Universal Scaling Laws in Quantum-Probabilistic Machine Learning by Tensor Network: Toward Interpreting Representation and Generalization Powers

Sheng-Chen Bai and Shi-Ju Ran*

Center for Quantum Physics and Intelligent Sciences, Department of Physics, Capital Normal University, Beijing 100048, China

(Received 15 October 2024; accepted manuscript online 19 November 2024)

The interpretation of representations and generalization powers has been a long-standing challenge in the fields of machine learning (ML) and artificial intelligence. This study contributes to understanding the emergence of universal scaling laws in quantum-probabilistic ML. We consider the generative tensor network (GTN) in the form of a matrix-product state as an example and show that with an untrained GTN (such as a random TN state), the negative logarithmic likelihood (NLL) L generally increases linearly with the number of features M, that is, $L \simeq kM + \text{const}$. This is a consequence of the so-called "catastrophe of orthogonality," which states that quantum many-body states tend to become exponentially orthogonal to each other as M increases. This study reveals that, while gaining information through training, the linear-scaling law is suppressed by a negative quadratic correction, leading to $L \simeq \beta M - \alpha M^2 + \text{const}$. The scaling coefficients exhibit logarithmic relationships with the number of training samples and quantum channels χ. The emergence of a quadratic correction term in the NLL for the testing (training) set can be regarded as evidence of the generalization (representation) power of the GTN. Over-parameterization can be identified by the deviation in the values of α between the training and testing sets while increasing χ. We further investigate how orthogonality in the quantum-feature map relates to the satisfaction of quantum-probabilistic interpretation and the representation and generalization powers of the GTN. Unveiling universal scaling laws in quantum-probabilistic ML would be a valuable step toward establishing a white-box ML scheme interpreted within the quantum-probabilistic framework.

DOI: 10.1088/0256-307X/41/12/120701

Introduction. Developing "white-box" (or interpretable) schemes of machine learning (ML) and artificial intelligence is key to ensuring robustness, stability, fairness, and privacy protection.[1–3] Characterizing and interpreting representation and generalization powers, that is, the ability of ML models to utilize both learned and unlearned samples, is a fundamental topic in academic research and real applications.

Tensor networks (TNs) are a promising candidate for developing "white-box" ML owing to their solid and systematic theoretical foundation in quantum information sciences and quantum mechanics (see Refs. [4–14] and a recent review[15]). TNs originate from quantum many-body simulations, and are regarded as interpretable and efficient numerical tools[16–21] with their representation power characterized by the scaling laws of entanglement entropy (see Refs. [22–25]). The representation power in a quantum many-body simulation can be understood as the scaling of the simulation accuracy of physical properties with parameter complexity. For example, the matrix product state (MPS)[21,26] provides an accurate representation of states that satisfy the one-dimensional (1D) area law of entanglement entropy, providing a solid foundation for the efficient simulation of 1D quantum systems with local interactions using the MPS.

In ML, representation and generalization powers can be characterized by the loss functions and prediction accuracies of the training and testing sets, respectively. Recent studies have determined the power laws in the scaling of the loss function in recursive neural networks[27] and transformer models,[28] which have been applied in different scenarios, such as multimodality,[29,30] computer-constrained environments,[31] data engineering,[32,33] and reinforcement learning.[34] Some pioneering studies have also revealed the scaling behaviors of TN-based ML,[10,14] preliminarily demonstrating the significance of this topic.

This study contributes to determining scaling laws in quantum-probabilistic ML, which provides interpretations of representation and generalization powers. Quantum-probabilistic ML models joint-probabilistic distributions using quantum many-body states that satisfy Born's quantum-probabilistic interpretation.[15] We consider the generative TN (GTN) in the MPS form as an example.[6] With a GTN that is irrelevant to the data (e.g., untrained or random), the negative logarithmic likelihood L universally obeys a linear scaling law with respect to the number of features M as follows:

$$L \simeq kM + b. \tag{1}$$

This suggests the exponentially-vanishing probabilities of the samples obtained from the GTN, which is a consequence of the COO of quantum many-body states. By

*Corresponding author. Email: sjran@cnu.edu.cn
© 2024 Chinese Physical Society and IOP Publishing Ltd

training the GTN, the acquired information suppresses the linear increase in NLL with a negative quadratic correction term as follows:

$$L \simeq \beta M - \alpha M^2 + \gamma. \tag{2}$$

See Fig. 1 for an illustration.

