Solving statistical mechanics on sparse graphs with feedback-set variational autoregressive networks

Feng Pan [1,2] Pengfei Zhou [1,2] Hai-Jun Zhou [1,2,*] and Pan Zhang [1,3,4,†]

[1]*CAS Key Laboratory for Theoretical Physics, Institute of Theoretical Physics,*
Chinese Academy of Sciences, Beijing 100190, China
[2]*School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China*
[3]*School of Fundamental Physics and Mathematical Sciences, Hangzhou Institute for Advanced Study, UCAS, Hangzhou 310024, China*
[4]*International Centre for Theoretical Physics Asia-Pacific, Beijing/Hangzhou, China*

We propose a method for solving statistical mechanics problems defined on sparse graphs. It extracts a small feedback vertex set (FVS) from the sparse graph, converting the sparse system to a much smaller system with many-body and dense interactions with an effective energy on every configuration of the FVS, then learns a variational distribution parametrized using neural networks to approximate the original Boltzmann distribution. The method is able to estimate free energy, compute observables, and generate unbiased samples via direct sampling without autocorrelation. Extensive experiments show that our approach is more accurate than existing approaches for sparse spin glasses. On random graphs and real-world networks, our approach significantly outperforms the standard methods for sparse systems, such as the belief-propagation algorithm; on structured sparse systems, such as two-dimensional lattices our approach is significantly faster and more accurate than recently proposed variational autoregressive networks using convolution neural networks.

I. INTRODUCTION

The underlying graphs of interacting systems are often quite sparse, and these include physics models defined on two-dimensional (2D) or three-dimensional (3D) lattices which encode physical interactions; social systems on networks which represent social relations [1]; mathematical and information systems on sparse factor graphs which induce constraints over variables as in the satisfiability problems [2], and in the low density error correcting codes [3,4]. The *statistical mechanics* problems defined on such systems are of great importance: they can, e.g., characterize phases and phase transitions in finite-dimensional physics model, describe dynamics of human behaviors, formulate posterior distribution in Bayesian inference, compute number of solutions of constraint satisfaction problems, and conduct message-passing algorithms for reconstructing signals in error-correcting codes, etc. Without loss of generality, in this paper we consider the prototype problem of statistical mechanics with n spins sitting on a sparse graph $\mathcal{G}$. The joint probability distribution of a configuration $\mathbf{s} \in \{+1, -1\}^n$ follows the Boltzmann distribution,

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

where Z denotes the partition function, β denotes the inverse temperature, and E is the energy function. It is well known that computing the partition function or, equivalently, the free energy,

$$F = -\frac{1}{\beta} \ln Z = -\frac{1}{\beta} \ln \sum_{\mathbf{s}} e^{-\beta E(\mathbf{s})} \tag{2}$$

belong to the class of #P complete problems, so it is hopeless to find polynomial algorithms to solve it, in general. In statistical physics, one usually applies approximate (mean-field) methods for estimating an upper bounds of the free energy, or subjects to sampling methods which generate samples and calculate observables based on the samples.

The sparse systems (e.g., on social networks) we are interested in are usually large so that the Markov-chain Monte Carlo (MCMC) based sampling methods are considered to be not scalable. For systems with *random sparsity*, such as on random graphs and real-world large networks, variational approximations based on the Bethe approximation [5], i.e., belief propagation (BP) [6], have been widely applied. The key idea in the Bethe approximation, the conditional independence assumption, is exact when the system is a tree, and is usually a good approximation in random sparse systems. However, the approximation does not perform well when the system contains many short loops. Approximations, such as Kikuchi loop expansions [7] have been proposed to partly take into account the effects of short loops, however, it is still far from being optimal when the system has loops with different lengths, such as lattices, which we call *structured sparse* systems. In this kind of systems, tensor network based renormalization and block decimation approaches are very powerful, but unfortunately they work poorly if system does not contain many intrinsic low-rank (or low-entanglement)

*zhouhj@itp.ac.cn
†panzhang@itp.ac.cn

structures, and they do not apply when system contains long-range interactions.

Recently a variational method using neural networks, the variational autoregressive network (VAN), is proposed for general statistical mechanics problems [8]. On small dense systems, VAN uses multilayered neural networks to describe a variational distribution and achieves much better performance in estimating free energy, obtaining observables, and in sampling, when compared with traditional variational mean-field methods. However, the VAN contains a huge number of parameters to optimize, hence, is not designed for large sparse systems at the first place. On structured sparse systems, such as (2D lattices, the issue can be eased by adopting convolution neural networks [9] as a good structure prior. But on random sparse systems it is still not clear how to propose efficient convolutionlike neural network structures for variational computations.

In this paper we extend the VAN approach to statistical mechanics problems on sparse graphs. Our approach converts a large sparse system to an equivalent but much smaller system composed of the vertices in a feedback vertex set (FVS) of the graph [10,11]. After the conversion, the reduced system is densely connected with many-body interactions with an effective energy on each configuration of FVS vertices. We then apply the VAN method to solve the statistical mechanics problem on the reduced system.

At a first glance, we convert a sparse system to a strongly correlated dense system which seems even harder than the original one. Indeed it is in the sense that the converted system loses all the internal sparse structures of the original system. But note that the key advantage of this converting is that the recently proposed neural-network based variational methods are particularly suitable for treating a small nonstructured dense system. We will show in what follows that the VAN method can utilize recently developed machine learning techniques, such as reinforcement learning, as well as advanced computational techniques, such as GPUs, to target the converted problem on the compressed dense graph and significantly outperforms existing methods operating on the original sparse system.

The rest of the paper is organized as follows. In Sec. II A we introduce the feedback set of sparse graphs and how to convert the statistical mechanics problems on a sparse graph to the problem defined on its feedback set. In Sec. II B we explain how to solve the converted statistical mechanics problem using variational autoregressive networks. In Sec. III we conduct extensive experiments on spin glasses and inference problems to evaluate our approach against mean-field methods, VAN with convolution neural networks, and the recently proposed expectation propagation based density consistency method. We conclude in Sec. IV.

II. FEEDBACK SET VARIATIONAL AUTOREGRESSIVE NETWORKS

Our method is composed of two steps, the first step is converting the statistical mechanics problem in a sparse graph to a problem with many-body interactions in a much smaller dense graph, using an approach based on the feedback set of the sparse graph which assigns each configuration of the feed-

back set an effective energy; the second step of our method is applying the recently proposed autoregressive variational network using effective energies obtained in the first step to solve the statistical mechanics on the dense graph. In the following text we will introduce the two steps in detail.

A. Compressing the sparsity using the feedback vertex set

A feedback vertex set Ω of a graph is a set of vertices intersecting with every loop of the graph. With all vertices of the FVS removed, the graph becomes loop free, hence, is composed of trees [10]. We denote the forest after removing all the vertices of the FVS as $\mathcal{T}$. The free energy of the whole system Eq. (2) can be rewritten in the following form by tracing over FVS variables and non-FVS variables separately,

$$F = -\frac{1}{\beta} \ln \sum_{\mathbf{s} \in \Omega} \sum_{\mathbf{t} \in \mathcal{T}} e^{-\beta E(\mathbf{s}, \mathbf{t})} = -\frac{1}{\beta} \ln \sum_{\mathbf{s} \in \Omega} e^{-\beta \widetilde{E}(\mathbf{s})}, \quad (3)$$

where $\mathbf{s} \in \{+1, -1\}^{|\Omega|}$ denotes an assignment of variables on FVS, and $\mathbf{t} \in \{+1, -1\}^{|\mathcal{T}|} = \{+1, -1\}^{n-|\Omega|}$ denotes an assignment of variables in the forest out of the FVS. $E(\mathbf{s}, \mathbf{t})$ denotes the energy of the original sparse system, and $\widetilde{E}(\mathbf{s})$ is the effective energy of a configuration $\mathbf{s}$ of the FVS vertices. The effective energy can be written as

$$\widetilde{E}(\mathbf{s}) = E_\Omega(\mathbf{s}) + E_\mathcal{T}(\mathbf{s}). \quad (4)$$

Here $E_\Omega(\mathbf{s})$ denotes the sum of interaction energies (of the original sparse system) involving only vertices of the FVS, and

$$E_\mathcal{T}(\mathbf{s}) = -\frac{1}{\beta} \ln \left(\sum_{\mathbf{t} \in \mathcal{T}} e^{-\beta E(\mathbf{s}, \mathbf{t})} \right) \quad (5)$$

is the (free) energy of the forest $\mathcal{T}$ given the configuration $\mathbf{s}$ of the FVS.

