Continuous-Mixture Autoregressive Networks Learning the Kosterlitz–Thouless Transition

Lingxiao Wang(王凌霄)[1,2], Yin Jiang(姜寅)[3*], Lianyi He(何联毅)[2*], and Kai Zhou(周凯)[1*]

[1]*Frankfurt Institute for Advanced Studies, Ruth-Moufang-Str. 1, 60438 Frankfurt am Main, Germany*
[2]*State Key Laboratory of Low-Dimensional Quantum Physics and Department of Physics, Tsinghua University, Beijing 100084, China*
[3]*Department of Physics, Beihang University, Beijing 100191, China*

(Received 5 August 2022; accepted manuscript online 21 November 2022)

We develop deep autoregressive networks with multi channels to compute many-body systems with *continuous* spin degrees of freedom directly. As a concrete example, we demonstrate the two-dimensional XY model with the continuous-mixture networks and rediscover the Kosterlitz–Thouless (KT) phase transition on a periodic square lattice. Vortices characterizing the quasi-long range order are accurately detected by the generative model. By learning the microscopic probability distributions from the macroscopic thermal distribution, the networks are trained as an efficient physical sampler which can approximate the free energy and estimate thermodynamic observables unbiasedly with importance sampling. As a more precise evaluation, we compute the helicity modulus to determine the KT transition temperature. Although the training process becomes more time-consuming with larger lattice sizes, the training time remains unchanged around the KT transition temperature. The continuous-mixture autoregressive networks we developed thus can be potentially used to study other many-body systems with continuous degrees of freedom.

DOI: 10.1088/0256-307X/39/12/120502

Machine learning techniques are attracting widespread interests in different fields of science and technology because of their power to extract and express structures inside complex data. In particular, physicists have employed them to perform tasks involving classification, regression, and pattern generation.[1,2] It was found that the neural networks can classify the phase structures of many-body systems.[3–7] As for the regression, it was successfully applied to the event selection in a large data set (e.g., from the LHCb),[8,9] detecting phase transitions in heavy ion collisions,[10–12] and the molecular structure prediction.[13] Furthermore, it also sheds light on innovating the methods of first-principle calculations of many-body systems,[14–19] in which boosting classical Markov Chain Monte-Carlo (MCMC) methods with machine learning is a potential direction.[20,21]

In addition, it is natural to apply neural networks to many-body systems on lattices, since they share similar discretized representation. In some pioneer attempts,[21–25] the training data were generated by the classical MCMC method. However, its expandability and efficiency are limited by the critical slowing down near the critical point[24] and the sign-problem.[23] Recently, a new method based on an autoregressive neural network was proposed and applied to discrete spin systems (e.g., the Ising model).[26,27] This new method uses a variational *Ansatz* that decomposes the system configuration's probability distribution into conditional distributions on its lattice sites. It was demonstrated that a higher computational accuracy can be achieved in solving several Ising-type systems.[28]

However, it still remains challenging to solve a many-body system which may exhibit a topological phase transition with continuous degrees of freedom (d.o.f.). Previous attempts failed when one applied methods that work well for discrete spin systems to continuous cases.[29] Differently from the classical phase transition, the topological phase transition occurs with topological defects emerging. This has been attracting the attention from various fields of physics.[30] Some related efforts using both supervised learning and unsupervised learning have been made to handle this problem.[31–37] The winding numbers were recognized by a neural network with supervised training for an one-dimensional insulator model.[34] By generating the configurations with MCMC sampling and supplying feature engineered vortex configurations as the input, neural networks could detect the topological phase transition from well-preprocessed configurations.[32] When the unsupervised learning was applied to identify the topological orders,[29,38,39] an efficient unbiased approach is still missing.

In this study, we propose continuous-mixture autoregressive networks (CANs) to solve the continuous spin systems efficiently, which can be further applied to continuous field systems. As a concise reference example, we study the two-dimensional (2D) XY model on a square lattice, which exhibits the so-called Kosterlitz–Thouless (KT) phase transition.[40–42] The CANs are introduced to recognize the topological phase transition with continuous variables in an unsupervised manner. In such neural networks, the microscopic state at each lattice site is modeled by a conditional probability, which constructs a

*Corresponding authors. Email: jiang_y@buaa.edu.cn; lianyi@mail.tsinghua.edu.cn; zhou@fias.uni-frankfurt.de
© 2022 Chinese Physical Society and IOP Publishing Ltd

joint probability for the whole configuration.[26,27] In this work, we introduce a generic framework with CANs and importance sampling (IS) to study the XY model. The vortices, serving as the signal of the KT transition, are automatically generated from the neural networks. Correspondingly, the KT transition temperature of the 2D XY model can be more accurately determined by calculating the helicity modulus.[42] As for the computing time cost, the training process becomes undoubtedly more time-consuming for larger lattice sizes. However, the training time remains unchanged around the transition point. Considering the advantages of the CANs, we propose further potential applications to other many-body systems with continuous d.o.f. in the final part of this work.