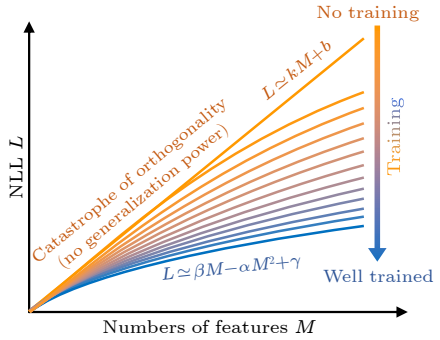

Fig. 1. Illustrative representation of how the gain of information alters the scaling law of NLL in quantum-probabilistic ML. By training the GTN, the linear scaling law with respect to the feature number M [Eq. (1)], which is the result of the "catastrophe of orthogonality (COO)" of quantum many-body states, is suppressed by the addition of a negative quadratic correction term [Eq. (2)].

The scaling coefficients in L exhibit logarithmic scaling behaviors against the virtual dimension χ of the GTN. The scaling laws of these coefficients are consistent with the logarithmic scaling of NLL versus χ, which explains how the powers of GTN improves as the number of quantum channels increases. The deviation in the scaling behaviors of the coefficients between the training and testing sets indicates over-parameterization. We further investigate how the orthogonality of the quantum-feature map (QFM) relates to the satisfaction of quantum-probabilistic interpretation and the scaling laws, as well as to the representation and generalization powers of GTN.

Generative Tensor Network for Quantum-Probabilistic ML. The central idea of quantum-probabilistic ML is to represent the joint-probabilistic distribution of M features [denoted as $P(\boldsymbol{x})$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_M)$] using a quantum many-body state (denoted as $|\Psi\rangle$) consisting of M spins (or qubits). The state is related to $P(\boldsymbol{x})$ via protective measurements, as follows:

$$P(\boldsymbol{x}) = |\langle \boldsymbol{x}|\Psi\rangle|^2, \tag{3}$$

where the product state $|\boldsymbol{x}\rangle \equiv |x_1, x_2, ..., x_M\rangle$ is obtained from the sample $\boldsymbol{x}$ by adopting the so-called QFM.

The definition of the QFM is flexible. Herein, we consider the following widely used definition:

$$|x_1, x_2, ..., x_M\rangle = \prod_{m=1}^{M} \left[\cos\frac{\theta\pi x_m}{2}|0\rangle + \sin\frac{\theta\pi x_m}{2}|1\rangle\right], \tag{4}$$

where θ is a hyperparameter that tunes the orthogonality, and $|0\rangle$ and $|1\rangle$ are the eigenstates of the Pauli matrix $\boldsymbol{\sigma}^z$.[4] With such a QFM, the number of features M in a sample equals the number of tensors in the MPS, which is analogous to the size of a quantum many-body system. The above formulation implies that with a well-trained $|\Psi\rangle$, the probability of any given sample $\boldsymbol{x}$ can be obtained by quantum measurements [Eq. (3)].

The complexity of $|\Psi\rangle$ scales exponentially with M. The use of TN to represent $|\Psi\rangle$ reduces the complexity from the exponential scale to polynomial scale. The TN state that correctly represents the joint-probability distribution is referred to as the GTN.

We consider MPS as the representation of GTN, which is written as

$$|\Psi\rangle = \sum_{s_1\ldots s_M}\sum_{a_1\ldots a_{M-1}} A^{[1]}_{s_1a_1}A^{[2]}_{s_2a_1a_2}\ldots A^{[M-1]}_{s_{M-1}a_{M-2}a_{M-1}}A^{[M]}_{s_Ma_{M-1}}|s_1\rangle\ldots|s_M\rangle, \tag{5}$$

with $s_m = 0, 1$; thus, $\dim(s_m) = d = 2$ (called the physical dimension). Indices $\{a_m\}$ are virtual indices. We assume that their dimensions are uniform, $\dim(a_m) = \chi$, which is denoted as a virtual dimension. The complexity of an MPS scales linearly with the number of spins M as follows:

$$\#|\Psi\rangle \sim O(Md\chi^2). \tag{6}$$

We train $|\Psi\rangle$ so that Eq. (3) can correctly provide the joint-probability distribution by updating the local tensors $\{\boldsymbol{A}^{[m]}\}$ of the MPS and minimizing the NLL:

$$L = -\frac{1}{N}\sum_n \ln P(\boldsymbol{x}^{(n)}), \tag{7}$$

where $\boldsymbol{x}^{(n)}$ denotes the nth sample in the provided training set. Taking the gradient-descent method as an example, the local tensors of the MPS are updated as follows:

$$\boldsymbol{A}^{[m]} \leftarrow \boldsymbol{A}^{[m]} - \eta\frac{\partial L}{\partial \boldsymbol{A}^{[m]}}, \tag{8}$$

where η is a small positive number known as the gradient step or learning rate.

