Geometry of learning neural quantum states

Chae-Yeun Park and Michael J. Kastoryano
Institute for Theoretical Physics, University of Cologne, Cologne 50937, Germany

(Received 11 November 2019; accepted 30 March 2020; published 26 May 2020)

Combining insights from machine learning and quantum Monte Carlo, the stochastic reconfiguration method with neural network *Ansatz* states is a promising new direction for high-precision ground-state estimation of quantum many-body problems. Even though this method works well in practice, little is known about the learning dynamics. In this paper, we bring to light several hidden details of the algorithm by analyzing the learning landscape. In particular, the spectrum of the quantum Fisher matrix of complex restricted Boltzmann machine states exhibits a universal initial dynamics, but the converged spectrum can dramatically change across a phase transition. In contrast to the spectral properties of the quantum Fisher matrix, the actual weights of the network at convergence do not reveal much information about the system or the dynamics. Furthermore, we identify a measure of correlation in the state by analyzing entanglement in eigenvectors. We show that, generically, the learning landscape modes with least entanglement have largest eigenvalue, suggesting that correlations are encoded in large flat valleys of the learning landscape, favoring stable representations of the ground state.

DOI: 10.1103/PhysRevResearch.2.023232

I. INTRODUCTION

Recently the fields of machine learning and quantum information science have seen a lot of crossbreeding. On the one hand, a number of promising results have been obtained suggesting the potential for performing quantum or classical machine learning tasks on a quantum computer [1]. In particular, the variational quantum eigensolver [2]—perhaps the most promising quantum algorithms for first-generation quantum computers—is based on the variational optimization of a cost function to be evaluated on a quantum device, providing a new playground for hybrid quantum-classical learning [3,4]. However, arguably the most significant advances have been in the field of classical variational algorithms for quantum many-body systems. A number of studies have shown that machine-learning-inspired sampling algorithms can reach state-of-the-art precision, including ground-state energy estimation [5–8], time evolution [5,9], identifying phase transitions [10–12], and decoding quantum error correcting codes [13,14] (for a recent review, see Ref. [15]).

A model that has gathered a particularly large amount of attention is the complex restricted Boltzmann machine (RBM) state *Ansatz* with stochastic reconfiguration optimization introduced by Carleo and Troyer [5]. The authors show that ground-state energy evaluations can outperform the state-of-the-art tensor network methods on benchmark problems.

At present, however, there is lacking a theoretical underpinning for explaining why the complex RBM—or any other machine-learning-inspired parametrization—is a good *Ansatz* for describing ground states of physical Hamiltonians or for accessing its features. This is sometimes referred to as the "black box" problem of machine-learning-inspired approaches, that theoretical understanding lags far behind the numerical state of the art. In particular, it is difficult to assess and quantify the role of entanglement in these new classes of wave functions. This is to be contrasted with the density matrix renormalization group (DMRG) [16], which was first developed as an extension of the numerical renormalization group. Subsequently, it was realized that the theoretical underpinning of DMRG was the theory of tensor network states, which connect the efficiency of simulation in one-dimensional systems with the amount and nature of entanglement in the spin chain. We are far from such a detailed understanding of machine-learning-inspired methods.

Thus it is natural that some studies have related complex RBM states to tensor network states [17,18]. But these studies are mostly based on constructing abstract mappings between RBM wave functions and tensor network states, and usually provide at best existence proofs.

In this paper, we aim to obtain a better understanding of the learning dynamics with complex RBM wave functions by analyzing the geometry induced in parameter space. Indeed, the stochastic reconfiguration method updates the variational parameters of the wave function with gradient descent of the energy, weighted by a "quantum Fisher matrix," which is the quantum analog of the Fisher information matrix. The Fisher information matrix is known to be the unique Riemannian metric associated to a probability space invariant under sufficient statistics [19]. Hence it is the natural candidate for associating an "information geometry" to a statistical model.

We analyze the spectral properties of the "quantum Fisher matrix" for various lattice spin models. We argue that the information geometry provides us with clues for both the expressibility of the *Ansatz* state and the underlying physics, provided the optimization converges. In particular, we identify a number of features which we believe to be universal for spin models:

(i) The spectrum of the quantum Fisher matrix becomes singular in phases connected to a product state (in the computational basis). The singularity is more pronounced the closer one gets to the product state.

(ii) Critical phases have a smooth and extended spectrum, which is also reminiscent of image recognition models in classical machine learning.

(iii) Kinks in the spectrum reveal symmetries in the state.

(iv) The eigenvalues are exponentially decaying in value. The largest eigenvalues have eigenvectors that are dominated by first moments; i.e., they do not contain much information about correlations in the system. This feature is accentuated the sharper the spectrum profile of the quantum Fisher matrix.

The above insight was extracted from extensive numerical data calculated using quantum spin Hamiltonians such as transverse field Ising and Heisenberg spin-XXZ models as well as coherent Gibbs states for the two-dimensional classical Ising model. Various Monte Carlo sampling strategies were used to optimize the results on large system sizes.

Importantly, we observe that the bare values of the variational parameters reveal very little information about the physical properties of the system, contrary to what is often claimed that "activations indicate regions of activity in the underlying data." We take this as evidence that there are many equivalent representations of the states in the vicinity of the ground state, suggesting that the optimizer preferentially chooses robust representation of the ground state. Robustness of the Monte Carlo methods might be related to the generalization property in supervised learning. Our study shows that the spectrum of the quantum Fisher matrix can be an essential diagnostic tool for further exploration with complex RBM wave functions as well as with other machine-learning-inspired wave functions.

A. Complex RBM and optimization by stochastic reconfiguration

The complex restricted Boltzmann machine (RBM) neural network quantum state specifies the amplitudes of a wave function $|\psi_\theta\rangle = \sum_x \psi_\theta(x)|x\rangle$ in some chosen computational basis $\{|x\rangle\}$ by the exponential family

$$\psi_\theta(x) = \sum_y e^{a\cdot x + b\cdot y + x^T wy}/\sqrt{Z}, \tag{1}$$

where the vectors $\{a, b\}$ and the matrix w contain complex parameters to be varied in the optimization, and y is a binary vector indexing "hidden" units. $Z = \sum_x |\psi_\theta(x)|^2$ is a constant guaranteeing normalization of the state ψ. The complex RBM can be visualized as a binary graph (V, E) between the visible nodes x and the hidden nodes y (see Fig. 1). To each edge $e \in E$ we associate a variational parameter w_e, and at each vertex $v \in V$ we associate a bias weight a or b to a visible (x) or hidden (y) binary degree of freedom. We will often express the variational parameters as a concatenated vector labeled $\theta = (a, b, \text{vec}(w))$. For classical RBMs, the normalization constant is the partition function of a joint probability distribution on the hidden and visible units. This is generally not true in the complex case.

The goal of variational Monte Carlo is to find the optimal parameters θ that minimize the energy of a given Hamiltonian in the state $|\psi_\theta\rangle$. The standard approach would be to use gradient descent, but this performs very poorly for spin Hamiltonians, as the updates tend to get stuck oscillating back and forth along steep wells of the energy landscape rather than falling down the more shallow directions. The stochastic reconfiguration (SR) method [20,21] for energy minimization is derived as a second-order iterative approximation to the imaginary time ground-state projection method (see Appendix A for a self-contained derivation). In SR the parameters of the *Ansatz* wave function are iteratively updated as

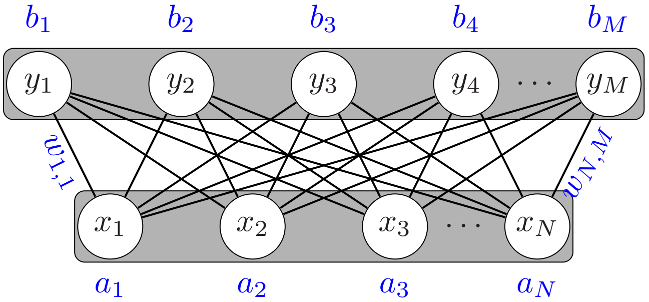

FIG. 1. Complex RBM consisting of one hidden and one visible layer. Visible, hidden biases, and weights are $a \in \mathbb{C}^N$, $b \in \mathbb{C}^M$, and $w \in \mathbb{C}^N \times \mathbb{C}^M$, respectively. x, y are binary vectors of length n and m, respectively.

$$\theta \to \theta - \eta S^{-1}\nabla_\theta\langle H\rangle, \tag{2}$$

where η is a constant specifying the rate of learning. The second-order effects which take curvature into account are determined by the matrix

$$S_{\alpha\beta} = \langle O_\alpha^\dagger O_\beta\rangle - \langle O_\alpha^\dagger\rangle\langle O_\beta\rangle \tag{3}$$

of the diagonal operators O_α, with $\alpha \in \theta$, which act, for instance, as

$$O_{w_{ij}}|x\rangle = \frac{\partial \log \psi_\theta(x)}{\partial w_{ij}}|x\rangle \tag{4}$$

in the computational basis $\{x\}$. We will call the matrix S the *quantum Fisher matrix*, because of its connection with information geometry as discussed in detail in the next section. The quantum Fisher matrix can be reformulated as a classical covariance matrix of the operators O_α, O_β,

$$S_{\alpha\beta} = \mathbb{E}[O_\alpha^\dagger O_\beta] - \mathbb{E}[O_\alpha^\dagger]\mathbb{E}[O_\beta], \tag{5}$$

and similarly

$$\partial_\alpha\langle H\rangle = \mathbb{E}[O_\alpha H_{\text{loc}}] - \mathbb{E}[O_\alpha]\mathbb{E}[H_{\text{loc}}], \tag{6}$$

where $\mathbb{E}[A] = \sum_x A(x)|\psi_\theta(x)|^2$ is the classical expectation of operator A in the state $|\psi_\theta(x)|^2$, and

$$H_{\text{loc}}(x) = \frac{\langle x|H|\psi_\theta\rangle}{\langle x|\psi_\theta\rangle} \tag{7}$$

is called the local energy.

For the RBM *Ansatz*, the diagonal operators O_α take on the following simple form:

$$O_{a_i}(x) = x_i. \tag{8}$$

$$O_{b_j}(x) = \tanh \chi_j(x), \tag{9}$$

$$O_{w_{ij}}(x) = x_i \tanh \chi_j(x), \tag{10}$$

where $\chi_j(x) = b_j + \sum_i w_{ij}x_i$, and indices i run over $[1,\dots,N]$ visible vertices and j run over $[1,\dots,M]$ hidden vertices. Thus the size of the quantum Fisher matrix is $N + M + NM$.

The SR method is computationally efficient when the following are true:

(1) The operators $O_\alpha(x)$ and $H_{\text{loc}}(x)$ can be computed efficiently for every point x.

(2) The probability distribution $|\psi_\theta(x)|^2$ can be sampled from for any values of θ, meaning that any single Monte Carlo update can be computed efficiently. In practice we require that each Monte Carlo update is independent of system size; i.e., updates are local.

(3) The sampling procedure converges rapidly (in subpolynomial time) to the desired state $|\psi_\theta(x)|^2$.

The complex RBM *Ansatz* guarantees that (1) and (2) hold whenever the number of hidden units is a constant multiple of the visible units. However, like essentially any sampling algorithm, provably guaranteeing (3) seems nearly impossible in any practically relevant problem. However, experience has shown that convergence often is rapid in practice, or can be curtailed, whenever one steers clear of frustration or the Fermionic sign problem. It is worth pointing out, though, that convergence of the sampler can depend sensitively on the chosen basis and the initial state, as evidenced in Sec. III C.

B. Natural gradient and SR

The SR method [20,21] can be interpreted geometrically [22], which makes a direct connection to Amari's natural gradient optimization [23]. Plain vanilla gradient descent optimizes a multivariate function $L(\theta)$ by updating the parameters in the direction of steepest descent:

$$\theta \to \theta - \eta \nabla_\theta L(\theta) \tag{11}$$

at a certain rate η.