Observe that, once a configuration $\mathbf{s}$ of the FVS is determined, what is left in the graph is a forest with boundary conditions fixed by $\mathbf{s}$. As a consequence, the summation over all variables in the forest as shown in Eq. (5) can be calculated in linear time straightforwardly, i.e., by taking a sequence from leaf vertices (a leaf vertex of the graph is a vertex which has only one attached edge) of the forest to the root vertices. This approach is known as "conditioning" in graphical model community [12]. A detailed example about calculating free energy of the Ising model based on this method can be found in Appendix B.

There are many possible FVS one can find for the graph $\mathcal{G}$. Finding the smallest one, which is called the minimum feedback vertex set problem, belongs to the class of *NP-hard* problems. In this paper we employ an heuristic algorithm called CoreHD, proposed in Ref. [11], which is a fast and simple heuristic method with $O(n)$ computation complexity. The sizes of the FVS for the graph instances studied in this paper are about 20%–33% of the original graph sizes. We note that it is possible to further reduce the FVS sizes through a more advanced message-passing method [10].

Equation (3) converts the partition-function calculation of a sparse system with size n to that of a system of much smaller size $|\Omega|$. The conversion comes with a price: the effective energy of the FVS indicates that there are complicated

many-body interactions, thus, good approximations on sparse graphs (e.g., Bethe approximation) fail completely in the FVS system. Indeed it is really difficult to find good approximations or variational Ansatz for this dense and many-body interacting system, in general, so we adopt a neural network based method.

B. Variational autoregressive networks for statistical mechanics on the FVS

The variational autoregressive network [8] has been proposed recently for statistical mechanics problems, in general, and it is particularly powerful for densely connected structureless problems. Analogous to the mean-field methods, the VAN maintains a variational distribution $q_\theta(\mathbf{s})$ and gradually updates parameters θ of the distribution to minimize the variational free energy. What is different from the canonical mean-field methods is that in the VAN, the variational distribution is represented using product of conditional probabilities,

$$q_\theta(\mathbf{s}) = \prod_{i=1}^{n} q_\theta(s_i | s_1, \dots, s_{i-1}), \tag{6}$$

with all the conditional distributions parametrized by a neural network. The advantage of the autoregressive distribution as well as the neural network representation is the representation power. Theoretically it is able to represent any Boltzmann distribution given enough parameters in the hidden layers of the neural network, known as the *representation theorem* [13] so is much more representative compared with mean-field Ansätze, such as Bethe and Thouless-Anderson-Palmer [14]. However the price that VAN pays for the representation power is that the variational free energy is no longer analytically computable as opposed to mean-field methods where the variational free energies and their derivatives over parameters can always be written out analytically. As a solution, the VAN generates many unbiased samples from the variational distribution with exact probability (known as direct sampling in Ref. [15]), which is easily computable using Eq. (6), then estimates the variational free energy using

$$F_q = \sum_s q_\theta(s) \left[\widetilde{E}(s) + \frac{1}{\beta} \ln q_\theta(s). \right] \tag{7}$$

Once the variational free energy is computed, in VAN we compute its gradients with respect to parameters θ using

$$\nabla_\theta F_q = \sum_s q_\theta(s) \left\{ \left[\widetilde{E}(s) + \frac{1}{\beta} \ln q_\theta(s) \right] \nabla_\theta \ln q_\theta(s) \right\}, \tag{8}$$

which is known as the REINFORCE algorithm in the machine learning literature [16]. A short introduction to VAN and the learning algorithm can be found in Appendix A (see also Ref. [8] for more details). Using the gradients, we update parameters of the VAN gradually until the variational free energy stops decreasing. After that, the learned VAN can be used to generate a massive number of configurations unbiasedly and parallelly using the direct sampling, and then observables, such as magnetizations and correlations can be computed straightforwardly using the configurations.

We refer to the algorithm as FVSVAN, it takes the sparse graph $\mathcal{G}$ and energy function E as inputs, then outputs the variational free energy and unbiased samples after the learning is finished. Given $\mathcal{G}$ and E, the FVSVAN is sketched as follows:

(1) Extract a FVS Ω from the graph $\mathcal{G}$ using the CoreHD algorithm [11].

(2) Construct VAN to express the variational distribution $q_\theta(\mathbf{s})$ for $\mathbf{s} \in \Omega$, initialize neural networks in VAN randomly.

(3) Do direct sampling to obtain unbiased samples $\{\mathbf{s}\}$.

(4) Compute effective energy $\widetilde{E}(\mathbf{s})$, probability $q(\mathbf{s})$ for each sample $\mathbf{s}$ using Eqs. (4) and (6), respectively, then compute variational free energy F_q according to Eq. (7).

(5) Compute gradients of the variational free energy with respect to parameters according to Eq. (8), then update the parameters (i.e., weights of neural network) of the variational distribution using the gradients.

(6) Go to step 3. if the variational free energy has not converged.

III. NUMERICAL EXPERIMENTS

To evaluate the ability of our approach in estimating the free energy and in computing physical quantities in sparse systems, and to evaluate our method against existing approaches, we perform extensive experiments on sparse graphs using spin glasses and inference problems.

There are two kinds of sparsities we usually encounter in statistical mechanics problems: random sparsity in, e.g., random graphs and real-world sparse networks, and structured sparsity in, e.g., 2D and 3D lattices. For random graphs and large real-world networks, the Bethe approximation based methods, such as the belief propagation are considered to be standards approaches. A recently proposed approach, density consistency (DC) [17], which is based on the expectation propagation [18] algorithm, has been shown to outperform BP in calculating physical quantities, so we also take this approach as a baseline for comparison in our experiments. On the lattice systems, recently proposed convolution based VAN [8] takes the structure prior to 2D graphs by adapting the convolution neural networks and largely outperforms mean-field approximations, so we mainly focus on the comparison to the VAN with convolution neural networks operating on the whole lattice. In our experiments we consider two kinds of problems. The first one is a representative model of spin glasses where the tasks include estimating the tree energy and computing the correlation functions; the second problem we consider is an inference problem on sparse graphs known as the *censored stochastic block model* [19] where the task is to solve the posterior distribution (which is the Boltzmann distribution) in the Bayesian inference and infer the planted partition using observations on each edge of the graph.