The formulation above explicitly connects the quantum probabilities to the learning process, that is, the minimization of L. This potentially makes quantum-probabilistic properties (such as fidelity and entanglement) analogous to those of ML. Below, we start with the quantum probability itself, specifically the COO, to exhibit the scaling laws of L for the training and testing sets.

Catastrophe of Orthogonality. The COO of quantum many-body states indicates that the fidelity between two "different" quantum states generally decreases exponentially with the number of spins M as follows:

$$|\langle\Psi|\Phi\rangle| \sim k^M, \tag{9}$$

with $0 < k < 1$.

We now focus on the COO of the MPSs. We utilize the orthogonal form of the MPS and place the orthogonal center on the right end. Subsequently, all tensors except the last are isometries, satisfying the orthogonal conditions $\sum_{s_ma_{m-1}} A^{[m]*}_{s_ma_{m-1}a_m}A^{[m]}_{s_ma_{m-1}a'_m} = I_{a_ma'_m}$, where $\boldsymbol{I}$ denotes the identity. The rightmost tensor satisfies the normalization condition as $\sum_{s_Ma_{M-1}} A^{[M]*}_{s_Ma_{M-1}}A^{[M]}_{s_Ma_{M-1}} = 1$; thus, the entire MPS satisfies the normalization condition (say, $|\langle\Psi|\Phi\rangle| = 1$ for $|\Psi\rangle = e^{i\zeta}|\Phi\rangle$ with ζ as an

arbitrary phase factor). Note that any MPS can be transformed into an orthogonal form by gauge transformations (that is, the transformations do not change the tensor obtained by contracting all the virtual indices).

$|\Psi\rangle$ and $|\Phi\rangle$ denote two different MPSs (for example, their local tensors, denoted by $\{\boldsymbol{A}^{[m]}\}$ and $\{\boldsymbol{B}^{[m]}\}$, are randomly generated); thus, $|\langle\Psi|\Phi\rangle|$ generally satisfies Eq. (9). To demonstrate this, we may compute $|\langle\Psi|\Phi\rangle|$ by contracting the shared indices from left to right, starting with $v_{a_1 a_1'} = \sum_{s_1} A^{[1]*}_{s_1 a_1} B^{[1]}_{s_1 a_1'}$. As $\boldsymbol{A}^{[1]}$ and $\boldsymbol{B}^{[1]}$ are isometries, we obtain $|\boldsymbol{v}| \leq 1$, where the equals sign holds for $\boldsymbol{A}^{[1]} = \boldsymbol{B}^{[1]}$. We may normalize $\boldsymbol{v}$ as $\boldsymbol{v} \leftarrow \boldsymbol{v}/z_1$ with $z_1 = |\boldsymbol{v}| < 1$ (assuming $\boldsymbol{A}^{[1]} \neq \boldsymbol{B}^{[1]}$).

Subsequently, we contract $\boldsymbol{v}$ with $\boldsymbol{A}^{[m]}$ and $\boldsymbol{B}^{[m]}$ for $m = 2, 3, \ldots, M$. After each contraction, we generally have $|\boldsymbol{v}| < 1$ because $\boldsymbol{A}^{[m]}$ and $\boldsymbol{B}^{[m]}$ are different isometries. Then, we normalize it as $\boldsymbol{v} \leftarrow \boldsymbol{v}/z_m$ with $z_m = |\boldsymbol{v}| < 1$. Finally, we obtain $|\langle\Psi|\Phi\rangle| = \tilde{z}\prod_{m=1}^{M-1} z_m$ with $\tilde{z} = |\sum v_{a_{M-1}a'_{M-1}} A^{[M]*}_{s_M a_{M-1}} B^{[M]}_{s_M a'_{M-1}}|$, which is also generally less than 1. Clearly, $|\langle\Psi|\Phi\rangle| \to 0$ exponentially with M because it is the product of M numbers that are generally less than 1.

