Deep Autoregressive Models for the Efficient Variational Simulation of Many-Body Quantum Systems

Or Sharir,[1,*] Yoav Levine,[1,†] Noam Wies,[1,‡] Giuseppe Carleo,[2,§] and Amnon Shashua[1,‖]
[1]The Hebrew University of Jerusalem, Jerusalem 9190401, Israel
[2]Center for Computational Quantum Physics, Flatiron Institute, 162 5th Avenue, New York, New York 10010, USA

(Received 12 May 2019; published 16 January 2020)

Artificial neural networks were recently shown to be an efficient representation of highly entangled many-body quantum states. In practical applications, neural-network states inherit numerical schemes used in variational Monte Carlo method, most notably the use of Markov-chain Monte Carlo (MCMC) sampling to estimate quantum expectations. The local stochastic sampling in MCMC caps the potential advantages of neural networks in two ways: (i) Its intrinsic computational cost sets stringent practical limits on the width and depth of the networks, and therefore limits their expressive capacity; (ii) its difficulty in generating precise and uncorrelated samples can result in estimations of observables that are very far from their true value. Inspired by the state-of-the-art generative models used in machine learning, we propose a specialized neural-network architecture that supports efficient and exact sampling, completely circumventing the need for Markov-chain sampling. We demonstrate our approach for two-dimensional interacting spin models, showcasing the ability to obtain accurate results on larger system sizes than those currently accessible to neural-network quantum states.

DOI: 10.1103/PhysRevLett.124.020503

Introduction.—The theoretical understanding and modeling of interacting many-body quantum matter has represented an outstanding challenge since the early days of quantum mechanics. At the heart of several problems in condensed matter, chemistry, nuclear matter, and more lies the intrinsic difficulty of fully representing the many-body wave function, in principle needed to exactly solve Schrödinger's equation. These mainly fall into two categories: on one hand, there are states traditionally used in stochastic variational Monte Carlo (VMC) calculations [1]. Chief examples are Jastrow wave functions [2], carrying high entanglement, but also with a limited variational freedom. On the other hand, more recently, tensor-network approaches have been put forward, based on non-stochastic variational optimization, and most chiefly on entanglement-limited variational wave functions [3–6].

In an attempt to circumvent the limitations of the approaches above, architectures based on artificial neural-networks (ANN) were proposed as variational wave functions [7]. Restricted Boltzmann machines (RBM), which represent relatively veteran machine-learning constructs, were shown to be capable of representing volume-law entanglement scaling in two dimensions (2D) [8–11]. Recently, other neural-network architectures have been explored. Most notably, convolutional neural networks (ConvNets)—leading deep learning architectures that stand at the forefront of empirical successes in various artificial intelligence domains—have been applied to both bosonic [12] and frustrated spin systems [13].

Despite the provable theoretical advantage of ConvNet architectures [14], however, early numerical studies have been limited to relatively shallow architectures, far from the very deep networks used in modern machine-learning applications. This practical limitation is mostly due to two main factors. First, it is computationally expensive to obtain quantum expectation values over ConvNet states using stochastic sampling based on Markov-chain Monte Carlo (MCMC) calculations, as is it is customary in VMC applications. Second, there is an intrinsic optimization bottleneck to be faced when dealing with a large number of parameters. However, both limitations are routinely faced when learning deep autoregressive models, recently introduced machine-learning techniques that have enabled previously intractable applications.

In this paper, we propose a pivotal shift in the use of neural-network quantum states (NQS) for many-body quantum systems that markedly sets a discontinuity with traditionally adopted VMC methods. Inspired by the latest advances in generative machine-learning models, we introduce variational states for which both the sampling and the optimization issues are substantially alleviated. Our model is composed of a ConvNet that allows direct, efficient, and *i.i.d.* sampling from the highly entangled wave function it represents. The network architecture draws upon successful autoregressive models for representing and sampling from probability distributions. These are widely employed in the machine-learning literature [15], and have been recently used for statistical mechanics applications [16], as well as

density matrix reconstructions from experimental quantum systems [17]. We generalize these autoregressive models to treat complex-valued wave functions, obtaining highly expressive architectures parametrizing an automatically normalized many-body quantum wave function.