Autoregressive Network Importance Sampling. To build a high-efficient sampler for continuous many-body systems, we begin with the free energy, $F = -(1/\beta)\ln Z$, where $\beta \equiv 1/(k_{\rm B}T)$, with T being the temperature and $k_{\rm B}$ the Boltzmann constant setting as 1 in natural units. The free energy is obtained from the partition function $Z \equiv \sum_{\boldsymbol{s}} \exp[-\beta E(\boldsymbol{s})]$, which contains all information of the system in a probability description. The summation runs over all possible configurations $\{\boldsymbol{s}\}$ of the system with the Boltzmann distribution to each of them,

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

Routinely, Monte Carlo (MC) algorithms are applied to generate the configurations by calculating proper relative importance among different configurations based on the energy function $E(\boldsymbol{s})$. For a discrete system, because the d.o.f. is countable, the sampling space is relatively accessible. However, for continuous d.o.f. many-body systems, one will inevitably encounter the "curse of dimensionality". Although one can design a similar cluster algorithm as in discrete systems, the physics insights are not explicit as, e.g., Swendsen–Wang algorithm in Ising systems.[43] Instead, recent advances in generative models have boosted the computational efficiency for continuous systems while the physical interpretation keeps intact.[24,44] In the following, as a deep generative model, the continuous-mixture autoregressive network is constructed to sample configurations for continuous many-body systems. Based on this neural network sampler together with importance sampling[45] (also see the Supplemental Materials), thermodynamic observables of the system are estimated efficiently and accurately.

In general, for generative models, the well-designed machines are trained to reach the target probability distribution p with a tunable distribution function q_θ parameterized by, e.g., neural networks with parameters $\{\theta\}$. For our case, the target function is set to be the Boltzmann distribution dictating the probability of configurations in thermal equilibrium, $p(\boldsymbol{s})$. Within a variational description, we first construct a variational *Ansatz* for the distribution, denoted by $q_\theta(\boldsymbol{s})$. It is parameterized with a set of variational parameters $\{\theta\}$, which can be tuned to drive $q_\theta(\boldsymbol{s})$ approaching $p(\boldsymbol{s})$. The Kullback–Leibler divergence between the variational and the target distributions,

$$D_{\rm KL}(q_\theta \| p) \equiv \mathbb{E}_{\boldsymbol{s}\sim q_\theta}[-\log p(\boldsymbol{s}) + \log q_\theta(\boldsymbol{s})], \tag{2}$$

provides a measure of the distance from q_θ to p. The corresponding variational free energy F_q can be derived from $D_{\rm KL}(q_\theta\|p) = \beta(F_q - F)$, where we obtain

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

Since $D_{\rm KL}(q_\theta\|p)$ is non-negative (also known as Gibbs–Bogoliubov–Feynman inequality), the variational free energy F_q gives an upper bound of the true free energy F. Thus the minimizations on $D_{\rm KL}(q_\theta\|p)$ and on the variational free energy F_q are equivalent. Meanwhile, as pointed out in previous works,[14,26,27] it is straightforward to map the variational parameters $\{\theta\}$ onto the weights of a DNN. Correspondingly, the variational free energy serves as the loss function. Using the log-derivative trick[46] we can obtain its gradient

$$\beta\nabla_\theta F_q = \mathbb{E}_{\boldsymbol{s}\sim q_\theta}\{[\beta E(\boldsymbol{s}) + \ln q_\theta(\boldsymbol{s})]\nabla_\theta \ln q_\theta(\boldsymbol{s})\}, \tag{4}$$

where the gradient term $\nabla_\theta \ln q_\theta(\boldsymbol{s})$ is weighted by the reward signal $\beta E(\boldsymbol{s}) + \ln q_\theta(\boldsymbol{s})$.

For configurations $\boldsymbol{s} = \{s_1, s_2, \ldots, s_N\}$ with continuous spins sit on the lattice with N sites, the variational distribution can be further decomposed into a product of the conditional probabilities,

$$q_\theta(\boldsymbol{s}) = \prod_{i=1}^{N} f_\theta(s_i|s_1, \ldots, s_{i-1}), \tag{5}$$

with the conditional probability for each site-i parameterized to provide the variational *Ansatz*.

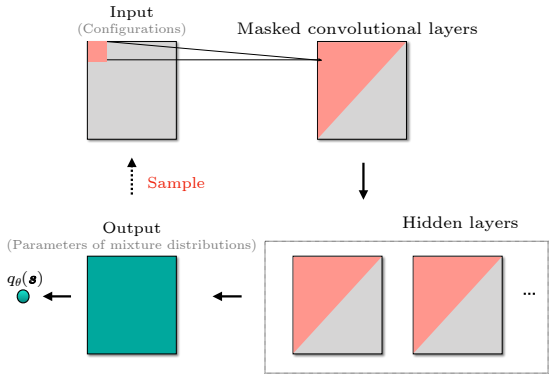

Fig. 1. The architecture of the continuous-mixture autoregressive networks (CANs) for the computation of a continuous spin system. It is a continuous-mixture PixelCNN structure,[47] in which the masked layers are added into the network to establish the autoregressive property.

A flexible and unbiased parametrization for the conditional probabilities is neural network. It can be realized by the scheme shown in Fig. 1 for the variational probability representation, and we will explain it in the following and dub as a continuous-mixture autoregressive network (CAN). Considering the symmetric properties of the system, we employ the PixelCNN[47] that can naturally preserve the locality and the translational symmetry on a square lattice. In addition, the autoregressive property is necessary in the approach to allow for using the above decomposition to evaluate each configuration's variational probability and further the variational free energy, and this is guaranteed by putting a mask on the convolution

kernel.[48] This also ensures that the weights are nonzero for half of the kernel, and each conditional probability $f_\theta(s_i|s_1,\ldots,s_{i-1})$ is independent of s_j with $j > i$ for a prechosen ordering. As shown in Fig. 1, the input layer takes in the configurations $\boldsymbol{s}$ on the lattice. After passing it through the masked convolution layers, the parameters of a mixture distribution at each site are obtained from the outputs. The mixture distribution makes it possible to reach a non-central continuous distribution of the d.o.f. at each site.