In systems where the landscape of the function $L(\theta)$ is very steep in certain directions and shallow in others, convergence can be very slow as the updates fluctuate back and forth in a deep valley but take a long time to "drift" down a shallow one. The natural gradient method proposes to update the parameters according to the natural (Riemannian) geometric structure of the information space, so that the landscape is made locally Euclidean before the update. Suppose the coordinate space is a curved manifold in the sense that the infinitesimal square length is given by the quadratic form

$$ds^2 = \sum_{\alpha\beta} g_{\alpha\beta}(\theta)d\theta_\alpha d\theta_\beta, \tag{12}$$

where the matrix $g(\theta)$ is the Riemannian metric tensor. Amari showed that the steepest descent direction of the function $L(\theta)$ in the Riemannian space is given by

$$-\tilde{\nabla}(\theta) = -g^{-1}(\theta)\nabla L(\theta). \tag{13}$$

The action of the inverse of g can be heuristically understood as "flattening" out the space locally. For general optimization problems, the Hessian is a natural choice for $g(\theta)$, as it reproduces Newton's second-order method. In machine learning applications, and with RBMs in particular, the Hessian is hard to construct from sampling. It also appears to be attracted to saddle points [24].

When the parameter space in question is naturally associated with a classical probability distribution, the "natural" geometry is chosen to be the Fisher information matrix as it is the unique metric that is invariant under sufficient statistics [19]. For pure parametrized quantum states, the natural Riemannian metric is derived from the Fubini-Study distance:

$$\gamma(\psi,\varphi) = \arccos\sqrt{\frac{\langle\psi|\varphi\rangle\langle\varphi|\psi\rangle}{\langle\psi|\psi\rangle\langle\varphi|\varphi\rangle}}. \tag{14}$$

Infinitesimal distances are given by

$$ds^2 = \gamma(\psi,\psi+\delta\psi)^2 = \frac{\langle\delta\psi|\delta\psi\rangle}{\langle\psi|\psi\rangle} - \frac{\langle\delta\psi|\psi\rangle}{\langle\psi|\psi\rangle}\frac{\langle\psi|\delta\psi\rangle}{\langle\psi|\psi\rangle}, \tag{15}$$

which reproduces the quantum Fisher matrix for parametrization θ as $ds^2 = \sum_{\alpha\beta} S_{\alpha\beta}d\theta_\alpha^* d\theta_\beta$.

In particular, when the wave function is positive in a given computational basis, the quantum state can be written as $|\psi\rangle = \sum_x \sqrt{p_\theta(x)}|x\rangle$, and the quantum Fisher matrix is

$$\begin{aligned} S_{\alpha\beta} &= \frac{1}{4}\left\langle \frac{\partial \log p_\theta(x)}{\partial\theta_\alpha}\frac{\partial \log p_\theta(x)}{\partial\theta_\beta}\right\rangle \\ &\quad - \frac{1}{4}\left\langle\frac{\partial \log p_\theta(x)}{\partial\theta_\alpha}\right\rangle\left\langle\frac{\partial \log p_\theta(x)}{\partial\theta_\beta}\right\rangle \qquad (16) \\ &= \frac{1}{4}\mathcal{F}_{\alpha\beta}, \qquad (17) \end{aligned}$$

where $\langle A\rangle = \mathbb{E}[A]$ and $\mathcal{F}$ is the Fisher information matrix associated to the probability distribution $p_\theta(x)$. Thus, the SR method reproduces the natural gradient method for positive wave functions. For this reason, we will call the S matrix associated to a pure quantum state the *quantum Fisher matrix*.

C. Spectral analysis of the quantum Fisher matrix

In this paper, we will argue that spectral properties of the quantum Fisher matrix reveal essential information about the physical properties of the system under study as well as the dynamics of optimization.

The quantum Fisher matrix is positive semidefinite, implying that its spectrum is real and there exists a set of orthonormal eigenvectors. The magnitude of an eigenvalue determines how steep the learning landscape is in that particular direction. The spectrum will generically be sloppy [25,26], with a spectral function bounded above by a decaying exponential.

It is often argued in the machine learning community that gradient descent algorithms favor regions in parameters space where most eigenvalues are close to zero [27,28]. This implies that at convergence, most directions in the landscape are nearly flat, suggesting that nearby points in parameter space encode much of the same physical properties. In classical supervised learning, the flatness of the landscape has been associated with the "generalization" ability of the learned model [29]; in the physics setting we interpret it to mean that the representation is robust.

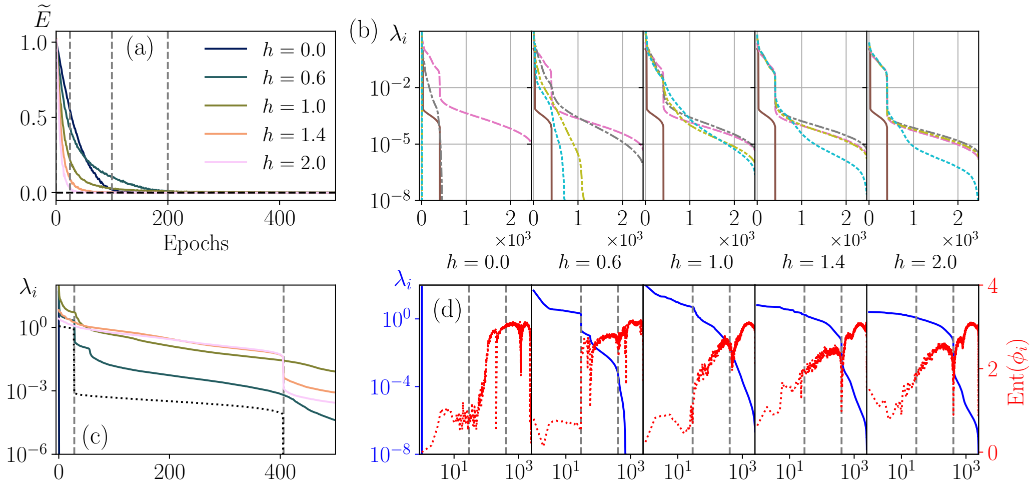

FIG. 2. Transverse field Ising model, variational ground-state energy optimization using the SR: (a) rescaled energy as a function of epochs for different values of $h \in [0.0, 0.6, 1.0, 1.4, 2.0]$ (from darkest to lightest). The energy is rescaled to have 0 at the exact ground-state energy and 1 at initialization. (b) Ordered eigenvalues of the quantum Fisher matrix [Eq. (3)] at epochs 0 (solid), 25 (dashed), 100 (dot-dashed), 200 (dot-dot-dashed), and 2000 (dotted). Results from $h = 0.0$ (the leftmost) to $h = 2.0$ (the rightmost) are shown in each subplot. The spectrum exhibits universal behavior for the first ~25 epochs. After that, the eigenvalues slowly approach a model-dependent final profile (see main text). (c) The 500 largest eigenvalues after convergence for different values of h as well as for randomly initialized RBM (black dotted curve). Color coding is the same as in (a). The two vertical gray dashed lines indicate $N = 28$ and $N(N+1)/2 = 406$. (d) Spectrum (blue solid) and entanglement in the eigenvectors (red dotted) on log-log scale. The eigenvectors corresponding to the dominant eigenvalues have significantly reduced entanglement, especially in the ferromagnetic phase. Hyperparameters $\eta = 0.01$ and $\epsilon = 0.001$ are used.

Because of the bipartite graph structure of the RBM *Ansatz*, it is natural to talk about correlations between the visible and hidden units. The quantum Fisher matrix is a square $(N + M + NM)$ matrix, with the first two blocks corresponding to the biases a, b, and the third block corresponding to the weights matrix w. The main w block describes the orientations in parameter space that can affect correlations in the model. We will see later that eigenvectors associated to eigenvalues of large magnitude are typically close to a product state between the visible and hidden part, meaning that they mostly just affect the first moments of the spin variables.

To measure correlations in the eigenvectors $\{\psi_\alpha\}$, we truncate the first two blocks of the eigenvectors associated with the biases and renormalize the "w" part to have Hilbert Schmidt norm 1. We then calculate the entanglement in the eigenstate ψ_α^w:

$$\mathrm{Ent}(\psi_\alpha) = S\big(\mathrm{Tr}_h\big[\psi_\alpha^w\big]\big), \tag{18}$$

where Tr_h is the partial trace over the hidden layer, and $S(\cdot)$ is the von Neumann entropy of the reduced density matrix.

II. RESULTS

In this section, we analyze the spectral properties of the quantum Fisher matrix during the learning process of finding the ground state of the transverse field Ising (TFI) model. The TFI Hamiltonian is given by

$$H = -\sum_{i=1}^{N} \sigma_z^i \sigma_z^{i+1} - h \sum_{i=1}^{N} \sigma_x^i, \tag{19}$$

where $\boldsymbol{\sigma}^i = \{\sigma_x^i, \sigma_y^i, \sigma_z^i\}$ are Pauli spin operators, and h is the external field. The system has $\mathbb{Z}_2$ symmetry $(\sigma_z^i \to -\sigma_z^i)$, which is explicitly broken for $h < 1$ in the thermodynamic limit $(N \to \infty)$. A second-order phase transition occurs at $h = 1$. At zero external field the model has two degenerate ground states $|0\rangle^{\otimes N}$ and $|1\rangle^{\otimes N}$, whereas in the limit of $h \to \infty$ the ground state is unique, given by $|+\rangle^{\otimes N}$.

We trained the RBM for this model with $N = 28$ and $\alpha = M/N = 3$. The spectral properties of the quantum Fisher matrix as well as the energy during the learning process obtained from the simulation are plotted in Fig. 2 (details of the simulation are described in Appendix B). Figure 2(a) confirms that the optimization procedure successfully finds the ground state for all values of h, albeit at different speeds. The quantum Fisher matrix is constructed approximately by Monte Carlo sampling and its full spectrum is evaluated every five epochs during learning. The eigenvalues at some representative epochs are plotted in decreasing order in Fig. 2(b).

The dynamics of the learning process proceeds in two distinct stages. The first stage is observed at the very beginning of the learning, lasting for roughly 25 epochs [30], and is the same for all values of h. The initial shape of the spectrum has two sharp drops located at N and $N(N+1)/2$ [see Fig. 2(c)]. This is a consequence of the random initialization with small weights. An analytic justification of this behavior is provided in Appendix C. The spectrum then gets pushed up until approximately the 25th epoch, revealing that more and more dimensions in the information space become relevant.

The second stage of learning then slowly transforms the distribution to that of the final converged state. We observe that the spectrum falls off very sharply (exponentially) in all cases examined [Fig. 2(b)], but the exact spectral profile depends strongly on the details of the model, yet not on the system size or on the specific values of the learned weights (see Appendix C for an in-depth discussion). We take this as evidence that the learned state not only minimizes the energy, but also closely matches the actual ground state of the model (that we also checked using the spin-spin correlation functions). The behavior of the spectrum of the quantum Fisher matrix for each phase of the TFI model is discussed in the next subsection.

A. Phases of the TFI model

1. The ferromagnetic phase ($h < 1.0$)

Let us start by considering the extreme case with $h = 0.0$. The quantum Fisher matrix after convergence becomes a pure state up to numerical precision. The singularity of the quantum Fisher matrix in this case can be explained from the properties of the ground state: When $h = 0.0$, the Hamiltonian Eq. (19) has two ground states $|0\rangle^{\otimes N}$ and $|1\rangle^{\otimes N}$. We first note that the optimization consistently found a solution with $a \approx 0$ and $b \approx 0$, leading to a $\mathbb{Z}_2$ symmetric state. Let us therefore assume that the solution we have exactly describes the $\mathbb{Z}_2$ symmetric ground state; i.e., $a = b = 0$. Then the ground state is $|0\rangle^{\otimes N} + |1\rangle^{\otimes N}$ leading to an RBM representation $|\psi_\theta(x)|^2 = 1/2$ for $x = x_0$ or $x = -x_0$ where $x_0 = [1 \cdots 1]$, and zero otherwise.

