Information Scrambling in Quantum Neural Networks

Huitao Shen,[1] Pengfei Zhang,[2,3,4] Yi-Zhuang You,[5] and Hui Zhai[2,*]
[1]Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[2]Institute for Advanced Study, Tsinghua University, Beijing 100084, China
[3]Institute for Quantum Information and Matter, California Institute of Technology, Pasadena, California 91125, USA
[4]Walter Burke Institute for Theoretical Physics, California Institute of Technology, Pasadena, California 91125, USA
[5]Department of Physics, University of California, San Diego, California 92093, USA

(Received 3 November 2019; revised manuscript received 19 April 2020; accepted 1 May 2020; published 21 May 2020)

The quantum neural network is one of the promising applications for near-term noisy intermediate-scale quantum computers. A quantum neural network distills the information from the input wave function into the output qubits. In this Letter, we show that this process can also be viewed from the opposite direction: the quantum information in the output qubits is scrambled into the input. This observation motivates us to use the tripartite information—a quantity recently developed to characterize information scrambling—to diagnose the training dynamics of quantum neural networks. We empirically find strong correlation between the dynamical behavior of the tripartite information and the loss function in the training process, from which we identify that the training process has two stages for randomly initialized networks. In the early stage, the network performance improves rapidly and the tripartite information increases linearly with a universal slope, meaning that the neural network becomes less scrambled than the random unitary. In the latter stage, the network performance improves slowly while the tripartite information decreases. We present evidences that the network constructs local correlations in the early stage and learns large-scale structures in the latter stage. We believe this two-stage training dynamics is universal and is applicable to a wide range of problems. Our work builds bridges between two research subjects of quantum neural networks and information scrambling, which opens up a new perspective to understand quantum neural networks.

DOI: 10.1103/PhysRevLett.124.200504

The neural network (NN) lies at the heart of the recent blossom of deep learning [1]. The NN distills information from the input, usually represented by a high-dimensional vector, and encodes it into a lower-dimensional output vector. Recently, quantum generalizations of NNs have been proposed and actively studied [2–16]. In a quantum NN, both the input and the output are quantum wave functions. The classical mapping is replaced by a quantum channel composed of unitary evolutions and measurements [17]. The quantum NN is considered as one of the promising applications for near-term noisy intermediate-scale quantum devices [18]. Moreover, it has been suggested that the quantum NN has more expressive power than its classical counterpart [14].

Similar to a classical NN, quantum information in the input wave function is distilled and encoded into the output in a quantum NN. This process is illustrated by the forward arrow in Fig. 1(a). Intriguingly, for a quantum NN, this process can also be viewed from the opposite direction. By deferring measurements until the end of the quantum channel [19], the information encoded in output qubits just before the measurement is spread into the entire system by unitary transformations, as illustrated by the backward arrow in Fig. 1(a). Such processes that the information of a small subsystem is scrambled to a larger region are known as the information scrambling. The subject of information scrambling is well studied in contexts such as thermalization, chaos and information dynamics in quantum many-body systems, and even black-hole physics [20–27].

Quantum NNs and quantum information scrambling so far are two separated research topics. The purpose of this Letter is to bridge the gap and make their connection: In a quantum NN, information encoding and the information scrambling are the same process viewed from opposite directions.

There have been information-theoretic studies of classical NNs [28–31]. However, in classical NNs, the mapping at every layer is usually not invertible and the information is generally not preserved. Due to the information loss during the process, the mutual information always decreases with the network depth. In contrast, the unitarity of quantum evolutions preserves the information perfectly. The mutual information between the input and the output of any unitary transformation is always maximal. In order to have nontrivial diagnosis in quantum NNs, the key is to consider the mutual information between subsystems of the input and the output. This naturally leads to

