


Mutual information of spin systems from autoregressive neural networks

Piotr Białas*

*Institute of Applied Computer Science, Jagiellonian University, ul. Łojasiewicza 11, 30-348 Kraków, Poland*Piotr Korcyl[†] and Tomasz Stebel[‡]*Institute of Theoretical Physics, Jagiellonian University, ul. Łojasiewicza 11, 30-348 Kraków, Poland* (Received 4 May 2023; revised 31 August 2023; accepted 4 October 2023; published 24 October 2023)

We describe a direct method to estimate the bipartite mutual information of a classical spin system based on Monte Carlo sampling enhanced by autoregressive neural networks. It enables us to study arbitrary geometries of subsystems, and it can be generalized to classical field theories. We demonstrate it on the Ising model for four partitionings, including a multiply connected even-odd division. We show that the area law is satisfied for temperatures away from the critical temperature: the constant term is universal, whereas the proportionality coefficient is different for the even-odd partitioning.

DOI: [10.1103/PhysRevE.108.044140](https://doi.org/10.1103/PhysRevE.108.044140)**I. INTRODUCTION**

The discovery of topological order in quantum many-body systems [1] initiated a very fruitful exchange of ideas between experts in the fields of solid-state physics and information theory. Many new theoretical tools developed to quantitatively describe the flow of information or the amount of information shared by different parts of the total system have been employed in studies of physical systems [2]. Among these tools, quantum entanglement entropy or mutual information and their various alternatives were found to be particularly useful [3–10]. With their help, it came to be understood that some new phases of matter have the same set of symmetries but differ in long-range correlations quantified by bipartite, tripartite, or higher information-theoretic measures such as mutual information [11,12]. Looking at this in the opposite way, it is expected that calculating mutual information can provide hints about the topological phase of the system, playing a role similar to order parameters in the usual Landau picture of phase transitions (see, for example, [13,14]). Indeed, such quantities are not only useful for theoretical understanding, but they are also measurable observables in experiments. For example, in Ref. [15] the mutual information measured in a quantum spin chain demonstrated the area law [16] governing the scaling of mutual information with the volume of the bipartite partition. The law [17] claims that the mutual information or entanglement entropy in the thermodynamic limit scales with the boundary area separating the two parts of the system, instead of the volume, as for the other extensive properties. The increased interest comes from the fact that black hole entropy was shown a long time ago [18,19] to follow a similar law: black hole entropy depends only on its surface

and not on the interior. Surprisingly, this connection was recently made explicit with the realization that a certain quantum spin model in zero spatial dimensions, called the SYK model [20], is related to a black hole via the gravity/gauge duality [21]. From this perspective, quantum entanglement entropy or quantum mutual information, which can be defined and calculated on both sides of this relationship, became even more attractive from a theoretical point of view. Therefore, there is a strong pressure to develop computational tools able to estimate such quantities, which in itself is, however, very difficult in the general case.

Although an important part of condensed-matter physics [16,22–25] entails studying entanglement entropy and quantum mutual information in quantum spin systems, in classical spin systems the Shannon and Rényi entropies were already found to follow the area law [26]. With the interface boundary sizes up to 64 spins, the coefficients of the area law were precisely determined, and their universality was verified using different lattice shapes. After that, the mutual information of two halves of the classical Ising model on an infinitely long cylinder was calculated using the transfer-matrix approach [27]. Both quantities were discussed with more precision in [28] using a different method, called the bond propagation algorithm. Similarities between quantum entanglement entropy and classical mutual information were studied in Ref. [29].

Recent developments in machine-learning algorithms have created new opportunities to study information theory observables. Quantum entanglement entropy can be calculated using an approximation of the ground state provided by neural networks. For example, in Refs. [30,31], autoregressive architectures were used to calculate variational Rényi entropies for one- and two-dimensional (1D and 2D) Ising and Heisenberg models. In Ref. [32], mutual information of classical spin systems was calculated using the method called machine-learning iterative calculation of entropy (MICE). It is based on the idea from Ref. [33] of exploiting the Donsker-Varadhan

*piotr.bialas@uj.edu.pl

†piotr.korcyl@uj.edu.pl

‡tomasz.stebel@uj.edu.pl

representation of Kullback-Leibler (KL) divergence. We discuss the MICE method further in Appendix F.

In this work, we propose a method to directly estimate classical bipartite mutual information. It is based on the incorporation of machine-learning techniques into Monte Carlo simulation algorithms. From a theoretical point of view, the algorithm we use belongs to the class of metropolized independent sampling [34] algorithms. Therefore, our calculations are stochastic and provably exact within their statistical uncertainties, assuming that all the modes of the target distribution are probed (no mode collapse). The main advantage of the method is its flexibility, i.e., it can be applied as follows: (i) to any geometry of the partitioning, (ii) to any statistical system with a finite number of degrees of freedom, and (iii) in an arbitrary number of space dimensions, provided that the target probability can be effectively trained (e.g., with no mode collapse).

II. METHOD

A. Mutual information for spin systems

We consider a classical system of spins that we divide into two arbitrary parts A, B . In this case, the Shannon mutual information is defined as

$$I = \sum_{\mathbf{a}, \mathbf{b}} p(\mathbf{a}, \mathbf{b}) \log \frac{p(\mathbf{a}, \mathbf{b})}{p(\mathbf{a})p(\mathbf{b})}, \quad (1)$$

where a particular configuration \mathbf{s} of the full model has parts \mathbf{a} and \mathbf{b} , and where the Boltzmann probability distribution of states, depending on inverse temperature β , is given by (we omit the explicit dependence of Z on β)

$$p(\mathbf{a}, \mathbf{b}) = \frac{1}{Z} e^{-\beta E(\mathbf{a}, \mathbf{b})}, \quad Z = \sum_{\mathbf{a}, \mathbf{b}} e^{-\beta E(\mathbf{a}, \mathbf{b})}, \quad (2)$$

and

$$p(\mathbf{a}) = \sum_{\mathbf{b}} p(\mathbf{a}, \mathbf{b}), \quad p(\mathbf{b}) = \sum_{\mathbf{a}} p(\mathbf{a}, \mathbf{b}) \quad (3)$$

are probability distributions of subsystems. We shall use the same symbol p for all probability distributions defined on different state spaces, distinguishing them by the arguments. In the above expressions, the summation was performed over all configurations of subsystems A or B . Inserting Eqs. (2) and (3) into Eq. (1), we obtain

$$I = \log Z - \sum_{\mathbf{a}, \mathbf{b}} p(\mathbf{a}, \mathbf{b}) [\beta E(\mathbf{a}, \mathbf{b}) + \log Z(\mathbf{a}) + \log Z(\mathbf{b})], \quad (4)$$

where $Z(\mathbf{a}) = \sum_{\mathbf{b}} e^{-\beta E(\mathbf{a}, \mathbf{b})}$ and $Z(\mathbf{b}) = \sum_{\mathbf{a}} e^{-\beta E(\mathbf{a}, \mathbf{b})}$. Please note that in typical Monte Carlo approaches, the partition functions Z , $Z(\mathbf{a})$, and $Z(\mathbf{b})$ are not available.

B. Neural importance sampling for MI

Below we argue that I can be obtained from a Monte Carlo simulation enhanced with autoregressive neural networks (ANNs). It was recently shown that ANNs can be used to approximate the Boltzmann probability distribution $p(\mathbf{a}, \mathbf{b})$ for spin systems, and they provide a means to sample from this

approximate distribution [35–38]. Let us call this approximate distribution $q_\theta(\mathbf{a}, \mathbf{b})$, where θ stands here for the parameters of the neural network that are to be tuned so that q_θ is as close to p as possible under an appropriate measure, typically the backward Kullback-Leibler divergence

$$D_{\text{KL}}(q_\theta | p) = \sum_{\mathbf{a}, \mathbf{b}} q_\theta(\mathbf{a}, \mathbf{b}) \log \left(\frac{q_\theta(\mathbf{a}, \mathbf{b})}{p(\mathbf{a}, \mathbf{b})} \right). \quad (5)$$

The formula Eq. (4) can be rewritten in terms of averages with respect to the distribution q_θ ,

$$\begin{aligned} I &= \log Z - \frac{1}{Z} \beta \langle \hat{w}(\mathbf{a}, \mathbf{b}) E(\mathbf{a}, \mathbf{b}) \rangle_{q_\theta(\mathbf{a}, \mathbf{b})} \\ &+ \frac{1}{Z} \langle \hat{w}(\mathbf{a}, \mathbf{b}) \log Z(\mathbf{a}) \rangle_{q_\theta(\mathbf{a}, \mathbf{b})} \\ &+ \frac{1}{Z} \langle \hat{w}(\mathbf{a}, \mathbf{b}) \log Z(\mathbf{b}) \rangle_{q_\theta(\mathbf{a}, \mathbf{b})}, \end{aligned} \quad (6)$$

where the importance ratios are defined as

$$\hat{w}(\mathbf{a}, \mathbf{b}) = \frac{e^{-\beta E(\mathbf{a}, \mathbf{b})}}{q_\theta(\mathbf{a}, \mathbf{b})}. \quad (7)$$

The crucial feature of ANN-enhanced Monte Carlo is that contrary to standard Monte Carlo, we can estimate directly the partition functions Z , $Z(\mathbf{a})$, and $Z(\mathbf{b})$ (see Appendix A for details). In this way, we have expressed the mutual information I only through averages with respect to the distribution q_θ . It can now be estimated by sampling from this distribution. The procedure of sampling configurations from the approximate probability distribution provided by neural networks together with reweighting observables with importance ratios was proposed in Ref. [39] and named neural importance sampling (NIS). This paper reports on an application of this technique to information theory observables.