Neural autoregressive quantum states.—We consider in the following a pure quantum system, constituted by N discrete degrees of freedom $\mathbf{s} \equiv (s_1, \ldots, s_N)$ (e.g., spins, occupation numbers, etc.) such that the wave-function amplitudes $\Psi(s)$ fully specify its state. Here we follow the approach introduced in [7], and represent $\ln[\Psi(\mathbf{s})]$ as a feed-forward ANN, parametrized by a possibly large number of network connections. Given an arbitrary set of quantum numbers, s, the output value computation of the corresponding NQS, known as its forward pass, can generally be described as a sequence of K matrix-vector multiplications separated by the applications of a nonlinear elementwise activation function σ: $\mathbb{C} \to \mathbb{C}$. More formally, the un-normalized log amplitudes are given by

$$\ln[\Psi(\mathbf{s})] = W_K \sigma[W_{k-1}\sigma(\cdots\sigma(W_1\mathbf{s}))], \tag{1}$$

where $\mathcal{W} \equiv \{W_i \in \mathbb{C}^{r_i \times r_{i-1}}\}_{i=1}^{K}$, $r_0 = N, r_K = 1, r_1, \ldots, r_{K-1}$ are known as the widths of the network, and K as the depth. In practice, specialized variants of Eq. (1) are commonly used; e.g., early applications have focused on shallow architectures ($k = 1$) such as restricted Boltzmann machines, for which the activation function is typically taken to be $\sigma(z) = \ln\cosh(z)$. Other, deeper, choices are often advantageous, such as convolutional networks, in which most of the matrices are restricted to act on a subset of the quantum numbers, computing convolutions with small filters.

Given an NQS representation of a many-body quantum state, estimating physical observables $\langle\Psi|\mathcal{O}|\Psi\rangle$ of a local operator $\mathcal{O}$ is in general analytically intractable, but can be realized numerically through a stochastic procedure, as done in VMC. Specifically, $\langle\Psi|\mathcal{O}|\Psi\rangle = \langle O^{\mathrm{loc}}\rangle_{\mathcal{P}}$, where $\langle\ldots\rangle_{\mathcal{P}}$ denote statistical expectation values over the Born probability density $\mathcal{P}(\mathbf{s}) \equiv |\Psi(\mathbf{s})|^2$, and $O^{\mathrm{loc}} \equiv \sum_{\mathbf{s}'}\langle\mathbf{s}|\mathcal{O}|\mathbf{s}'\rangle\Psi(\mathbf{s}')/\Psi(\mathbf{s})$ is the corresponding statistical estimator. In the vast majority of VMC applications, including NQS so far, a MCMC algorithm is typically used to generate samples from $\mathcal{P}(\mathbf{s})$. While MCMC is a rather flexible technique, it comes with a large computational cost, especially for deep ANNs. Additionally, though MCMC asymptotically generates samples that are correctly distributed, in practice it can be plagued by very large autocorrelation times, and lack of ergodicity, that can severely affect the quality of the samples being generated.

In light of these limitations, we propose here a specialized network architecture that instead supports efficient and exact sampling. Our approach is an extension of neural autoregressive density estimators (NADE) [15] to quantum applications, resulting in what we dub neural autoregressive quantum states (NAQS). To start with, first consider the task of representing a probability distribution with NADE models. These models build on the so-called autoregressive property, which entails a decomposition of the full probability distribution as a product of conditionals, i.e., $P(s_1, \ldots, s_N) = \prod_{i=1}^{N} p_i(s_i|s_{i-1}, \ldots, s_1)$. The power of these models comes from the observation that, for every i, the conditional probabilities p_i can be individually represented as an ANN receiving as input the variables $s_1, \ldots, s_{i-1}$ and outputting a vector $\mathbf{v}_i \equiv (v_{i,s_1}, v_{i,s_2}, \ldots, v_{i,s_M})$ representing the un-normalized probabilities for s_i to take one of the M possible discrete values s_j, conditioned on given $s_1, \ldots, s_{i-1}$. It is crucial that each output vector $\mathbf{v}_i$ does not depend on the value of s_i or any of the variables appearing with a larger index, $s_{i+1}, \ldots, s_N$, for a prechosen ordering. To ensure that each network outputs a valid conditional distribution, it is then sufficient to take the exponent of each entry and normalize it according to the l_1 norm, i.e., $p_i(s_i|s_{i-1}, \ldots, s_1) = \exp(v_{i,s_i})/\sum_{s'}|\exp(v_{i,s'})|$, also known as a *Softmax* operation.

