[44] M. Mezard and T. Mora, Journal of Physiology-Paris **103**, 107 (2009).

[45] 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.

[46] D. A. Moore, in *NIPS Workshop on Advances in Approximate Bayesian Inferece* (2016).

[47] S.-H. Li and L. Wang, (2018), arXiv:1802.02840.

[48] D. P. Kingma and J. Ba, (2014), arXiv:1412.6980.

Gradient estimator and variance reduction

The gradient of the variational free energy (3) can be written as

$$\beta \nabla_\theta F_q = \nabla_\theta \sum_{\mathbf{s}} \left[q_\theta(\mathbf{s}) \cdot (\beta E(\mathbf{s}) + \ln q_\theta(\mathbf{s})) \right]$$

$$= \sum_{\mathbf{s}} \left[\nabla_\theta q_\theta(\mathbf{s}) \cdot (\beta E(\mathbf{s}) + \ln q_\theta(\mathbf{s})) + q_\theta(\mathbf{s}) \nabla_\theta \ln q_\theta(\mathbf{s}) \right]$$

$$= \mathbb{E}_{\mathbf{s} \sim q_\theta(\mathbf{s})} \left[\nabla_\theta \ln q_\theta(\mathbf{s}) \cdot \underbrace{(\beta E(\mathbf{s}) + \ln q_\theta(\mathbf{s}))}_{R(\mathbf{s})} \right], \quad (6)$$

which is Eq. (5) in the main texts. The above equation is easy to understand: the contribution of each sample $\mathbf{s} \sim q_\theta(\mathbf{s})$ to the gradient is $\nabla_\theta \ln q_\theta(\mathbf{s})$ weighted by the reward signal $R(\mathbf{s})$. This ensures that when $R(\mathbf{s})$ is large, the optimizer will try to reduce the probability of generating such configuration, hence reduce the variational free energy. Learning probability distribution using the score function gradient estimator (5) is also known as the REINFORCE algorithm [42] in reinforcement learning [20].

We also notice that in deriving the last equation we have used

$$\mathbb{E}_{\mathbf{s} \sim q_\theta(\mathbf{s})} [\nabla_\theta \ln q_\theta(\mathbf{s})] = \nabla_\theta \sum_{\mathbf{s}} q_\theta(\mathbf{s}) = \nabla_\theta 1 = 0. \quad (7)$$

For the same reason, one can subtract any $\mathbf{s}$-independent constant in the last equation without affecting the expectation, that is

$$\nabla_\theta F_q = \frac{1}{\beta} \mathbb{E}_{\mathbf{s} \sim q_\theta(\mathbf{s})} \left[\nabla_\theta \ln q_\theta(\mathbf{s}) \cdot (R(\mathbf{s}) - b) \right]. \quad (8)$$

The bias b is useful to reduce the variance of the gradient, and is known as *variance reduction* in the context of the reinforcement learning learning literature [8]. In this work we consider only a very simple strategy [19] by setting the baseline to

$$b = \mathbb{E}_{\mathbf{s} \sim q_\theta(\mathbf{s})} R(\mathbf{s}). \quad (9)$$

Zero variance condition and exact free energy

The variational free energy (3) is an estimator over the variational distribution,

$$F_q = \mathbb{E}_{\mathbf{s} \sim q_\theta(\mathbf{s})} \left[E(\mathbf{s}) + \frac{1}{\beta} \ln q_\theta(\mathbf{s}) \right]. \quad (10)$$
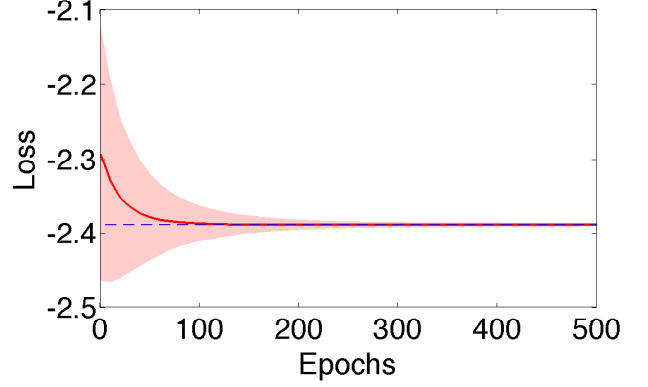