A. Spin glasses on random graphs

We first consider the Ising $\pm J$ spin glasses on random graphs, the so-called Viana-Bray spin glass model [20] with distribution of couplings following $P(J_{ij} = 1) = P(J_{ij} = -1) = 1/2$. We evaluate the performance of our approach by comparing correlations obtained by our method to those obtained using MCMC for a very long time ($5 \times 10^5 n$ steps,
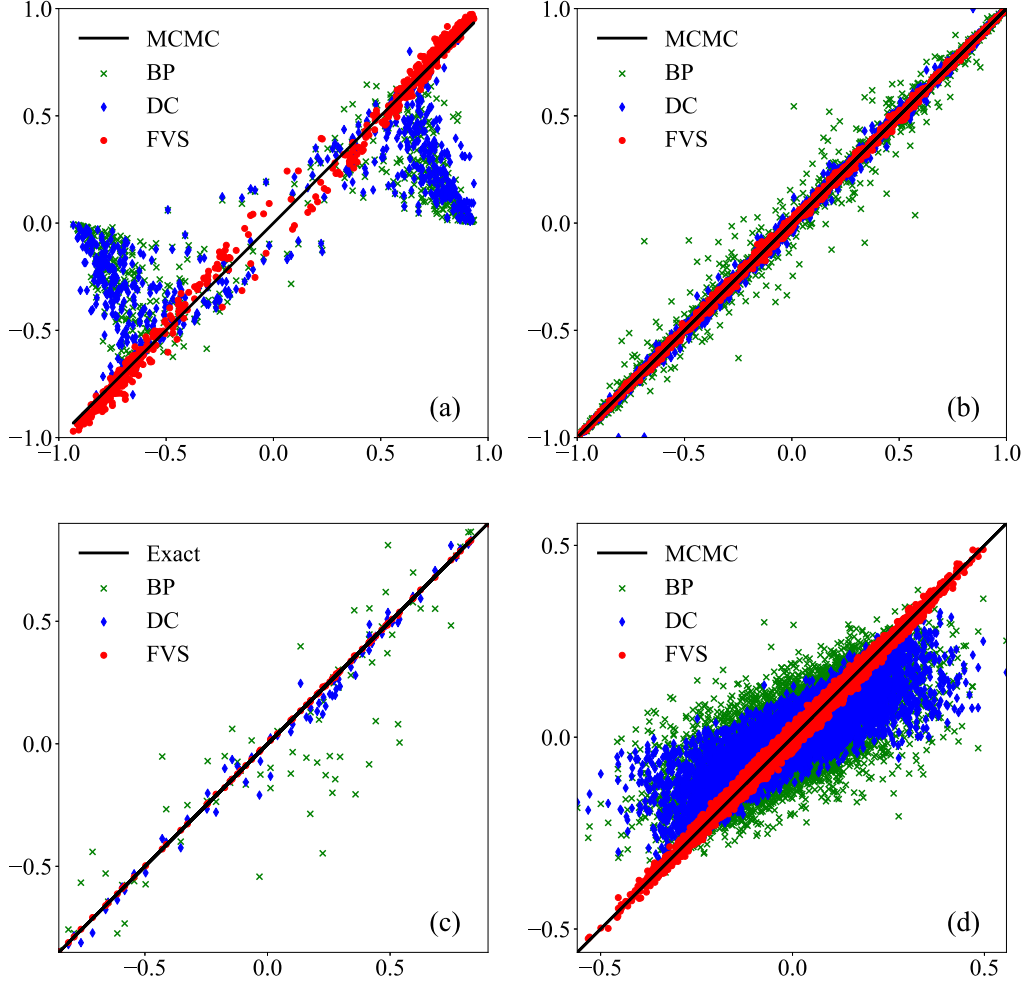
FIG. 1. Correlations of spin glass models obtained by our method (FVS), BP, and dDC compared with data given by MCMC running for a long time ($5 \times 10^5 n$ steps) on various of graphs. (a) The Viana-Bray spin glass model [20] on a random regular graph, $n = 300$ spins, degree 4, $\beta = 0.8$, couplings $J_{ij} \in \{+1, -1\}$ with $P(J_{ij} = 1) = P(J_{ij} = -1) = \frac{1}{2}$; (b) the Erdös-Rényi random graphs, $n = 1100$ spins, average degree 3, $\beta = 0.8$, and couplings are Gaussian random variables with zero mean and unit variance; (c) the model is the same as that of (b) but on the real-world karate club network [21] with $n = 34$ variables, average degree 2.29, and $\beta = 0.54$; (d) the same as (c) but on the real-world political blogs network [22] with $n = 1490$ variables, average degree 11.21, and $\beta = 0.1$.

and we observed that longer time simulation gives almost identical results), which is considered to be very accurate. We also compare our results against correlations obtained by the BP and DC algorithms. The results are shown in Fig. 1.

In Fig. 1(a) the graph is a regular random graph with the degree of every node being 4. We set $\beta = 0.8$ which is beyond the spin glass transition, and the system is at the spin glass phase. We can see from the figure that the correlations given by our method are very close to the MCMC results, whereas the BP and DC results deviate significantly from the MCMC data. In Fig. 1(b) the spin glass model has couplings J following Gaussian distribution with unit variance, and on an Erdös-Rényi random graph. We can see that the results of BP and DC of are much better than those in Fig. 1(a), but they are still much less accurate than our method. In Figs. 1(c) and 1(d) we show the results of our algorithm on two classic real-world networks, the Karate club network [21], and the network of political blogs [22]. On the Karate club we see that our method is almost identical to the exact results (exact

results are given by enumerating all the configurations of the FVS, which is fortunately possible in this case because the system size is small). The figure also indicates that the BP and DC results are much worse than those in random graphs. The reason is that although the average degree of the Karate club $c = 2.29$ is small, the network contains many short loops, which make Bethe approximation less accurate than in random graphs even with DC corrections. This effect turns out to be more severe for BP and DC on the political blogs networks as shown in Fig. 1(d) due to the power-law degree distribution and relatively large average degree $c = 11.21$. But we can see the FVS method is less affected by the large degree and gives significantly more accurate correlation estimates than the other two methods.

B. Spin glasses on 2D lattices

In this section we consider statistical mechanics on the graphs with structured sparsity. A classical example is the 2D ferromagnetic Ising model. Without external fields, the
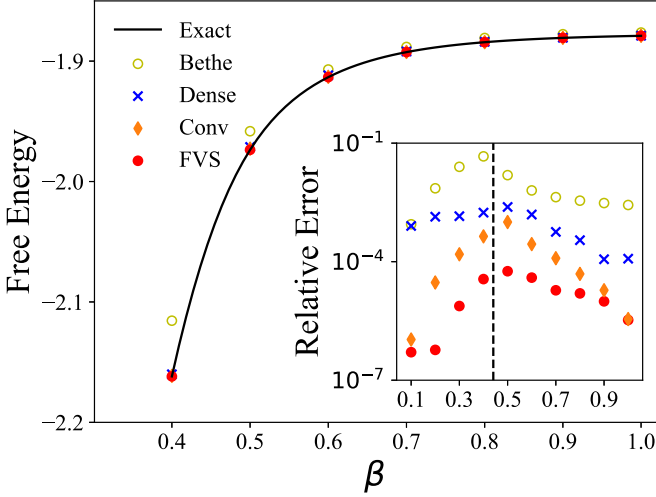
FIG. 2. Free energy (per spin) of the Ising model on the 16×16 lattice with open boundary condition obtained by our method (FVS), original two versions proposed in Ref. [8] (Dense, Conv) and belief propagation (Bethe) with their relative errors to the exact solution [23]. The vertical dashed line in the inset represents the phase transition point of an infinite system ($\beta = 0.4406\,868$).

2D Ising model is on a planar graph, so we can compute an exact solution of the free energy using, e.g., the Kac-Ward formula [23] and use it for evaluating our algorithm. On this problem we set a baseline using the variational autoregressive network using convolution neural networks [8]. Convolution networks are originally proposed to extract relevant features from 2D data, such as images [9,24] and have been shown to give much better results than VAN without using convolution networks on the 2D Ising model [8].

In Fig. 2 we plot the variational free energy given by our method at different temperatures compared with other methods. The relative error to the exact solution is plotted in the inset of this figure. We can see clearly that in all

temperature regimes, our FVS method produces errors that are several orders of magnitude smaller than all the other methods, including the VAN based on convolution (Conv) neural networks, particularly, on the paramagnetic-ferromagnetic transition point with $\beta \approx 0.4406\,868$ where the system has long-range correlations. Moreover, our FVS based method employs much less parameters (22 532) than the convolution networks (714 113, details about model parameters are listed in Appendix C), indicating that our method is more effective than convolutions in characterizing the internal structure of the sparse statistical physics systems.