Then the configuration probability $q_\theta(\boldsymbol{s})$ can be derived and the variational free energy can be further calculated via Eq. (3) after such a forward process, in which the independent configurations are sampled from the last step mixture distribution. Thus, training the neural networks with the variational free energy to be the loss here is the key to perform the variational calculation, where we use a classical back-propagation algorithm. On the square lattice with N sites, the convergence is reached if the change of the variational energy per site is much smaller than the superior limit, $\delta F_q \ll 4J/N$,[49] where J denotes the coupling strength.

Application of Autoregressive Network to KT Transition. To demonstrate the above method on a practical example of a many-body system with continuous spin d.o.f. at each lattice site, we choose the two-dimensional (2D) XY model. The energy function or Hamiltonian is expressed in terms of spins on the lattice with nearest-neighbor interactions,

$$H = -J\sum_{\langle i,j\rangle} s_i s_j = -J\sum_{\langle i,j\rangle}\cos(\phi_i-\phi_j), \tag{6}$$

where $\langle i,j\rangle$ indicates that the sum is taken over all nearest-neighbor pairs and the angle $\phi_i \in [0,2\pi)$ denotes the spin orientation at site i. The Mermin–Wagner theorem forbids long-range order (LRO) in 2D systems with continuous d.o.f., since the strong fluctuations in 2D break the order.[50] Nevertheless, the formation of topological defects (i.e., vortices and anti-vortices) in the XY model distinguishes phases with and without quasi-LRO, which characterizes the global properties of this many-body system.

Because the orientation of each spin varies continuously, the conditional probability $f_\theta(s_i|s_{<i})$ for the spin at site i must also be a continuous distribution in a finite range. We propose a mixture of the beta distribution $\mathrm{Beta}(a,b)$ as the prior probability to ensure that the continuous variables can distribute randomly in a finite interval, which can flexibly cover even an uniform distribution on site. Intuitively, in Bayesian statistics, the beta distribution is the conjugate prior probability for the Bernoulli distribution which has been proven to be a proper prior distribution for the Ising model case.[26] Although one can choose any bias such as Gaussian-like distributions, based on our numerical attempts and also studies from flow-based models,[51] Gaussian-like prior is centralized and thus weakening the model's representation ability especially in the cases with complicated likelihood. The beta distribution $\mathrm{Beta}(a,b)$ is continuously defined in a finite interval with two positive shape parameters a, b. Thus, the conditional probabilities at each site are parameterized by mixing building blocks of the following beta component,

$$B(s_i;a_i,b_i) = \frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)} s_i^{a_i-1}(1-s_i)^{b_i-1}. \tag{7}$$

Here, $\Gamma(a)$ is the gamma function and $s_i = \phi_i/2\pi \in [0,1]$. We introduce multi-channel PixelCNNs to produce the parameters of the conditional probabilities, thus the network outputs are $\boldsymbol{a} \equiv (a_1,a_2,\ldots,a_M)$ and $\boldsymbol{b} \equiv (b_1,b_2,\ldots,b_M)$ for constructing a mixture of beta distribution with $M = C/2\times N$. C is the number of channels, and the size of the 2D lattice is N.

Now we study the KT transition and calculate the thermodynamic observables of the 2D XY model using the above devised CANs. In the calculation, the width and depth of the network we adopted in CANs are set to be $(32,3)$, with configurations containing $N = L^2$ spins on the lattice to be the input. The multi-channel features from the output are used to construct a mixture of the beta distributions in evaluating the probability of the input configuration, which makes the networks more expressive.[52] The *Adam* optimizer is applied to minimize the loss function. The codes are written with PyTorch library and open on GitHub.[53] In the following, we demonstrate the results of the $L = 16$ case for thermodynamics calculation.

Thermodynamic Observables Estimation. The thermodynamic observables of the 2D XY model have been computed with the MCMC method at a high accuracy.[40,42,53,54] To validate our method, we compare the results from our methods with computations from MCMC. Figure 2 shows the energy per site as a function of β obtained from CANs, CANs-IS and MCMC with 1000 configurations. Here, the MCMC calculation was implemented with a classical algorithm,[55,56] and its details can be found in the Supplementary Materials. The CANs results are calculated based on direct sampled configurations from the network. The CANs-IS calculations follow the approach of importance sampling described in the Supplemental Materials. It should be mentioned, as a baseline, the CANs in the plot have 20 channels for the network adopted and the CANs-IS has the same setup. More detailed comparisons can be found in the Supplemental Materials.