More rigorous deductions about the COO can be made if we impose certain restrictions on the MPSs, such as translational invariance ($\boldsymbol{A}^{[m]} = \boldsymbol{A}$ and $\boldsymbol{B}^{[m]} = \boldsymbol{B}$ for any m). The exponential decay of $|\langle\Psi|\Phi\rangle|$ can be immediately observed because the dominant eigenvalue of the transfer matrix ($M_{aa',bb'} = \sum_s A_{sab} B_{sa'b'}$) is typically less than 1. This corresponds to the "shrinkage" of the norm in the contractions discussed above. Considering that the GTNs are obtained by updating the local tensors with ML data, we will not seek further for more rigorous deductions, but will assume the satisfaction of the COO with the GTNs.

Universal-Scaling Laws of Negative Logarithmic Likelihood. From Eqs. (3) and (9), the probability of obtaining sample ($\boldsymbol{x}$) from an irrelevant GTN should be exponentially small. Substituting this into Eq. (7), we immediately obtain the linear-scaling law of the NLL, as given by Eq. (1). Figure 2(a) demonstrates such a scaling law on the Fashion-MNIST dataset [35] for the interclass NLL, where we consider that the categories of the samples differ from those of the (trained) GTNs. Using linear fitting, we obtain $k = 0.1837$, 0.1764, and 0.1758 for the training set ($k = 0.1835$, 0.1762, and 0.1756 for the testing set) for $\chi = 32$, 64, and 128, respectively. This indicates that the scaling law of the NLL does not change with varying χ.

Figure 2(b) shows that the intraclass NLL, in which the categories of the samples correspond to those of the GTNs, obeys the corrected scaling law given by Eq. (2). We obtain the quadratic coefficients $\alpha = 3.653 \times 10^{-5}$, 3.252×10^{-5}, and 3.041×10^{-5} for the training set ($\alpha = 3.663 \times 10^{-5}$, 3.582×10^{-5}, and 3.603×10^{-5} for the testing set) for $\chi = 32$, 64, and 128, respectively. As $M \sim O(10^2)$, we obtain $M^2 \sim O(10^4)$ or $O(10^5)$. Thus, the contribution of the quadratic term is approximately $O(1)$, which cannot be ignored, particularly for large M. For the linear term, we obtain $\beta = 7.365 \times 10^{-2}$, 6.582×10^{-2}, and 6.006×10^{-2} for the training set ($\beta = 7.617 \times 10^{-2}$, 7.277×10^{-2}, and 7.196×10^{-2} for the testing set). The magnitudes of the constant terms are approximately $O(1)$ or $O(10^{-1})$. This indicates that the contributions of all three terms cannot be ignored. The universality of the scaling laws is further demonstrated using various datasets.

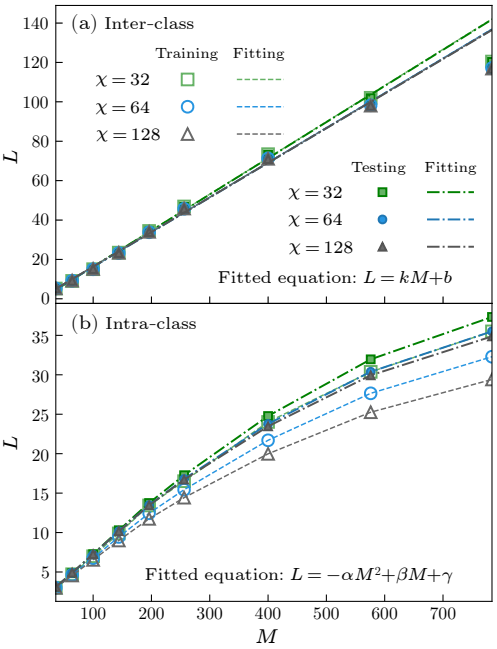

Fig. 2. (a) Linear scaling of the "interclass" NLL [Eq. (1)], where the samples and GTNs correspond to different categories while computing NLL. (b) The emergence of the negative quadratic correction in the scaling of the "intraclass" NLL [Eq. (2)], where the samples and GTNs correspond to the same category. Herein, we choose the Fashion-MNIST dataset. The data for $M < 784$ are obtained by cropping the middle section of the images.