Moreover, we have $O(x_0) = [x_0, y_0, x_0 \otimes y_0]$ and $O(-x_0) = [-x_0, -y_0, x_0 \otimes y_0]$ where $y_0 := [\tanh \chi_1(x_0), \ldots, \tanh \chi_m(x_0)]$. This gives

$$\mathbb{E}[O] = (0 \;\; 0 \;\; x_0 \otimes y_0), \tag{20}$$

$$\begin{aligned}\mathbb{E}[O^\dagger O] &= \frac{1}{2}[O(x_0)^\dagger O(x_0) + O(-x_0)^\dagger O(-x_0)] \\ &= \begin{pmatrix} x_0^\dagger x_0 & x_0^\dagger y_0 & 0 \\ y_0^\dagger x_0 & y_0^\dagger y_0 & 0 \\ 0 & 0 & (x_0 \otimes y_0)^\dagger (x_0 \otimes y_0) \end{pmatrix}. \end{aligned} \tag{21}$$

Thus, the quantum Fisher matrix is

$$\begin{aligned} S &= \begin{pmatrix} x_0^\dagger x_0 & x_0^\dagger y_0 & 0 \\ y_0^\dagger x_0 & y_0^\dagger y_0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\ &= (x_0 \;\; y_0 \;\; 0)^\dagger (x_0 \;\; y_0 \;\; 0), \end{aligned} \tag{22}$$

which is rank 1. We note that the above argument does not depend on the details of the weights w, rather only on its magnitude $|w|$, so that any set of RBM weights that accurately model the ground state will exhibit the same behavior. The SR optimization typically favors small weights.

As the external field h increases, the number of terms of the ground state in the computation basis increases; thus we also expect that rank of S to increase as $\mathbb{E}[O^\dagger O] = \sum_x |\psi_\theta(x)|^2 O(x)^\dagger O(x)$. This is consistent with the results from our numerical data in Fig. 2(b). Importantly, rank deficiency is observed throughout the ferromagnetic phase, albeit much more pronounced in the vicinity of $h = 0$. We interpret this behavior as a signature that the phase is connected to a product state in the physical basis. For values of h close to one, the rank deficiency can only be seen at large system sizes and after many training epochs.

2. The critical point ($h = 1.0$)

At the critical point, the distribution of eigenvalues after convergence is smooth and decreasing exponentially. This behavior is also seen in many classical image processing tasks in machine learning [28,31], suggesting that it might be signature of (critical) long-range order. Indeed, each element of the quantum Fisher matrix can be expanded in terms of correlation functions, all of which are sizable in the critical case. This eigenvalue distribution is characteristic of "sloppy model universality," which has been shown to reflect systems with certain forms of scale invariance [25,26], further corroborating the claim. We will see in Sec. III B that this behavior is seen in many other systems and reveals that the RBM is fine tuning a solution with the help of a large number of hidden units.

3. The paramagnetic phase ($h > 1.0$)

In this case, we see that the energy converges rapidly and the eigenvalues almost do not change after the initial learning stage. In particular the second jump in the spectrum of the initial random RBM survives until the end. When $h = 2.0$, the jump is located at $N + N(N-1)/2 = 406$, revealing that the quantum Fisher matrix has no support on the antisymmetric subspace (see Appendix B). Precisely, the 406th eigenvalue has magnitude $\approx 4.08 \times 10^{-2}$ and the next one has magnitude $\approx 1.38 \times 10^{-3}$ in our numerical data.

To understand the stepwise behavior, we first focus on the randomly initialized RBM case, i.e., at epoch 0. As we initialize the parameters of the RBM with small random Gaussian values [sampled from $\mathcal{N}(0, \sigma^2)$ where $\sigma \sim 10^{-2}$], the classical probability distribution $|\psi_\theta(x)|^2$ would be similar to the case when all parameters are zero. When $a = b = w = 0$, the RBM gives $|\psi_\theta(x)|^2 = \mathbb{1}/2^N$, i.e., the identity distribution. We can then perturbatively expand the quantum Fisher matrix in terms of the parameters. The derivation up to $O(\sigma^3)$ is given in Appendix C. Our derivation gives N eigenvalues of $O(1)$ associated with the visible biases block of the matrix and $N(N-1)/2$ eigenvalues of order $O(\sigma^2)$ in the weights block of the quantum Fisher matrix. This explains the first and the second jumps in the eigenvalue distribution of the random RBM.

The randomly initialized RBM also hints at the fact that the quantum Fisher matrix throughout the paramagnetic phase strongly retains properties of the $h \gg 1$ limit with product state $|+\rangle^N$. We can compare the spectra of the quantum Fisher matrix for $h = 2.0$ and the randomly initialized case in Fig. 2(c). It shows that the second step is preserved but the first step disappears. This is because the first step depends on the details of weights, but the second one is the consequence of the symmetry. We make a detailed comparison between the quantum Fisher matrix for the paramagnetic phase and randomly initialized RBM in Appendix D. We there show that the converged matrix has larger diagonal elements in the w part of the matrix than the random RBM case which also support eigenvalues between N to $N(N+1)/2$.

Throughout the phase diagram of the TFI, the spectrum of the quantum Fisher matrix at convergence has two special points at N and at $N(N+1)/2$, as seen in Fig. 2(c). The location of these points is independent of the number of hidden units, suggesting that they originate from the $\mathbb{Z}_2$ nature of the physical system, and the overall bipartite structure of the RBM, rather than any details of the RBM graph.

B. Eigenvectors

Above we have argued the eigenvalues of the quantum Fisher matrix reveal signatures of the phase of matter being simulated. We now ask whether the eigenvectors can teach us anything about how correlations are conveyed in the learning landscape. In particular, since the complex RBM is constructed from a bipartite graph with no connections among the hidden and visible units, we know that all correlations have to be mediated by weights. Entanglement in the information manifold is therefore completely contained in the weights block of the Fisher matrix.

In Fig. 2(d) we plot the entanglement between the visible and hidden units of the w part of each eigenvector [see Eq. (18)]. We observe that the first N eigenvectors have very little entanglement when $0 \leqslant h \leqslant 1$. This suggests that the directions of largest curvature are almost exclusively associated with the biases, or first moments, of the distribution. Note that this does not imply that the values of the w weights are small, as representations of the first moments are distributed over the biases and the weights. Rather it is a reminder that the actual values of the weights of the network reveal little information of the correlations in the system, as is manifest in Fig. 7 in Appendix C. This behavior is less pronounced for $h > 1$ as the quantum Fisher matrix behaves more like a random matrix whose eigenvectors are expected to have a more homogeneous amounts of entanglement.

The entanglement increases in the bulk of the spectrum. Interestingly, this means that the directions in parameter space that encode information about correlations are typically dense, smooth, and flat. In the context of classical ML, these properties are akin to good generalization ability of the learning models, whereas in the present physics context, we interpret it to meant that the algorithm preferentially learns stable configurations, where changes (even large) in most directions in configuration space will not affect the physically observable properties of the system. Similar conclusions have been alluded to in the context of sloppy models universality in statistical mechanics [25,26].

C. Predictions

From the spectral analysis of the quantum Fisher matrix for the transverse field Ising model, we make the following predictions, which we expect to hold more generally for ferromagnetic quantum spin models:

(1) The spectral profile is universal within a phase of the model and is only weakly dependent on system size away from phase transition points. The spectrum of the quantum Fisher matrix is therefore a good indicator of the existence of a phase transition if it is possible to find two points in phase space with vastly different spectral profiles.

(2) The first N eigenvectors are close to product states and hence do not encode correlations in the system. They mostly pertain to first moments of the distribution.

(3) A rank-deficient quantum Fisher matrix is evidence that the state is in a phase connected to a product state in the chosen computational basis. A smoothly decaying spectrum is a sign that the system contains significant correlation, often a critical phase with polynomial decaying correlation functions.

(4) Kinks in the spectrum reveal symmetries in the model. In the case of the TFI, the persistent kink at $N(N+1)/2$ is a sign that the symmetric and antisymmetric subspaces are strictly separated everywhere except at the critical point.

III. FURTHER EXPERIMENTS

In this section, we study three further models to test whether the predictions made in Sec. II C extend to more general spin systems. The first model is the two-dimensional transverse field model, which is not known to be exactly solvable. The second is the coherent Gibbs state, whose quantum Fisher matrix is evaluated exactly without having recourse to learning. These two models exhibit Z_2 symmetry breaking as in the one-dimensional transverse Ising model that we studied above. For these models, we find the similar quantitative behaviors of the Fisher matrix, which strongly suggest the universality of our predictions. Our last example is the XXZ model, where we explore the Fisher matrix in all three phases.

A. Two-dimensional transverse Ising model

We consider the Hamiltonian defined in a $L\times L$ two-dimensional lattice given as

$$H = -\sum_{\langle i,j\rangle} \sigma_z^i \sigma^j - h \sum_i \sigma_x^i, \tag{23}$$

where the first summation is over all nearest neighbors $\langle i, j\rangle$ of the lattice. The essential physics is the same as the one-dimensional model; i.e., the system is in the ferromagnetic phase when $h < h_c$ and paramagnetic phase when $h > h_c$. However, the critical point h_c is only approximately known $\approx 3.00 \pm 0.05$ as the system is not exactly solvable in this case [32,33].

For the system size $L = 5$ that we can directly compare with the exact diagonalization, we simulated the system and plot the normalized energy and the spectral profiles of the Fisher matrix in Figs. 3(a) and 3(b). We clearly see the rank deficiency for $h = 0.0$ and 1.5, smooth spectrum at $h \approx h_c$, and kinks when $h = 4.5$ and 6.0, which confirms the universality of our predictions. In addition, Fig. 3(c) verifies that the kinks are located at $N(N+1)/2$ and Fig. 3(d) indicates low entanglement between hidden and visible layers in leading eigenvectors.