The energy estimation from CANs-IS and MCMC on the 2D XY model reach a good consistency as shown in Fig. 2, where the inset plot on vortices density demonstrates that the well-trained CANs sampler can be implemented to detect the KT phase transition unsupervisedly. The number of free vortices (which equals the number of free anti-vortices) is given by $n = v/(2L^2)$, where the vorticity is defined as $v = (1/2\pi)\oint_C \nabla\phi(\boldsymbol{r})\cdot d\boldsymbol{r}$ in the continuum limit. The rapid decrease of the number of free vortices per site with decreasing temperature at around $\beta \sim 1$ indicates the existence of a topological phase transition. In the low temperature regime ($\beta > 1$), both results from CANs and CANs-IS agree well with that from MCMC, despite the statistical error from thermal fluctuations. In the high temperature regime ($\beta < 1$), the computation based on direct CANs sampling does not match the results from MCMC, but can be corrected efficiently by the importance sampling without any further sampling efforts. Notice that the deviation between CANs and MCMC on

energy estimation shows consistency to the comparison on the number of free vortices (or anti-vortices) shown in the Supplemental Materials. Since a larger number of free vortices and anti-vortices indicates a larger entropy in the XY model, configurations with more vortices own a lower free energy. Thus we see that the energy per site from CANs is larger than that from MCMC at $\beta < 0.7$, because the entropy from the free vortices and anti-vortices balances the free energy. The situation is reversed at $0.7 < \beta < 1$ under similar picture with vortices d.o.f.

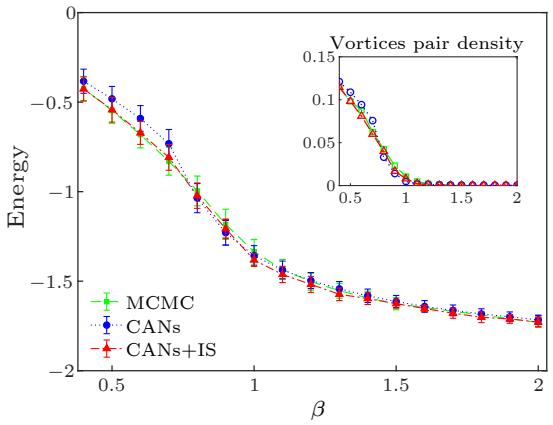

Fig. 2. The energy per site of the 2D XY model with lattice size $L = 16$. The bar on each point denotes the standard deviation. As a comparison, we show the results from CANs, CANs+IS and MCMC. The upper inset shows the number of vortex pairs per site (y-axis) as a function of inverse temperature (x-axis).

KT Phase Transition and Vortices. The KT transition is a transition from bound vortex-antivortex pairs at low temperatures to unpaired vortices and anti-vortices at high temperature. In previous works,[40,54,57] the KT transition temperature for the 2D XY model was calculated. In our method, to recognize the KT transition point quantitatively, we consider the spin stiffness ρ_s, which reflects the change of the free energy in response to an infinitesimally slow twist $\delta\phi$ on the spins. In the continuum limit, it is $\rho_s = [\partial^2 F(\delta\phi)/\partial(\delta\phi)^2]|_{\delta\phi=0}$. In practice, we use the related quantity, i.e., helicity modulus $\gamma(L)$,[40,42,53,54] which is equivalent to ρ_s in the limit of $\delta\phi \to 0$. It can be expressed as

$$\gamma(L) = -\frac{E}{2L^2} - \frac{J\beta}{L^2}\left\langle \left(\sum_{\langle i,j \rangle} \sin(\phi_i - \phi_j)\boldsymbol{e}_{ij} \cdot \boldsymbol{x} \right)^2 \right\rangle, \quad (8)$$

where L is the size of the square lattice, $\boldsymbol{e}_{ij}$ is the vector pointing from site j to site i, and $\boldsymbol{x}$ is an arbitrary unit vector in the 2D plane. It is predicted with the Kosterlitz renormalization group[41] that $\gamma(L \to \infty)$ jumps from the value $2k_B T_c/\pi$ to zero at the critical temperature, and hence the helicity modulus gives a reliable estimation of the KT transition point.

In Fig. 3, we show the helicity modulus of the 2D XY model with lattice size $L = 16$. The crossing point with the curve $2k_B T/\pi$ gives the KT transition temperature. The triangle markers are the numerical results evaluated from CANs-IS (red color) and CANs (blue), and the dot-dashed lines are the corresponding interpolation curves. We observe that the crossing point is located at $\beta_c \simeq 1.1796$ for both CANs and CANs-IS, this is also consistent with the vorticity evaluation in Fig. 2. Since the helicity modulus depends on the correlation function that needs a higher order statistics than the energy (i.e., the mean square of the energy), here we only consider a large lattice size that can give a precise enough evaluation of γ. For different system sizes $L = 4, 8, 16$, the crossing point behaves as $\beta_c \simeq 1.3125, 1.2646, 1.1796$. Noticeably, it is found that increasing statistics could also improve the estimation here. For instance, for different sample statistics $N = 1000, 10000, 100000$ in the same system size ($L = 16$), the crossing point behaves as $\beta_c \simeq 1.1796, 1.1528, 1.1412$. The converging result is consistent with the estimation ($\beta_{KT} \approx 1.12$ for $L \to \infty$) from the standard Monte Carlo simulations.[32,54]

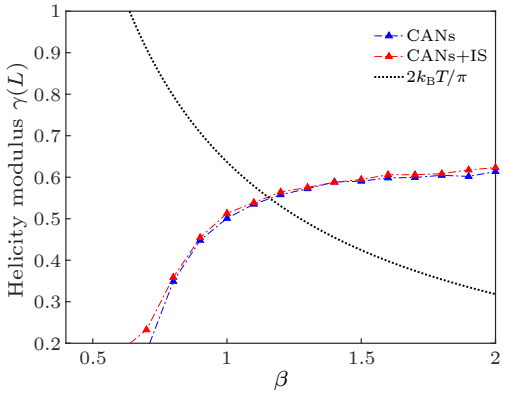

Fig. 3. The helicity modulus of the 2D XY model with lattice size $L = 16$ from CANs and CANs-IS. The cross point with the curve $2k_B T/\pi$ gives the transition point $\beta_c \simeq 1.1796$.