Autoregressive neural networks rely on the product rule, i.e., factorization of q_θ into the product of conditional probabilities

$$q_\theta(\mathbf{a}, \mathbf{b}) = \prod_{i=1}^{L^2} q_\theta(s^i | s^1, s^2, \dots, s^{i-1}). \quad (8)$$

Due to the fact that the labeling of spins in Eq. (8) is arbitrary, we can choose it in such a way that we first enumerate all spins from part A , $\mathbf{a} = (s^1, s^2, \dots, s^{n_A})$, and only afterward all spins from part B , $\mathbf{b} = (s^{n_A+1}, s^{n_A+2}, \dots, s^{n_A+n_B})$. We then obtain

$$q_\theta(\mathbf{s}) \equiv q_\theta(\mathbf{a}, \mathbf{b}) = q_\theta(\mathbf{a}) q_\theta(\mathbf{b} | \mathbf{a}), \quad (9)$$

with

$$q_\theta(\mathbf{a}) = \prod_{i=1}^{n_A} q_\theta(s^i | s^1, s^2, \dots, s^{i-1}) \quad (10)$$

and

$$q_\theta(\mathbf{b} | \mathbf{a}) = \prod_{i=1}^{n_B} q_\theta(s^{n_A+i} | s^{n_A+1}, s^{n_A+2}, \dots, s^{n_A+i-1}, \mathbf{a}). \quad (11)$$

These features of ANN support the fact that we can readily estimate $\log Z(\mathbf{a})$ and $\log Z(\mathbf{b})$ required by Eq. (6) directly. In this respect, the ANN approach differs from *normalizing flows* used to approximate continuous probability distributions

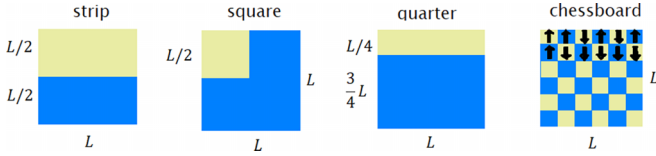


FIG. 1. Four considered partitioning geometries. Periodic boundary conditions are applied. The small blocks in the chessboard partitioning represent single spins.

and employed recently in the context of lattice field theory [40–47]. There, the probability is calculated for the whole field configuration at once, and the conditional probability $q_\theta(\mathbf{b}|\mathbf{a})$ as well the marginal distribution $q_\theta(\mathbf{a})$ are not so easily available.

III. RESULTS

We leave the technical details of the evaluation of the terms in Eq. (6) to Appendix A and now provide an example of its application. We demonstrate it on the Ising model on a periodic $L \times L$ lattice, with ferromagnetic, nearest-neighbor interactions, defined by the Hamiltonian

$$E(\mathbf{a}, \mathbf{b}) = - \sum_{\langle i,j \rangle} s^i s^j, \quad (12)$$

where $s^i \in \{1, -1\}$. We consider the following divisions and respective mutual information observables:

(i) “Strip” geometry—the system is divided into equal rectangular subsystems.

(ii) “Square” geometry—subsystem A is the square of size $\frac{1}{2}L \times \frac{1}{2}L$.

(iii) “Quarter” geometry— A is rectangular of size $\frac{1}{4}L \times L$.

We show them schematically in the three sketches on the left in Fig. 1. We note that all of them have the same length of the border between the subsystems, i.e., $2L$ (as we used periodic boundary conditions), although they may differ in their volumes. In the discussion below, we refer to those three partitionings as “block” partitionings. In addition, we also consider the following division:

(iv) “Chessboard” partitioning, where the system is divided using even-odd labeling of spins.

In this case, the boundary between spins is L^2 , i.e., every spin of the system is at the boundary between parts A and B . Note that in the chessboard partitioning, the subsystems are not simply connected, as is usually considered in the literature.

We investigate a wide range of system sizes, reaching $L = 66$ for $\beta > 0.3$ and $L = 130$ for $\beta \leq 0.3$.¹ We combine results obtained by the variational autoregressive network (VAN) approach of Ref. [35] and our recently proposed modification called the hierarchical autoregressive network (HAN) algorithm [37]. Both approaches are applied to several divisions,

¹The reason behind this choice is that for larger β , training of the neural networks is less efficient, and for these temperatures we did not reach a sufficient quality of training. This could be done in principle using additional tricks, such as, e.g., pretraining, which we used for simulations of the Potts model [38].

TABLE I. The system sizes L which were considered for a given partitioning type. Note that HAN requires a total system size $L = 2^n + 2$, where n would be the number of levels in the hierarchy.

Partitioning	VAN	HAN
Strip	8,12,16,	10,18,34,
	20,24,28	66,130 (for $\beta \leq 0.3$)
Square		10,18,34,
		66,130 (for $\beta \leq 0.3$)
Quarter	8,12,16,	
	20,24,28	
Chessboard	8,12,16,20,	
	24,28,32	

as summarized in Table I. We describe the details of the VAN and HAN architectures and the quality of their training in Appendix B. In Appendix C we discuss the details of mutual information calculation for the chessboard partitioning. Simulations are performed at 13 values of the inverse temperature: from 0.1 to 0.4 with a step 0.05, 0.44, and from 0.5 to 0.9 with a step 0.1. For each β we collect the statistics of at least 10^6 configurations, and we estimate the statistical uncertainties using the jackknife resampling method. In all cases, the relative total uncertainty of I combining the systematic and statistical uncertainties is of the order of 1%, or smaller—see Appendix A [in particular, the discussion around Eq. (A12)].

A. Area law

In Fig. 2 we show I as a function of L for the three block partitionings and for three representative inverse temperatures, β : 0.1 (high-temperature regime), 0.44 [very close to critical temperature, $\beta_c = 1/2 \ln(1 + \sqrt{2}) \approx 0.44069$], and 0.9 (low-temperature regime). We clearly observe a linear dependence on the system size with a strongly β -dependent slope and intercept. Similar behavior can be observed for the

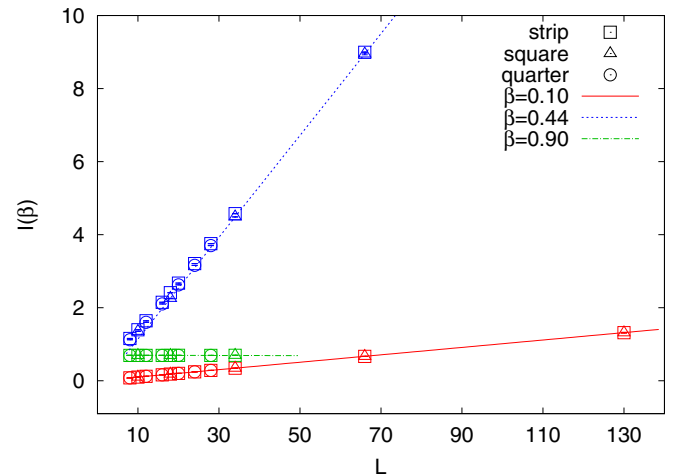


FIG. 2. General dependence of I on L for various geometries and three representative inverse temperatures: $\beta = 0.10$ in the disordered phase, $\beta = 0.44$ close to the phase transition, and $\beta = 0.90$ in the ordered phase. Lines correspond to attempted fits using the area law Ansatz.