B. Coherent Gibbs state of the two-dimensional classical Ising model

We next consider the RBM representation of the coherent Gibbs state of the two-dimensional classical Ising model. Recall the classical Ising model

$$H(x) = -J \sum_{\langle i,j\rangle} x_i x_j, \tag{24}$$

where x is the configuration of the spin and $\langle i, j\rangle$ are nearest neighbors on a two-dimensional lattice. For convenience, we

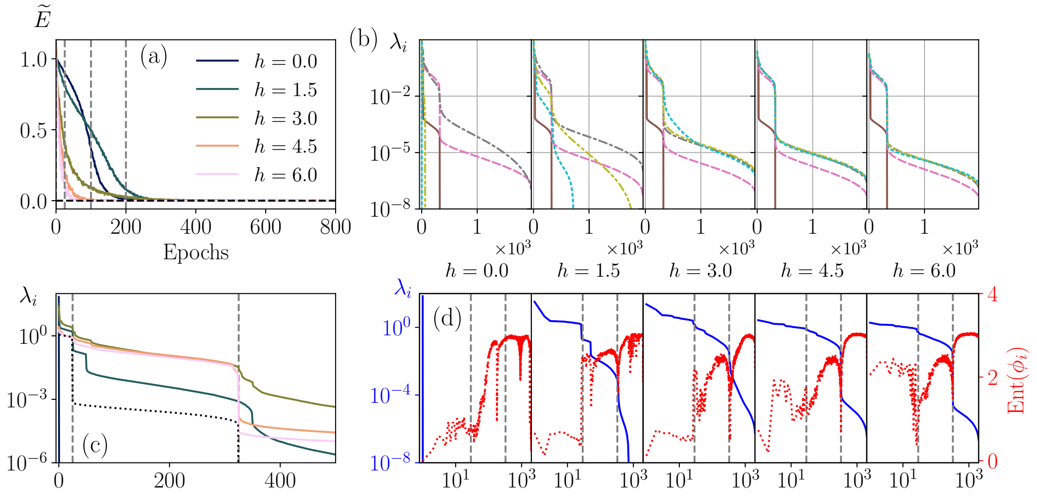

FIG. 3. Two-dimensional transverse field Ising model in 5×5 lattice: (a) Rescaled energy as a function of epochs for $h = [0.0, 1.5, 3.0, 4.5, 6.0]$ (from darkest to lightest). (b) Ordered eigenvalues of the quantum Fisher matrix [Eq. (3)] at epochs 0 (solid), 25 (dashed), 100 (dot-dashed), 200 (dot-dot-dashed), and 2000 (dotted). The results from $h = 0.0$ (leftmost) to $h = 6.0$ (rightmost) are shown in each subplot. (c) The 500 largest eigenvalues after convergence and for randomly initialized RBM (black dotted curve). The same color coding as in (a) is used. Two gray lines indicate $N = 25$ and $N(N+1)/2 = 325$. (d) Spectrum (blue solid) and entanglement in the eigenvectors (red dotted) on log-log scale. Hyperparameters $\eta = 0.002$ and $\epsilon = 0.001$ are used.

set $J = 1$. We consider a system in thermal equilibrium with inverse temperature $\beta = 1/T$. At high temperature $\beta < \beta_c$, the system exhibits a disordered paramagnetic phase characterized by zero magnetization $\langle x \rangle = 0$, whereas it shows a $\mathbb{Z}_2$ symmetry broken ferromagnetic phase with nonzero magnetization at sufficiently low temperature $\beta > \beta_c$ [34]. The phase transition takes place at $\beta = \beta_c \approx 0.44$ in the thermodynamic limit and is second-order. We thus have polynomial decay of the correlation function $\langle x_i x_j \rangle_c \sim 1/\mathrm{dist}(i, j)^\alpha$ at the critical point.

The coherent Gibbs state for the model with inverse temperature β is given by

$$|\varphi(\beta)\rangle = \sum_{\{x\}} \frac{e^{-\beta H(x)/2}}{\sqrt{Z}} |x\rangle \tag{25}$$

in a chosen computational basis $\{x\}$, and $Z = \sum_{\{x\}} e^{-\beta H(x)}$ is the normalization factor, which is the same as the partition function of the classical model. A key observation is that correlation functions of spin-z operators are exactly the same as that of the classical model, i.e., $\langle \varphi(\beta) | \sigma_z^i \sigma_z^j | \varphi(\beta) \rangle = \langle x_i x_j \rangle_{x \sim p(x)}$ where $p(x) = e^{-\beta H(x)}/Z$ is the Boltzmann distribution. Thus we also have polynomially decaying quantum correlation functions for this state at $\beta = \beta_c$. We also note that even though this state is artificially constructed, the state is a ground state of a Hamiltonian that is local in a two-dimensional lattice [35].

It is known that coherent Gibbs states of Ising-type models can be represented exactly as an RBM [36] by associating each edge of the lattice to one hidden unit (we provide a self-contained derivation in Appendix E). In particular, the coherent Gibbs state of an Ising-type model defined on a graph $G = (V, E)$ can be described using the RBM with parameters $a = b = 0$ and a $|V|$ by $|E|$ sparse weight matrix w.

Using this mapping, we construct the quantum Fisher matrix of the RBM representation for coherent Gibbs states. To sample from the distribution, we have employed the Wolff algorithm [37] instead of the usual local update scheme in this case as it is more efficient close to the transition point. The spectral profiles of the quantum Fisher matrix for different values of β are shown in Fig. 4(a).

The figure shows a very similar shape to that of the TFI case when they are deep in the ferromagnetic or paramagnetic phase. The eigenvalues exhibit a collapsing distribution in the ferromagnetic phase for large β and get progressively more

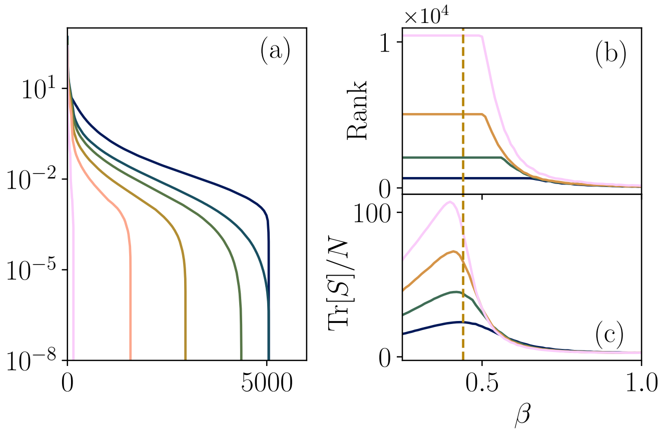

FIG. 4. (a) Eigenvalue distributions of the quantum Fisher matrix for coherent Gibbs states of two-dimensional classical Ising model. The inverse temperature $\beta \in [0.10, 0.50, 0.52, 0.55, 0.6, 0.9]$ (from darkest to lightest) is used. We used a $L \times L$ lattice with $L = 10$, so $N = 100$. The number of hidden units M is given by the number of edges in the graph, which is 180 (open boundary condition is used). The step is exactly located at $N(N+1)/2 = 5050$. (b) The rank of the quantum Fisher matrix and (c) the trace of the quantum Fisher matrix as functions of β from $L = 6$ (lower dark curves) to $L = 12$ (upper light curves).

singular as we increase β. Compare this behavior to the TFI for $h < h_c$ depicted in Fig. 2. In the paramagnetic phase ($\beta < \beta_c$), we see a stepwise distribution where the step is exactly located at $N(N+1)/2$, very much like the TFI model at large h. Thus for coherent Gibbs states that are deep in each phase, we get the same qualitative behavior of the quantum Fisher matrix in both models.

In contrast to the learned TFI case in Sec. II, the drop-off at $N(N+1)/2$ survives also at criticality. This can be understood by the fact that the quantum Fisher matrix is constructed from the exact coherent Gibbs state, which is exactly symmetric in the exchange of spins. Hence the quantum Fisher matrix has zero support on the antisymmetric subspace also at criticality. In Fig. 4(c) we have plotted the *quantum Fisher information*, which is simply the trace of the quantum Fisher matrix for different values of β. We see that the quantum Fisher information reaches a maximum in the vicinity of the phase transition point, hence acting as an order parameter reminiscent of the magnetic susceptibility. A more detailed analysis of the quantum Fisher information as a witness of phase transitions for this and other models will be presented elsewhere.

C. The XXZ model

We now consider the Heisenberg XXZ model

$$H = \sum_{i=1}^{N} \sigma_x^i \sigma_x^{i+1} + \sigma_y^i \sigma_y^{i+1} + \Delta \sigma_z^i \sigma_z^{i+1}. \qquad (26)$$

This model is exactly solvable using the Bethe *Ansatz*. The solution shows three distinct phases: (1) a gapped ferromagnetic phase for $\Delta \leqslant -1.0$, (2) a critical phase for $-1.0 < \Delta \leqslant 1.0$, and (3) a gapped antiferromagnetic phase for $\Delta > 1.0$. The ground state when $\Delta \leqslant 1.0$ is a superposition between $|0\rangle^{\otimes N}$ and $|1\rangle^{\otimes N}$. It is also known that the ground state is in $J_z := \sum_i \sigma_z^i = 0$ subspace for $\Delta > -1.0$. In the critical phase ($-1.0 < \Delta \leqslant 1.0$), the Hamiltonian is gapless in the thermodynamic limit, and the correlation length diverges. The phase transition at $\Delta = -1.0$ is first order, and an infinite order Kosterlitz-Thouless transition takes place at $\Delta = 1.0$.

We will again look at the spectral properties of the Fisher information matrix in this model for $\Delta = -1.0$, 0.0, and 1.0. For $\Delta = 0.0$ and 1.0, we have restricted the wave function to the $U(1)$ symmetric subspace $J_z = 0$ by applying the swap update rule in MCMC. Figure 5(a) shows the convergence of sampled energy over SR iterations. We see that SR successfully finds the ground states in all cases, but the initial drift starts later in the XXX case ($\Delta = 1.0$). Slow initial learning when $\Delta = 1.0$ is also checked in the spectrum of the quantum Fisher matrix shown in Fig. 5(b) where the spectrum begins to change slowly compared to other cases. We suspect that the SU(2) symmetry of the Hamiltonian is related to slow learning in the initial stage. When we compare the quantum Fisher matrices and the gradient of energies, which are two main ingredients of SR, for different values of Δ, quantum Fisher matrices do not differ much as they depend only on the parameters of the RBM, but the gradient of the energy $\nabla_\theta \langle H \rangle$ is much smaller when $\Delta = 1.0$ than other cases.

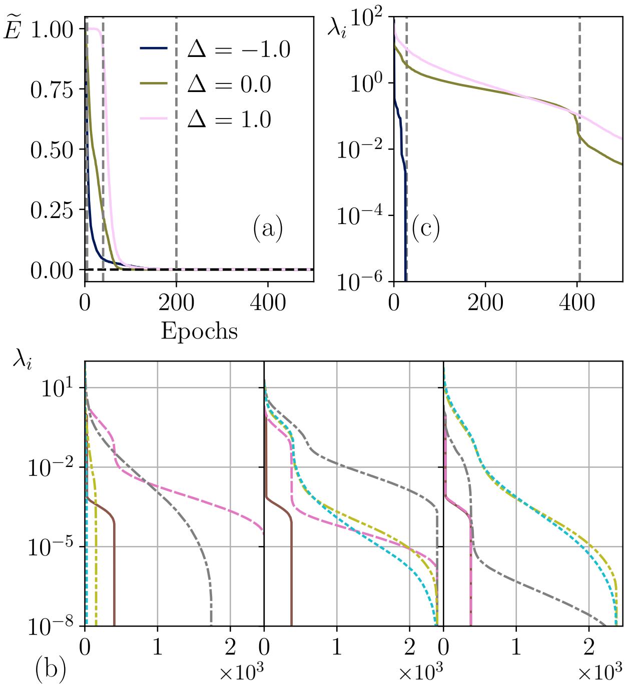

FIG. 5. (a) Rescaled energy as a function of epochs for the XXZ model with $\Delta = -1.0$, 0.0, and 1.0 (from darkest to lightest). (b) Spectra of the quantum Fisher matrix at epochs 0 (solid), 5 (dashed), 40 (dot-dashed), 200 (dot-dot-dashed), and 2000 (dotted) when $\Delta = -1.0$ (left), 0.0 (middle), and 1.0 (right). (c) Spectra of converged Fisher matrices. The same colors with (a) are used for Δ. Hyperparameters $\eta = 0.02$ and $\epsilon = 0.001$ are used for SR.

We plot the converged spectra in Fig. 5(c). Using this, we can extract some information for the converged ground state when $\Delta = -1.0$. As the first-order phase transition occurs at this point, the system has two different types of ground states: one that is a superposition of $|0\rangle^{\otimes N}$ and $|1\rangle^{\otimes N}$ from $\Delta \leqslant -1.0$ and the other one living in a subspace $J_z = 0$ from $\Delta > -1.0$. As the converged spectrum is singular, we can expect that the ground state found in our simulation is ferromagnetic. We indeed have calculated $\langle J_z^2 \rangle$ from Monte Carlo samples, and it gives $\langle J_z^2 \rangle / N^2 \approx 0.984$, which means a large portion of the state is in $|0\rangle^{\otimes N}$ and $|1\rangle^{\otimes N}$. When $\Delta = 0.0$ and 1.0, we see broader converged spectra. We note that there is a small step at $\sim N(N+1)/2$ when $\Delta = 0.0$ even though the whole spectrum is dense. In comparison, a smoother spectrum is obtained when $\Delta = 1.0$.