Even though it is possible to use N separate networks for each of the N conditional probabilities, and each accepting a variable number of inputs, in practice it is more common to use a single ANN that accepts N inputs and outputs N probability vectors. In this case, the autoregressive property is enforced by masking the inputs $s_i, \ldots, s_N$ for the ith output vector, i.e., ensuring that the contributions of higher-ordered spins to the output of the network vanish. PixelCNN [18] is such an architecture, and is built as a sequence of *masked* convolutional layers, whose filters are restricted to having 0's at positions "ahead." For example, in a one-dimensional system, a filter of width R, where R is odd, would be constrained to have $(w_1, \ldots, w_{(R-1)/2}, 0, \ldots, 0)$, and thus the ith output of each layer depends uniquely on the indices at $s_1, \ldots, s_{i-1}$.

A chief advantage of networks with the autoregressive property is that directly drawing samples according to $P(\mathbf{s})$ is conceptually straightforward. One can sample each s_i in sequence, according to its given conditional probability that depends just on the previously sampled $(s_1, \ldots, s_{i-1})$. Carefully exploiting the intrinsic sparseness of the network weights further leads to a very efficient algorithm for sampling [19]. Remarkably, the complexity of sampling a full string $s_1 \ldots s_N$ in a PixelCNN architecture can be reduced to the complexity of just a single forward pass.

Our NAQS model for representing wave functions is based on the same NADE principles so-far described. Specifically, just as probability functions can be factorized into a product of conditional probabilities, we represent a normalized wave function as a product of normalized conditional wave functions, such that

$$\Psi(s_1, \ldots, s_N) = \prod_{i=1}^{N} \psi_i(s_i|s_{i-1}, \ldots, s_1), \tag{2}$$

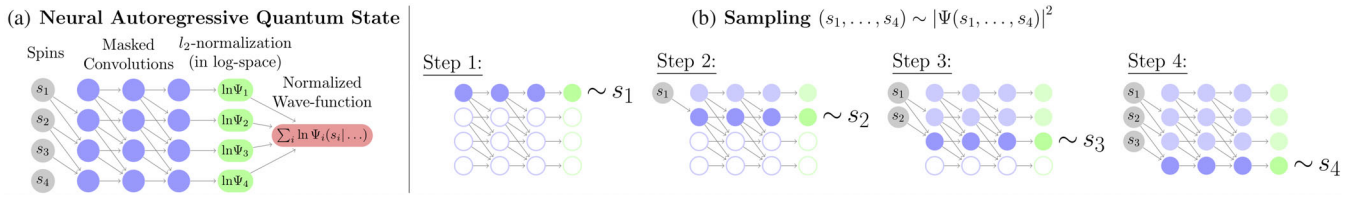

FIG. 1. Neural autoregressive quantum states are neural networks that represent a normalized wave function, $\Psi(s_1, \ldots, s_N)$, by factoring it to a sequence of normalized conditional wave functions, denoted by $\Psi_i(s_i|s_{i-1}, \ldots, s_1)$ for the ith particle, in a manner similar to that of a neural autoregressive density estimator [see Eq. (3)]. (a) Illustration of a deep one-dimensional-convolutional NAQS model following the PixelCNN [18] architecture. Each column of nodes represents a layer in the network, starting with the input layer representing the N-particle configuration $(s_1, \ldots, s_N)$. Each internal node in the graph is a complex vector computed according to its layer type. Namely, masked convolutions are limited to having local connectivity, where a node at the jth row is only connected to nodes with connections to s_i where $i < j$. All inputs to a node at the lth layer are multiplied by a matrix $W^{(l)}$, shared across all rows in the same layer, and followed by applying a nonlinear elementwise function σ: $\mathbb{C} \to \mathbb{C}$. (b) Depicts the exact sampling algorithm for NAQS, where empty nodes represent unused nodes, and filled but faded nodes represent cached results from previous steps. The quantum number of each particle is generated sequentially, by computing its respective conditional wave function, and sampling according to the squared magnitude. Notice that only a single row is processed at each step, and so sampling a complete configuration has the same runtime as a single forward pass.