Figure 5. Evolution of mean and variance of loss function during training process of VAN on an SK model with $n = 20$ spins, $\beta = 0.3$. The red area denotes variance, the red line in the middle denotes mean Eq. (10), and the blue dashed line denotes the exact free energy value of the SK model. The VAN uses 2 layers and totally 8200 trainable parameters.

We can see that if exact learning is achieved, i.e. $q_\theta(\mathbf{s}) = p_{\text{Boltzmann}}(\mathbf{s})$, we have

$$E(\mathbf{s}) + \frac{1}{\beta} q_\theta(\mathbf{s}) = -\frac{1}{\beta} \log(Z) \quad (11)$$

which means that the quantity $E(s) + \frac{1}{\beta} \ln q_\theta(\mathbf{s})$ has *zero variance*.

On the other hand, if the variance is zero, then the distribution q_θ must be a Boltzmann distribution. To prove this, we make use that zero variance implies the quantity in the square bracket to be a constant,

$$E(\mathbf{s}) + \frac{1}{\beta} \ln q_\theta(\mathbf{s}) = C. \quad (12)$$

Solving the equation gives

$$q_\theta(\mathbf{s}) = e^{\beta C - \beta E(\mathbf{s})}. \quad (13)$$

However, notice that the $e^{\beta C}$ does not necessarily be equal to $1/Z$, the normalization of the original Boltzmann distribution, due to *modal collapse* where not all the modes (pure states) of the original Boltzmann distribution are captured by our model.

Nevertheless, if one can ensure that modal collapse never happens, then low variance indeed indicates a good estimate to the true free energy. In this work we propose to use temperature annealing to avoid modal collapse, and our results in Fig. 3 gave an evidence that modal collapse does not happen. Therefore, we can use the variance to practically indicate the closeness between q_θ and the exact distribution without knowing the latter. As an example to illustrate this, in Fig. 5 we plot evolution of F_q from Eq. (10) and the variance during training of a VAN on an SK model with $n = 20$ spins. The figure shows that when the variance (the area) decreases during

training, the variational free energy (the red line) converges to the true free energy (blue dashed line).

Inverse Ising problem and conventional mean-field methods

It is well known that the Ising model

$$p(\mathbf{s}) = \frac{1}{Z} e^{\sum_{ij} J_{ij} s_i s_j + \sum_i h_i s_i} \tag{14}$$

is the maximum entropy model when the first and second moment of the distribution $p(\mathbf{s})$ are constrained. The inverse Ising model asks to reconstruct couplings $\{J_{ij}\}$ and external field $\{h_i\}$ of the underlying Ising model when the magnetizations $\{m_i\}$ and correlations $\{C_{ij}\}$ are given, where

$$m_i = \sum_{\mathbf{s}} p(\mathbf{s}) s_i, \tag{15}$$

$$C_{ij} = \sum_{\mathbf{s}} p(\mathbf{s}) s_i s_j. \tag{16}$$

The maximum likelihood inference gives a simple condition that the magnetization $m(\mathbf{h}, \mathbf{J})$ and correlation $C(\mathbf{h}, \mathbf{J})$ of the learned model should match the given magnetization and correlation. If they do not match, the difference between two quantities provide gradient for learning the external fields and couplings.

The main difficulty of reconstruction is that computing exact correlations and magnetizations of learned model (i.e. with leaned external fields and couplings) are intractable. Various of mean-field methods have been proposed for estimating the modal correlations. In our method, we estimate correlations and external fields using configurations sampled from the learned autoregressive networks, thanks to efficient direct sampling of our model.