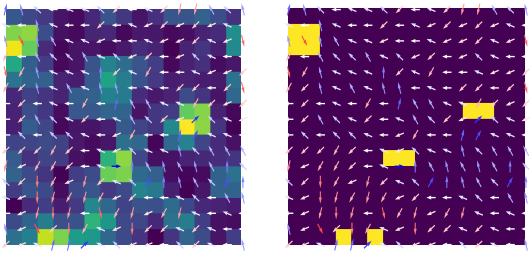

Fig. 4. Probabilities analyzed with CANs and corresponding vortices for a random 2D XY model configuration sampled from well-trained CANs at $\beta = 1.0$.

It is non-trivial to capture the vortex structure of the XY model at high temperature.[32] The vorticity on a discrete 2D lattice can be calculated with

$$v_{[i,j]} = \frac{1}{2\pi} \sum_{[i,j]} \mathrm{Mod}(\phi_{ij}, \pi), \quad (9)$$

where $[i, j]$ labels a plaquette which is an elementary square with 4 corner lattice sites; ϕ_{ij} represents the angle differences in a given direction (e.g., clockwise) along the square; $v_{[i,j]}$ is interpreted as a half-vortex or number of the vortices pair. To investigate whether the trained CANs is capable of capturing and recognizing the vortex structure of XY model, we randomly sample one configuration with size $L = 16$ directly from the trained CANs

at $\beta = 1.0$ as shown in Fig. 4, with the colored arrows indicating the angular of spins at each site. Given the configuration, its joint probability $q_\theta(\boldsymbol{s})$ and also each conditional component on the sites can be calculated as well from the well-trained CANs. The left panel of Fig. 4 shows the conditional probability distribution calculated on each plaquette,

$$\delta q_{[i,j]} = \sum_{[i,j]} q_\theta(s_{ij}), \tag{10}$$

where $q_\theta(s_{i,j})$ is the conditional probability differences in the same given direction. As a comparison, the vortices distribution of this configuration is calculated with Eq. (9), as shown in the right panel of Fig. 4.

Clearly the conditional probability profile calculated from the CANs in the left panel of Fig. 4 shows good consistency with the physical vortices distribution displayed in the right panel. This is reasonable because the conditional probability difference indicates the local energy difference induced by the vortex as the effective d.o.f. The well-trained CANs captured these underlying vortex structure by achieving efficient estimation for the free energy and thus the probability for the system configuration, which also gives rise to efficient estimation for its thermodynamic property. One should be aware that the temperature effect attenuates the long-range correlation exponentially,[41] which will slightly weaken the expressive ability of CANs for such a finite-size system.

Computational Efficiency of the Approach. We observe that, for different temperatures being irrelevant of approaching the critical point or not, the computational time for the training of CANs is almost the same. As a reference it is 0.2 s for one training step on a Nvidia RTX 2080 GPU for our calculation shown above. This indicates that the critical slowing down (CSD) is hopefully avoided. Even though the burden due to the increasing auto-correlation time[58] in MCMC is not notable in topological phase transitions, the relation between the training time cost and the lattice size should be mentioned here.

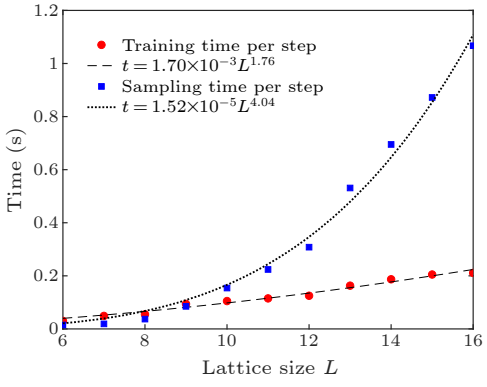

Fig. 5. The computational efficiency of CANS for the 2D XY model at different lattice sizes. The training time and sampling time per step for a typical 2-channel CAN.

The training time per step for CANs with the default network setup for different lattice sizes at the KT transition temperature is displayed in Fig. 5. We find that as a function of the lattice size L it can be well fitted as $t_{\text{train}}(L) = aL^b$, with $a = 1.70 \times 10^{-3}$ and $b = 1.76$. As a comparison, the auto-correlation time of a Monte Carlo algorithm with only local updates[59] should be proportional to L^2. In addition, the training for CANs is one-off and the following configuration sampling procedure can naturally and efficiently take on the parallel advantage of GPU for a large ensemble generation.