One should also ask about the behavior of quantum Fisher matrix in the antiferromagnetic phase. However, we found that usual MCMC does not produce unbiased samples in the antiferromagnetic phase, so SR does not converge to the real ground state [38]. As a consequence, we checked the optimization using the exactly constructed quantum Fisher matrix for small enough systems from the probability distribution $|\psi_\theta(x)|^2$. The result obtained from the exact simulation for the system size $N = 20$ is shown in Appendix F. One observation is that we see a dense converged spectrum when $\Delta = 2.0$ despite the system being gapped. Thus the gap of the system alone does not implies a dense spectrum of the quantum Fisher matrix.

IV. IMPLICATION FOR OPTIMIZATION

In this section, we use the insight gained about the structure of the quantum Fisher matrix to construct an optimization method for quantum spin systems. This method allows for significant savings in evaluation time for solving the inverse linear problem in the SR. Precisely, in each step of SR, we need to solve the linear equation

$$Sv = \nabla_\theta \langle H\rangle \tag{27}$$

for a given quantum Fisher matrix S. Even when the matrix S is well conditioned, the complexity of solving this equation scales as $O(D^2)$ where D is the dimension of the S matrix or number of parameters. As D itself scales like $O(\alpha N^2)$, the time cost is quartic in N. This is one of the main reasons why second-order methods, including natural gradient descent, are not widely used in classical large-scale deep learning applications.

Our optimization method can be seen as an extension of RMSProp [39]. The method provides a significant advantage in computation time as it does not involve solving a large system of linear equations. However, the method is not always a good approximation of the natural gradient, but rather depends decisively on the structure of the quantum Fisher matrix.

Before describing our method, we briefly review RMSProp for classical machine learning and how it is related to the Fisher information metric from the viewpoint of Ref. [40]. For convenience, the original RMSProp is described in Appendix G. This algorithm improves a naive stochastic gradient descent by using v_t, the running average of the squared gradients, to rescale the instantaneous gradient for updating weights. An observation in Ref. [40] is that v_t is a diagonal approximation of the uncentered covariance matrix of gradients when the learning is in the steady state. When the function we want to optimize f is the logarithmic likelihood (which is typical in classical machine learning), v_t recovers the diagonal part of the Fisher information metric at stationarity. The additional square root and ϵ prefactor in the last step are added to correct for "poor conditioning" [41]. This provides a plausible argument for why such a simple algorithm works incredibly well. One can also argue that other popular and efficient optimizers such as Adagard, Adadelta, and Adam similarly use a type of diagonal approximation of the Fisher information metric [40].

We now describe our variant of RMSProp applied to the ground-state optimization problem. Using the same principle as above, one may use $\langle O\rangle$ to estimate the diagonal part of the uncentered quantum Fisher matrix $\tilde{S}_{\alpha,\alpha} = \langle O_\alpha^\dagger O_\alpha\rangle$. The details of the algorithm are outlined in Algorithm I. A distinguishing property of this algorithm to the original RMSProp is that it uses different vectors for a gradient decent direction and estimating the curvature: v_t is calculated by $\langle O\rangle$, but the gradient of the energy is used for update in the last step. The algorithm suggested here is also different from the method used in Refs. [42,43] that put the energy gradient directly to the classical optimizers.

Algorithm 1. RMSProp for ground-state calculation. Hyperparameters $\beta = 0.9$ and $\epsilon = 10^{-8}$ are used in our example.

```
Require: $\eta$: Learning rate
Require: $\beta$: Exponential decay rate
Require: $\theta_0$: Initial parameter vector
$t \leftarrow 0$ (Initialize time step)
$v_0 \leftarrow 0$ (Initialize second moment vector)
while $\theta_t$ is not converged
    $t \leftarrow t + 1$
    $g_t \leftarrow$ Gradient of the energy
    $O_t \leftarrow \langle O\rangle$
    $v_t = \beta v_{t-1} + (1-\beta) O_t^* \odot O_t$
    $\theta_t = \theta_{t-1} - \eta g_t \odot 1/(\sqrt{v_t} + \epsilon)$
end while
```

We have tested the proposed version of RMSProp using different learning rates η for the TFI. The results for the ferromagnetic phase and the critical case ($h = 0.0$ to 1.0) are shown in Fig. 6. For small h, we see that RMSProp gets easily stuck in local minima unlike SR. When $h = 0.0$ and 0.2, the figure shows that the energy converges to that of the ground state for some learning rate η. However, such a convergence is probabilistic. For $h = 0.0$, 0.2, and 0.4, we ran the same simulation several times and found that, for any η, some instances converge to the ground state whereas others get stuck in local minima. In contrast, SR works properly for a wide range of hyperparameters and h, for which the energy converges to the ground state regardless of the choice of the learning rates $\eta = [0.005, 0.01, 0.02]$.

For larger h such as $h = 0.6, 0.8$, the proposed RMSProp shows better convergence behaviors for most values of η, but it still shows stepwise dynamics. In the critical case $h = 1.0$, the learning curves of RMSProp are smooth and insensitive to the choice of the learning rate, suggesting that the system no longer gets stuck in problematic local minima.

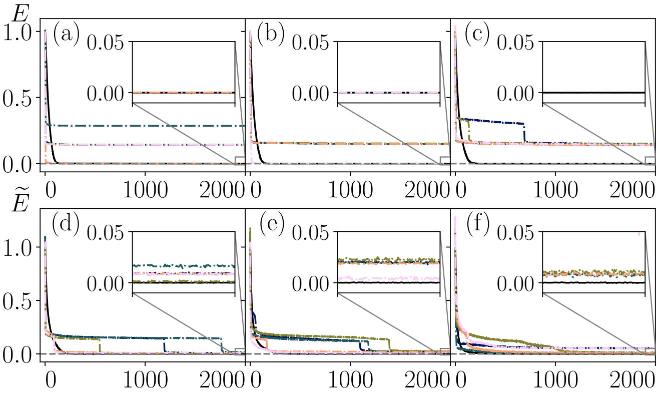

FIG. 6. Epochs versus rescaled energies obtained from the RMSProp (dot-dashed) with different learning rates and the SR with $\eta = 0.01$ (black solid). The TFI with the transverse fields from (a) $h = 0.0$ to (f) 1.0 are used. For the RMSProp, we used learning rates $\eta = 1.4\times10^{-3}$ (the darkest) to 2.2×10^{-3} (the lightest) with the interval 0.2×10^{-3}.

Our results suggest that preserving the singular nature of the quantum Fisher matrix is essential for ensuring convergence to the ground-state energy. Indeed, the converged quantum Fisher matrices studied in Appendix D show that the diagonal of the Fisher matrices give rank $N + M = 112$ for $h = 0.0$ and full rank ($NM + N + M = 2464$) for other values of h. In contrast, the real ranks of the quantum Fisher matrices (measured by counting the number of eigenvalues larger than 10^{-10}) are given as 1, 78, 242, 726, 1698, 2464 for $h = 0.0$, 0.2, 0.4, 0.6, 0.8, and 1.0, respectively.

We still note that even though the rank provides a plausible argument for the behavior of the learning curves, it does not for the converged energies; the converged energies for $h = 0.8$ and 1.0 are slightly larger than the ground-state energies. Moreover, the convergence behavior in the paramagnetic phase ($h > 1.0$) is more complicated and cannot be solely explained from the quantum Fisher matrix. A partial reason is that the path taken by RMSProp deviates from that of the SR in initial stage of learning (see Appendix G). Detailed investigations in this regime remain for future work.

V. CONCLUSION

We have initiated a detailed study of the quantum information geometry of learning ground states of spin chains in the artificial neural network framework. We have focused on complex restricted Boltzmann states and the stochastic reconfiguration method, which implements a quantum version of Amari's natural gradient update scheme. Our main result is that the eigenvalues and eigenvectors of the quantum Fisher matrix reflect both the learning dynamics, which is unsurprising, as well as the intrinsic static phase information of the model under study, which is rather surprising. In particular, we found that in the entire noncritical ferromagnetic phase of a number of models, the spectrum of the quantum Fisher matrix has reduced rank. The matrix becomes highly singular in regions of the phase that are close to product states. In critical phases, the spectrum becomes smooth with more and more eigenvectors contributing to the information geometry landscape.

We have identified a universal behavior of the leading eigenvectors of the quantum Fisher matrix: they all convey little entanglement, as measured by the entanglement entropy between the visible and hidden layers. This, in combination with the insight that critical models have smooth spectra, suggests that correlations in the complex RBM *Ansatz* are preferentially represented in the bulk of the information geometry space. Our interpretation of this key dynamical feature of RBM learning is that the model preferentially chooses stable representations, where the entropy of the landscape dominates over the energy. A similar phenomenon is classical supervised machine learning is frequently observed in discussion of "generalization." Finally, we explored strategies for diagonal approximations of the quantum Fisher matrix and found that their success crucially depends on the phase of the model under study. We therefore do not expect any diagonal approximation of the quantum Fisher matrix to be effective in general.

ACKNOWLEDGMENTS

We thank S. Trebst and D. Gross for helpful discussions. We acknowledge support from the DFG (CRC TR 183) and the ML4Q excellence cluster. Source codes for the current paper can be found in CYP's github repository [44]. The numerical simulations were performed on the CHEOPS cluster at RRZK Cologne.

APPENDIX A: STOCHASTIC RECONFIGURATION

For the reader's convenience, we derive the stochastic reconfiguration method of Sorella [20,21]. The main idea of stochastic reconfiguration (SR) is to modify the parameters of a trial wave function in such a way that it approaches the ground state along a path dictated by the projection $\mathbb{1} - \epsilon H$, where ϵ is chosen such that $\mathbb{1} - \epsilon H \geqslant 0$.

Let $|\psi_\theta\rangle$ be a state in our *Ansatz* class, with θ its vector of parameters. From now on, we will suppress the parameters θ. Then, for sufficiently small ϵ, we can write

$$(\mathbb{1} - \epsilon H)|\psi\rangle = e_0|\psi\rangle + \sum_\alpha e_\alpha|\psi_\alpha\rangle + |\psi^\perp\rangle, \tag{A1}$$

where $|\psi_\alpha\rangle = \frac{\partial}{\partial\theta_\alpha}|\psi\rangle$, $\{e_\alpha\}$ are coefficients, and $|\psi^\perp\rangle$ is a state in the orthogonal subspace. Note the identity $|\psi_\alpha\rangle = O_\alpha|\psi\rangle$, where the operators O_α are defined as

$$O_\alpha|x\rangle = \frac{\partial \log(\langle x|\psi\rangle)}{\partial\theta_\alpha}|x\rangle, \tag{A2}$$

where $|x\rangle$ is the computational basis.

We can now obtain a system of linear equations for the e_α coefficients by multiplying Eqn. (A1) by $\langle\psi|$ and by $\langle\psi_\alpha|$ to get

$$1 - \epsilon\langle H\rangle = e_0 + \sum_\alpha e_\alpha\langle O_\alpha\rangle, \tag{A3}$$

$$\langle O_\alpha^\dagger\rangle - \epsilon\langle O_\alpha^\dagger H\rangle = e_0\langle O_\alpha^\dagger\rangle + \sum_\beta e_\beta\langle O_\alpha^\dagger O_\beta\rangle. \tag{A4}$$

The averages are taken in the states $|\psi\rangle$. We can then solve for e_0 to get

$$\sum_\beta S_{\alpha,\beta}e_\beta = -\epsilon R_\alpha, \tag{A5}$$

where the matrix S is given by

$$S_{\alpha,\beta} = \langle O_\alpha^\dagger O_\beta\rangle - \langle O_\alpha^\dagger\rangle\langle O_\beta\rangle, \tag{A6}$$

and the vector R_α is given by

$$R_\alpha = \langle O_\alpha^\dagger H\rangle - \langle O_\alpha^\dagger\rangle\langle H\rangle. \tag{A7}$$

We can now identify the coefficients e_α as the update coefficients for the variables θ_α, up to an overall constant e_0, which can be interpreted as the learning rate. The SR update scheme can then be summarized as

$$\theta_\alpha \to \theta_\alpha - \eta\sum_\beta(S + \epsilon\mathbb{1})^{-1}_{\alpha,\beta}R_\beta, \tag{A8}$$

for some learning rate η. Here ϵ is regularization constant that is typically $\sim 10^{-3}$.