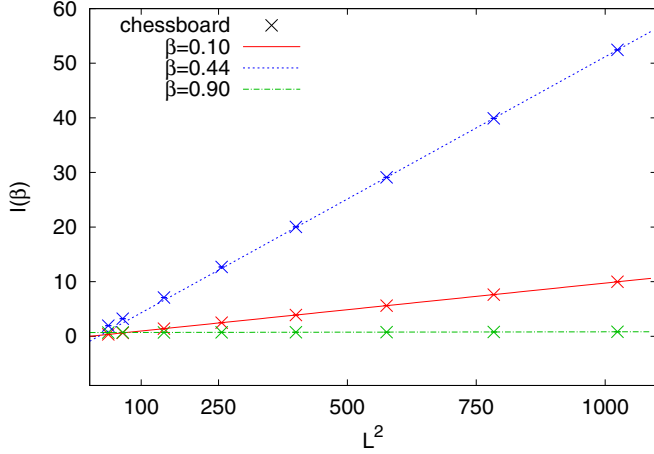


FIG. 3. General dependence of I on L^2 for the chessboard partitioning at three inverse temperatures: $\beta = 0.10, 0.44$, and 0.90 . Lines correspond to attempted fits using the area law *Ansatz*. Statistical uncertainties are shown but are much smaller than the symbol size.

chessboard partitioning when one plots I as a function of L^2 ; see Fig. 3. The emerging picture, supporting the conjectured area law, is that I can be written in a compact form as

$$I_{\text{geom}}(\beta, L) = \alpha_{\text{geom}}(\beta)B(L) + r_{\text{geom}}(\beta), \quad (13)$$

where the variable $B(L)$ corresponds to the length of the boundary between the two parts A and B : $B(L) = L^2$ for the chessboard partitioning and $B(L) = 2L$ for other partitionings considered in this work. Equation (13) has two parameters, α and r , which, as we denoted, depend on β and can differ for different partitioning geometries.

We expect that Eq. (13) is valid as long as finite volume effects in I are small. In general, they may depend on three length scales present in the system: the size of the whole system L , the size of the smallest subsystem, and the correlation length ξ determined by the temperature of the system. Finite volume effects should vanish when ξ is smaller than the rest of the scales, and they may depend on the geometry of the partitioning. A closer look at the data indeed reveals additional contributions to mutual information which spoil the area law (13). To see this, we plot in Fig. 4 the $\sqrt{\chi^2/\text{DOF}}$ of the fits, which were performed assuming relation Eq. (13). Three regions of inverse temperatures can be easily defined. At small β the *Ansatz* Eq. (13) describes all system sizes since $\sqrt{\chi^2/\text{DOF}} \approx 1$, suggesting that finite volume effects are smaller than the statistical uncertainties. A similar situation occurs for large β , where again the fit involved all available system sizes. On the contrary, in the region close to the phase transition, where the correlation length ξ is the largest, the fits have clearly bad quality, $\sqrt{\chi^2/\text{DOF}} \gg 1$. This picture is further confirmed by the fact that the quality of the fit improves when we discard smaller system sizes, $L < L_{\min}$, as shown for the strip partitioning in the inset.

The values of $\sqrt{\chi^2/\text{DOF}}$ close to $\beta_c \approx 0.44$ are particularly large for the chessboard partitioning. However, since the errors of the individual points are much smaller than for the rest of the partitionings [due to the simplified calculation of $Z(\mathbf{a})$; see Appendix C], we refrain from drawing conclusions

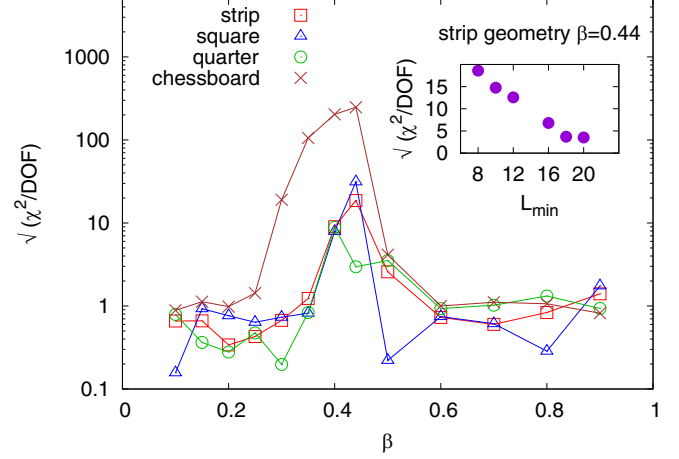


FIG. 4. Quality of fits involving all available data points and assuming the area law functional form of the mutual information for the four geometries shown in Fig. 1: $\sqrt{\chi^2/\text{DOF}}$ plotted as a function of the inverse temperature. In the inset, we show how the fit improves as we restrict the system sizes included in the fit to be bigger than or equal to L_{\min} .

about the size of the finite-size effects in this partitioning compared to block partitionings.

B. Discussion of $\alpha(\beta)$ and $r(\beta)$ coefficients

For β far from the critical inverse temperature, we can reliably describe our data with the area law *Ansatz* Eq. (13). This allows us to extract the coefficients α and r for the four different partitionings and compare them. We show the results in Fig. 5. In the main figures, we show the dependence on β , whereas in the insets we show the difference between the “strip” and “square” partitionings (the differences between other partitionings look similar). We have shown only the values that were obtained from fits with $\sqrt{\chi^2/\text{DOF}} \lesssim 1$ so as to be sure that the postulated dependence Eq. (13) is indeed reflected in the data. This means $\beta \leq 0.35$ and $\beta \geq 0.5$ for block partitionings, and $\beta \leq 0.25$ and $\beta \geq 0.5$ for chessboard. The uncertainties of data points in Fig. 5 contain (i) the propagation of statistical uncertainties of $\hat{I}_{N,M}(\beta, L)$ through the extrapolation fits ($M \rightarrow \infty$ and the dependence on L), and (ii) the systematic uncertainty of the fits estimated by taking the maximal difference of the outcomes when using different fit *Ansätze* or fit intervals. Both types of uncertainties were added in quadrature to obtain the final error bars shown in Fig. 5.

In the left panel, we show the coefficient α as a function of β . The qualitative behavior of $\alpha(\beta)$ seems to be universal: it goes to 0 at $\beta = 0$ and $\beta \rightarrow \infty$ and rises around the critical temperature. The data clearly show that the chessboard partitioning yields different values from the three other possibilities, which seem to be compatible with each other. The differences shown in the inset are indeed compatible with zero within their uncertainties. Therefore, we conclude that in this range of β the partitioning does not influence the mutual information (when block partitionings are considered). In the right panel, we show the value of r . In that case, all four

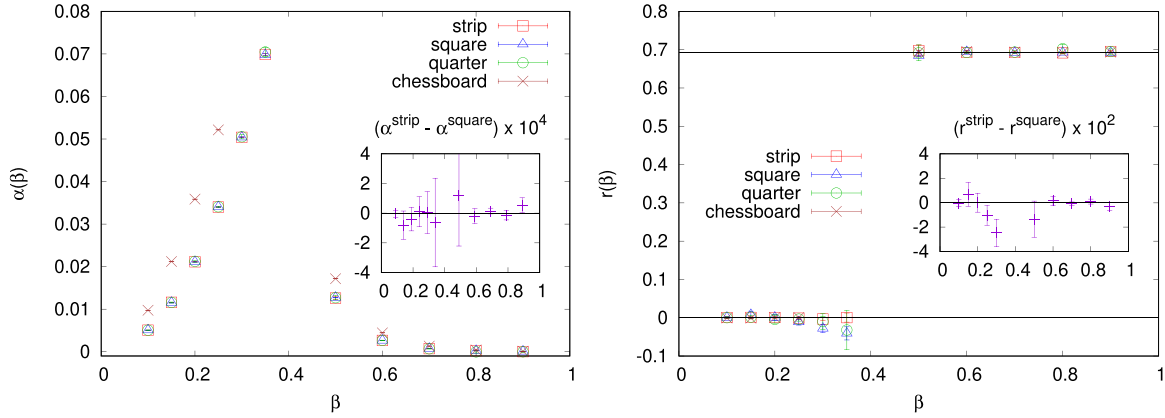


FIG. 5. Coefficients α and r for four geometries shown in Fig. 1 as a function of the inverse temperature β . The inset shows the difference between α and r from strip and square geometries. Both geometries yield results compatible within their statistical and systematical uncertainties.

partitionings give the same result: 0 for $\beta < \beta_c$ and $\ln 2$ for $\beta > \beta_c$.²

Comparing our results to those from Ref. [28], where the bond propagation and transfer matrix methods were used to calculate mutual information for cylinderlike geometry in the limit of infinite length of the cylinder, we may make several observations. First, we note that our $r(\beta)$ follows the behavior $r(\beta < \beta_c) = 0$ and $r(\beta > \beta_c) = \ln 2$ obtained in Ref. [28] in the limit of infinite cylinder's circumference. The deviations from this behavior, which we observe close to β_c , should be attributed to the finite-size effects. Finite volume effects are also responsible for the fact that we cannot reproduce the value of $r(\beta_c)$ at the critical point, Ref. [28]: $r(\beta = \beta_c) = 0.254\,392\,5(5)$. With regard to the α coefficient, in Ref. [28] only the critical value was calculated, $\alpha(\beta = \beta_c) = 0.376\,926\,26(7)$. The low- β leading behavior of α was obtained in Ref. [32] using the partitioning, which we called “strip”: $\alpha(\beta) \approx \frac{1}{2}\beta^2$ for $\beta \rightarrow 0$, and our results follow this behavior exactly.