In conclusion, we have proposed continuous-mixture autoregressive networks (CANs) for an efficient variational calculation of many-body systems with continuous d.o.f. Specifically, we show that a CAN can be trained as an efficient sampler to the 2D XY model, and it can correctly reveal the topological phase transition for the system. CANs are able to learn to construct microscopic configurations of the system, in which vortices and anti-vortices emerge automatically. Thanks to the efficient probability estimation for the generated configurations by CANs, thermodynamic observables, e.g., the energy and the vorticity, can be evaluated correctly by CANs with importance sampling. The autoregressive structure of the neural networks is beneficial to study the long-range correlations even beyond the phase transition point. Moreover, a straightforward determination of the KT transition point in CANs is shown to be consistent with previous works using the standard Monte Carlo methods, which demonstrate the potential ability of the approach to be adopted into other related studies. It can be generalized directly to investigate more latent topological structures, such as a coupled XY model[57] and a twisted bilayer graphene,[60] where novel long-range correlations may emerge.

Although the increasing time cost with the lattice size is unavoidable, it provides an economical way for the critical point searching in the limit of a large ensemble of configurations, because of the handy GPU usage for the approach. With the help of CANs, the CSD problem occurring in MCMC is expected to be explicitly alleviated, which may reduce the computational difficulties in more complicated many-body systems, e.g., lattice simulation for the possible critical end point in the QCD phase diagram. The ϕ^4 model could be a practical step,[24] where the computational accuracy and practicability should be rigorously examined. In summary, the deep learning approach, especially with well-designed deep neural networks, can be applied to calculate physical systems high-efficiently, which will help us further understand the properties of different many-body systems.

Acknowledgments. We thank Professor Junwei Liu for useful discussions. This work was supported by the BMBF under the ErUM-Data project (K.Z.), the AI grant of SAMSON AG, Frankfurt (K.Z.), Xidian-FIAS International JRC support (L.W.), the National Natural Science Foundation of China [Grant No. 11875002 (Y.J.) and 11775123 (L.H. and L.W.)], the Zhuobai Program of Beihang University (Y.J.), and the National Key R&D Program of China [Grant No. 2018YFA0306503 (L.H.)]. K.Z. also thanks the donation of NVIDIA GPUs from NVIDIA corporation.

References

[1] Buchanan M 2019 *Nat. Phys.* **15** 1208

[2] Carleo G, Cirac I, Cranmer K, Daudet L, Schuld M, Tishby N, Vogt-Maranto L, and Zdeborová L 2019 *Rev. Mod. Phys.* **91** 045002

[3] Wang L 2016 *Phys. Rev. B* **94** 195105

[4] Carrasquilla J and Melko R G 2017 *Nat. Phys.* **13** 431

[5] Pang L G, Zhou K, Su N, Petersen H, Stöcker H, and Wang X N 2018 *Nat. Commun.* **9** 210

[6] Fujimoto Y, Fukushima K, and Murase K 2018 *Phys. Rev. D* **98** 023019

[7] Fujimoto Y, Fukushima K, and Murase K 2020 *Phys. Rev. D* **101** 054016

[8] Metodiev E M and Thaler J 2018 *Phys. Rev. Lett.* **120** 241602

[9] Kasieczka G, Plehn T, Butter A, Cranmer K, Debnath D, Dillon B M, Fairbairn M, Faroughy D A, Fedorko W, Gay C, Gouskos L, F K J, Komiske P, Leiss S, Lister A, Macaluso S, Metodiev E, Moore L, Nachman B, Nordström K, Pearkes J, Qu H, Rath Y, Rieger M, Shih D, Thompson J, and Varma S 2019 *SciPost Phys.* **7** 014

[10] Steinheimer J, Pang L G, Zhou K, Koch V, Randrup J, and Stoecker H 2019 *J. High Energ. Phys.* **2019(12)** 122

[11] Jiang L J, Wang L X, and Zhou K 2021 *Phys. Rev. D* **103** 116023

[12] Zhao Y S, Wang L, Zhou K, and Huang X G 2022 *Phys. Rev. C* **106** L051901

[13] Smith J S, Isayev O, and Roitberg A E 2017 *Chem. Sci.* **8** 3192

[14] Carleo G and Troyer M 2017 *Science* **355** 602

[15] Nagy A and Savona V 2019 *Phys. Rev. Lett.* **122** 250501

[16] Hartmann M J and Carleo G 2019 *Phys. Rev. Lett.* **122** 250502

[17] Pfau D, Spencer J S, Matthews A G D G, and Foulkes W M C 2020 *Phys. Rev. Res.* **2** 033429

[18] Vicentini F, Biella A, Regnault N, and Ciuti C 2019 *Phys. Rev. Lett.* **122** 250503

[19] Yoshioka N and Hamazaki R 2019 *Phys. Rev. B* **99** 214306

[20] Shen H T, Liu J W, and Fu L 2018 *Phys. Rev. B* **97** 205140

[21] Mori Y, Kashiwa K, and Ohnishi A 2018 *Prog. Theor. Exp. Phys.* **2018**

[22] Alexandru A, Bedaque P F, Lamm H, and Lawrence S 2017 *Phys. Rev. D* **96** 094505

[23] Broecker P, Carrasquilla J, Melko R G, and Trebst S 2017 *Sci. Rep.* **7** 8823

[24] Pawlowski J M and Urban J M 2020 *Mach. Learn.: Sci. Technol.* **1** 045011

[25] Zhou K, EndrŐ G, Pang L G, and Stöcker H 2019 *Phys. Rev. D* **100** 011501

[26] Wu D, Wang L, and Zhang P 2019 *Phys. Rev. Lett.* **122** 080602

[27] Sharir O, Levine Y, Wies N, Carleo G, and Shashua A 2020 *Phys. Rev. Lett.* **124** 020503