In this paper, we consider models with no external field, thus the task is to reconstruct couplings from correlations. To avoid influence of measurement noise in correlation data, we test in small systems and compute exact correlations by enumerating all the configurations. In this way, the magnetization is apparently exactly zero, due to the $\mathbb{Z}_2$ symmetry. We compare performance of methods against to several well-known mean-field methods. These include naïve mean-field method (NMF), Sessak–Monasson small correlation expansion method and Bethe approximation. In the NMF, the correlations are computed using naïve mean-field approximation and linear response relation [43], and coupling are given by

$$J_{ij}^{\text{NMF}} = \delta_{ij} - (C^{-1})_{ij}. \tag{17}$$

In the Sessak–Monasson small correlation expansion, a perturbation expansion of entropy in terms of the connected correlation is carried out, and the reconstructed couplings are given by

$$J_{ij}^{\text{SM}} = -(C^{-1})_{ij} + J_{ij}^{\text{IP}} - \frac{C_{ij}}{1 - C_{ij}^2}, \tag{18}$$

where

$$J_{ij}^{\text{IP}} = \frac{1}{4} \ln \left[\frac{(1 + C_{ij})^2}{(1 - C_{ij})^2} \right] \tag{19}$$

is known as independent-pair approximation. The Bethe approximation [38, 39] is rather simple,

$$J^{\text{Bethe}} = \frac{1}{2} \operatorname{arcsinh} \left[2(C^{-1})_{ij} \right]. \tag{20}$$

despite its relation to the susceptibility propagation [44] which considers both belief propagation and linear response relation have been explored in [44]. We refer to [34] for an overview of these mean-field methods.

After all, the performance of reconstruction is characterized by the reconstruction error between the inferred couplings J_{infer} and true couplings J_{true}, defined as

$$\Delta_J = \frac{1}{N} \sqrt{\sum_{ij} \left(J_{ij}^{\text{infer}} - J_{ij}^{\text{true}} \right)^2}. \tag{21}$$

Details on network structure and training for Ising model

For 2-d Ising model, we set the lattice size to be 16×16, and specify the network's depth (the number of layers) and width (the number of channels in a layer). We test convolution layers and densely connected layers respectively. For convolution layers, we specify the kernel radius (kernel radius $\times 2 + 1$ = edge length of the kernel). To cover a lattice with edge length L, depth and kernel radius should satisfy

$$\text{depth} \times \text{kernel radius} + 1 \geq L. \tag{22}$$

We test a "shallow" network with depth = 3, and a "deep" network with depth = 6 and residue blocks [45]. The result shown in the main text is chosen according to a lower free energy between them. In practice, we find that the "shallow" one gives lower free energy for high temperature, otherwise the "deep" one works better. We set width = 64, because it is the elbow point when we plot the relative error of the free energy versus the number of parameters. The values of depth, width and kernel radius are summarized in Table I. The settings ensure the number of parameters in those networks are within the same magnitude, and capable under our computation resources.

	Shallow	Deep
Conv	3/64/6	6/64/3
Dense	3/4/−	6/2/−

Table I. Depth/width/kernel radius of employed networks.

To implement $\mathbb{Z}_2$ symmetry, we create a mixture model of the network and itself with input inversed. The probability of the configuration $\mathbf{s}$ is $q_{\mathbb{Z}_2}(\mathbf{s}) = \frac{1}{2}(q_\theta(\mathbf{s}) + q_\theta(-\mathbf{s}))$ [46, 47],

where $q_\theta(\mathbf{s})$ is the probability given by the network in (4). In sampling, we first generate a batch of samples from the network, then randomly inverse them by probability 1/2.

We use the conventional Adam optimizer [48] to minimize the variational free energy. To avoid mode collapse, we start training at infinite temperature ($\beta = 0$), and slowly increase β until the desired temperature is reached. Moreover, we clip the norm of the gradient to increase the stability of training.

The code is avaliable at https://github.com/wdphy16/stat-mech-van