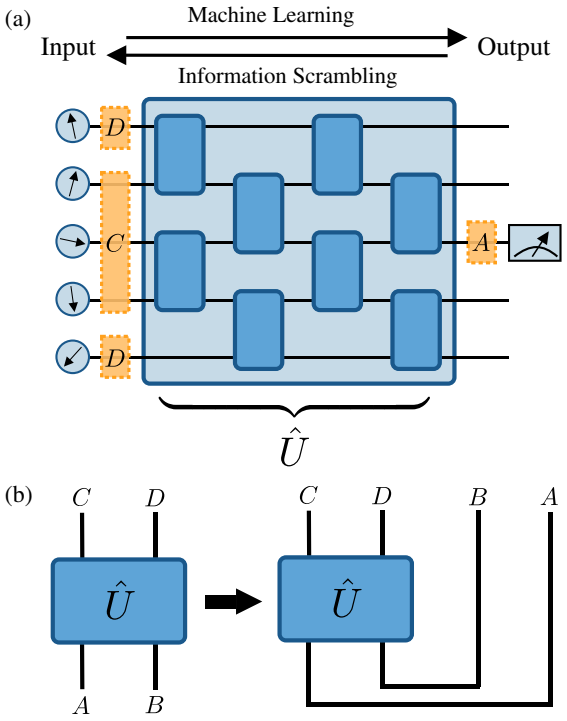

FIG. 1. (a) Schematic of a quantum circuit with brick-wall geometry. Here, the network has $n = 5$ qubits and depth $l = 4$. All two-qubit gates form a giant unitary transformation $\hat{U}$ that distills the information from the input qubits and encodes it into one output qubit. The inverse process is that the information of one output qubit is scrambled into input qubits by $\hat{U}^\dagger$. A is the output subsystem, and C and D are input subsystems in the definition of the tripartite information. (b) Illustration for the operator-state mapping in the definition of tripartite information. Each leg may represent multiple qubits.

the tripartite information—a quantity that characterizes the information scrambling [32,33].

Here, we study the training dynamics of quantum NNs using the tripartite information. We simultaneously monitor both the network performance and the tripartite information during training and observe empirical relations between them. Based on the behavior of these two quantities, the training process can be decomposed into two stages, which we call the "local construction stage" and the "global relaxation stage." In the following, we present a detailed analysis of the training dynamics and provide evidence to support our claim.

Tripartite information of quantum neural networks.—Consider a unitary operator $\hat{U}$ in the n-qubit Hilbert space

$$\hat{U} = \sum_{i,j=1}^{2^n} U_{ij}|i\rangle\langle j|,$$

where $\{|i\rangle, i = 1, \ldots, 2^n\}$ denotes a complete set of bases in the Hilbert space. It can be regarded as a tensor with n input and n output legs. As illustrated in Fig. 1(b), we divide the output legs (qubits) to two non-overlapping subsystems A and B and similarly divide the input legs (qubits) to C and D.

The operator can be mapped to a state in the $2n$-qubit Hilbert space as

$$|U\rangle = \sum_{i,j=1}^{2^n} U_{ij}/\sqrt{2^n}|i\rangle|j\rangle.$$

Since $|U\rangle$ is a pure state, the entanglement entropy of its subsystem is well defined, e.g., $S(A) \equiv -\mathrm{tr}(\rho_A \log_2 \rho_A)$ with $\rho_A \equiv \mathrm{tr}_{B,C,D}(|U\rangle\langle U|)$ being the reduced density matrix of subsystem A. The mutual information between the output subsystem A and the input subsystem C is $I(A, C) \equiv S(A) + S(C) - S(A \cup C)$. A similar definition can be made for $I(A, D)$ and $I(A, C \cup D)$. The tripartite information of the unitary $\hat{U}$ is defined as [32,33]

$$I_3(A, C, D) \equiv I(A, C) + I(A, D) - I(A, C \cup D). \quad (1)$$

Because $C \cup D$ are all input qubits, it can be proved that $I(A, C \cup D) = 2|A|$, where $|A|$ is the number of qubits in subsystem A. Therefore, it is crucial to consider the mutual information between subsystems of both input and output qubits.

The strong subadditivity of the entanglement entropy leads to $I_3(A, C, D) \leq 0$ for a unitary gate. The absolute value of the tripartite information $I_3(A, C, D)$ measures how much information of subsystem A is shared by C and D simultaneously after the unitary transformation, and thus quantifies how scrambled a unitary is. For example, for an identity unitary transformation $U_{ij} = \delta_{ij}$, if A is entirely contained in C or D, it is straightforward to show that $I_3(A, C, D) = 0$. As an opposite limit, consider a state $|U\rangle$ with U_{ij} sampled from the uniform Haar random ensemble. The reduced density matrix of a small subsystem $A \cup C$ satisfying $|A|+|C| < n/2$ is approximately identity. This leads to $I(A, C) = 0$. Similarly, $I(A, D) = 0$. Therefore $I_3(A, C, D) = -2|A|$, which is the minimal value for I_3 [33].