IV. SUMMARY

In this work, we have provided a numerical demonstration that the Shannon bipartite mutual information (MI) can be readily obtained from the neural importance sampling (NIS) algorithm for the classical Ising model on a square lattice. Our approach allows for studies of different partitionings, and we provided comparisons of the mutual information estimated for four geometries. We successfully exploited the hierarchical algorithm (HAN) to reach larger system sizes than achievable using standard variational autoregressive networks (VANs). The crucial property of our approach is that it provides unbiased results (unlike the MICE [32] approach, which is variational) assuming all modes of the target probability distribution can be probed.

²For $\beta \rightarrow \infty$, the space of states reduces to the two states—one with all spins up and the second with all spins down—both having the same probability 1/2. Plugging such probability distribution into Eq. (1), one obtains $I(\beta = \infty) = \ln 2$.

It is important to note that the condition that the probability distribution modeled by the neural network contains all the modes of the target distribution is of central importance for the NIS approach (and, more generally, for all MCMC methods). Obviously, when mode collapse is present and some part of the target distribution support is not sampled, the reweighting procedure cannot compensate for the missing contributions to the partition function, which may lead to systematic bias. It is known that generative neural networks are sensitive to the mode collapse [35,48–50]. In the case of the Ising model considered here, where analytical results for the partition function are known, the possibility of mode collapse can be very much excluded by checking that the NIS and exact result agree.

We discussed the validity and universality of the area law for different partitionings. We found that at low and high temperatures the area law is satisfied, whereas such an *Ansatz* does not describe our data in the vicinity of the phase transition, supposedly due to inherent finite volume effects.

V. OUTLOOK

Our proposal can be applied to many other spin systems with short- and long-ranged interactions, such as various classes of spin glass. However, the main difficulty one encounters when dealing with spin-glass systems is the efficiency of training these complicated Boltzmann distributions, and in particular, the avoidance of mode collapse issues. As such, the applicability of our method depends on the progress in the neural network sampler's efficiency. Also, the long-ranged interactions prohibit the application of the HAN algorithm, limiting, at the moment, the available system sizes to $L \sim 30$.

We believe that by exploiting the Feynman path integral quantization prescription, one may use the approach based on autoregressive networks to estimate also the entanglement entropy in quantum spin systems. In such a picture, a D -dimensional quantum system is described by a $D + 1$ statistical system, where the machine-learning enhanced Monte Carlo is applicable. In particular, thanks to the replica method [51], one can directly express the Rényi entropies in terms

of partition functions [52] readily obtainable in the NIS. In this approach, systems in any space-time dimension can be studied, with the obvious limitation that performance is limited by the total number of sites in the system, hence reducing the available volumes in higher dimensions. Still, due to its straightforwardness, it should be seen as a valuable alternative to studying information-theoretic properties of one-, two-, or three-dimensional quantum systems. In particular, the method based on path integral quantization can be considered as complementary to variational approaches based on autoregressive networks such as Refs. [30,31] where the approximation of the ground-state wave function is constructed. First, with an ergodic sampling it provides unbiased results (see the discussion in Appendix F); second, it gives access to the entanglement entropy for thermal states [3,53].

One should also observe that our proposal in principle can work unaltered for systems with continuous degrees of freedom. Neural generative networks have already been discussed in the context of ϕ^4 classical field theory [41] as well as U(1), SU(N) [54], and the Schwinger model gauge theories [55]. In all these cases, one would introduce a conditional normalizing flow or another proposal that would give access to conditional probabilities (see, for example, Ref. [56]). For instance, a hierarchical construction similar to the HAN algorithm employing normalizing flows could be used to simulate the ϕ^4 classical field theory. In this way, conditional probabilities would be naturally introduced and could be used to calculate the mutual information for some specific partitioning. The combination of the proposed method of measuring information-theoretic quantities with the recent advancements [46] in machine-learning enhanced algorithms for simulating four-dimensional lattice quantum chromodynamics (LQCD) can open new ways of investigating this phenomenologically important theory, for instance by studying quantum correlations and entanglement of the QCD vacuum.

ACKNOWLEDGMENTS

Computer time allocation “plngn” on the Prometheus and ARES supercomputers hosted by AGH Cyfronet in Kraków, Poland was used through the Polish PLGRID consortium. T.S. kindly acknowledges the support of the Polish National Science Center (NCN) Grants No. 2019/32/C/ST2/00202 and No. 2021/43/D/ST2/03375 and support of the Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University Grant No. LM/23/ST. P.K. acknowledge that this research was partially funded by the Priority Research Area Digiworld under the program Excellence Initiative—Research University at the Jagiellonian University in Kraków. P.K. and T.S. thank Alberto Ramos and the University of Valencia for their hospitality during the stay when part of this work was performed and discussed. We also acknowledge very fruitful discussions with Leszek Hadasz on entanglement entropy.

APPENDIX A: DETAILS OF THE METHOD

To calculate the mutual information according to the definition Eq. (6), we need to estimate the averages $\langle \dots \rangle_{q_\theta}$, which

we do using the Monte Carlo approach,

$$\langle O \rangle_{q_\theta} \approx \frac{1}{N} \sum_{i=1}^N O(\mathbf{s}_i), \quad \mathbf{s}_i \sim q_\theta, \quad (\text{A1})$$

where the configurations \mathbf{s}_i are drawn from the distribution q_θ . In particular, the partition function can be approximated as

$$\begin{aligned} Z &= \sum_{\mathbf{a}, \mathbf{b}} q_\theta(\mathbf{a}, \mathbf{b}) \frac{e^{-\beta E(\mathbf{a}, \mathbf{b})}}{q_\theta(\mathbf{a}, \mathbf{b})} \equiv \langle \hat{w}(\mathbf{a}, \mathbf{b}) \rangle_{q_\theta(\mathbf{a}, \mathbf{b})} \\ &\approx \frac{1}{N} \sum_{i=1}^N \hat{w}(\mathbf{a}_i, \mathbf{b}_i) \equiv \hat{Z}_N, \quad (\mathbf{a}_i, \mathbf{b}_i) \sim q_\theta. \end{aligned} \quad (\text{A2})$$

In a similar way,

$$Z(\mathbf{a}) = \sum_{\mathbf{b}} q_\theta(\mathbf{b}|\mathbf{a}) \frac{e^{-\beta E(\mathbf{a}, \mathbf{b})}}{q_\theta(\mathbf{b}|\mathbf{a})} \equiv \langle \hat{w}(\mathbf{b}|\mathbf{a}) \rangle_{q_\theta(\mathbf{b}|\mathbf{a})}, \quad (\text{A3})$$

where we introduced $\hat{w}(\mathbf{b}|\mathbf{a}) = e^{-\beta E(\mathbf{a}, \mathbf{b}_i)} / q_\theta(\mathbf{b}|\mathbf{a})$, and $\langle \dots \rangle_{q_\theta(\mathbf{b}|\mathbf{a})}$ is an average over the conditional probability. This can be approximated as

$$Z(\mathbf{a}) \approx \hat{Z}_M(\mathbf{a}) = \frac{1}{M} \sum_{i=1}^M \frac{e^{-\beta E(\mathbf{a}, \mathbf{b}_i)}}{q_\theta(\mathbf{b}_i|\mathbf{a})}, \quad \mathbf{b}_i \sim q_\theta(\mathbf{b}|\mathbf{a}), \quad (\text{A4})$$

where we have denoted by M the number of configurations used to estimate that average. In general, M is independent of N ; for practical reasons, we always take $M < N$. To calculate $Z(\mathbf{b})$ we need $q_\theta(\mathbf{a}|\mathbf{b})$, which is not readily available. Therefore, we define a new distribution \tilde{q}_θ obtained by feeding a permutation of spins that swaps \mathbf{a} and \mathbf{b} into q_θ . For VAN this is obtained by reversing the order of spins,

$$\tilde{q}_\theta(s^1, s^2, \dots, s^{n_A+n_B}) \equiv q_\theta(s^{n_A+n_B}, s^{n_A+n_B-1}, \dots, s^1). \quad (\text{A5})$$

We can use \tilde{q}_θ together with factorization Eq. (8) to obtain $\tilde{q}_\theta(\mathbf{a}|\mathbf{b})$ and $Z(\mathbf{b})$,

$$Z(\mathbf{b}) \approx \hat{Z}_M(\mathbf{b}) = \frac{1}{M} \sum_{i=1}^M \frac{e^{-\beta E(\mathbf{a}_i, \mathbf{b})}}{\tilde{q}_\theta(\mathbf{a}_i|\mathbf{b})}, \quad \mathbf{a}_i \sim \tilde{q}_\theta(\mathbf{a}|\mathbf{b}). \quad (\text{A6})$$

Please note that in general $\tilde{q}_\theta(\mathbf{s}) \neq q_\theta(\mathbf{s})$, but for a well-trained network these two distributions should be close. Furthermore, the approximation Eq. (A6) is valid for any \tilde{q}_θ .