In Fig. 3 we compare as well the learning process as well as running time (in seconds) of our method and the VAN running on the whole graph [8]. The figure illustrates that our method converges faster, and each learning step (epoch) takes significantly less time in computation, hence, is much more efficient than VAN operating on the whole graph.

C. Inference problems on sparse graphs

Apart from statistical mechanics problems, our method can also be adopted to statistical inference problems where the Bayesian inference formula corresponds to the Boltzmann distribution [25,26]. In this paper we pick up the inference problem in the censored block model (CBM) [19] as an example. The CBM is a variant of the famous stochastic block model [27] in the field of community detection in networks. It asks to recover the hidden (or planted) group assignment (i.e., a configuration of binary variables) from censored edge measurements on a sparse graph. Given a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with n nodes carrying binary variables $\sigma_i \in \{-1, +1\}$, each number represents a group, we have a planted partition $\hat{\sigma}$ which represents the ground truth of community. Our task here is to recover the planted configuration from the information of censored edges, which are depicted by edge labels $J_{ij} = \pm 1, (i, j) \in \mathcal{E}$ drawn from a distribution,

$$P(J_{ij}|\hat{\sigma}_i, \hat{\sigma}_j) = (1 - \alpha)\delta(J_{ij} - \hat{\sigma}_i\hat{\sigma}_j) + \alpha\delta(J_{ij} + \hat{\sigma}_i\hat{\sigma}_j), \quad (9)$$
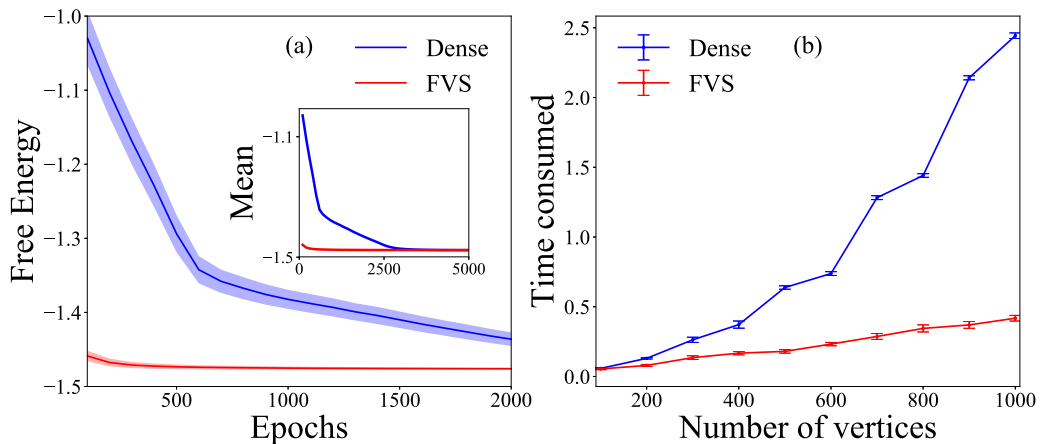

FIG. 3. (a) Evolution of the variational free energy through training. Solid lines represent mean of $E(\mathbf{s}) + \frac{1}{\beta} \ln q_\theta(\mathbf{s})$, and shaded areas represent standard derivation. In the inset, a longer timescale of 5000 epochs is illustrated; (b) Time used for one epoch (training step) in seconds. Each point is the average value over 100 instances. Here FVS is our approach whereas Dense represents densely connected VAN in Ref. [8].
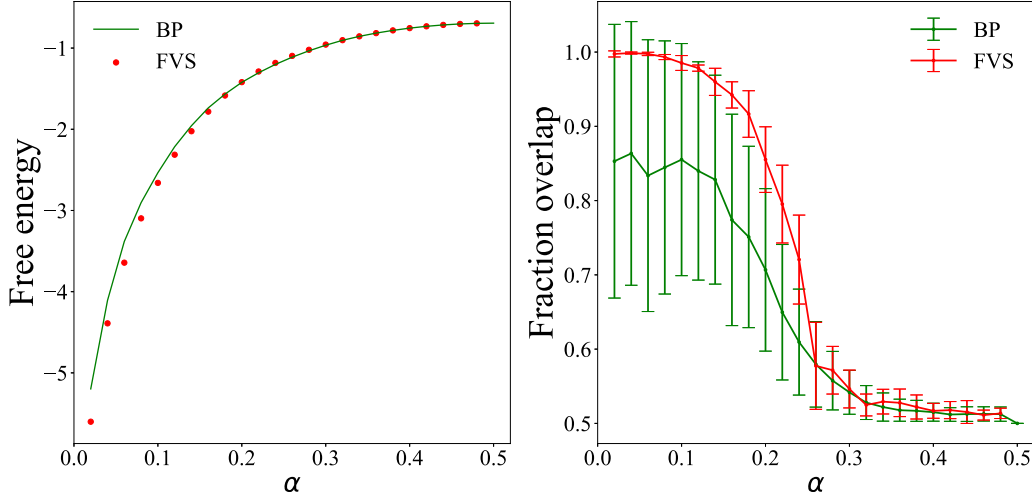
FIG. 4. Inference in censored block model on a small-world network with $n = 1000$ and $p = 0.2$ versus the noise parameter α of censoring. (a) Free energy (per node, the smaller the better) of BP and our method (FVS). (b) Fraction overlap (11) comparison, the FVS points are averaged over ten instances whereas the BP results are averaged over 300 instances due to the large error bars.

where α is the noise parameter, ranging from 0 (noiseless) to 0.5 (informationless); $\delta(x)$ is the Kronecker symbol such that $\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ if $x \neq 0$. The corresponding posterior distribution is

$$P(\boldsymbol{\sigma}|\boldsymbol{J}) = \frac{1}{Z} \exp \left(\beta_0 \sum_{(ij) \in E} J_{ij} \sigma_i \sigma_j \right), \quad (10)$$

which is exactly the form of Ising spin glass with $\beta_0 = 0.5 \ln (1 - \alpha)/\alpha$ acting as inverse temperature. In this case, we use our FVSVAN method to estimate the posterior distribution and then infer the group assignment using the posterior distribution. The performance of algorithms can be evaluated by the fraction of correctly inferred group labels, the so-called *fraction overlap*,

$$F(\hat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = \frac{1}{n} \max \left(\sum_i \delta(\hat{\sigma}_i - s_i), \sum_i \delta(\hat{\sigma}_i + s_i) \right), \quad (11)$$

where s_i represents the inferred assignment of nodes i. It has been shown in Ref. [19] that on large random graphs, the belief propagation algorithm as well as the nonbacktracking [28] and the Bethe-Hessian based algorithm are optimal in the sense that they work all the way down to the theoretical detectability transition, analogous to the stochastic block model [29].

We conduct experiments using the censored block model defined on small-world networks (the Watts-Strogatz model [30]), and the results are shown in Fig. 4. We can see from the left panel of the figure that the variational free energy computed by the FVSVAN is always smaller than that of BP. The right panel indicates that the accuracy of inference [fraction overlap (11)] given by our method is much higher than BP. In the regime where the noise parameter α is large, the system is in the nondetectable phase [19,29] and the edges carry almost no information about the planted partition. In this case we can see that both BP and our method reports overlap of 0.5, basically a random guess on the group assignment.

Although when the noise parameter α is low where the inference of the planted partition is supposed to be easy because edge measurements J_{ij} carry enough information about the ground-true planted partition. However, what we can observe from Fig. 4 is that the overlap obtained by BP is quite low even when the noise level is low, whereas our method can successfully recover the planted configuration, reporting high overlap. The reason for this phenomenon is that, although the graph is sparse, BP still suffers from diverging issue due to small loops in the small-world networks, thus is not able to give an optimal estimate of marginals about group assignments and produces large error bars. Clearly our method can overcome the issue of BP and give an accurate reconstruction when the edge measurements contain information about the planted partition.