Having introduced the tripartite information for a general unitary transformation, we now turn to tripartite information of a quantum NN. Here, we only consider parameterized quantum circuits with brick-wall geometry. As shown in Fig. 1(a), each brick represents an independent two-qubit unitary gate in the SU(4) group, and it is parameterized using its 15 Euler angles [34]. During training, these parameters are optimized with classical optimization algorithms. All these two-qubit gates form a quantum circuit represented by a giant unitary transformation $\hat{U}$.

The datasets to be studied in this work have several important features. First, the input wave functions all have time reversal symmetry, and consequently can be represented as real vectors. Therefore, we restrict two-qubit gates to SO(4) with six Euler angles each. Second, the

output target is either a real number within $[-1, 1]$ or a binary label within $\{0, 1\}$: only one readout qubit is needed at the end of the quantum circuit. For simplicity, we always let n be odd and fix the readout qubit to be the qubit at the center, i.e., the $[(n+1)/2]$th qubit.

To define tripartite information, we always fix the output subsystem A to be the central readout qubit. To respect the symmetry that A is located at the center, we always choose C to be the central $|C|$ input qubits in the circuit and D to be the remaining input qubits. Note that, under this definition, D in general contains two disconnected regions. The tripartite information $I_3(A, C, D)$ characterizes how much information of the output qubit is scrambled on the input side between the central region C and the outer region D.

Magnetization learning.—The first task is to learn the average magnetization of a many-body wave function of n half-spins in a supervised manner. The dataset consists of N input-target pairs $\{(|G^\alpha\rangle, M_z^\alpha), \alpha = 1, \ldots, N\}$, where the input wave function $|G^\alpha\rangle$ is the ground state wave function of the parent Hamiltonian with random long-ranged spin-spin interactions:

$$\hat{H} = \sum_{i,j=1}^{n} (J_{ij}\sigma_i^z\sigma_j^z + K_{ij}\sigma_i^x\sigma_j^x) + \sum_{i=1}^{n} (g_i\sigma_i^x + h\sigma_i^z), \quad (2)$$

where σ_i^μ represents the μth Pauli matrix on the ith qubit; $\mu = x, y, z$; and $i = 1, \ldots, n$. J_{ij}, K_{ij}, g_i, and h are all random numbers. The target is the average magnetization computed as $M_z^\alpha \equiv \langle G^\alpha|\hat{M}_z|G^\alpha\rangle$, where the magnetization operator is

$$\hat{M}_z \equiv \sum_{i=1}^{n} \sigma_i^z/n.$$

In sampling the random Hamiltonian, we ensure $J_{ij} \leq 0$ such that the ground state wave functions are either "ferromagnetic" or "paramagnetic" measured under $\hat{M}_z$. h is a small pinning field randomly drawn from a distribution with zero mean, which is used to trigger the spontaneous Z_2 symmetry breaking in the ferromagnetic phase.

The quantum NN takes the input wave function $|G^\alpha\rangle$ and applies the unitary transformation $\hat{U}$ on it. The magnetization is read out by measuring σ^x of the central qubit. We choose to measure σ^x instead of σ^z because the quantum NN may learn some shortcut that is unable to generalize if the measurement and the target physical observable are under the same basis. This is essentially a regression task, and the loss function to be minimized is the absolute error of the magnetization:

$$\mathcal{L} = \frac{1}{N}\sum_{\alpha=1}^{N} |\langle G^\alpha|\hat{U}^\dagger\sigma_{(n+1)/2}^x\hat{U}|G^\alpha\rangle - M_z^\alpha|. \quad (3)$$

We simulate the above hybrid quantum-classical quantum NN training algorithm. The distributions of random parameters in the Hamiltonian [Eq. (2)] are chosen such that M_z^α in the dataset roughly distributes uniformly within $[-1, 1]$. All two-qubit unitaries in the quantum NN are initialized randomly. The parameters are optimized with the gradient descent algorithm in Ref. [35]. The gradients can be computed directly, thanks to the linearity of the quantum channel, and are measurable in a realistic quantum NN [7,9,36].