TABLE I. Parameters used for the simulations.

Model	Lattice size	Monte Carlo update	η	ϵ
1D TFI	28	Spin flip	0.01	0.001
2D TFI	5×5	Spin flip	0.002	0.001
XXZ	28	Swap	0.02	0.001

APPENDIX B: NUMERICS

For numerical simulation, we set the ratio between the numbers of hidden units and visible units of the complex RBM to $\alpha = M/N = 3$. Thus the RBM has $(\alpha + 1)N + \alpha N^2$ parameters overall (N and αN for biases and αN^2 for the weight matrix w). To sample from the RBM, the Markov chain Monte Carlo (MCMC) method enhanced with parallel tempering was employed [45]. We used 16 parallel Markov chains with linearly divided temperatures from 1/16 to 1. For each Markov chain, we used local spin flip updates for the transverse field Ising models (1D and 2D) and total magnetization-conserving swap updates for the XXZ model. To directly compare the results from variational Monte Carlo with exact diagonalization, we have used the size of system $N = 28$ for 1D models, $L\times L$ with $L = 5$ for 2D TFI, and imposed the periodic boundary condition. In our case, SR has two hyperparameters: the learning rate [η in Eq. (2)] and the regularization ϵ. These hyperparameters in our simulation results are summarized in Table I.

APPENDIX C: QUANTUM FISHER MATRIX OF RANDOM RBM

We provide an explanation of the stepwise structure of the spectrum of the quantum Fisher matrix upon small random initialization of the weights. The quantum Fisher matrix is broken up into three main sectors: $[a, b, w]$, corresponding to the visible biases, the hidden biases, and the weights.

As in the main text, we use $N = |a|$ and $M = |b|$ to indicate the number of visible and hidden units, respectively. In our simulations, the weights are initialized to be Gaussian distributed with an average magnitude of order $\sigma = 10^{-2}$. We therefore make the following assumption about the initial state: *the classical probability distribution associated with the initial quantum state is close to the identity, and in particular is separable*. This implies that each spin has zero expectation value at initialization $\langle x_j\rangle = 0$ for all j, and that $\langle x_j x_k\rangle \propto \delta_{jk}$ for all jk.

As the entries of the visible biases block are

$$S_{a_i,a_j} = \langle x_i x_j\rangle - \langle x_i\rangle\langle x_j\rangle = \delta_{ij}, \tag{C1}$$

we get the identity matrix for the a part. The covariance between the visible and hidden units involves the term $\langle x_i \tanh[\chi_j(x)]\rangle$. Recall that the argument of the hyperbolic tangents are

$$\chi_j(x) = b_j + \sum_i w_{ij} x_i, \tag{C2}$$

where b_j are the hidden biases and w_{ij} are the weights connecting the hidden and visible units. Under the assumption that all parameters are small, we approximate $\tanh[\chi_j(x)] \approx \chi_j(x)$. Then

$$\langle x_i \tanh[\chi_j(x)]\rangle \approx \langle x_i \chi_j(x)\rangle = b_j\langle x_i\rangle + \sum_k w_{kj}\langle x_i x_k\rangle$$
$$\approx \sum_k w_{kj}\delta_{ik} = w_{ij}. \tag{C3}$$

Likewise, we can obtain the full unary part ($[a, b]$) of the S matrix as

$$S_{\text{un}} = \begin{pmatrix} \mathbb{1}_N & w \\ w^\dagger & w^\dagger w \end{pmatrix}. \tag{C4}$$

We can easily see this is rank N as the first N row generates the remaining rows. This explains the first N eigenvalues, which are $O(1)$.

Next, the w part of the quantum Fisher matrix is given by

$$(S_w)_{ij,i'j'} = \langle x_i \tanh[\chi_j(x)]^* x_{i'} \tanh[\chi_{j'}(x)]\rangle$$
$$- \langle x_i \tanh[\chi_j(x)]^*\rangle\langle x_{i'} \tanh[\chi_{j'}(x)]\rangle, \tag{C5}$$

where i, i' label the visible units and j, j' label the hidden units. Using the expansion

$$\langle x_i \tanh[\chi_j(x)]^* x_{i'} \tanh[\chi_{j'}(x)]\rangle$$
$$\approx b_j^* b_{j'}\delta_{ii'} + \sum_{kk'} w_{ki'}^* w_{k'j'}\langle x_i x_k x_j x_{k'}\rangle, \tag{C6}$$

we have

$$S_w(b) = (\mathbb{1} \otimes w^\dagger)X(\mathbb{1} \otimes w) + \mathbb{1}_n \otimes |b\rangle\langle b|, \tag{C7}$$

where w is the $N\times M$ matrix of weights, $|b\rangle = \sum_j b_j|j\rangle$ is a vector form of the bias b, and $X = \sum_{ijkl} x_{ikjl}|ik\rangle\langle jl|$ with $x_{ikjl} = \langle x_i x_k x_j x_l\rangle - \langle x_i x_k\rangle\langle x_j x_l\rangle$. Using the assumption of small initial weights, we have

$$x_{ikjl} = \delta_{ij}\delta_{kl} + \delta_{il}\delta_{jk} - 2\delta_{ikjl}. \tag{C8}$$

Then the X matrix is approximately

$$X = \sum_{jk}(|jk\rangle\langle jk| + |jk\rangle\langle kj|) - 2\sum_j |jj\rangle\langle jj|$$
$$= \mathbb{1} + V - 2\sum_j |jj\rangle\langle jj|, \tag{C9}$$

where $V = \sum_{jk} |jk\rangle\langle kj|$ is the swap operator. The rank of X is given by $N(N-1)/2$. Moreover, X is the projector that preserves the symmetric states except the copied state, i.e., $X(|ab\rangle + |ba\rangle) \propto |ab\rangle + |ba\rangle$ when $a \neq b$ but $X|aa\rangle = 0$.

When $b = 0$, the whole covariance matrix is given by $S = S_{\text{un}} \oplus S_w$, and the matrix S_w [Eq. (C7)] has rank $N(N-1)/2$. This explains the small subleading eigenvalues of order $O(\sigma^2)$.

However, the block-diagonal assumption breaks down when we have nonzero bias in the hidden layer ($b \neq 0$) as we have off-diagonal blocks between the unary and w part. An additional $\mathbb{1} \otimes |b\rangle\langle b|$ also enters into S_w. Still, it is not difficult to see that this does not change the overall rank. A precise calculation gives

$$S(b) = \begin{pmatrix} \mathbb{1}_N & w & \mathbb{1} \otimes \langle b| \\ w^\dagger & w^\dagger w & w \otimes \langle b| \\ |b\rangle \otimes \mathbb{1} & |b\rangle \otimes w & S_w(0) + \mathbb{1} \otimes |b\rangle\langle b| \end{pmatrix} \tag{C10}$$

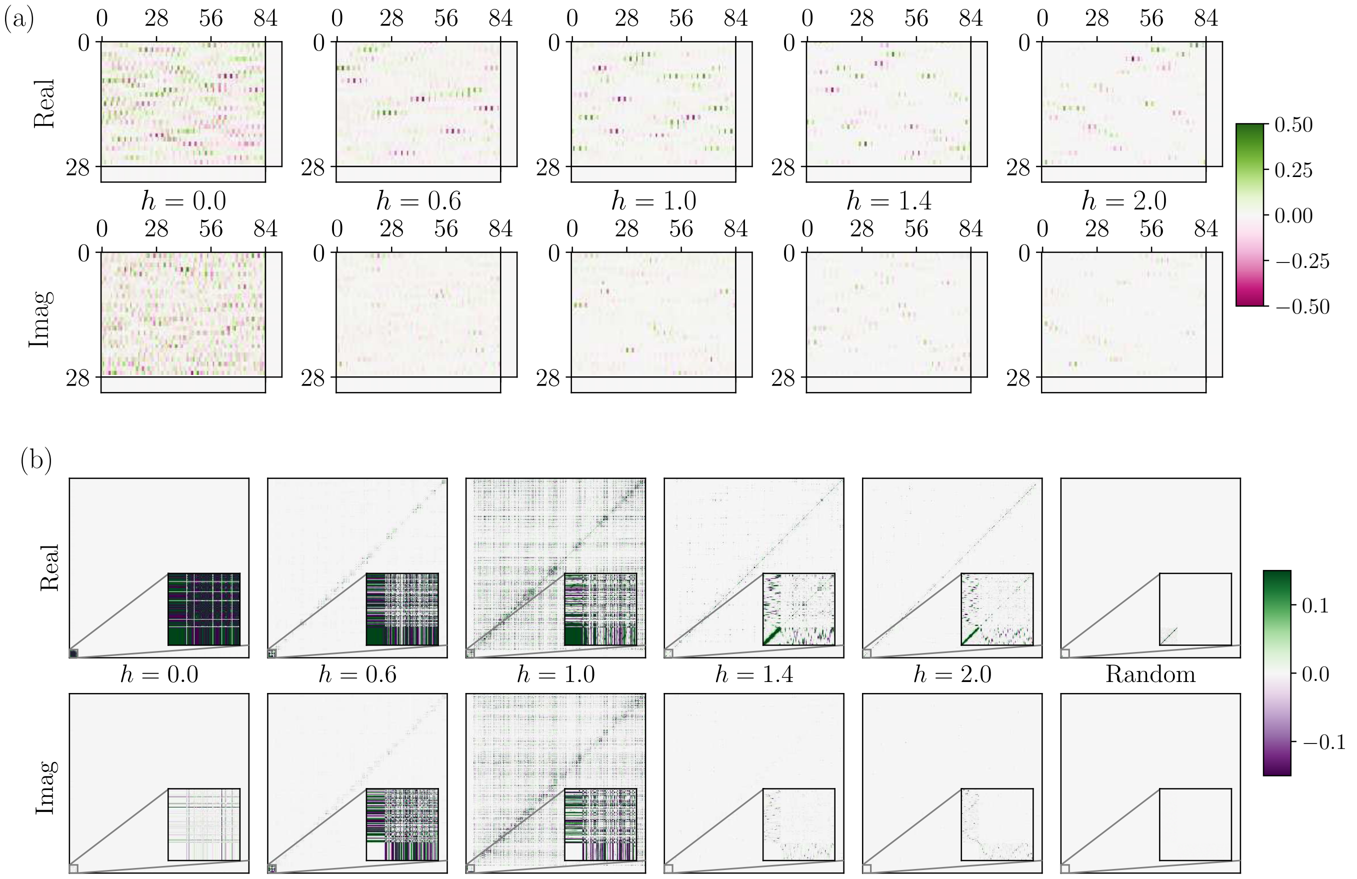

FIG. 7. (a) Converges weights (a, b, w) for the TFI model with different values of h. The large rectangle shows the weights w, whereas the small strips show the biases a and b, which are much weaker in magnitude than the leading weights. (b) Real and imaginary parts of the quantum Fisher matrix after convergence for the TFI as well as randomly initialized RBM. Insets show the correlation between unary variables. The whole matrix is order $N + M + NM = 2464$, and the unary part is order $N + M = 112$. The covariance between visible units are small left bottom corner of the unary part.

up to third-order corrections. It is simple to see that first N rows still generate the next M rows. Moreover, applying $|b\rangle$ to the first N rows gives the additional terms in the last NM rows so the rank of the S matrix from the w part also does not change. Thus we have exactly the same rank even when we turn on hidden biases b.