IV. CONCLUSION AND DISCUSSIONS

We have introduced a new approach for solving statistical mechanics on sparse graphs. Our approach treats the random sparsity (such as random graphs and real-world networks) and the structured sparsity (such as lattices) in the same way by extracting a small feedback vertex set, mapping the original problem to a statistical mechanics problem on the FVS with an effective energy, then solving it using the neural network based variational method. We have illustrated using extensive numerical experiments that our method significantly outperforms existing methods on both spin glass problems and statistical inference problems on various of sparse graphs.

In this paper we have adopted the feedback set to compress the sparse structures in graphs. We note that recently there is an interesting approach, the graph convolution neural network [31], which adapts convolution neural networks from lattices to sparse graphs for some machine learning tasks, such as the semisupervised classifications. We have tested that naively applying the graph convolution (basically the message passing rule using neighborhood information) in the VAN

works much worse than our FVSVAN approach. It would be interesting to study how to effectively combining the graph convolution approaches and the variational approach for solving statistical mechanics.

A potential application of our approach will be combinatorial optimization problems and constraint satisfaction problems defined on sparse graphs, such as vertex cover [32] and satisfiability problems [2], which can be converted to a statistical mechanics problem with inverse temperature $\beta \to \infty$. Directly setting a large β may induce some instabilities in VAN, a possible solution to this might be introducing the tropic algebra as proposed in Ref. [33]. We will put this into future work. So far, another limitation of our method is that due to the weakness of optimization methods, although the expressive power is enough, it is difficult for our method to handle systems whose FVS set is larger than 2000. Breaking this limitation requires more advanced developments in optimization of neural networks.

APPENDIX A: VARIATIONAL AUTOREGRESSIVE NETWORKS

The variational mean-field methods approximate the Boltzmann distribution using some tractable variational distributions from certain families. Due to the limitations of computing the variational free energy, the variational distributions are usually not very representative. The recent developments of deep neural networks are very helpful to offer much more representative variational distributions. The universal approximation theorem [9] ensures that a neural network containing a large-enough hidden layer can approximate any continuous function.

This idea has recently been used in VAN for building a neural-network-based variational mean-field method for statistical mechanics problems [8], which factorizes the variational distribution by writing the joint probability as the product of conditional probabilities. In the VAN, we aim to approximate the Boltzmann distribution using product of conditional probabilities, each of which is represented by a neural network. This is to say that if any conditional probability could be faithfully represented by a neural network, a universal approximator, then the Boltzmann distribution can be fit perfectly by VAN. This is conceptually important, especially when distinguished with traditional mean-field methods where the variational approximations, e.g., the product distribution or the Bethe approximation, have a fixed number of

parameters, hence, are not universal approximators,

$$q(\boldsymbol{s}) = \prod_{i=1}^{N} q(s_i|\boldsymbol{s}_{<i}), \tag{A1}$$

where $\boldsymbol{s}_{<i} \equiv (s_1, \ldots, s_{i-1})$ represents the state of all variables with indices less than i. Here we see that, for the variable with index i, its state s_i depends on the variables with indices $j < i$ but not on those of the variables with indices $k > i$. This is called the autoregressive property in the machine learning community. From the viewpoint of graphical models, Eq. (A1) connects all variables through their indices to a complete directed graph (suppose each variable is represented by a vertex). This connection pattern is universal for all graphical models, but it makes no conditional independence assumptions about the variables.

Conditional probabilities $q(s_i|\boldsymbol{s}_{<i})$ then will be parametrized by carefully designing neural networks to represent the many-body interactions of the variables. Without loss of generality, for the simplest case of the one layer autoregressive network, the outputs can be written as

$$s_i' = \sigma\left(\sum_{j<i} W_{ij} s_j + b_i\right),$$

where $\sum_{j<i} W_{ij}$ is achieved by adding mask at weights W_{ij}. Since the sigmoid function $\sigma(\cdot)$ limits the output to be in the range of (0,1), the output s_i' can be naturally interpreted as a probability. In this case, the conditional probability can be written as

$$q(s_i|\boldsymbol{s}_{<i}) = (s_i')^{s_i}(1-s_i')^{(1-s_i)}, \tag{A2}$$

which is a Bernoulli distribution with $s_i' = p(s_i = +1|\boldsymbol{s}_{<i})$. And the variational joint distribution becomes

$$q_\theta(\boldsymbol{s}) = \prod_{i=1}^{N} q_\theta(s_i|\boldsymbol{s}_{<i}), \tag{A3}$$

where θ denotes the set of all the network parameters, such as weights and biases.

In variational methods, the parameters θ are adjusted to make the variational probability distribution $q_\theta(\boldsymbol{s})$ as close to the equilibrium Boltzmann distribution as possible. The distance between two probability distributions is usually quantified by the Kullback-Leibler (KL) divergence as [34]

$$D_{\mathrm{KL}}(q\|p) = \sum_s q(\boldsymbol{s}) \ln\left(\frac{q(\boldsymbol{s})}{p(\boldsymbol{s})}\right) = \beta(F_q - F), \tag{A4}$$

where

$$F_q = \sum_s q_\theta(\boldsymbol{s})\left[E(\boldsymbol{s}) + \frac{1}{\beta} \ln q_\theta(\boldsymbol{s})\right] \tag{A5}$$

is the variational free energy, and F is the true free energy. Since the KL divergence is positive, minimizing the KL divergence is equivalent to minimizing F_q to its lower bound F. Thus, the variational free energy is set to be the loss function of the VAN. Here $E(\boldsymbol{s})$ is the energy function of the given model, and $\ln q_\theta(\boldsymbol{s})$ can be calculated combining Eqs. (A2)
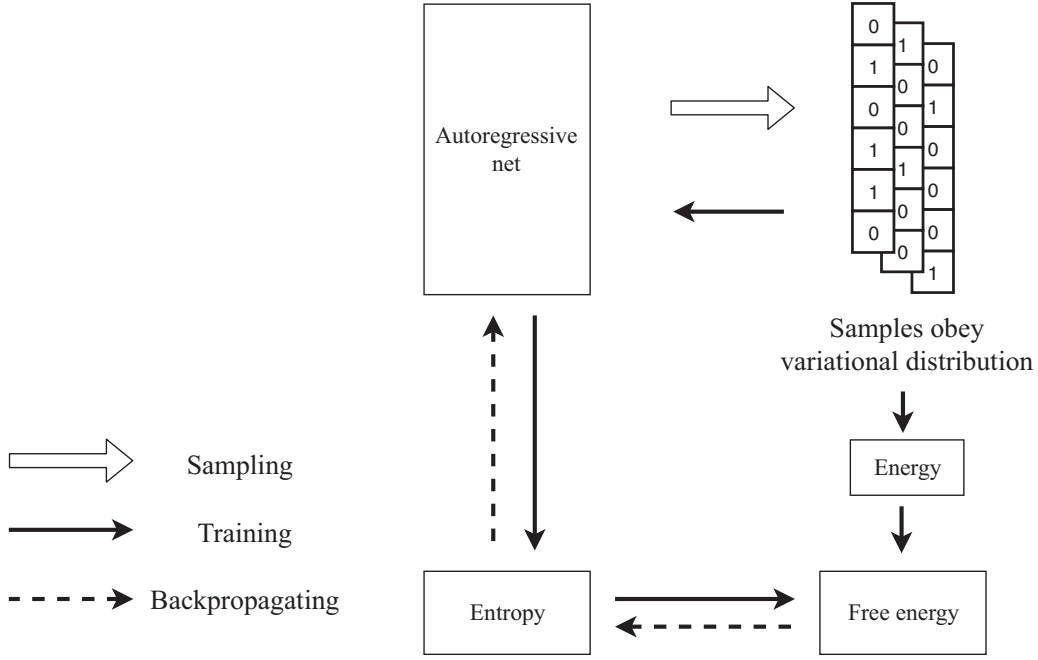
FIG. 5. Workflow of a variational autoregressive network.