Two-stage training.—In Fig. 2(a), we show the training loss and the tripartite information, both averaged over different initializations, as functions of the training epoch. Averaging over different initializations reduces the volatility within a single training instance and makes the correlation between the two quantities clearer. At the early stage of the training, the rapid improvement of the quantum NN performance, characterized by a fast decrease of the training loss, is accompanied by an almost linear increase of the tripartite information. In other words, the quantum NN becomes less scrambled compared with the initial random unitary. This training stage terminates when the tripartite information reaches its local maximum. In the

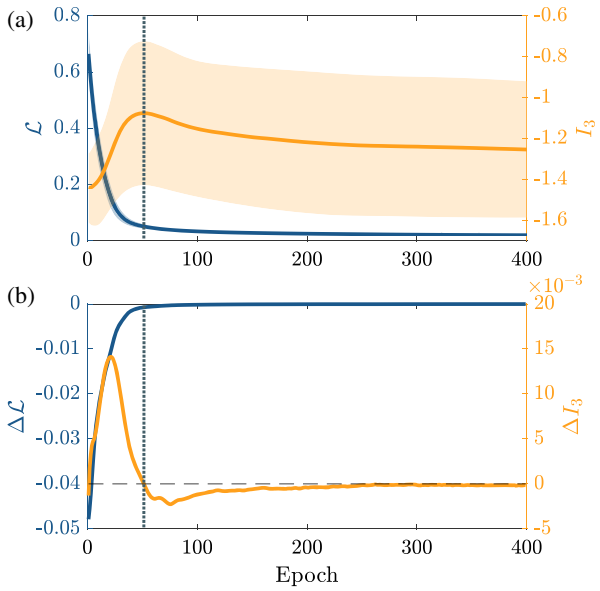

FIG. 2. (a) Training loss $\mathcal{L}$ and tripartite information $I_3(A, C, D)$ as functions of the training epoch. The shaded area represents one standard deviation. (b) Finite difference of training loss $\Delta\mathcal{L}$ and tripartite information ΔI_3 as functions of the training epoch. The dotted vertical line indicates the boundary between two training stages, which is determined as the maximum of the averaged I_3 given by $\Delta I_3 = 0$. All results are averaged over 20 different random initializations. The network has $n = 9$ qubits and depth $l = 6$. The training and validation dataset contains $N = 2500$ and 500 wave function-magnetization pairs, respectively, sampled from random Hamiltonian ensemble, where random parameters are distributed uniformly within $J_{ij}/J \in [-1, 0]$, $K_{ij}/J \in [-1, 1]$, $g_i/J \in [-6, 6]$, and $h/J \in [-0.04, 0.04]$. J is the energy unit. The learning rate is $\lambda = 10^{-2}$. Here and in the rest of the Letter, the input subsystem size $|C| = 5$.

next stage, the tripartite information decreases again, meaning that the network scrambles information faster. The network performance also improves but with a much slower rate compared with that in the first stage. In Fig. 2(b), we plot the finite difference of the two metrics $\Delta\mathcal{L}$ and ΔI_3 together and use a dashed line to indicate the maximum of I_3 given by $\Delta I_3 = 0$. One can see clearly that $\Delta\mathcal{L}$ also drops to negligible small values around the dashed line, meaning a much slower decreasing rate of $\mathcal{L}$ in the later stage.

We call the training stage before I_3 reaching the maximum the local construction stage, and the latter stage where I_3 decreases is the global relaxation stage. The reason for the names will be clear after we study the training dynamics in detail below. The empirical observation that quantum NN performance and the information scrambling are closely correlated is the main finding of this work. This correlation has been observed in all our numerical simulations with different network initializations, training algorithms, system sizes, and network depths [38]. We also train quantum NNs for learning the staggered magnetization from the ground state of random antiferromagnetic and even frustrated Hamiltonians, as well as the winding number of a product quantum state. Despite the very different natures of these tasks, the empirical correlation between the NN performance and the tripartite information still holds. All details were presented in [36].

Local construction stage.—We claim that, during the first stage when the tripartite information linearly increases, the quantum NN learns local features of the input wave function. For the magnetization learning task, because of the existence of ferromagnetic domain, there is some probability that any single spin is aligned relatively well with remaining spins in the system. Simply outputting any single-spin magnetization of the input wave function is actually a reasonable guess so that the training loss can decrease rapidly. For such networks where only local features are extracted, information does not need to be scrambled into the whole system. Therefore, the tripartite information increases during this stage.