where $\psi_i(s_i|s_{i-1}, \ldots, s_1)$ are such that, for any fixed $(s_1, \ldots, s_{i-1}) \in \{1, \ldots, M\}^{i-1}$, they satisfy the normalization condition $\sum_{s'} |\psi_i(s'|s_{i-1}, \ldots, s_1)|^2 = 1$. If this condition holds, then a strong normalization condition for the full wave function follows (see proof in Supplemental Material [20]).

Claim 1 Let Ψ: $[M]^N \to \mathbb{C}$ such that $\Psi(s_1, \ldots, s_N) = \prod_{i=1}^N \psi_i(s_i|s_{i-1}, \ldots, s_1)$, where $\{\psi_i\}_{i=1}^N$ are normalized conditional wave functions. Then, Ψ is normalized, i.e., $\sum_{s_1,\ldots,s_N} |\Psi(s_1, \ldots, s_N)|^2 = 1$.

As in the NADE case, we represent a conditional wave function with an ANN accepting $(s_1, \ldots, s_{i-1})$ and outputting a complex vector $\mathbf{v}_i \equiv (v_{i,s_1}, v_{i,s_2}, \ldots, v_{i,s_M}) \in \mathbb{C}^M$ for each of the M possible values taken by the local quantum numbers s_i. To obtain a normalized conditional wave function, we take its exponent and normalize it according to the l_2-norm, i.e., $\psi_i(s_i|s_{i-1}, \ldots, s_1) = \hat{v}_{i,s_i} \equiv \exp(v_{i,s_i})/\sqrt{\sum_{s'} |\exp(v_{i,s'})|^2}$. Given this parametrization, the full wave-function log amplitude $\ln \Psi(s_1 \ldots s_N)$ is easily obtained, once all the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_N$ have been computed, as given by

$$\ln \Psi(\mathbf{s}) = \sum_{i=1}^{N} \left(v_{i,s_i} - \frac{1}{2} \ln \sum_{s'} |\exp(v_{i,s'})|^2 \right). \quad (3)$$

As in the probabilistic autoregressive model, we can represent the entire NAQS by a single neural network outputting N complex vectors, as illustrated in Fig. 1(a). Though our proposed architecture can work with either complex or real parameters, we have found that using the latter work better, where we represent each complex conditional log amplitude using two real values, log magnitude and phase.

Moreover, there is a special relationship between a NAQS and its induced Born probability, since $|\Psi(s_1, \ldots, s_N)|^2 = \prod_{i=1}^N |\psi_i(s_i|s_{i-1}, \ldots, s_1)|^2$, implying that $|\psi_i(s)|^2$ is a valid conditional probability. Thus, the induced Born probability of a NAQS has the exact same structure of a NADE model. Specifically, taking the squared magnitude of its output vectors, i.e., $\forall\, i, s', \bar{v}_{i,s'} = |\hat{v}_{i,s'}|^2$, transforms NAQS into a standard NADE representation of this distribution, which importantly includes its efficient and exact sampling method. In contrast to standard MCMC sampling employed for correlated wave functions, NAQS thus allows for direct, efficient sampling with the computational complexity of a single forward pass, as depicted in Fig. 1(b).

Optimization.—The NAQS representation of many-body wave functions can be used in practice for several applications. These include, for example, ground-state search [7], quantum-state tomography [21], dynamics [7], and quantum circuits simulation [22]. Here we more specifically focus on the task of finding the ground state of a given Hamiltonian $\mathcal{H}$. In this context, we denote by $\Psi_{\mathcal{W}}$ the wave function represented by a NAQS of a fixed architecture that is parametrized by $\mathcal{W}$, and we wish to find $\mathcal{W}$ values that minimize the energy, i.e., $\mathcal{W}^* = \mathrm{argmin}_{\mathcal{W}} E(\mathcal{W})$, where $E(\mathcal{W}) \equiv \langle \Psi_{\mathcal{W}} | \mathcal{H} | \Psi_{\mathcal{W}} \rangle = \mathbb{E}_{\mathbf{s}\sim|\Psi_{\mathcal{W}}|^2}[E_{\mathrm{loc}}(\mathbf{s}; \mathcal{W})]$, $E_{\mathrm{loc}}(\mathbf{s}; \mathcal{W}) \equiv \sum_{\mathbf{s}'} \mathcal{H}_{\mathbf{s},\mathbf{s}'} [\Psi_{\mathcal{W}}(\mathbf{s}')/\Psi_{\mathcal{W}}(\mathbf{s})]$, and $\mathcal{H}$ is usually a highly sparse matrix, and so computing E_{loc} for a given sample takes at most $O(N)$ forward passes.