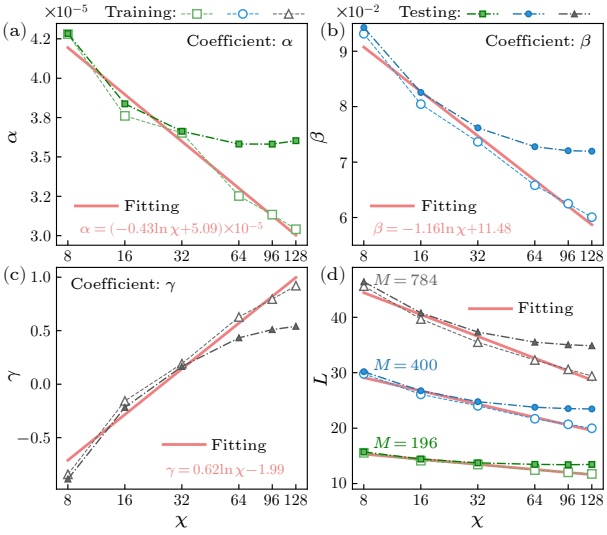

Fig. 3. (a)–(c) Coefficients in the corrected scaling laws [α, β, and γ in Eq. (2)] of the intraclass NLL and their logarithmic scaling against the virtual dimension χ. (d) Intraclass NLL L and its logarithmic scaling against χ. Deviations between the curves of the training and testing sets are observed, which indicate over-parameterization. We consider the Fashion-MNIST dataset and fix $M = 784$. The fittings (red solid lines) are given based on the training data.

We focus on the intraclass NLL, unless otherwise mentioned. In Fig. 3(a), we change the virtual dimension χ of the GTN by fixing $M = 784$ and show the values of α in Eq. (2) for the intraclass NLL. Our results suggest that the logarithmic-scaling behavior of α against χ is:

$$\alpha \simeq p_\alpha \ln \chi + q_\alpha. \tag{10}$$

Similar logarithmic behaviors are observed for the coefficients β and γ, as shown in Fig. 3(b) and 3(c).

Substituting the logarithmic relationships of α, β, and γ [refer to Eq. (10)] into the corrected scaling law, we obtain

$$L \simeq (-p_\alpha M^2 + p_\beta M + p_\gamma) \ln \chi + (-q_\alpha M^2 + q_\beta M + q_\gamma). \tag{11}$$

This finding suggests that the logarithmic scaling of L against χ is as follows:

$$L \simeq p_L \ln \chi + q_L, \tag{12}$$

which is verified in Fig. 3(d). This indicates a logarithmic improvement in the ML power versus the number of quantum channels, which is consistent with the results reported in Ref. [6]. We further discuss this issue below.

Another important observation shown in Fig. 3(a)–3(c) is the deviation between the training- and testing-set curves. For $\chi > 30$, the coefficients in the corrected scaling law for the training set maintain the logarithmic-scaling behavior given by Eq. (12). However, those for the testing set begin to deviate, leading to a higher NLL, as shown in Fig. 3(d). These results imply that by further increasing the parameter complexity of the GTN by approximately $\chi > 30$, the representation power improves by obeying the expected scaling law; however, the improvement in the generalization power decelerates.

Note that representation power refers to how well the GTN models the joint-probabilistic distribution of the training samples quantified by the NLL of the training set. This representation power is analogous to that used to approximate quantum many-body states (e.g., the ground state of a given quantum many-body Hamiltonian). Generalization power indicates how well the GTN models the joint-probabilistic distribution of unlearnt samples quantified by the NLL of the testing set. The deviation in growth between the representation and generalization powers can signal over-parameterization. Such a sign can be clearly observed with the deviation in the scaling of the coefficients, which is consistent with the observation of NLL shown in Fig. 3(d). Interestingly, the logarithmic-scaling law for L versus χ can be obtained from the scaling laws of L versus M, which originate from the quadratically corrected COO.

Orthogonality, Quantum-Probabilistic Interpretation, and Their Relations to Generalization Power. To gain a deeper understanding of the representation and generalization powers of the GTN, we investigate the orthogonality of the QFM [referring to the orthogonality of the states obtained by the QFM of Eq. (4)] and the satisfaction of the quantum-probabilistic interpretation.

Consider a simple case of binary images with $\chi = N$ (where N is the number of training samples) that can provide fundamental information on quantum-probabilistic ML with small N or sufficiently large χ. Taking $\theta = 1$ in Eq. (4), we obtain the orthogonality $\langle x|y\rangle = 0$ for any two distinct samples $x \neq y$. With a well-trained or constructed GTN, for example, $|\Psi\rangle = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} |\boldsymbol{x}^{(n)}\rangle$ (which can be written exactly as an MPS with $\chi = N$), we obtain $P(\boldsymbol{x}^{(n)}) = 1/N$ for any training sample.