To support the above claim, we compute two-point correlations between input qubits and the readout qubit:

$$C_2(i) \equiv \frac{1}{N}\sum_{\alpha=1}^{N}\langle G^\alpha|\sigma_i^z \hat{U}^\dagger \sigma^x_{(n+1)/2}\hat{U}|G^\alpha\rangle. \tag{4}$$

If one views $\hat{U}$ as a time evolution operator, then $C_2(i)$ is simply a two-point function between two different places and two different times. In Fig. 3(a), we plot C_2 as a function of different input qubits and training epochs in the early training stage. As can be seen, they increase rapidly and then saturate to large values. The increasing correlation indicates that the quantum NN is establishing the correspondence between local input features and the output qubit. During this stage, the tripartite information also increases, and the two-point correlation function saturates when the tripartite information reaches the maximum. All these observations are consistent with our claim that, during the first local construction stage, local features are extracted from the input.

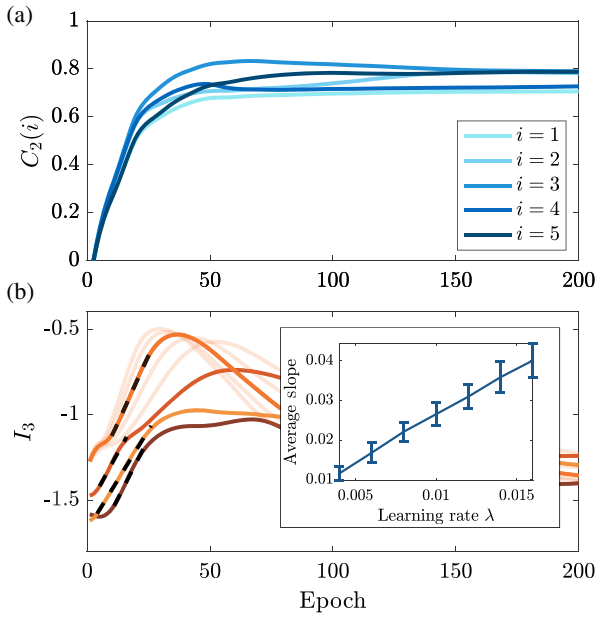

FIG. 3. (a) Two-point correlation function $C_2(i)$ as a function of the training epoch and the input qubit i for a typical initialization. (b) Tripartite information $I_3(A, C, D)$ as a function of the training epoch for different initializations and learning rates. All solid lines are trained under learning rate $\lambda = 10^{-2}$. The transparent orange lines are trained with the same initialization as the solid orange line but with learning rates $\lambda = 6, 8, 12$, and 14×10^{-3}. The average slope for the four initializations shown here is plotted in the inset as a function of the learning rates. The error bars represent the standard deviations of fitted slopes for the fixed learning rate but different initializations.

Before concluding this section, we point out another interesting observation that the linear increasing slope of the tripartite information is nearly a constant that is independent of the initialization, shown in Fig. 3(b). Of course, this slope depends on the learning rate of the gradient descent algorithm. As shown in the inset, the I_3-independent slope scales linearly with the learning rate.

Global relaxation stage.—We now turn to the second stage where the tripartite information decreases and the training loss decreases with a much slower rate. We claim that, during this stage, the quantum NN learns global features of the wave function. To provide evidence for this claim, we test the quantum NN in an artificial test dataset $\{(|\psi_D^\alpha\rangle, M_z^\alpha), \alpha = 1, \ldots, N_D\}$, constructed according to the following process. First, we sample ground states $|G^\alpha\rangle$ from the random Hamiltonian of Eq. (2). Next, we apply the following unitary transformation to flip a region of spins:

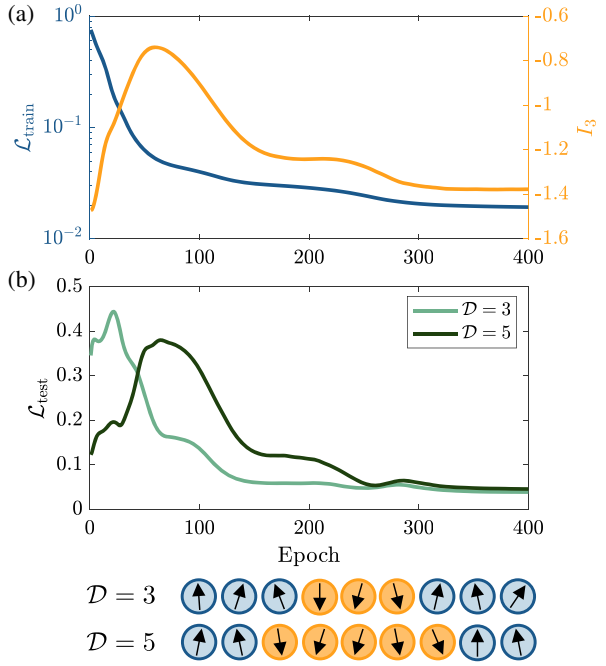

FIG. 4. (a) Training loss and tripartite information as functions of the training epoch for a typical initialization. (b) Loss functions on the artificial test dataset with "ferromagnetic domain" of size $\mathcal{D} = 3$ and 5 for the same training instance as Fig. 4(a).