For the remaining terms, we use the standard way to calculate the average, i.e., the mean energy is

$$\begin{aligned} &\frac{1}{Z} \langle \hat{w}(\mathbf{a}, \mathbf{b}) E(\mathbf{a}, \mathbf{b}) \rangle_{q_\theta(\mathbf{a}, \mathbf{b})} \\ &\approx \frac{1}{N \hat{Z}_N} \sum_{i=1}^N \hat{w}(\mathbf{a}_i, \mathbf{b}_i) E(\mathbf{a}_i, \mathbf{b}_i), \quad (\mathbf{a}_i, \mathbf{b}_i) \sim q_\theta. \end{aligned} \quad (\text{A7})$$

In that case, the statistical uncertainty of the result is governed by the square root of the number of samples generated, N . The mean logarithm of $Z(\mathbf{a})$ is estimated as

$$\begin{aligned} &\frac{1}{Z} \langle \hat{w}(\mathbf{a}, \mathbf{b}) \log Z(\mathbf{a}) \rangle_{q_\theta(\mathbf{a}, \mathbf{b})} \\ &\approx \frac{1}{N \hat{Z}_N} \sum_{i=1}^N \hat{w}(\mathbf{a}_i, \mathbf{b}_i) \log \hat{Z}_M(\mathbf{a}_i), \quad (\mathbf{a}_i, \mathbf{b}_i) \sim q_\theta \end{aligned} \quad (\text{A8})$$

and analogously for $\log Z(\mathbf{b})$.

The procedure of calculating Eq. (A8) is the following: (i) we draw N configurations of the entire system using the probability distribution q_θ encoded by the neural network; (ii) we estimate the full partition function \hat{Z}_N according to (A2); (iii) for each of the N configurations, we freeze the subsystem \mathbf{a} and generate additional $M - 1$ configurations with the random part \mathbf{b} using the conditional probability $q_\theta(\mathbf{b}|\mathbf{a})$; (iv) for each of the N configurations, we calculate $\hat{Z}_M(\mathbf{a})$ using Eq. (A4) using the M configurations with the same frozen part \mathbf{a} ; and (v) we calculate the average over N configurations of the logarithm of $\hat{Z}_M(\mathbf{a})$ according to Eq. (A8). Steps (iii)–(v) can be analogously applied to calculate $\langle w \log Z(\mathbf{b}) \rangle_{q_\theta}$.

Adding up the terms according to Eq. (6), we obtain the estimator of the mutual information $\hat{I}_{N,M}$, which yields the exact value in the following infinite-statistics limit:

$$I = \lim_{N \rightarrow \infty} \lim_{M \rightarrow \infty} \hat{I}_{N,M}. \quad (\text{A9})$$

At finite N and M our estimator $\hat{I}_{N,M}$ is biased due to the non-linearity of the log function. As was shown in the Appendix of Ref. [41], $\log \hat{Z}_N$ has a bias due to the finite value of N , which is given by

$$\begin{aligned} \mathcal{B}[\log \hat{Z}_N] = & -\frac{1}{2N} \frac{\langle \hat{w}(\mathbf{a}, \mathbf{b})^2 \rangle_{q_\theta(\mathbf{a}, \mathbf{b})} - \langle \hat{w}(\mathbf{a}, \mathbf{b}) \rangle_{q_\theta(\mathbf{a}, \mathbf{b})}^2}{\langle \hat{w}(\mathbf{a}, \mathbf{b}) \rangle_{q_\theta(\mathbf{a}, \mathbf{b})}^2} \\ & + \mathcal{O}\left(\frac{1}{N^2}\right), \end{aligned} \quad (\text{A10})$$

where the numerator is equal to the variance of Z . This is a systematic bias of the observable and would be zero for a perfectly trained network, i.e., when $q_\theta = p \Leftrightarrow \hat{w}(\mathbf{a}, \mathbf{b}) = \text{const}$. For N sufficiently large, this bias can be neglected as it decreases with $1/N$ and is usually smaller than the statistical noise, which decreases only as $\sim 1/\sqrt{N}$. In practical terms, with our statistics of $N \approx 10^6$ we are always working in this regime. Also, $\log \hat{Z}_M(\mathbf{a})$ is affected by a similar bias for a finite value of M . Repeating the calculation for that observable, one can obtain, in analogy to Eq. (A10),

$$\begin{aligned} \mathcal{B}[\log \hat{Z}_M(\mathbf{a})] = & -\frac{1}{2M} \frac{\langle \hat{w}(\mathbf{b}|\mathbf{a})^2 \rangle_{q_\theta(\mathbf{b}|\mathbf{a})} - \langle \hat{w}(\mathbf{b}|\mathbf{a}) \rangle_{q_\theta(\mathbf{b}|\mathbf{a})}^2}{\langle \hat{w}(\mathbf{b}|\mathbf{a}) \rangle_{q_\theta(\mathbf{b}|\mathbf{a})}^2} \\ & + \mathcal{O}\left(\frac{1}{M^2}\right). \end{aligned} \quad (\text{A11})$$

In the practical implementation, it is difficult to afford $M \sim N \sim 10^6$ so typically $M \sim 10^2 \ll N$. Therefore, neglecting $\mathcal{B}[\log \hat{Z}_N]$, the final systematic bias of $\hat{I}_{N,M}$ is given by

$$\begin{aligned} \mathcal{B}[\hat{I}_{N,M}] = & \frac{1}{2ZM} \left\langle \hat{w}(\mathbf{a}, \mathbf{b}) \right. \\ & \times \left(\frac{\langle \hat{w}(\mathbf{b}|\mathbf{a})^2 \rangle_{q_\theta(\mathbf{b}|\mathbf{a})} - \langle \hat{w}(\mathbf{b}|\mathbf{a}) \rangle_{q_\theta(\mathbf{b}|\mathbf{a})}^2}{\langle \hat{w}(\mathbf{b}|\mathbf{a}) \rangle_{q_\theta(\mathbf{b}|\mathbf{a})}^2} + \mathbf{a} \leftrightarrow \mathbf{b} \right) \Bigg\rangle_{q_\theta} \\ & + \mathcal{O}\left(\frac{1}{M^2}\right). \end{aligned} \quad (\text{A12})$$

We expect that a positive bias may affect I , which decreases as $\sim 1/M$ for large enough M . In our calculation, we take $M = 64, 128, 256$ and in some cases $M = 512$ and 1024 , and

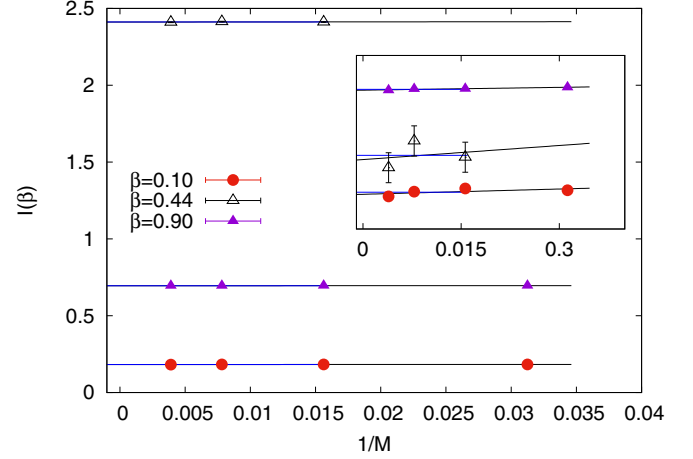


FIG. 6. Dependence on M at $L = 18$ for strip partitioning for three representative values of $\beta = 0.1, 0.44$, and 0.9 . Two extrapolations are shown: constant and linear. Their difference is taken as a systematic uncertainty. The inset shows the same data shifted by a constant amount to show the details.

we perform an extrapolation to $M = \infty$. We attach a systematic uncertainty in that step by comparing the result of a constant extrapolation to the data at the two largest values of M with the linear $a + b^2/M$ extrapolation to the data at the three largest values of M . In fact, in most cases we observe that $\hat{I}_{N,M}$ is equal within statistical errors for all values of M . We provide representative examples of such extrapolations in Fig. 6.