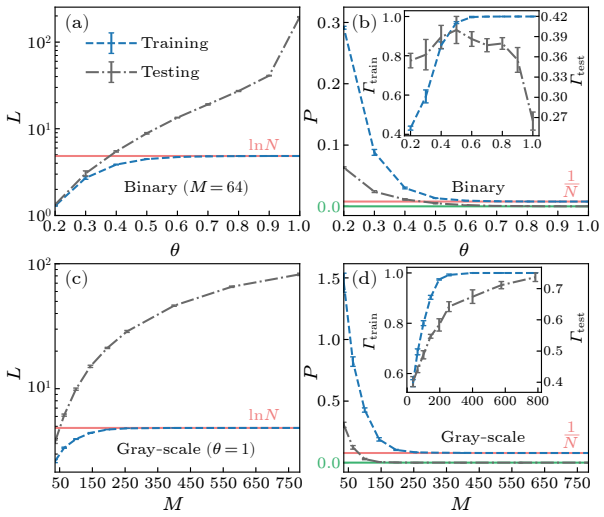

Fig. 4. (a) NLL L and (b) average probability per sample P for the training and testing sets of the binarized Fashion-MNIST versus θ in the QFM [Eq. (4)]. We fix $M = 64$. (c), (d) L and P versus M for the original (grayscale) dataset with $\theta = 1$. In the insets of (b) and (d), we show the classification accuracy of the GTN classifier against θ and M, respectively. The error bars indicate the standard deviations over five independent simulations. We take $N = \chi = 128$.

For any sample (for example, a testing sample denoted as $\boldsymbol{y}$) that does not belong to the training set, we obtain $P(\boldsymbol{y}) = 0$, owing to orthogonality. We can consider χ as the number of channels in the GTN. N orthogonal states must occupy N quantum channels. For $\chi = N$, all quantum channels in the GTN are taken by the training samples, which provide an ideal equal distribution of probabilities and a minimal NLL $L = \ln N$. No channel is left for the test samples, providing vanishing probabilities and divergent NLL. This is illustrated in Figs. 4(a) and 4(b) using the binarized Fashion-MNIST dataset.

In this binary case with $\theta = 1$, the normalization condition for the full Hilbert-space quantum probabilities is satisfied, that is, $\sum_{\boldsymbol{x}} P(\boldsymbol{x}) = 1$ (because $\sum_{\boldsymbol{x}} P(\boldsymbol{x}) = \langle\Psi|\Psi\rangle$) with the summation of all possible samples. Thus, the obtained NLL is always larger than the theoretical minimum $\ln N$. However, for $0 < \theta < 1$, the product states obtained using QFM from different samples are generally nonorthogonal to each other. Consequently, reducing θ causes an "impertinent" decrease in NLL, essentially owing to the violation of the normalization condition of quantum probabilities, that is, $\sum_{\boldsymbol{x}} P(\boldsymbol{x}) > 1$. Consequently, the NLL can become lower than its theoretical minimum $L < \ln N$ [see Fig. 4(a)].

The NLL for the testing set also decreases with θ, which is also "impertinent," but may be a positive signal on the generalization power (we will show this later). Figure 4(b) shows an unusually large average probability P for both the training and testing sets, while reducing $\theta \to 0$.

We now adopt the GTN classification (GTNC) scheme[8] to demonstrate the representation and generalization powers from another perspective. The idea is to use G GTNs (where G is the number of categories) to parameterize the classification boundaries of the samples in the Hilbert space of quantum states (the GTNs). Given a sample $\boldsymbol{y}$, the GTNC predicts its classification as follows:

$$g^{\text{pred}}(\boldsymbol{y}) = \text{argmax}_g |\langle \boldsymbol{y} | \Psi^{(g)} \rangle|, \tag{13}$$

where $|\Psi^{(g)}\rangle$ is the GTN trained using the training samples in the gth category.

The inset of Fig. 4(b) shows the classification accuracy against θ for the training and testing sets of the binarized Fashion-MNIST dataset. For $\theta = 1$, the theoretical training accuracy is 100%, whereas the testing accuracy is low. This is a typical example of over-parameterization, which can be the result of insufficient training data or redundant parameters (note that, here, $N = \chi$).