[28] Ou Z 2019 arXiv:1808.01630v4 [cs.LG]

[29] Cristoforetti M, Jurman G, Nardelli A I, and Furlanello C 2017 arXiv:1705.09524 [hep-lat]

[30] Thouless D J, Duncan F, Haldane M, and Kosterlitz J M 2016 *Scientific Background: Topological Phase Transitions and Topological Phases of Matter*, in *The Nobel Prize in Physics, Advanced Information* (The Royal Swedish Academy of Sciences)

[31] Wang C and Zhai H 2017 *Phys. Rev. B* **96** 144432

[32] Beach M J S, Golubeva A, and Melko R G 2018 *Phys. Rev. B* **97** 045207

[33] Suchsland P and Wessel S 2018 *Phys. Rev. B* **97** 174435

[34] Zhang P, Shen H, and Zhai H 2018 *Phys. Rev. Lett.* **120** 066401

[35] Carvalho D, García-Martínez N A, Lado J L, and Fernández-Rossier J 2018 *Phys. Rev. B* **97** 115453

[36] Hu H Y, Li S H, Wang L, and You Y Z 2020 *Phys. Rev. Res.* **2** 023369

[37] Fukushima K, Funai S S, and Iida H 2019 arXiv:1908.00281 [cs.LG]

[38] Rodriguez-Nieva J F and Scheurer M S 2019 *Nat. Phys.* **15** 790

[39] Scheurer M S and Slager R J 2020 *Phys. Rev. Lett.* **124** 226401

[40] Gupta R, DeLapp J, Batrouni G G, Fox G C, Baillie C F, and Apostolakis J 1988 *Phys. Rev. Lett.* **61** 1996

[41] Kosterlitz J M 1974 *J. Phys. C: Solid State Phys.* **7** 1046

[42] Weber H and Minnhagen P 1988 *Phys. Rev. B* **37** 5986

[43] Swendsen R H and Wang J S 1987 *Phys. Rev. Lett.* **58** 86

[44] Blücher S, Kades L, Pawlowski J M, Strodthoff N, and Urban J M 2020 *Phys. Rev. D* **101** 094507

[45] Nicoli K, Kessel P, Strodthoff N, Samek W, Müller K R, and Nakajima S 2019 arXiv:1903.11048 [cond-mat.stat-mech]

[46] Williams R J 1992 *Mach. Learn.* **8** 229

[47] van den Oord A, Kalchbrenner N, and Kavukcuoglu K 2016 *Proceedings of the 33rd International Conference on International Conference on Machine Learning* vol 48 pp 1747–1756

[48] Germain M, Gregor K, Murray I, and Larochelle H 2015 *Proceedings of the 32nd International Conference on Machine Learning* vol 37 pp 881–889

[49] Chung S G 1999 *Phys. Rev. B* **60** 11761

[50] Wagner H and Schollwoeck U 2010 *Scholarpedia* **5** 9927

[51] Wehenkel A and Louppe G 2020 arXiv:2006.00866 [cs,stat]

[52] Salimans T, Karpathy A, Chen X, and Kingma D P 2017 arXiv:1701.05517 [cs,stat]

[53] Hasenbusch M 2005 *J. Phys. A* **38** 5869

[54] Komura Y and Okabe Y 2012 *J. Phys. Soc. Jpn.* **81** 113001

[55] Tobochnik J and Chester G V 1979 *Phys. Rev. B* **20** 3761

[56] Teitel S and Jayaprakash C 1983 *Phys. Rev. B* **27** 598

[57] Bighin G, Defenu N, Nándori I, Salasnich L, and Trombettoni A 2019 *Phys. Rev. Lett.* **123** 100601

[58] Goodman J and Sokal A D 1989 *Phys. Rev. D* **40** 2035

[59] Kusnezov D and Sloan J H 1993 *Nucl. Phys. B* **409** 635

[60] Julku A, Peltonen T J, Liang L, Heikkilä T T, and Törmä P 2020 *Phys. Rev. B* **101** 060505

Supplemental Materials: Continuous-mixture autoregressive networks learning the Kosterlitz-Thouless transition

Lingxiao Wang,[1,2,*] Yin Jiang,[3,†] Lianyi He,[2,‡] and Kai Zhou[1,§]

[1]*Frankfurt Institute for Advanced Studies, Ruth-Moufang-Str. 1, 60438 Frankfurt am Main, Germany*
[2]*State Key Laboratory of Low-Dimensional Quantum Physics and Department of Physics, Tsinghua University, Beijing 100084, China*
[3]*Department of Physics, Beihang University, Beijing 100191, China.*

(Dated: November 19, 2022)