The common approach for solving the optimization problem above with a NQS is to estimate the gradient of $E(\mathcal{W})$ with respect to $\mathcal{W}$, and use variants of stochastic gradient descent (SGD) to find the minimizer of $E(\mathcal{W})$. Estimating the gradient can be done by first employing a variant of the log-derivative trick, i.e.,

$$\frac{\partial E}{\partial \mathcal{W}} = \mathbb{E}_{\mathbf{s}\sim|\Psi_{\mathcal{W}}|^2} \left[2\mathrm{Re} \left([E_{\mathrm{loc}}(\mathbf{s})^* - E^*] \frac{\partial \ln \Psi_{\mathcal{W}}}{\partial \mathcal{W}} \right) \right]. \quad (4)$$

Now, while we can efficiently compute the log derivative of $\Psi_{\mathcal{W}}$, exactly computing the expected value is intractable, but we can still approximate it by computing its value over a finite batch of samples $\{\mathbf{s}_{(i)}\}_{i=1}^B$. The quality of this

TABLE I. Estimates of the ground-state energies of the transverse-field Ising model for different values of Γ on a 12×12 lattice, and the corresponding estimates of $\langle \sigma_z \rangle$, as obtained by either NAQS or QMC.

Γ	NAQS Energy	QMC Energy	NAQS $\langle \lvert\sigma_z\rvert \rangle$	QMC $\langle \lvert\sigma_z\rvert \rangle$
2.0 J	−2.4096022(2)	−2.40960(3)	0.78326(2)	0.78277(38)
2.5 J	−2.7476550(5)	−2.74760(3)	0.57572(3)	0.57566(63)
3.0 J	−3.1739005(5)	−3.17388(4)	0.16179(4)	0.16207(54)
3.5 J	−3.6424799(3)	−3.64243(4)	0.11094(3)	0.11011(30)
4.0 J	−4.1217979(2)	−4.12178(4)	0.09725(2)	0.09728(24)

approximation depends on the batch size, B, but also on the degree of correlations between the individual samples. The advantages of our direct sampling method supported by NAQS over MCMC are twofold in this context: (i) Faster sampling: each individual sample can be generated with fewer network passes, and generating a batch of samples is embarrassingly parallel, as opposed to the sequential nature of MCMC; (ii) faster convergence: the generated samples are exact and *i.i.d.*, and so result in more accurate estimates of the gradient at each step.

Experiments.—As a first benchmark for our approach, we consider a case where MCMC sampling can be strongly biased. A paradigmatic quantum system exhibiting this issue is found in the ferromagnetic phase of the transverse field Ising model. The Hamiltonian for this model is given by $H = -J \sum_{\langle i,j \rangle} \sigma_z^i \sigma_z^j - \Gamma \sum_i \sigma_x^i$, where the summation runs over pairs of lattice edges. Here we study the case of a 2D square lattice with open boundary conditions, and for varying strengths of the transverse field. The system is in a ferromagnetic phase when the transverse magnetic field Γ is weak with respect to the coupling constant, and specifically in 2D when $\Gamma < \Gamma_c \simeq 3.044J$ [23].

In order to verify the correctness of the model proposed in Sec. II, we begin by comparing the ground-state energy and system magnetization obtained for a 12×12 system with those obtained by an unbiased quantum Monte Carlo (QMC) simulation. Using our FlowKet open-source library [24], we employ a NAQS model following the PixelCNN architecture, using the ADAM [25] SGD variant with the gradient estimator of Eq. (4) (see further technical details in Supplemental Material [20]). Table I shows that our model achieves very high accuracy for both magnetization and energy densities for different transverse field values across the phase diagram: when the system is in the ferromagnetic phase, the normal phase, and near the phase transition.