APPENDIX B: NEURAL NETWORK ARCHITECTURES

In this work, we use two algorithms employing autoregressive neural networks: variational autoregressive networks (VANs) [35] and their modification, called hierarchical autoregressive networks (HANs) [37]. Both architectures provide access to the conditional probabilities Eq. (8). The training of the neural networks consists in generating N_{batch} spin configurations $\{\mathbf{s}_1, \dots, \mathbf{s}_{N_{\text{batch}}}\}$ from the probability distribution q_θ currently encoded in the neural weights and calculating the variational estimate of the free energy:

$$F_q = \frac{1}{\beta} \sum_{k=1}^{N_{\text{batch}}} q_\theta(\mathbf{s}_k) [\beta H(\mathbf{s}_k) + \log q_\theta(\mathbf{s}_k)], \quad (\text{B1})$$

which, up to an additive constant, corresponds to the backward Kullback-Leibler divergence between q_θ and p . With this loss function, the weights θ are updated according to the gradient back-propagation algorithm with an ADAM optimizer [57].

We use dense (fully connected) neural networks with two layers. The autoregressive property is enforced by multiplying half of the weights by 0. The neural network for the VAN approach is constructed as

$$\mathcal{N}_\theta = \sigma \circ l_2 \circ \text{ReLU} \circ l_1, \quad (\text{B2})$$

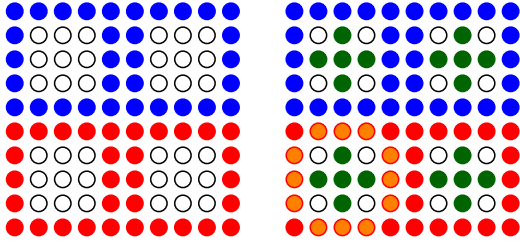


FIG. 7. Sketch of the partitioning in HAN of an $L \times L = 10 \times 10$ lattice. Border spins are shown in blue and red, and the first set of interior spins is shown in green. The probabilities of the green spins depend conditionally on all spins surrounding them (colored orange for one set of interior spins).

where layer l acts on the input vector $x \in \mathbb{R}^n$ in the following way:

$$y_i = \sum_{j < i} W_{ij} x_j + b_i \quad \text{for } y = l(x). \quad (\text{B3})$$

The autoregressive property of the neural network is guaranteed by the fact that the above sum runs only up to $i - 1$.

The activation functions act pointwise on their arguments. We use two types of functions: the rectifier $\text{ReLU}(x) = \max(0, x)$ and the sigmoid $\sigma(x) = 1/(1 + e^{-x})$. Weights W_{ij} 's and biases b_i 's are parameters of the neural network and collectively denoted as θ .

The neural network acts on a spin configuration $\{s^1, \dots, s^{L^2}\}$ as follows: $\mathcal{N}_\theta(\{s^1, \dots, s^{L^2}\}) = \{\hat{s}^1, \dots, \hat{s}^{L^2}\}$, where each \hat{s}^i is interpreted as the conditional probability of the i th spin to be pointed up given all the previous $i - 1$ spins were fixed: $q_\theta(s^i = +1 | s^{i-1}, \dots, s^1) = \hat{s}^i$. The autoregressive condition Eq. (B3) guarantees that the conditional probability of the i th spin depends only on the previous $i - 1$ spins.

In the VAN approach, the conditional probabilities of all L^2 spins are generated by a single neural network \mathcal{N}_θ described above. It has L^2 input neurons and L^2 output neurons. To fix all the L^2 spins, one needs to evoke the neural network L^2 times: at the i th invocation one calculates \hat{s}_i , which is then used to draw a value of s_i . This leads to an unfavorable $\sim L^6$ rise in the numerical cost of the generation of samples (which is the main numerical cost of the algorithm). Additionally, the larger the system is, the more configurations are needed to train the network to the given quality [measured, for example, by the effective sample size (ESS) [34]]. This makes the effective numerical cost grow even faster than $\sim L^6$ [43].

To mitigate the above-mentioned problems, we use a version of the HAN algorithm proposed in Ref. [37], which introduces a specific enumeration of the spins (see Fig. 7): we first fix the frames (denoted in blue and red) which surround all the remaining spins, then iteratively we fix the ‘‘crosses’’ inside the frames (green crosses on the figure) to end up with single spins. There are several advantages of this division, which come from the fact that, for nearest-neighbor interactions, the probability of a group of spins depends conditionally on the values of spins on a closed contour enclosing that set (a result known also as the Hammersley-Clifford theorem in the literature [58,59]). First, instead of one network for the whole system, one can use several smaller networks that fix

the spins at a given level of hierarchy (levels of hierarchy in our Fig. 7 are denoted by different colors³). The networks generating the spins in the crosses depend conditionally only on the surrounding spins (denoted by orange for one cross). Therefore, those networks are much smaller than the single network in the VAN approach, hence the numerical cost is significantly reduced. What is more, at a given level of the hierarchy, the crosses can be generated in parallel. As was shown in [37], the numerical cost of HAN is much smaller than VAN; it scales⁴ as $\sim L^3$. Smaller networks are also easier to train, so with the same number of epochs one reaches a much higher ESS with HAN.

In practice, we are limited to $L < 30$ in VAN simulations, whereas with HAN we can reach sizes $L = 66$ (or even $L = 130$ for some temperatures). On the other hand, as the crosses in HAN can only have specific numbers of spins (in order to close the recurrence), this algorithm can be used for simulations with $L = 10, 18, 34, 66, 130, \dots$, whereas VAN can be applied to any L value. Also, VAN is much more elastic concerning the possible divisions into **a** and **b** subsystems: with the proper enumeration of the spins, *any* division is possible in VAN. In HAN, with the implementation described here, only strip and square partitionings are possible. Each division that requires a specific enumeration of the spins requires also new training.

In this manuscript, we used both VAN and HAN algorithms: the latter was used to obtain mutual information for $L = 10, 18, 34, 66, 130$, with strip and square geometries. All other values of L were obtained using VAN. To show the numerical cost of the method, we provide the runtime of the $L = 66$ system simulated with the HAN algorithm (one of the largest simulated): we trained the hierarchy of neural networks for 220 000 epochs, which took 110 h on a NVIDIA V100 graphic card to reach the ESS of 0.508. The model had 2 200 000 parameters. The consecutive measurement of the mutual information for $M = 256$ took 20 h. After repeating such measurements for $M = 128$ (10 h of running) and $M = 64$ (5 h of running), we performed the extrapolation in M . This allowed us to achieve a total uncertainty of I of 0.15% at $\beta = 0.44$, which is the worst case.

³Note that, in principle, red and blue spins could be treated as belonging to the same level of the HAN hierarchy. However, here we explicitly distinguish them because we need to conditionally sample red spins based on blue spins and calculate the conditional probabilities needed for the mutual information observable. Therefore, red spins have a separate network from the blue ones.

⁴We note that the $\sim L^6$ scaling for the VAN approach is due to the sampling procedure: the neural network has to be evaluated L^2 times to fix all the spins and naively each evaluation requires $\sim L^4$ floating point operation as the network has L^2 neurons per layer. However, this counting assumes that operations using all neural network weights are done during each evaluation of the network. In a more efficient implementation of autoregressive networks, one would perform multiplication with only the weights which are necessary to fix the given spin. It is easy to check that the scaling of the numerical cost of the VAN algorithm is then $\sim L^4$ [60]. By the same argument, the scaling for the HAN algorithm with optimal implementation is L^2 . This is because the first neural network in the HAN hierarchy describes the spins at the border of the lattice, which has size $\sim L$.

APPENDIX C: CHESSBOARD FACTORIZATION

The consequence of the Hammersley-Clifford theorem is that once all four nearest-neighbors $n(s^i)$ of the spin s^i are fixed, its probability is given by

$$p_{\text{chess}}(s^i = \pm 1, n(s^i)) = [1 + \exp(\mp 2\beta \mathcal{S}_i)]^{-1}, \quad (\text{C1})$$

where $\mathcal{S}_i = \sum_{j \in n(s^i)} s^j$, and j runs over four nearest neighbors of spin i .

In [36] we proposed to use this property to reduce the number of spins that need to be generated by the network by a factor of 2. The idea is to divide the spin configuration into subsystems according to the chessboard pattern: spins at white fields are fixed by the network using the usual VAN algorithm (we call them \mathbf{a}). The other half of the system (called \mathbf{b}) can be drawn from probabilities Eq. (C1).