I. SAMPLING USING AUTOREGRESSIVE NETWORKS

For a well-trained autoregressive network, it's direct to sample configurations from it, while the unavoidable numerically non-zero loss implies that the samples follow a distribution $q_\theta(\mathbf{s})$ which approximates but is not exactly the same as $p(\mathbf{s})$. To reach unbiased estimations of the thermodynamic observables, we adopt the importance sampling(IS) approach [1, 2], and constructs an unbiased estimator for each observable $O(\mathbf{s})$ as following,

$$\langle O\rangle \simeq \sum_{\mathbf{s}\sim q_\theta(\mathbf{s})} O(\mathbf{s})w(\mathbf{s}), \tag{1}$$

where the weight is $w(\mathbf{s}) = (p(\mathbf{s})/q_\theta(\mathbf{s}))/A$ with a normalization factor $A = \sum_{\mathbf{s}\sim q_\theta(\mathbf{s})} p(\mathbf{s})/q_\theta(\mathbf{s})$, and the configurations $\{\mathbf{s}\}$ sampled from the trained autoregressive network are following $q_\theta(\mathbf{s})$ which can be well evaluated. The above importance sampling corrected estimator can be directly derived by inserting the reference distribution of $q_\theta(\mathbf{s})$ into the evaluation of the observable, $\langle O\rangle = \int O(\mathbf{s})(p(\mathbf{s})/q_\theta(\mathbf{s}))q_\theta(\mathbf{s})d\mathbf{s}$.

After insertion the estimation can be viewed as a weighted integration with respect to $q_\theta(\mathbf{s})$, and the integral can naturally be evaluated with the summation of all weighted observables as Eq.(1) shown. In the main text, we denote calculations from the CANs together with importance sampling to be "CANs-IS", while use "CANs" to denote calculations with direct sampling purely from CANs.

II. TRAINING AND PARAMETER SET-UPS

Before more detailed comparisons, the information on the training process should be added here. We adopt the Adam algorithm as the optimizer with parameters $\epsilon = 10^{-8}, \beta_1 = 0.5, \beta_2 = 0.999$. The initial learning rate is set as $\eta = 10^{-3}$ and adjusts in *ReduceLROnPlateau* scheme. Besides, the annealing strategy for inverse temperature was also adopt with a rate of 0.998 at step-t, $\beta_t = \beta * (1-0.998^t)$. The gradient clip was also applied to prevent exploding gradients. The stopping condition should be when variational free energy is converging, but in our practical computations, they all converge before epoch$= 10,000$. Thus, we trained the CANs at different inverse temperatures with 10,000 steps.

The advantage of the variational CANs is that the free energy per site can be directly estimated. It is presented Fig. 1 for three different lattice sizes, $L = 4, 8, 16$. The results for $L = 8$ and $L = 16$ indicate that the free energy converges rapidly with increasing lattice sizes, which ensures that the size effect can be avoided.

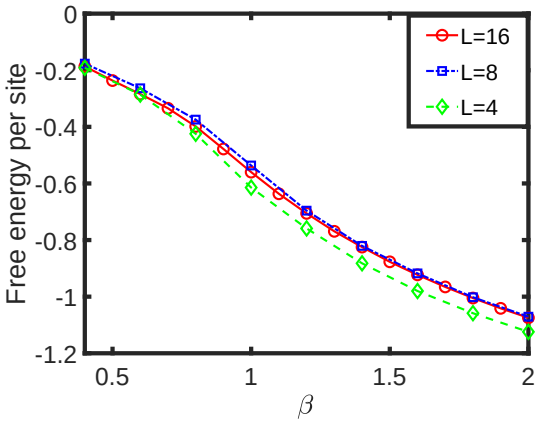

FIG. 1: The variational free energy F_θ per site on a square lattice from CANs.

For the MCMC calculations mentioned in the main text, the equilibration steps is set to be 50,000 for one Markov Chain, and the production steps is 50,000,000 from where we take sample per 10,000 steps. We adopt the classical Metropolis-Hastings algorithm to update the configurations which means the proposal for each step is that randomly rotating all spins iteratively, see more details in Refs [3, 4].

*Electronic address: lwang@fias.uni-frankfurt.de
†Electronic address: jiang_y@buaa.edu.cn
‡Electronic address: lianyi@mail.tsinghua.edu.cn
§Electronic address: zhou@fias.uni-frankfurt.de

TABLE I: The energy and the number of free vortices (or anti-vortices) per site extracted from CANs and MCMC. The results from CANs are obtained from an ensemble average with 1000 configurations.

	β	0.4	0.6	0.8	1.0	1.2
Energy	MCMC	**-0.424**	**-0.682**	**-0.996**	**-1.336**	**-1.502**
	CANs(2)	-0.231	-0.585	-1.165	-1.395	-1.513
	CANs	-0.386	-0.589	-1.037	**-1.358**	**-1.496**
	CANs+IS	**-0.423**	**-0.672**	**-1.022**	-1.387	-1.518
Vortices	MCMC	**0.114**	**0.080**	**0.042**	**0.010**	**0.002**
	CANs(2)	0.138	0.087	0.012	0.001	0.000
	CANs	0.121	0.094	0.033	0.005	0.001
	CANs+IS	**0.114**	**0.081**	**0.040**	**0.008**	**0.002**

Bibliography

[1] T. Müller, B. McWilliams, F. Rousselle, M. Gross, and J. Novák, ACM Transactions on Graphics (TOG) **38**, 1 (2019).

[2] K. Nicoli, P. Kessel, N. Strodthoff, W. Samek, K.-R. Müller, and S. Nakajima, arXiv preprint arXiv:1903.11048 (2019).

[3] J. Tobochnik and G. V. Chester, Phys. Rev. B **20**, 3761 (1979).

[4] S. Teitel and C. Jayaprakash, Phys. Rev. B **27**, 598 (1983).