and (A3),

$$\ln q_\theta(\boldsymbol{s}) = \sum_{i=1}^{N} [s_i \ln s_i' + (1 - s_i) \ln(1 - s_i')]. \quad \text{(A6)}$$

Minimizing F_q requires optimization over the network parameters θ. Since the objective function to optimize is a possibly nonconvex function over parameters, so there is no method that guarantees to work perfectly. In our paper the score function estimator [16] is adopted to calculate the gradient of the variational free energy with respect to the network parameters,

$$\nabla_\theta F_q = \sum_{\boldsymbol{s}} q_\theta(\boldsymbol{s}) \left\{ \left[E(\boldsymbol{s}) + \frac{1}{\beta} \ln q_\theta(\boldsymbol{s}) \right] \nabla_\theta \ln q_\theta(\boldsymbol{s}) \right\}. \quad \text{(A7)}$$

Then, the parameters can be updated using a gradient based optimization, such as the stochastic gradient descent or ADAM [35] in order to make the variational free energy as small as possible. After training, the variational distribution $q_\theta(\boldsymbol{s})$ will be a good approximation to the Boltzmann distribution $p(\boldsymbol{s})$.

The workflow of a VAN is shown in Fig. 5. At first, an autoregressive network needs to be constructed according to factorization of the joint probability. Then we need to get samples from the autoregressive network following the variational distribution through vertex by vertex sampling in a given ordering, the probability of next vertex given former vertices is given by Eq. (A2). Then we can calculate energy $E(\boldsymbol{s})$, entropy $\ln q_\theta(\boldsymbol{s})$ and variational free energy F_q using the samples and estimate the gradients of F_q using backpropagation and update network parameters θ according to Eq. (A7).

APPENDIX B: EXAMPLE OF FREE ENERGY COMPUTATION ON THE ISING MODEL

We consider the Ising spin glass as a representative model in which the energy function is defined as

$$E(\boldsymbol{\sigma}) = - \sum_{(ij) \in \mathcal{E}} J_{ij} \sigma_i \sigma_j - \sum_i \sigma_i \theta_i, \quad \text{(B1)}$$

where $\boldsymbol{\sigma}$ is a spin state of the system, $\mathcal{E}$ denotes the set of edges, J_{ij} is the coupling constant between two vertices i and j of the Ising model, and θ_i is the external field acting on vertex i.

Given a configuration $\boldsymbol{s}$ of the FVS, the resulting effective external fields acting on the vertices out of the FVS are computed as

$$h_i^0 = \sum_{j \in \Omega \cap \partial i} \beta J_{ij} s_j, \quad \text{(B2)}$$

where ∂i is the set of neighbors of vertex i. Then the energy of the forest $\mathcal{T}$ complementary to the FVS will be

$$E(\boldsymbol{t}) = - \sum_{(ij) \in \mathcal{E}_{\mathcal{T}}} J_{ij} t_i t_j - \sum_{i \in \mathcal{T}} t_i h_i. \quad \text{(B3)}$$

Here $\mathcal{E}_{\mathcal{T}}$ denotes the set of edges of the forest $\mathcal{T}$, and $h_i = \theta_i + h_i^0$ is the overall external field on vertex i of the forest.

When performing leaf removal (which removes all leaves of this graph recurrently), the whole forest will be hierarchized to levels with leaves on the top and roots on the bottom. Our leaf removal ordering is constructed by appending these levels from top to bottom to ensure that after removing vertices in upper level, vertices in the next level will all be leaves.

When a leaf i is removed from the forest, the partition function of $\mathcal{T}$ can be written as

$$
\begin{aligned}
Z(\boldsymbol{s}) &= \sum_{\boldsymbol{t} \in \mathbf{T}} e^{-\beta E(\boldsymbol{s}, \boldsymbol{t})} \\
&= \sum_{t_i} e^{\beta t_i (h_i + J_{ij} t_j)} \sum_{\boldsymbol{t}_{\backslash i}} e^{-\beta E_{\backslash i}(\boldsymbol{t}_{\backslash i})} \\
&= 2 \cosh \beta (J_{ij} t_j + h_i) \sum_{\boldsymbol{t}_{\backslash i}} e^{-\beta E_{\backslash i}(\boldsymbol{t}_{\backslash i})} \\
&= 2\sqrt{\cosh \beta (J_{ij} + h_i) \cosh \beta (-J_{ij} + h_i)} \exp\left[\frac{t_j}{2} \ln\left(\frac{\cosh \beta (J_{ij} + h_i)}{\cosh \beta (-J_{ij} + h_i)}\right)\right] \sum_{\boldsymbol{t}_{\backslash i}} e^{-\beta E_{\backslash i}(\boldsymbol{t}_{\backslash i})} \\
&= 2\sqrt{\cosh \beta (J_{ij} + h_i) \cosh \beta (-J_{ij} + h_i)} \sum_{\boldsymbol{t}_{\backslash i}} e^{\beta h'_j t_j} e^{-\beta E_{\backslash ij}(\boldsymbol{t}_{\backslash i})} \\
&= 2\sqrt{\cosh \beta (J_{ij} + h_i) \cosh \beta (-J_{ij} + h_i)} Z_{\backslash i},
\end{aligned}
\tag{B4}
$$

where $E_{\backslash i}(\boldsymbol{t}_{\backslash i})$ represents energy with vertex i and its interaction excluded, $E_{\backslash ij}(\boldsymbol{t}_{\backslash i}) = E_{\backslash i}(\boldsymbol{t}_{\backslash i}) - h_j t_j$ and h'_j is the external field that combines the original external field of vertex j and an external field vertex i exert on j.

Thus, when a leaf vertex i is removed from the graph, it gives a factor of $2\sqrt{\cosh \beta (J_{ij} + h_i) \cosh \beta (-J_{ij} + h_i)}$ to the partition function, and changing the external field of its neighbor j is expressed as

$$
h'_j = h_j + \frac{1}{2\beta} \ln\left(\frac{\cosh \beta (J_{ij} + h_i)}{\cosh \beta (-J_{ij} + h_i)}\right).
\tag{B5}
$$

For the root vertex k, the factor is $2 \cosh \beta h_k$ since it has no neighbors when removed.

Multiplying these factors together and combining Eqs. (4) and (5), the effective energy of the FVS can be analytically expressed as

$$
\begin{aligned}
\widetilde{E}(\mathbf{s}) = E_\Omega(\mathbf{s}) + \frac{1}{\beta} \Bigg[&\sum_{(i,j) \in \mathcal{E}_{\mathcal{T}}} \ln \sqrt{4 \cosh \beta (J_{ij} + h_i)} \\
&+ \sum_{(i,j) \in \mathcal{E}_{\mathcal{T}}} \ln \sqrt{4 \cosh \beta (-J_{ij} + h_i)} \\
&+ \sum_{i \in \mathcal{R}} \ln[2 \cosh(\beta h_i)] \Bigg].
\end{aligned}
\tag{B6}
$$

Here $\mathcal{R}$ denotes set of roots of the remaining forest, and h_i is the effective field acting on vertex i when j is a neighbor of i.

APPENDIX C: EXPERIMENTAL SETTINGS

In the first experiment, we compare correlations of three methods based on spin glass models. Since connected correlations calculation of MCMC is straightforward, here we only introduce how to calculate connected correlations in the other two methods.