In order to quantify the behavior of our model in a region of broken symmetry, we consider the case of a transverse field deep in the ferromagnetic region, namely, $\Gamma = 2J$. The principal component analysis visualization in Fig. 2 shows that for this value of Γ the MCMC chains initialized at one of the oriented states composing the ground state are stuck at that specific orientation and cannot come around to sampling spin configurations that correspond to the opposite

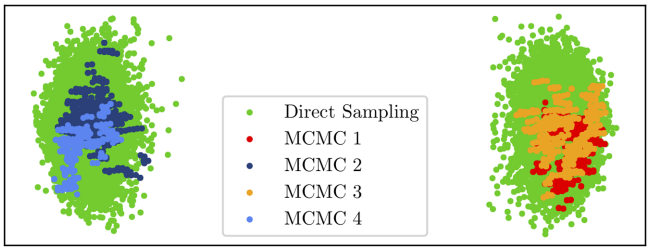

FIG. 2. An illustration of the two modes of the ground state, by taking the first two principal components of the generated samples. The green points correspond to our direct sampling method, and the other colors represent different MCMC chains. The plot was generated by training a NAQS on the transverse-field Ising model with $\Gamma = 2J$, below the critical value, on a 12×12 lattice until convergence to the ground state, and then sampling from the trained NAQS using either our direct sampling method, or four separate MCMC samplers.

orientation. In contrast, spin configurations sampled directly from the distribution by using our proposed technique include equally probable configurations from both orientations. The ergodicity breaking in local MCMC is also directly quantifiable by the expectation value of the total magnetization $m \equiv \langle \sum_i \sigma_i^z \rangle$, for which we expect $m = 0$ on any finite lattice. Indeed, the *i.i.d.* sampling enabled by our model correctly explores the two relevant ferromagnetic states (in agreement with the visualization of Fig. 2) and reaches a value close to a total zero magnetization, in stark contrast with MCMC estimation that effectively computes $\langle \lvert\sigma_z\rvert \rangle \approx 0.78$ rather than m. As expected, directly estimating $\langle \lvert\sigma_z\rvert \rangle$ with our sampling method correctly recovers it to a high precision; see Table I.

The limitation of the MCMC procedure in providing independent samples is not only conceptually relevant, but it can also have consequences on the quality of the resulting ground-state approximations. In Fig. 3, we show

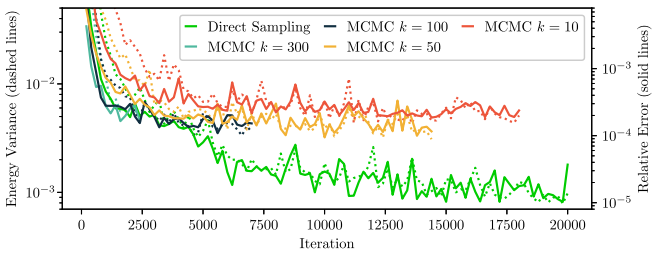

FIG. 3. Comparing the effects of the sampling method, either MCMC or direct sampling, on the training procedure for the transverse-field Ising model with $\Gamma = 3J$, close to the critical value, on a large (21×21) lattice. When using MCMC, samples are taken every $k \in \{10, 50, 100, 300\}$ steps in the chain, where increasing k reduces the correlation between samples at the expense of increased computational cost. The solid lines shows the relative error to the minimal energy found for this system in our experiments, and dashed lines shows the energy variance. Since MCMC takes a considerable time to complete just a single iteration, we have restricted the training to a maximum of 100 hours.

TABLE II. Ground-state energies for the antiferromagnetic Heisenberg model with open boundary conditions, as obtained by a state-of-the-art PEPS model [26], our NAQS model, and the exact QMC estimation, as reported in Liu *et al.* [26].