APPENDIX D: FURTHER PROPERTIES OF THE QUANTUM FISHER MATRIX

In this section, we investigate further properties of the quantum Fisher matrix. We use the same numerical data as in the main text; the TFI with system size $N = 28$.

1. Converged weights

Converged parameters of neural networks are often claimed to reveal features of the data or system under study [5,46]. We compare the converged weights and the quantum Fisher matrix for different values of h in Fig. 7. We find that, in contrast with the spectral information of the quantum Fisher matrix, it is difficult to infer any information from the converged weights of the network. For example, converged weights for $h = 0.6$, 1.0, and 1.4 are not sensibly different, whereas the quantum Fisher matrices reveal essential features of the phase of the system.

This brings to light one the of the key subtleties of RBM *Ansätze*, which is the extreme redundancy of representation. Let us illustrate this fact by constructing three completely different solutions of the RBM parameters that (approximately) represent the same quantum state $|0\rangle^{\otimes N} + |1\rangle^{\otimes N}$. As a first solution, consider the one obtained from our numerical simulation Fig. 7(a). This solution is fully complex, i.e., real and imaginary parts of the weights are both nonzero. On the other hand, a real solution can be found from the coherent Gibbs states for classical Ising model as discussed in Appendix E. The state is obtained by letting $J_{ij} = -1$ and $\beta \to \infty$ for a classical Ising model defined on any graph that does not have an isolated vertex. We note that the parameters obtained using this scheme are real as $e^{-\beta J_{i,j}} \geqslant 1$ (see Appendix E for details). Finally, it is also possible to represent this state only using pure imaginary parameters. By letting $a = 0$, $b = (i\pi/2, \dots, i\pi/2)$, and the weight w as

$$w_{i,j} = \begin{cases} i\pi/4, & \text{if } j = i + 1 \\ 0, & \text{otherwise} \end{cases}. \tag{D1}$$

It is clear from these examples that inferring information of quantum states solely from the activation parameters of the RBM is very ambiguous.

2. Nonzero elements of Fisher information matrix

We investigate the rank of the quantum Fisher matrix more closely. Let us first focus on the ferromagnetic phase ($h < 1.0$). In the main text, we have shown that the rank of the quantum Fisher matrix increases as h increases. A question we are interested in is how nonzero elements are distributed in unary and w parts of the matrix. To answer this question, we use the quantum Fisher matrix itself after convergence plotted in Fig. 7(b). When $h = 0.0$, we see that the Fisher information matrix has nonzero elements only in the unary part. In contrast, the w part of the matrix shows nonzero elements (especially in diagonal part) when $h = 0.6$. To see this clearly, we have counted the number of diagonal elements of the quantum Fisher matrix that are larger than 10^{-4}. It shows there are $N + M = 112$ such diagonal elements when $h = 0.0$ but $N + M + NM = 2464$ for all larger $h = 0.2, 0.4, 0.6, 0.8$. As the rank of the full matrix is small even for larger h, the nonzero elements in the w part in this case imply the eigenvectors with dominant eigenvalues have compelling w part. In addition, this provides an argument why RMSProp, which is studied in Sec. IV, works badly for small h.

Next, we consider the paramagnetic phase ($h > 1.0$). In the main text, we have shown that the Fisher information matrix when $h = 2.0$ shows a step at $N(N + 1)/2$. The whole shape of the spectrum remains similar for smaller h even though the location of step can be little shifted. Compared to the randomly initialized RBM, we see larger diagonal elements in w part. As Fig. 2 shows that eigenvalues between Nth to $N(N + 1)/2$ are much larger for the converged Fisher information matrix than the random RBM, we expect that the w part of the matrix contributes to these eigenvalues. To test this, we have diagonalized only the w part of the quantum Fisher matrix when $h = 2.0$ where we could observe a step at $N(N - 1)/2$. Thus despite that the whole spectrum does not show a clear step at the Nth eigenvalue, we may still consider that N eigenvalues are from the unary part and $N(N - 1)/2$ are from the w part. We also found that all diagonal elements of the quantum Fisher matrix are larger than 10^{-2} when $h \geqslant 1.0$, so the diagonal approximation of the quantum Fisher matrix is full rank.

3. System-size dependence of the spectral profile

When we use the same parameter $\alpha = M/N$ and the Hamiltonian, we observe that spectra of the converged Fisher information matrix behave almost the same for varying N. In Fig. 8 we show the spectra of the converged quantum Fisher matrix for different values of $N = [28, 32, 36, 40]$ using the TFI with different values of $h = [0.0, 0.6, 1.0, 1.4, 2.0]$. We clearly see that eigenvalue distributions for the same h vary only little with the change of the system size N. Still, it is not easy to make an exact correspondence between the results from different N as the order of the quantum Fisher matrix is given by $\alpha N^2 + (\alpha + 1)N$, which is not monomial. Thus there is no single constant scale factor we can use for rescaling the results. Still, this suggests that the spectrum of the quantum Fisher matrix can be used as a faithful diagnostic tool with small systems to infer qualitative behavior on larger systems.

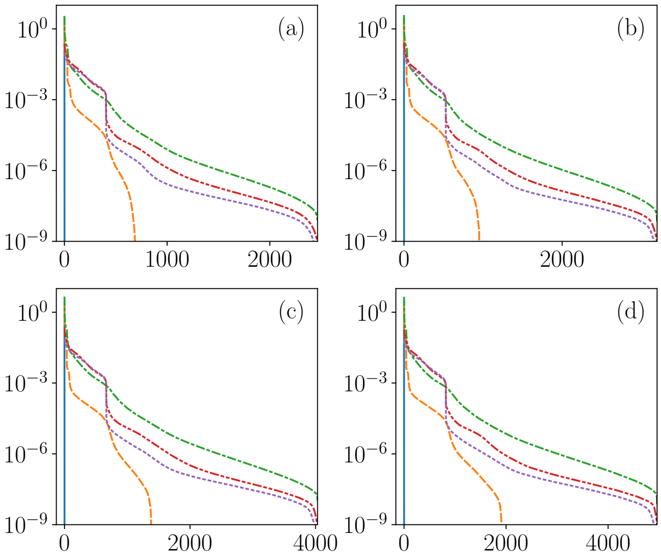

FIG. 8. Normalized eigenvalues λ_i/N of the converged quantum Fisher matrix for the TFI with system sizes $N = 28$ to 40 [from (a) to (d)]. The transverse fields $h = 0.0$ (solid), 0.6 (dashed), 1.0 (dot-dashed), 1.4 (dot-dot-dashed), and 2.0 (dotted) are used. The shapes of the distributions are independent to N.

APPENDIX E: COHERENT GIBBS STATES FOR CLASSICAL ISING MODELS

We consider a classical Ising model defined on a graph $G = (V, E)$ where $V = \{i\}$ is the set of vertices and $E = \{(i, j)\}$ is the set of edges. We assign binary values $x_i = 1$ or -1 to each vertex and interaction strengths $J_{i,j} \in \mathbb{R}$ to each edge $e = (i, j) \in E$. The Hamiltonian of this model is given by

$$H(x) = \sum_{(i,j)\in E} J_{i,j} x_i x_j. \tag{E1}$$

Then our objective is finding parameters of the RBM $[a, b, w]$ that describe coherent Gibbs states for the given β, i.e., solving the equation

$$\psi_\theta(x) = e^{a\cdot x} \prod_{j=1}^{M} 2\cosh \chi_j(x) = c \exp[-\beta H(x)/2] \tag{E2}$$

for all $x = \{-1, 1\}^N$. Here $\chi_j(x) = \sum_i w_{ij} x_i + b_j$ and c is a constant that can be freely chosen as our RBM does not use a specific normalization.

As the $H(x)$ is symmetric under overall flip ($x \to -x$), we first consider $\mathbb{Z}_2$ symmetric RBM that has zero biases, i.e., $a = b = 0$. Then we can simplify the equation to

$$\prod_{j=1}^{M} 2\cosh\left(\sum_k w_{kj} x_k\right) = c \prod_{(i,j)\in E} \exp[-\beta J_{i,j} x_i x_j/2]. \tag{E3}$$

We can find such a w easily by letting $M = |E|$ and equating each term using a column of w in the left-hand side to the term in the right-hand side using an edge. In other words, we solve

$$2\cosh\left(\sum_k w_{ke}x_k\right) = c_e \exp[-\beta J_{i,j}x_i x_j/2] \tag{E4}$$

for all $e \in E$ where c_e is a constant assigned to each edge e that gives $c = \prod_{e\in E} c_e$. Setting all $w_{ke} = 0$ if $k \neq i, j$, we then need to solve the coupled equations

$$2\cosh(w_{ie} + w_{je}) = c_e e^{-\beta J_e/2}, \tag{E5}$$

$$2\cosh(w_{ie} - w_{je}) = c_e e^{\beta J_e/2}. \tag{E6}$$

These equations can be solved for any $\beta J_{i,j}$ as w is a complex matrix.

For the two-dimensional Ising model we consider in the main text, $J_{i,j} = -1$ for all edges $(i, j) \in E$ that connect any neighboring vertices in 2D lattice. In this case, we can easily get a real solution $w_{ie} = w_{je} = \cosh^{-1}[e^{\beta}]/2$.

APPENDIX F: THE XXZ MODEL USING EXACT WAVE FUNCTIONS

In the main text, we studied the Heisenberg XXZ model using variational quantum Monte Carlo. There the observables such as the quantum Fisher matrix and the energy gradient are calculated from the samples obtained from MCMC. In this section, we study the same system using exactly constructed wave functions instead of MCMC. A modified step of each iteration of SR is as follows. First, we calculate all components of the wave function $\psi_\theta(x) = e^{a\cdot x}\prod_j 2\cosh\chi_j$ in the computational basis. Then we obtain the normalization factor by calculating the exponential sum $Z = \sum_{\{x\}} |\psi_\theta(x)|^2$. Using this result, the energy gradient and the Fisher information matrix are also calculated by computing Eqs. (5) and (6) exactly, and parameters are updated accordingly. As we do not sample from the distribution, the algorithm is not stochastic anymore. Thus we would call this method exact reconfiguration (ER) instead of SR. We note that ER is extremely expensive in computation since we need to calculate several exponential sums for each iteration.

Using ER, we have simulated the XXZ model with the system size $N = 20$, which is tractable using current CPUs. The result is shown in Fig. 9. There are two noteworthy features: First, the converged spectrum when $\Delta = -1$ shows a broader spectrum as compared to Fig. 5. We conjecture that this is related to the fact that the ground state found using ER has more component in $J_z = 0$ subspace compared to SR case. Indeed, we have $\langle J_z^2\rangle/N^2 \approx 0.963$, which is slightly smaller than what is found in the SR case in the main test. Second, the converged quantum Fisher matrix shows a smooth spectrum when $\Delta = 2.0$ even though the system has a gapped antiferromagnetic ground state. It implies that a smooth spectrum of the converged quantum Fisher matrix is not sufficient to infer criticality.