For the belief propagation, we adopt the scheme of Ref. [36] by propagating cavity fields $h_{i \to j}$. As long as cavity fields reach their fixed point, we can calculate physical quan-

tities based on them. The magnetization of vertex i is

$$
m_i = \sum_{j \in \partial i} \operatorname{arctanh}[\tanh(\beta J_{ij}) \tanh(\beta h_{j \to i})],
\tag{C1}
$$

where ∂i represents all neighbors of vertex i. And connected correlation of i and j is

$$
\langle s_i s_j \rangle_c = \frac{e^{\beta J_{ij}} \cosh \beta h_{ij}^+ - e^{-\beta J_{ij}} \cosh \beta h_{ij}^-}{e^{\beta J_{ij}} \cosh \beta h_{ij}^+ + e^{-\beta J_{ij}} \cosh \beta h_{ij}^-} - m_i m_j,
\tag{C2}
$$

where $h_{ij}^+ = h_{i \to j} + h_{j \to i}$ and $h_{ij}^- = h_{i \to j} - h_{j \to i}$.

For our method, calculations can be a little obscure since we only have samples of FVS vertices, but correlations of all edges are needed. But we can start from the definition of correlations,

$$
\begin{aligned}
\langle s_i s_j \rangle &= \sum_{s_i s_j} s_i s_j \, p(s_i s_j) \\
&= \sum_{s_i s_j} s_i s_j \frac{\sum_{\boldsymbol{s} \backslash s_i s_j} e^{-\beta E(\boldsymbol{s})}}{Z} \\
&= \sum_{s_i s_j} s_i s_j \frac{\sum_{\boldsymbol{s}_{\text{fvs}}} e^{-\beta \tilde{E}(\boldsymbol{s}_{\text{aug}})}}{\sum_{\boldsymbol{s}_{\text{fvs}}} e^{-\beta \tilde{E}(\boldsymbol{s}_{\text{fvs}})}} \\
&= \frac{Z_{++} + Z_{--} - Z_{+-} - Z_{-+}}{\sum_{\boldsymbol{s}_{\text{fvs}}} e^{-\beta \tilde{E}(\boldsymbol{s}_{\text{fvs}})}},
\end{aligned}
\tag{C3}
$$

where $\boldsymbol{s}_{\text{aug}} = \boldsymbol{s}_{\text{fvs}+ij}$ is the state of FVS vertices plus vertex i and j, $Z_{++} = \sum_{\boldsymbol{s}_{\text{fvs}}} e^{-\beta \tilde{E}(\boldsymbol{s}_{\text{fvs}}|s_i=+1, s_j=+1)}$. Then both the numerator and the denominator can be calculated using our method. $\langle s_i \rangle$ can be calculated similarly as

$$
\langle s_i \rangle = \frac{Z_+ - Z_-}{\sum_{\boldsymbol{s}_{\text{fvs}}} e^{-\beta \tilde{E}(\boldsymbol{s}_{\text{fvs}})}},
\tag{C4}
$$

where $Z_+ = \sum_{\boldsymbol{s}_{\text{fvs}}} e^{-\beta \tilde{E}(\boldsymbol{s}_{\text{fvs}}|s_i=+1)}$. Connected correlations of entire graph are calculated with expressions above.

In the second experiment, our method, two architectures from original version of VAN, and the belief propagation are implemented on a 16×16 2D ferromagnetic Ising model with open boundary conditions. For dense and convolution

TABLE I. Hyperparameters of Fig. 2.

	Dense	Convolution	FVS
Batch	1×10^3	1×10^3	1×10^4
Net depth	3	3	2
Net width	4	64[a]	3
Maximum step	1×10^4	1×10^4	1×10^4
Learning rate	1×10^{-3}	1×10^{-3}	1×10^{-3}
Number of parameters[b]	1577216	714113	22532

[a]For convolution, this number represents channels.

[b]This quantity is not a hyperparameter, but we list it here for comparison.

architectures, we borrow hyperparameter settings from Ref. [8]. All hyperparameters used in this experiment are listed in Table I. Since all these methods have been fine-tuned for their best performance, there are some differences on the hyperparameter settings.

In the third experiment in the main text, two methods are implemented on random regular graphs to compare their running times and convergence speed. Hyperparameters, such as batch size, depth of the neural network, and learning rate used in Fig. 3(a) are all the same for two methods, whereas the numbers of trainable parameters differs a lot: 63 503 for the FVS and 1002 000 for the dense VAN, for a random graph of size $N = 1000$. We see our method massively decreases the number of hyperparameters. The hyperparameters of Fig. 3(b) are similar to Fig. 3(a) with the only difference on the number of vertices.

In the fourth experiment, we run BP and our method on the censored block model defined on small-world networks. The

belief propagation is written as

$$\psi_{\sigma_i}^{i \to j} = \frac{\prod_{k \in \partial i \backslash j} \sum_{\sigma_k} e^{\beta_0 J_{ik} \sigma_i \sigma_k} \psi_{\sigma_k}^{k \to i}}{Z^{i \to j}}, \quad (C5)$$

where $\psi_{\sigma_i}^{i \to j}$ is the probability of node i being as spin or group σ_i with node j removed from the graph. After Eq. (C5) converged or iterated 2000 iterations, we evaluate the marginal of every variable as

$$\psi_{\sigma_i}^{i} = \frac{\prod_{k \in \partial i} \sum_{\sigma_k} e^{\beta_0 J_{ik} \sigma_i \sigma_k} \psi_{\sigma_k}^{k \to i}}{Z^i}. \quad (C6)$$

For our method, one-layer VAN will be enough to approximate the posterior distribution Eq. (10), which contains $l(l + 1)/2$ parameters (l is the size of FVS). After training, samples of FVS will be obtained, but they cannot used to directly infer ground truth. This is because unlike other methods, there is no spontaneous symmetry breaking of VAN, which means samples of VAN can be divided into two equally probable part with opposite orientation so all marginals will tend to be 0. Thus, manual symmetry breaking needs to be performed to these samples to break the $\mathbb{Z}_2$ symmetry of the censored block model. We pick a sample $\boldsymbol{\sigma_l}$ with lowest energy as a target and flip all other samples according to inner products,

$$\widetilde{\sigma}_j = \text{sgn}(\boldsymbol{\sigma}_j \cdot \boldsymbol{\sigma}_l) * \boldsymbol{\sigma}_j. \quad (C7)$$

After flipping, marginals can be calculated. For FVS vertices, marginals can be calculated from VAN samples, whereas for non-FVS vertices, Eq. (C4) can be used. Then the inferred assignment of node i will be

$$s_i = \text{sgn}(\langle \widetilde{\sigma}_i \rangle), \quad (C8)$$

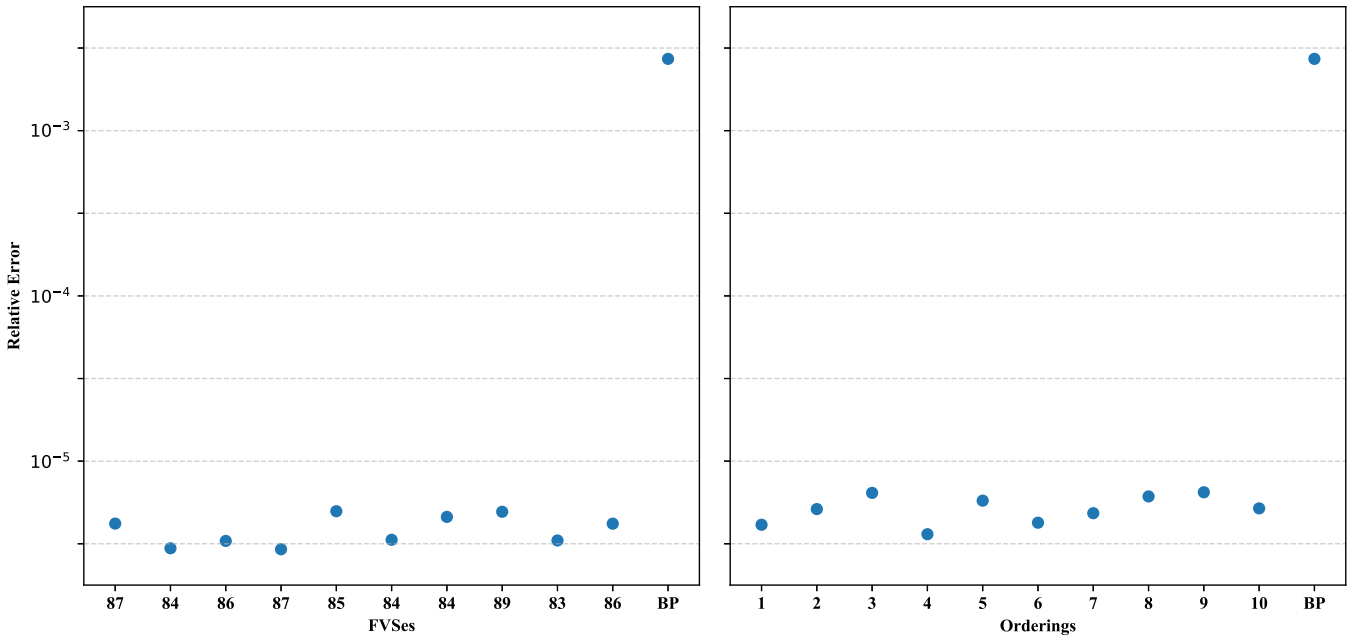then the fraction overlap is calculation according to Eq. (11).