Such factorization makes calculating the mutual information using chessboard partitioning particularly simple. We note that due to symmetry reasons, $\langle \hat{w} \log Z(\mathbf{a}) \rangle_{q_\theta} = \langle \hat{w} \log Z(\mathbf{b}) \rangle_{q_\theta}$, hence only the former needs to be calculated. For this purpose, we write

$$E(\mathbf{a}, \mathbf{b}) = - \sum_{(i,j)} s^i s^j = - \sum_{i \in \mathbf{b}} s^i \mathcal{S}_i(\mathbf{a}), \quad (\text{C2})$$

where we explicitly denoted that \mathcal{S}_i for $i \in \mathbf{b}$ depends only on the subsystem \mathbf{a} (since any spin from \mathbf{b} has nearest neighbors only from \mathbf{a}). Then

$$\begin{aligned} \log Z(\mathbf{a}) &= \log \sum_{\mathbf{b}} e^{\beta \sum_{i \in \mathbf{b}} s^i \mathcal{S}_i(\mathbf{a})} \\ &= \log \prod_{i \in \mathbf{b}} \left(\sum_{s^i \in \{-1, 1\}} e^{\beta s^i \mathcal{S}_i(\mathbf{a})} \right) \\ &= \sum_{i \in \mathbf{b}} \log [2 \cosh \beta \mathcal{S}_i(\mathbf{a})], \end{aligned} \quad (\text{C3})$$

which means that $\log Z(\mathbf{a})$ can be exactly calculated for any \mathbf{a} . In other words, in the chessboard division, there is no need for Eq. (A4); the observable $\log Z(\mathbf{a})$ can be exactly determined for any configuration and is not biased. This means that we need much less statistics to determine mutual information.

APPENDIX D: SYMMETRIES

To obtain good quality training of the autoregressive neural network, it is crucial to impose global symmetries of the system through the symmetrization of the loss function [35–38]. This can be achieved by defining a symmetrized probability for each generated configuration

$$\bar{q}_\theta(\mathbf{s}) = \frac{1}{S} \sum_{i=1}^S q_\theta(h_i(\mathbf{s})), \quad (\text{D1})$$

where h_i , $i = 1, \dots, S$ are the symmetry operators defined as transformations of the configuration space which keep the energy unchanged. This symmetrized probability replaces q_θ in the Kullback-Leibler (KL) loss function:

$$\bar{D}_{\text{KL}}(q_\theta | p) = \sum_{k=1}^{N_{\text{batch}}} \bar{q}_\theta(\mathbf{s}_k) \log \left(\frac{\bar{q}_\theta(\mathbf{s}_k)}{p(\mathbf{s}_k)} \right), \quad (\text{D2})$$

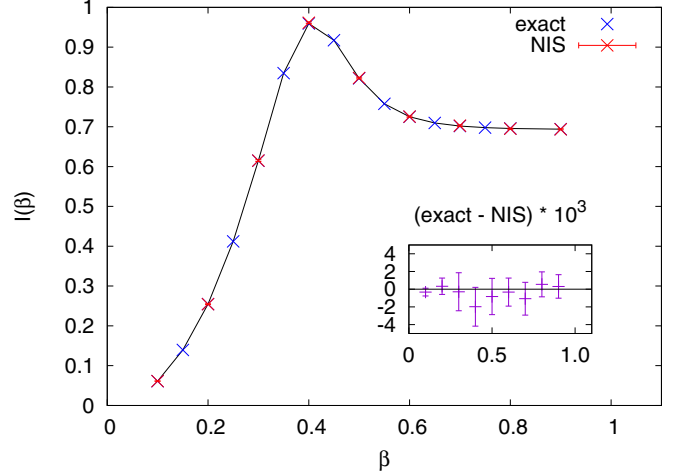


FIG. 8. Comparison of I for strip partitioning calculated exactly from Eq. (4) and with the NIS approach through Eq. (6) for $L = 6$. NIS data have statistical uncertainties, which are smaller than the symbol size. Differences between the exact values and NIS are magnified in the inset.

and in the definition of importance ratios Eq. (7). See Appendix B of Ref. [38] for a more detailed discussion.

In the case of $Z(\mathbf{a})$ or $Z(\mathbf{b})$, the calculation of the conditional probabilities $q_\theta(\mathbf{b}|\mathbf{a})$ and $q_\theta(\mathbf{a}|\mathbf{b})$ is needed. Since the partitioning itself breaks the symmetry, symmetrization is not possible. Hence, also at the level of conditional probabilities, one cannot impose any symmetry and so one is left with the nonsymmetrized version of q_θ .

APPENDIX E: NUMERICAL CALCULATION OF I FOR $L = 6$

For very small lattice sizes, the number of states is small enough to calculate the mutual information directly from Eq. (4). We were able to perform this calculation for $L = 6$, where the number of states is 2^{36} . Comparison with the results using VAN is shown in Fig. 8. We found perfect agreement, which we show by plotting the difference between the two methods in the inset.

APPENDIX F: COMPARISON WITH MICE

We now discuss the differences between the NIS and MICE [32] methods. In the MICE approach, one uses the fact that mutual information can be treated as the KL divergence [33], which in turn satisfies the Donsker and Varadhan theorem [61]: the I is an upper bound of some variable \mathcal{M}_θ over the set of functions parametrized by a neural network. One then trains the network to maximize \mathcal{M}_θ . With such a construction, the MICE method is variational: it provides an approximation of I that is in general smaller than the true value. However, as is typical in variational approaches, without knowing the exact result one cannot deduce the systematic uncertainty of the approximation. Applying MICE to the 2D Ising model, the authors of Ref. [32] obtained global entropy with an accuracy below 5%. The bias for the MI, for which analytic values are not known, may be larger.

The NIS approach that we discuss in this manuscript circumvents the above-mentioned inaccuracy of the variational approach using a reweighting procedure: due to the fact that

the VAN/HAN procedures use explicit q_θ probabilities, one can remove the difference between p and q_θ by calculating the weights $\hat{w}(a, b)$ and $\hat{w}(a|b)$, as discussed in Appendix A, and correcting the final outcome. Therefore, our method provides,

in the limit of large statistics, the exact result assuming the ergodicity condition of the algorithm is satisfied. The combined statistical and systematic uncertainty of MI obtained with NIS is less than 0.1%.