Lattice	PEPS	NAQS	QMC
10×10	$-0.628601(2)$	$-0.628627(1)$	$-0.628656(2)$
16×16	$-0.643391(3)$	$-0.643448(1)$	$-0.643531(2)$

the training procedure for the transverse-field $\Gamma = 3J$, close to the critical value on a larger system (21×21). The same NAQS architecture was trained once with the *i.i.d.* sampling procedure and once with MCMC chains of varying lengths. The optimization advantage obtained when relying on independent samples clearly emerges from those figures—this procedure is much quicker and results in a significantly more accurate ground-state energy and lower energy variance $\langle H^2 \rangle - \langle H \rangle^2$.

As a further benchmark, we also apply our method to a more complex system, the two-dimensional antiferromagnetic Heisenberg model with open boundary conditions, whose Hamiltonian is given by $H = \sum_{\langle i,j \rangle} \sigma_x^i \sigma_x^j + \sigma_y^i \sigma_y^j + \sigma_z^i \sigma_z^j$. We evaluate our approach by comparing the ground-state energy obtained for 10×10 and 16×16 systems with those obtained by QMC simulations, as well as other variational methods. We find that NAQS meaningfully improve upon the accuracy of the best-known variational methods for this problem. Namely, for 10×10, a relative error of $8.7 \times 10^{-5} \pm 0.6 \times 10^{-5}$ was reported in Liu *et al.* [26] using a PEPS model, whereas with our approach we were able to obtain $3.5 \times 10^{-5} \pm 0.4 \times 10^{-5}$. See Table II for exact results. While the PEPS results can be, in principle, further improved, increasing the accuracy comes with a very significant computational requirements [27] due to the unfavorable computational scaling with respect to the bond dimension. Moreover, though not directly comparable, it is noteworthy that the relative error of the ground-state energy with periodic boundary conditions obtained by NQS with MCMC sampling is significantly less accurate than ours (Carleo and Troyer [28] and Choo *et al.* [29] report relative error greater than 2×10^{-4}).

Discussion.—In this work, we have shown a scheme to facilitate the practical employment of contemporary deep learning architectures to the modeling of many-body quantum systems. This constitutes a striking improvement over currently used RBM methods that are limited to only hundreds of parameters, and very shallow networks. A further practical advantage we gain is the ability to make use of the substantial body of knowledge regarding optimization of these architectures that is accumulating in the deep learning literature. We empirically demonstrate that by employing common deep learning optimization methods such as SGD, our direct sampling approach allows us to train very large convolutional networks (20 layers, 21×21 lattice, $\sim 1M$ parameters). Our presented experiments demonstrate that even for relatively simple systems MCMC sampling can fail, while the *i.i.d.* sampling enabled by our model succeeds. Relying on the theoretically promising results regarding convolutional network capabilities in representing highly entangled systems [14], namely, systems satisfying volume law, we view the enabling of their optimization as an integral step in reaching currently unattainable insight on a vast variety of quantum many-body phenomena.

This work is supported by ISF Center Grant No. 1790/12 and by the European Research Council (TheoryDL project). Y. L. is supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities. QMC simulations for the 2D transverse-field Ising model have been performed using the open-source ALPS Library [30].

*or.sharir@cs.huji.ac.il
†yoavlevine@cs.huji.ac.il
‡noam.wies@cs.huji.ac.il
§gcarleo@flatironinstitute.org
‖shashua@cs.huji.ac.il

[1] W. L. McMillan, Ground state of liquid He_4, Phys. Rev. **138**, A442 (1965).
[2] R. Jastrow, Many-body problem with strong forces, Phys. Rev. **98**, 1479 (1955).
[3] M. Fannes, B. Nachtergaele, and R. F. Werner, Finitely correlated states on quantum spin chains, Commun. Math. Phys. **144**, 443 (1992).
[4] D. Perez-García, F. Verstraete, M. M. Wolf, and J. I. Cirac, Matrix product state representations, Quantum Inf. Comput. **7**, 401 (2007).
[5] F. Verstraete and J. I. Cirac, Renormalization algorithms for quantum-many body systems in two and higher dimensions, arXiv:cond-mat/0407066.
[6] G. Vidal, Class of Quantum Many-Body States that can be Efficiently Simulated, Phys. Rev. Lett. **101**, 110501 (2008).
[7] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, Science **355**, 602 (2017).
[8] D.-L. Deng, X. Li, and S. Das Sarma, Quantum Entanglement in Neural-Network States, Phys. Rev. X **7**, 021021 (2017).
[9] J. Chen, S. Cheng, H. Xie, L. Wang, and T. Xiang, Equivalence of restricted Boltzmann machines and tensor network states, Phys. Rev. B **97**, 085104 (2018).
[10] I. Glasser, N. Pancotti, M. August, I. D. Rodriguez, and J. I. Cirac, Neural-Network Quantum States, String-Bond States, and Chiral Topological States, Phys. Rev. X **8**, 011006 (2018).
[11] R. Kaubruegger, L. Pastori, and J. C. Budich, Chiral topological phases from artificial neural networks, Phys. Rev. B **97**, 195136 (2018).