FIG. 6. Relative errors on free energy given by our method with different FVSes (left panel) and different orderings (right panel) compared to the BP which are shown as the right most ones in the figures. Labels in the left panel represent sizes of the FVSes whereas labels in the right panel represent ordering 1–10. The physical model here is the 16×16 2D ferromagnetic Ising model with the open boundary condition and the inverse temperature $\beta = 1.0$.

APPENDIX D: DIFFERENCES OF FVS SELECTIONS AND ORDERINGS

For our approach, the FVS plays an important part since all the following procedures depend on the initial choice of FVS. One may wonder the selection of FVS (which vertices and how many vertices is contained in the FVS) and the ordering of vertices, may affect the final result. Here we provide two experiments to show differences.

Similar to Fig. 2, the 16×16 2D ferromagnetic Ising model with the open boundary condition is chosen as our physical model and the inverse temperature β is set to be 1. We run coreHD [11] ten times to get ten different FVSes then adopt our approach on these FVSes. Besides, we permute a FVS to get ten different orderings and adapt our approach accordingly to get free energies. The results are shown in Fig. 6, we also list the result of BP for comparison. As can be seen, relative error variances of both experiments are extremely low, and under all circumstances results of our approach are around 1000 times better than the BP result. These results clearly indicate that choices and orderings of FVSes will not make a huge difference in our method.

[1] M. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, 2018).

[2] M. Mézard, G. Parisi, and R. Zecchina, Analytic and algorithmic solution of random satisfiability problems, Science **297**, 812 (2002).

[3] M. P. Fossorier, M. Mihaljevic, and H. Imai, Reduced complexity iterative decoding of low-density parity check codes based on belief propagation, IEEE Trans. Commun. **47**, 673 (1999).

[4] M. Mézard and A. Montanari, *Information, Physics and Computation* (Oxford University Press, Oxford, 2009).

[5] H. Bethe, Statistical theory of superlattices, Proc. R. Soc. London, Ser. A **150**, 552 (1935).

[6] J. S. Yedidia, W. T. Freeman, and Y. Weiss, Understanding belief propagation and its generalizations, in *Exploring Artificial Intelligence in the New Millennium* (Morgan Kaufmann Publishers Inc., San Francisco, CA, 2003)pp. 239–269.

[7] R. Kikuchi, A theory of cooperative phenomena, Phys. Rev. **81**, 988 (1951).

[8] D. Wu, L. Wang, and P. Zhang, Solving Statistical Mechanics Using Variational Autoregressive Networks, Phys. Rev. Lett. **122**, 080602 (2019).

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016).

[10] H. J. Zhou, Spin glass approach to the feedback vertex set problem, Eur. Phys. J. B **86**, 455 (2013).

[11] L. Zdeborová, P. Zhang, and H. J. Zhou, Fast and simple decycling and dismantling of networks, Sci. Rep. **6**, 37954 (2016).

[12] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufman Publishers, Inc., San Francisco, 1988).

[13] B. C. Csáji *et al.*, Approximation with artificial neural networks, Faculty of Sciences, Etvs Lornd University, Hungary **24**, 7 (2001).

[14] D. J. Thouless, P. W. Anderson, and R. G. Palmer, Solution of'solvable model of a spin glass', Philos. Mag. **35**, 593 (1977).

[15] Bishop and C. M, *Pattern Recognition and Machine Learning* (Springer, New York, 2006).

[16] R. J. Williams, Simple statistical gradient-following algorithms for connectionist reinforcement learning, Mach. Learn. **8**, 229 (1992).

[17] A. Braunstein, G. Catania, and L. Dall'Asta, Loop Corrections in Spin Models through Density Consistency, Phys. Rev. Lett. **123**, 020604 (2019).

[18] T. P. Minka, Expectation propagation for approximate bayesian inference, in *UAI'01 Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, Seattle, WA, 2001* (Kaufmann, San Francisco, CA, 2001), pp. 362–369.

[19] A. Saade, M. Lelarge, F. Krzakala, and L. Zdeborová, Spectral detection in the censored block model, in *2015 IEEE International Symposium on Information Theory (ISIT), Hong Kong, 2015* (IEEE, Piscataway, NJ, 2015), pp. 1184–1188.

[20] L. Viana and A. J. Bray, Phase diagrams for dilute spin glasses, J. Phys. C **18**, 3037 (1985).

[21] W. W. Zachary, An information flow model for conflict and fission in small groups, J. anthropological Res. **33**, 452 (1977).

[22] L. A. Adamic and N. Glance, The political blogosphere and the 2004 us election: divided they blog, in *Proceedings of the 3rd International Workshop on Link Discovery, Chicago* (ACM, New York, 2005), pp. 36–43.

[23] M. Kac and J. C. Ward, A combinatorial solution of the two-dimensional ising model, Phys. Rev. **88**, 1332 (1952).

[24] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, Pixel recurrent neural networks, in *International Conference on Machine Learning, New York, 2016* (PMLR, New York, 2016), pp. 1747–1756.

[25] Y. Iba, The nishimori line and bayesian statistics, J. Phys. A **32**, 3875 (1999).

[26] L. Zdeborová and F. Krzakala, Statistical physics of inference: Thresholds and algorithms, Adv. Phys. **65**, 453 (2016).

[27] Y. J. Wang and G. Y. Wong, Stochastic blockmodels for directed graphs, J. Am. Stat. Assoc. **82**, 8 (1987).

[28] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, Spectral redemption in clustering sparse networks, Proc. Natl. Acad. Sci. USA **110**, 20935 (2013).

[29] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications, Phys. Rev. E **84**, 066106 (2011).

[30] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, Nature (London) **393**, 440 (1998).

[31] T. N. Kipf and M. Welling, Semi-supervised classification with graph convolutional networks, in *ICLR 2017 : International Conference on Learning Representations 2017, Toulon, France, 2017* (2017).

[32] P. Zhang, Y. Zeng, and H. J. Zhou, Stability analysis on the finite-temperature replica-symmetric and first-step replica-symmetry-broken cavity solutions of the

random vertex cover problem, Phys. Rev. E **80**, 021122 (2009).

[33] J.-G. Liu, L. Wang, and P. Zhang, Tropical tensor network for ground states of spin glasses, arXiv:2008.06888.

[34] D. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, UK, 2003).

[35] D. P. Kingma and J. L. Ba, Adam: A method for stochastic optimization, in *ICLR 2015 : International Conference on Learning Representations, San Diego, CA, 2015* (2015).

[36] H. J. Zhou, *Spin Glass and Message Passing* (Science, Beijing, 2015).