-
- [1] B. Zeng, X. Chen, D.-L. Zhou, and X.-G. Wen, *Quantum Information Meets Quantum Matter* (Springer, New York, 2019).
- [2] K. Okunishi, T. Nishino, and H. Ueda, Developments in the tensor network—from statistical mechanics to quantum entanglement, *J. Phys. Soc. Jpn.* **91**, 062001 (2022).
- [3] P. Calabrese and J. Cardy, Entanglement entropy and quantum field theory, *J. Stat. Mech.* (2004) P06002.
- [4] F. C. Alcaraz and M. S. Sarandy, Finite-size corrections to entanglement in quantum critical systems, *Phys. Rev. A* **78**, 032319 (2008).
- [5] J. Eisert, M. Cramer, and M. B. Plenio, Colloquium: Area laws for the entanglement entropy, *Rev. Mod. Phys.* **82**, 277 (2010).
- [6] F. C. Alcaraz, M. I. Berganza, and G. Sierra, Entanglement of low-energy excitations in conformal field theory, *Phys. Rev. Lett.* **106**, 201601 (2011).
- [7] J. Bhattacharya, M. Nozaki, T. Takayanagi, and T. Ugajin, Thermodynamical property of entanglement entropy for excited states, *Phys. Rev. Lett.* **110**, 091602 (2013).
- [8] J. Eisert, Entanglement and tensor network states, *Model. Simul.* **3**, 520 (2013).
- [9] M. Goldstein and E. Sela, Symmetry-resolved entanglement in many-body systems, *Phys. Rev. Lett.* **120**, 200602 (2018).
- [10] L. Capizzi, P. Ruggiero, and P. Calabrese, Symmetry resolved entanglement entropy of excited states in a CFT, *J. Stat. Mech.* (2020) 073101.
- [11] A. Kitaev and J. Preskill, Topological entanglement entropy, *Phys. Rev. Lett.* **96**, 110404 (2006).
- [12] M. Levin and X.-G. Wen, Detecting topological order in a ground state wave function, *Phys. Rev. Lett.* **96**, 110405 (2006).
- [13] J. Iaconis, S. Inglis, A. B. Kallin, and R. G. Melko, Detecting classical phase transitions with renyi mutual information, *Phys. Rev. B* **87**, 195134 (2013).
- [14] P. Fromholz, G. Magnifico, V. Vitale, T. Mendes-Santos, and M. Dalmonte, Entanglement topological invariants for one-dimensional topological superconductors, *Phys. Rev. B* **101**, 085136 (2020).
- [15] M. Tajik *et al.*, Experimental verification of the area law of mutual information in quantum field simulator, *Nat. Phys.* **19**, 1022 (2023).
- [16] M. M. Wolf, F. Verstraete, M. B. Hastings, and J. I. Cirac, Area laws in quantum systems: Mutual information and correlations, *Phys. Rev. Lett.* **100**, 070502 (2008).
- [17] M. Srednicki, Entropy and area, *Phys. Rev. Lett.* **71**, 666 (1993).
- [18] J. D. Bekenstein, Black holes and entropy, *Phys. Rev. D* **7**, 2333 (1973).
- [19] S. W. Hawking, Particle creation by black holes, *Commun. Math. Phys.* **43**, 199 (1975); Erratum: Particle creation by black holes, **46**, 206(E) (1976).
- [20] S. Sachdev and J. Ye, Gapless spin-fluid ground state in a random quantum heisenberg magnet, *Phys. Rev. Lett.* **70**, 3339 (1993).
- [21] J. M. Maldacena, The large N limit of superconformal field theories and supergravity, *Int. J. Theor. Phys.* **38**, 1113 (1999); The large N limit of superconformal field theories and supergravity, *Adv. Theor. Math. Phys.* **2**, 231 (1998).
- [22] L. Amico, R. Fazio, A. Osterloh, and V. Vedral, Entanglement in many-body systems, *Rev. Mod. Phys.* **80**, 517 (2008).
- [23] J.-M. Stéphan, S. Furukawa, G. Misguich, and V. Pasquier, Shannon and entanglement entropies of one- and two-dimensional critical wave functions, *Phys. Rev. B* **80**, 184421 (2009).
- [24] F. C. Alcaraz and M. A. Rajabpour, Universal behavior of the shannon mutual information of critical quantum chains, *Phys. Rev. Lett.* **111**, 017201 (2013).
- [25] F. C. Alcaraz and M. A. Rajabpour, Universal behavior of the shannon and rényi mutual information of quantum critical chains, *Phys. Rev. B* **90**, 075132 (2014).
- [26] J.-M. Stéphan, G. Misguich, and V. Pasquier, Rényi entropy of a line in two-dimensional Ising models, *Phys. Rev. B* **82**, 125455 (2010).
- [27] J. Wilms, M. Troyer, and F. Verstraete, Mutual information in classical spin models, *J. Stat. Mech.* (2011) P10011.
- [28] H. W. Lau and P. Grassberger, Information theoretic aspects of the two-dimensional Ising model, *Phys. Rev. E* **87**, 022128 (2013).
- [29] A. Pizzi, D. Malz, A. Nunnenkamp, and J. Knolle, Bridging the gap between classical and quantum many-body information dynamics, *Phys. Rev. B* **106**, 214303 (2022).
- [30] M. Hibat-Allah, M. Ganahl, L. E. Hayward, R. G. Melko, and J. Carrasquilla, Recurrent neural network wave functions, *Phys. Rev. Res.* **2**, 023358 (2020).
- [31] Z. Wang and E. J. Davis, Calculating Rényi entropies with neural autoregressive quantum states, *Phys. Rev. A* **102**, 062413 (2020).
- [32] A. Nir, E. Sela, R. Beck, and Y. Bar-Sinai, Machine-learning iterative calculation of entropy for physical systems, *Proc. Natl. Acad. Sci. USA* **117**, 30234 (2020).
- [33] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, Mutual information neural estimation, in *Proceedings of the 35th International Conference on Machine Learning*, edited by J. Dy and A. Krause, Proceedings of Machine Learning Research, Vol. 80 (PMLR, 2018), pp. 531–540.
- [34] J. S. Liu, Metropolized independent sampling with comparisons to rejection sampling and importance sampling, *Statist. Comput.* **6**, 113 (1996).
- [35] D. Wu, L. Wang, and P. Zhang, Solving statistical mechanics using variational autoregressive networks, *Phys. Rev. Lett.* **122**, 080602 (2019).
- [36] P. Białas, P. Korcyl, and T. Stebel, Analysis of autocorrelation times in neural Markov chain Monte Carlo simulations, *Phys. Rev. E* **107**, 015303 (2023).

- [37] P. Białas, P. Korcyl, and T. Stebel, Hierarchical autoregressive neural networks for statistical systems, *Comput. Phys. Commun.* **281**, 108502 (2022).
- [38] P. Białas, P. Czarnota, P. Korcyl, and T. Stebel, Simulating first-order phase transition with hierarchical autoregressive networks, *Phys. Rev. E* **107**, 054127 (2023).
- [39] K. A. Nicoli, S. Nakajima, N. Strodthoff, W. Samek, K.-R. Müller, and P. Kessel, Asymptotically unbiased estimation of physical observables with neural samplers, *Phys. Rev. E* **101**, 023304 (2020).
- [40] M. S. Albergo, G. Kanwar, and P. E. Shanahan, Flow-based generative models for markov chain Monte Carlo in lattice field theory, *Phys. Rev. D* **100**, 034515 (2019).
- [41] K. A. Nicoli, C. J. Anders, L. Funcke, T. Hartung, K. Jansen, P. Kessel, S. Nakajima, and P. Stornati, Estimation of thermodynamic observables in lattice field theories with deep generative models, *Phys. Rev. Lett.* **126**, 032001 (2021).
- [42] M. S. Albergo, D. Boyda, D. C. Hackett, G. Kanwar, K. Cranmer, S. Racanière, D. Jimenez Rezende, and P. E. Shanahan, Introduction to normalizing flows for lattice field theory, [arXiv:2101.08176](https://arxiv.org/abs/2101.08176).
- [43] L. Del Debbio, J. Marsh Rossney, and M. Wilson, Efficient modeling of trivializing maps for lattice ϕ^4 theory using normalizing flows: A first look at scalability, *Phys. Rev. D* **104**, 094507 (2021).
- [44] D. Albandea, P. Hernández, A. Ramos, and F. Romero-López, Improved topological sampling for lattice gauge theories, *PoS LATTICE2021*, 183 (2022).
- [45] P. Białas, P. Korcyl, and T. Stebel, Gradient estimators for normalising flows, [arXiv:2202.01314](https://arxiv.org/abs/2202.01314).
- [46] R. Abbott *et al.*, Sampling QCD field configurations with gauge-equivariant flow models, *PoS LATTICE2022*, 036 (2023).
- [47] P. Białas, P. Korcyl, and T. Stebel, Training normalizing flows with computationally intensive target probability distributions, [arXiv:2308.13294](https://arxiv.org/abs/2308.13294).
- [48] D. C. Hackett, C.-C. Hsieh, M. S. Albergo, D. Boyda, J.-W. Chen, K.-F. Chen, K. Cranmer, G. Kanwar, and P. E. Shanahan, Flow-based sampling for multimodal distributions in lattice field theory, [arXiv:2107.00734](https://arxiv.org/abs/2107.00734).
- [49] S. Ciarella, J. Trinquier, M. Weigt, and F. Zamponi, Machine-learning-assisted Monte Carlo fails at sampling computationally hard problems, *Mach. Learn.* **4**, 010501 (2023).
- [50] L. Illing Midgley, V. Stimper, G. N. C. Simm, B. Schölkopf, and J. M. Hernández-Lobato, Flow annealed importance sampling bootstrap, [arXiv:2208.01893](https://arxiv.org/abs/2208.01893).
- [51] C. G. Callan, Jr. and F. Wilczek, On geometric entropy, *Phys. Lett. B* **333**, 55 (1994).
- [52] S. Humeniuk and T. Roscilde, Quantum Monte Carlo calculation of entanglement Renyi entropies for generic quantum systems, *Phys. Rev. B* **86**, 235116 (2012).
- [53] E. Itou, K. Nagata, Y. Nakagawa, A. Nakamura, and V. I. Zakharov, Entanglement in four-dimensional SU(3) gauge theory, *Prog. Theor. Exp. Phys.* **2016**, 061B01 (2016).
- [54] R. Abbott *et al.*, Gauge-equivariant flow models for sampling in lattice field theories with pseudofermions, *Phys. Rev. D* **106**, 074506 (2022).
- [55] M. S. Albergo, D. Boyda, K. Cranmer, D. C. Hackett, G. Kanwar, S. Racanière, D. J. Rezende, F. Romero-López, P. E. Shanahan, and J. M. Urban, Flow-based sampling in the lattice Schwinger model at criticality, *Phys. Rev. D* **106**, 014514 (2022).
- [56] L. Wang, Y. Jiang, L. He, and K. Zhou, Continuous-mixture autoregressive networks learning the Kosterlitz-Thouless transition, *Chin. Phys. Lett.* **39**, 120502 (2022).
- [57] D. Kingma and J. Ba, Adam: A method for stochastic optimization, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [58] J. M. Hammersley and P. Clifford, *Markov fields on finite graphs and lattices*, 1971 (unpublished).
- [59] P. Clifford, Markov random fields in statistics, in *Disorder in Physical Systems, A Volume in Honour of John M. Hammersley* (Oxford University Press, Oxford, UK, 1990).
- [60] B. Uria, M.-A. Côté, K. Gregor, I. Murray, and H. Larochelle, Neural autoregressive distribution estimation, *J. Mach. Learn. Res.* **17**, 1 (2016).
- [61] M. Donsker and S. Varadhan, Asymptotic evaluation of certain markov process expectations for large time. IV, *Commun. Pure Appl. Math.* **36**, 183 (1983).