By reducing θ, the normalization condition of the full Hilbert-space quantum probabilities is violated, and the training accuracy decreases with θ. However, nonvanishing probabilities appear for the testing samples, which improve the testing accuracy (see the peak of the testing accuracy around $\theta = 0.5$). These results imply a balance between the violation of the quantum-normalization condition and generalization power of the quantum states for ML tuned by the orthogonality of the QFM.

This implication is meaningful for processing nonbinary data, which are encountered more frequently than binary data. For example, each feature (pixel) in an 8-bit grayscale image can take $D = 256$ distinct values. If orthogonality is rigorously required such that $\langle x|y\rangle = 0$ for $x \neq y$, each pixel should be mapped to a spin with D levels, which is infeasible and impractical. Our results provide an interpretation of why the QFM with two-level ($d = 2$) spins works well on grayscale images, although this clearly leads to a violation of the quantum-normalization condition.

In Figs. 4(c) and 4(d), we show the NLL L and average probability P against M for the original (grayscale) Fashion-MNIST dataset. For $\theta = 1$, we still do not have orthogonality; in general, $|\langle x|y\rangle| > 0$ for $x \neq y$. This explains why the NLL of the training set is lower than the theoretical minimum $\ln N$. However, $L \to \ln N$ and $P \to 1/N$ for a large M, suggesting the restoration of orthogonality (and the normalization condition) owing to the COO. In the inset of Fig. 4(d), we show that the testing accuracy increases with M, implying that for $D \ll d$, enhancing the orthogonality generally improves the generalization power. [In the supplemental material, additional results are provided to demonstrate the universality of the scaling laws. The scaling against the number of training samples and training epoch is investigated. By designing a multi-spin quantum feature map, we further show the influence to the representation and generalization powers by tuning orthogonality.]

Summary. The main contribution of this study is the proposal of universal scaling laws in quantum-probabilistic ML, where joint-probabilistic distributions are encoded as GTNs in the form of MPSs. By introducing a negative quadratic correction, we demonstrate that the gain of representation or generalization power suppresses the linear-scaling law of negative logarithmic likelihood, which is a result of the COO of quantum many-body states. The relationships among the scaling laws, orthogonality of the QFM, and normalization conditions of quantum probabilities are discussed, providing a systematic understanding of the representation and generalization powers of quantum-probabilistic ML.

Scaling laws play an important role in investigating quantum many-body physics, such as the complexity of approximating quantum many-body states and quantum criticality (see Refs. [22–25]). Our work on the scaling laws of the GTN is a step toward revealing potential connections between quantum many-body computation and quantum-inspired ML, two distinct fields that consider quantum-probabilistic interpretation as an essential foundation. We expect further connections to be uncovered in the future, advancing quantum-inspired ML as a novel data-driven approach for exploring quantum many-body physics.

Acknowledgements. SCB is grateful to Yi-Cheng Tang, Rui Hong, Peng-Fei Zhou, and Ying Lu for helpful discussions. This work was supported in part by the Beijing Natural Science Foundation (Grant No. 1232025), the Ministry of Education Key Laboratory of Quantum Physics and Photonic Quantum Information (Grant No. ZYGX2024K020), and Academy for Multidisciplinary Studies, Capital Normal University.

References

[1] Gilpin L H, Bau D, Yuan B Z, Bajwa A, Specter M, and Kagal L 2018 *IEEE 5th International Conference on data science and advanced analytics (DSAA)* (IEEE) p. 80
[2] Zhang Q S and Zhu S C 2018 *Front. Inf. Technol. & Electron. Eng.* **19** 27
[3] Carvalho D V, Pereira E M, and Cardoso J S 2019 *Electronics* **8** 832
[4] Stoudenmire E and Schwab D J 2016 *Supervised learning with tensor networks, in Advances in Neural Information Processing Systems* **29** (Curran Associates, Inc.)
[5] Liu D, Ran S J, Wittek P, Peng C, García R B, Su G, and Lewenstein M 2019 *New Journal of Physics* **21** 073059
[6] Han Z Y, Wang J, Fan H, Wang L, and Zhang P 2018 *Phys. Rev. X* **8** 031012
[7] Cheng S, Wang L, Xiang T, and Zhang P 2019 *Phys. Rev. B* **99** 155131
[8] Sun Z Z, Peng C, Liu D, Ran S J, and Su G 2020 *Phys. Rev. B* **101** 075135
[9] Bai S C, Tang Y C, and Ran S J 2022 *Chin. Phys. Lett.* **39** 100701
[10] Liu J, Li S, Zhang J, and Zhang P 2023 *Phys. Rev. E* **107** L012103