[12] H. Saito and M. Kato, Machine learning technique to find quantum many-body ground states of bosons on a lattice, J. Phys. Soc. Jpn. **87**, 014001 (2018).
[13] K. Choo, G. Carleo, N. Regnault, and T. Neupert, Symmetries and Many-Body Excitations with Neural-Network Quantum States, Phys. Rev. Lett. **121**, 167204 (2018).
[14] Y. Levine, O. Sharir, N. Cohen, and A. Shashua, Quantum Entanglement in Deep Learning Architectures, Phys. Rev. Lett. **122**, 065301 (2019).
[15] B. Uria, M.-A. Cote, K. Gregor, I. Murray, and H. Larochelle, Neural autoregressive distribution estimation, J. Mach. Learn. Res. **17**, 1 (2016).
[16] D. Wu, L. Wang, and P. Zhang, Solving Statistical Mechanics Using Variational Autoregressive Networks, Phys. Rev. Lett. **122**, 080602 (2019).
[17] J. Carrasquilla, G. Torlai, R. G. Melko, and L. Aolita, Reconstructing quantum states with generative models, Nat. Mach. Intell. **1**, 155 (2019).
[18] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves *et al.*, Conditional image generation with pixelcnn decoders, in Advances in Neural Information Processing Systems (NeurIPS, 2016), pp. 4790–4798.
[19] P. Ramachandran, T. L. Paine, P. Khorrami, M. Babaeizadeh, S. Chang, Y. Zhang, M. A. Hasegawa-Johnson, R. H. Campbell, and T. S. Huang, Fast generation for convolutional autoregressive models, in *Proceedings of ICLR 2017 Workshop Track*, https://openreview.net/forum?id=rkdF0ZNKl.
[20] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.124.020503 for proofs and technical details of experiments.
[21] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, Neural-network quantum state tomography, Nat. Phys. **14**, 447 (2018).
[22] B. Jnsson, B. Bauer, and G. Carleo, Neural-network states for the classical simulation of quantum computing, arXiv:1808.05232.
[23] H. W. J. Blte and Y. Deng, Cluster Monte Carlo simulation of the transverse Ising model, Phys. Rev. E **66**, 066110 (2002).
[24] O. Sharir, Y. Levine, N. Wies, G. Carleo, A. Shashua, Deep autoregressive models for the efficient variational simulation of many-body quantum systems, arXiv:1902.04057
[25] D. Kingma and J. Ba, Adam: A method for stochastic optimization, in *3rd International Conference on Learning Representations (ICLR)* (ICLR, 2014).
[26] W.-Y. Liu, S.-J. Dong, Y.-J. Han, G.-C. Guo, and L. He, Gradient optimization of finite projected entangled pair states, Phys. Rev. B **95**, 195154 (2017).
[27] L. He, H. An, C. Yang, F. Wang, J. Chen, C. Wang, W. Liang, S. Dong, Q. Sun, W. Han, W. Liu, Y. Han, and W. Yao, Peps++: Towards extreme-scale simulations of strongly correlated quantum many-particle models on sunway taihulight, IEEE Trans. Parallel Distributed Syst. **29**, 2838 (2018).
[28] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, Science **355**, 602 (2017).
[29] K. Choo, T. Neupert, and G. Carleo, Study of the two-dimensional frustrated j1-j2 model with neural-network quantum states, Phys. Rev. B **100**, 125124 (2019).
[30] B. Bauer *et al.*, The ALPS project release 2.0: Open source software for strongly correlated systems, J. Stat. Mech. (2011) P05001.