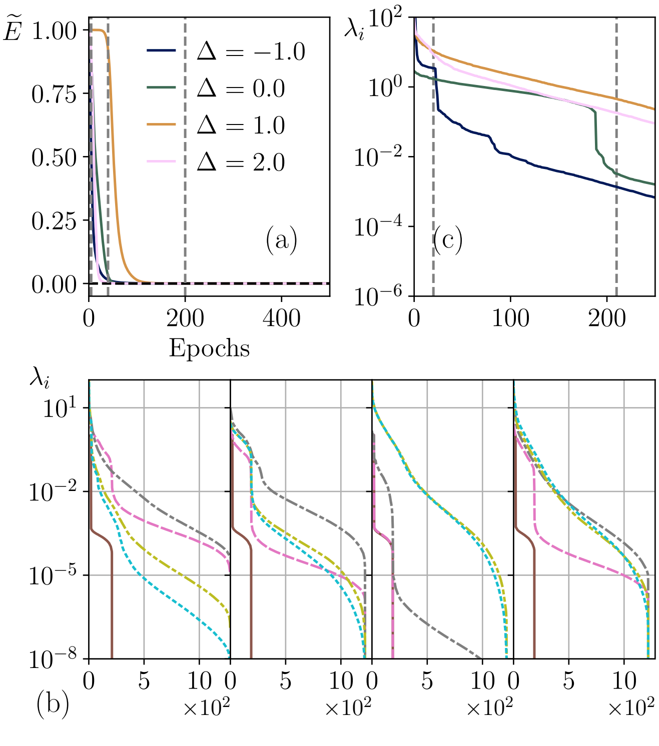

FIG. 9. Numerical results of the XXZ model with size $N = 20$ using exactly constructed wave functions. (a) Normalized energy $\widetilde{E} = (\langle E\rangle - E_{\rm ed})/(E_0 - E_{\rm ed})$ as a function of epochs. The interaction strengths from $\Delta = -1.0$ (the darkest) to 2.0 (the lightest) are used. (b) Dynamics of the spectrum of the Fisher information matrix at epochs 0 (solid), 5 (dashed), 40 (dot-dashed), 200 (dot-dot-dashed), and 2000 (dotted). Interaction strengths from $\Delta = -1.0$ (the leftmost) to 2.0 (the rightmost) with the interval 1.0 are used. (c) Spectrum of converged Fisher information matrix. The same colors with (a) are used to indicate Δ.

APPENDIX G: RMSPROP IN THE PARAMAGNETIC PHASE

We study in this Appendix RMSProp introduced in Sec. IV for the paramagnetic phase of TFI. The learning curves for five different values of h are shown in Fig. 10. We can see that the learning curves are more complex than those from the ferromagnetic and the critical cases. Specifically, we have

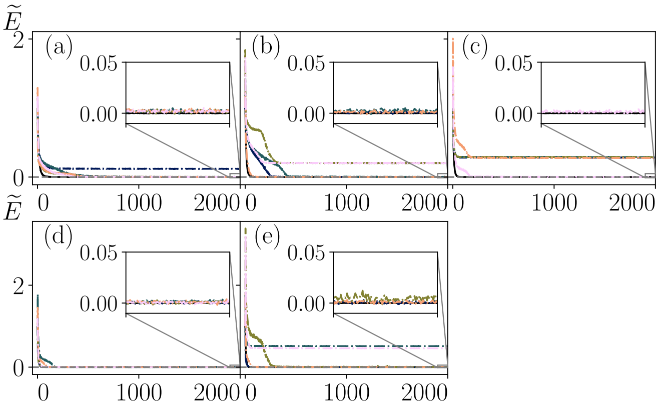

FIG. 10. Rescaled energy $\widetilde{E}$ as a function of epochs for TFI in the paramagnetic phase using the RMSProp (dot-dashed) and the SR with the learning rate $\eta = 0.01$ (black solid). Results from the transverse field (a) $h = 1.2$ to (e) 2.0 are shown. Learning rates 1.4×10^{-3} (the darkest) to 2.2×10^{-3} (the lightest) are used for the RMSProp.

three distinct observations, as follows. First, there is a spike of the rescaled energy that goes up in the initial stage of learning. In addition, the size of the spike grows with h. This means that an initial direction that the optimizer selects is different from the optimal direction. Second, the properties of the quantum Fisher matrix are not very relevant to the learning dynamics of the RMSProp. In Appendix D, we have shown that the properties of the quantum Fisher matrix do not change much within the paramagnetic phase. However, the learning curves from the RMSProp do not show a similarity between different values of h. Third, the converged energy can be as low as that of the SR case. This is interesting as the optimizer sometimes finds the proper solution even though the learning dynamic shows poor behavior.

From these observations, we suspect that RMSProp takes a different learning pathway than SR in the paramagnetic phase. To understand the applicability and details of the learning dynamics of the algorithm better, more detailed investigations such as tracking the path of optimization are required. We leave such a detailed investigation of this optimizer and the comparison to other optimizers for future work.

Algorithm 2. RMSProp. Here $\odot$ is the element-wise product of two vectors.

```
Require:  $\eta$: Learning rate
Require:  $\beta$: Exponential decay rate
Require:  $\theta_0$: Initial parameter vector
$t \leftarrow 0$ (Initialize time step)
$v_0 \leftarrow 0$ (Initialize second moment vector)
while$\theta_t$ is not converged
    $t \leftarrow t + 1$
    $g_t \leftarrow \langle \nabla_\theta f(\theta_{t-1}) \rangle$
    $v_t = \beta v_{t-1} + (1-\beta) g_t \odot g_t$
    $\theta_t = \theta_{t-1} - \eta g_t \odot 1/(\sqrt{v_t} + \epsilon)$
end while
```

[1] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Quantum machine learning, Nature (London) **549**, 195 (2017).

[2] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, A variational eigenvalue solver on a photonic quantum processor, Nat. Commun. **5**, 4213 (2014).

[3] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, The theory of variational hybrid quantum-classical algorithms, New J. Phys. **18**, 023023 (2016).

[4] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets, Nature (London) **549**, 242 (2017).

[5] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, Science **355**, 602 (2017).

[6] Y. Nomura, A. S. Darmawan, Y. Yamaji, and M. Imada, Restricted Boltzmann machine learning for solving strongly correlated quantum systems, Phys. Rev. B **96**, 205152 (2017).

[7] 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).

[8] K. Choo, T. Neupert, and G. Carleo, Two-dimensional frustrated J_1-J_2 model studied with neural network quantum states, Phys. Rev. B **100**, 125124 (2019).

[9] G. Carleo, F. Becca, M. Schiró, and M. Fabrizio, Localization and glassy dynamics of many-body quantum systems, Sci. Rep. **2**, 243 (2012).

[10] E. P. Van Nieuwenburg, Y.-H. Liu, and S. D. Huber, Learning phase transitions by confusion, Nat. Phys. **13**, 435 (2017).

[11] J. Carrasquilla and R. G. Melko, Machine learning phases of matter, Nat. Phys. **13**, 431 (2017).

[12] P. Broecker, J. Carrasquilla, R. G. Melko, and S. Trebst, Machine learning quantum phases of matter beyond the fermion sign problem, Sci. Rep. **7**, 8823 (2017).

[13] R. Sweke, M. S. Kesselring, E. P. L. van Nieuwenburg, and J. Eisert, Reinforcement learning decoders for fault-tolerant quantum computation, arXiv:1810.07207 (2018).

[14] P. Andreasson, J. Johansson, S. Liljestrand, and M. Granath, Quantum error correction for the toric code using deep reinforcement learning, Quantum **3**, 183 (2019).

[15] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, Rev. Mod. Phys. **91**, 045002 (2019).

[16] S. R. White, Density Matrix Formulation for Quantum Renormalization Groups, Phys. Rev. Lett. **69**, 2863 (1992).

[17] 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).

[18] M. Collura, L. Del'Anna, T. Felser, and S. Montangero, On the descriptive power of Neural-Networks as constrained Tensor Networks with exponentially large bond dimension, arXiv:1905.11351 (2019).

[19] N. N. Cencov, *Statistical Decision Rules and Optimal Inference*, Translations of Mathematical Monographs, Vol. 53 (American Mathematical Society, Providence, RI, 2000).

[20] S. Sorella, Green Function Monte Carlo with Stochastic Reconfiguration, Phys. Rev. Lett. **80**, 4558 (1998).

[21] S. Sorella, Generalized Lanczos algorithm for variational quantum Monte Carlo, Phys. Rev. B **64**, 024512 (2001).

[22] G. Mazzola, A. Zen, and S. Sorella, Finite-temperature electronic simulations without the Born-Oppenheimer constraint, J. Chem. Phys. **137**, 134112 (2012).

[23] S.-I. Amari, Natural gradient works efficiently in learning, Neural Comput. **10**, 251 (1998).

[24] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, Identifying and attacking the saddle point problem in high-dimensional nonconvex optimization, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Montral, Canada, 2014), pp. 2933–2941.

[25] J. J. Waterfall, F. P. Casey, R. N. Gutenkunst, K. S. Brown, C. R. Myers, P. W. Brouwer, V. Elser, and J. P. Sethna, Sloppy-Model Universality Class and the Vandermonde Matrix, Phys. Rev. Lett. **97**, 150601 (2006).

[26] B. B. Machta, R. Chachra, M. K. Transtrum, and J. P. Sethna, Parameter space compression underlies emergent theories and predictive models, Science **342**, 604 (2013).

[27] L. Sagun, L. Bottou, and Y. LeCun, Eigenvalues of the Hessian in deep learning: Singularity and beyond, arXiv:1611.07476 (2016).

[28] V. Papyan, The full spectrum of deepnet Hessians at scale: Dynamics with SGD training and sample size, arXiv:1811.07062 (2018).

[29] S. Hochreiter and J. Schmidhuber, Flat minima, Neural Comput. **9**, 1 (1997).

[30] The duration of the first phase appears to depend on the hyperparameters (learning rate, regularization), but not on the system size.

[31] R. Grosse and R. Salakhudinov, Scaling up natural gradient by sparsely factorizing the inverse fisher matrix, in *International Conference on Machine Learning*, Vol. 37 (JMLR: W&CP, Lille, France, 2015), pp. 2304–2313.

[32] C. Hamer, Finite-size scaling in the transverse Ising model on a square lattice, J. Phys. A: Math. Gen. **33**, 6683 (2000).

[33] S. Suzuki, J.-I. Inoue, and B. K. Chakrabarti, *Quantum Ising Phases and Transitions in Transverse Ising Models*, Lecture Notes in Physics, Vol. 862 (Springer, Berlin, 2012).

[34] R. J. Baxter, *Exactly Solved Models in Statistical Mechanics* (Academic Press, San Diego, 1982).

[35] D. Perez-Garcia, F. Verstraete, J. I. Cirac, and M. M. Wolf, PEPS as unique ground states of local Hamiltonians, Quantum Inf. Comp. **8**, 0650 (2008).

[36] X. Gao and L.-M. Duan, Efficient representation of quantum many-body states with deep neural networks, Nat. Commun. **8**, 662 (2017).

[37] U. Wolff, Collective Monte Carlo Updating for Spin Systems, Phys. Rev. Lett. **62**, 361 (1989).

[38] Even though this problem can be solved by applying a local basis transformation that makes the Hamiltonian stoquastic, we did not use such a technique as we want to see how RBM encodes a quantum state without post-manipulation of the problem.

[39] G. Hinton, Neural networks for machine learning lecture notes, http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf (2012).

[40] J. Martens, New insights and perspectives on the natural gradient method, arXiv:1412.1193 (2014).

[41] J. Martens, Deep learning via Hessian-free optimization, in *International Conference on Machine Learning* (Omnipress, Haifa, Israel, 2010), pp. 735–742.

[42] J. Kessler, F. Calcavecchia, and T. D. Kühne, Artificial neural networks as trial wave functions for quantum Monte Carlo, arXiv:1904.10251 (2019).

[43] L. Yang, Z. Leng, G. Yu, A. Patel, W.-J. Hu, and H. Pu, Deep learning-enhanced variational Monte Carlo method for quantum many-body physics, Phys. Rev. Res. **2**, 012039 (2020).

[44] https://github.com/chaeyeunpark/yannq.

[45] 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).

[46] D. Sehayek, A. Golubeva, M. S. Albergo, B. Kulchytskyy, G. Torlai, and R. G. Melko, Learnability scaling of quantum states: Restricted Boltzmann machines, Phys. Rev. B **100**, 195125 (2019).