[11] Meng Y M, Zhang J, Zhang P, Gao C, and Ran S J 2023 *SciPost Phys.* **14** 142

[12] Chen H and Barthel T 2024 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46** 7825

[13] Pozas-Kerstjens A, Hernández-Santana S, Pareja Monturiol J R, Castrillón López M, Scarpa G, González-Guillén C E, and Pérez-García D 2024 *Quantum* **8** 1425

[14] Convy I, Huggins W, Liao H R, and Whaley K B 2022 *Mach. Learn.: Sci. Technol.* **3** 015017

[15] Ran S J and Su G 2023 *Intelligent Computing* **2** 0061

[16] Verstraete F, Murg V, and Cirac J I 2008 *Adv. Phys.* **57** 143

[17] Cirac J I and Verstraete F 2009 *J. Phys. A: Math. Theor.* **42** 504004

[18] Orús R 2014 *Annals of Physics* **349** 117

[19] Orús R 2019 *Nat. Rev. Phys.* **1** 538

[20] Ran S J, Tirrito E, Peng C, Chen X, Tagliacozzo L, Su G, and Lewenstein M 2020 *Tensor Network Contractions: Methods and Applications to Quantum Many-Body Systems* (Springer, Cham)

[21] Cirac J I, Pérez-García D, Schuch N, and Verstraete F 2021 *Rev. Mod. Phys.* **93** 045003

[22] Vidal G, Latorre J I, Rico E, and Kitaev A 2003 *Phys. Rev. Lett.* **90** 227902

[23] Pollmann F, Mukerjee S, Turner A M, and Moore J E 2009 *Phys. Rev. Lett.* **102** 255701

[24] Schuch N, Wolf M M, Verstraete F, and Cirac J I 2008 *Phys. Rev. Lett.* **100** 030504

[25] Tagliacozzo L, de Oliveira T R, Iblisdir S, and Latorre J I 2008 *Phys. Rev. B* **78** 024410

[26] Pérez-García D, Verstraete F, Wolf M M, and Cirac J I 2007 *Quantum Inf. Comput.* **7** 401

[27] Hestness J, Narang S, Ardalani N, Diamos G, Jun H, Kianinejad H, Patwary M M A, Yang Y, and Zhou Y H 2017 arXiv:1712.00409 [cs.LG]

[28] Kaplan J, McCandlish S, Henighan T, Brown T B, Chess B, Child R, Gray S, Radford A, Wu J, and Amodei D 2020 arXiv:2001.08361 [cs.LG]

[29] Henighan T, Kaplan J, Katz M, Chen M, Hesse C, Jackson J, Jun H, Brown T B, Dhariwal P, Gray S, Hallacy C, Mann B, Radford A, Ramesh A, Ryder N, Ziegler D M, Schulman J, Amodei D, and McCandlish S 2020 arXiv:2010.14701 [cs.LG]

[30] Zhai X, Kolesnikov A, Houlsby N, and Beyer L 2022 *Scaling vision transformers, in Proceedings of the IEEECVF Conference on Computer Vision and Pattern Recognition (CVPR)* p. 1204

[31] Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, Rutherford E, de Las Casas D, Hendricks L A, Welbl J, Clark A, Hennigan T, Noland E, Millican K, van den Driessche G, Damoc B, Guy A, Osindero S, Simonyan K, Elsen E, Rae J W, Vinyals O, and Sifre L 2022 arXiv:2203.15556 [cs.CL]

[32] Sorscher B, Geirhos R, Shekhar S, Ganguli S, and Morcos A 2022 *Advances in Neural Information Processing Systems* **35** (Curran Associates, Inc.) p. 19523

[33] Muennighoff N, Rush A, Barak B, Le Scao T, Tazi N, Piktus A, Pyysalo S, Wolf T, and Raffel C A 2023 *Advances in Neural Information Processing Systems* **36** (Curran Associates, Inc.) p. 50358

[34] Gao L, Schulman J, and Hilton J 2023 *Proceedings of the 40th International Conference on Machine Learning* **202** (PMLR) p. 10835

[35] Xiao H, Rasul K, and Vollgraf R 2017 arXiv:1708.07747