$$|\psi^{\alpha}_{\mathcal{D}}\rangle = \prod_{\frac{n-\mathcal{D}+1}{2} \le i \le \frac{n+\mathcal{D}}{2}} \sigma^{x}_{i} |G^{\alpha}\rangle. \tag{5}$$

For paramagnetic wave functions $|G^{\alpha}\rangle$, this transformation leaves these wave functions paramagnetic. However, for ferromagnetic wave functions $|G^{\alpha}\rangle$, the transformation creates a ferromagnetic domain wall of size $\mathcal{D}$, as sketched in Fig. 4. In order to accurately compute the magnetization of such wave functions, the quantum NN must be able to learn structures larger than the domain wall size $\mathcal{D}$. In [36], we present an argument on why, in this task, long string operators should exist in $\hat{U}^{\dagger}\sigma^{x}_{(n+1)/2}\hat{U}$ when it is expanded under the basis of the product of local Pauli matrices.

In Fig. 4(b), we show losses on test datasets with $\mathcal{D} = 3$ and 5 as functions of the training epoch. In the later stage of training, although the training loss decreases slowly, the tripartite information can decrease rather drastically, accompanied by a rapid decreasing of losses on both test datasets. Moreover, the larger the domain wall size is, the later the test loss begins to decrease. This means that the information scrambling is associated with the performance improvement on wave functions with large domain structures. This naturally explains why the unitary has to become more scrambled. Since such data are rare in the training dataset, it also explains why the training loss improvement is slow. Finally, we note that, in Fig. 2, the standard deviation of I_3 is quite large in the later stage. This is consistent with the chaotic nature of the information scrambling because it is now known that the quantum many-body chaos and the information scrambling are two closely related concepts.

Discussion and outlook.—In summary, we apply a metric of quantum information scrambling—the tripartite information—to diagnose the training process of quantum NNs. We find strong correlation between this metric and the loss function, and we identify a two-stage training dynamics of quantum NN. We show that the quantum NN establishes local correlations in the early stage and builds up global structures in the later stage. Such two-stage dynamics is reminiscent of physical processes such as annealing of ferromagnetism and the operator growth in many-body quantum chaos. We believe this two-stage dynamics is universal for a wide range of quantum machine learning problems. We also believe that the profound connection between the information scrambling and the quantum NN could find broader applications in quantum machine learning, such as revealing the underlying mechanism of quantum machine learning and guiding the quantum NN architecture design.

We thank Yingfei Gu for discussions and an anonymous referee for the suggestion to average results from different initializations. H. S. thanks IASTU for hosting his visit to Beijing, where key parts of this work were done. H. S. thanks Guangyu Du for suggestions on the data presentation. P. Z. acknowledges support from the Walter Burke Institute for Theoretical Physics at Caltech. This work is supported by Beijing Outstanding Young Scientist Program (H. Z.), MOST under Grant No. 2016YFA0301600 (H. Z.), and NSFC Grant No. 11734010 (H. Z.).

*hzhai@tsinghua.edu.cn

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT, Cambridge, MA, 2016).
[2] M. Benedetti, J. Realpe-Gómez, R. Biswas, and A. Perdomo-Ortiz, Quantum-Assisted Learning of Hardware-Embedded Probabilistic Graphical Models, Phys. Rev. X **7**, 041052 (2017).
[3] E. Torrontegui and J. J. García-Ripoll, Unitary quantum perceptron as efficient universal approximator, Europhys. Lett. **125**, 30004 (2019).
[4] M. Benedetti, D. Garcia-Pintos, O. Perdomo, V. Leyton-Ortega, Y. Nam, and A. Perdomo-Ortiz, A generative modeling approach for benchmarking and training shallow quantum circuits, npj Quantum Inf. **5**, 45 (2019).
[5] E. Farhi and H. Neven, Classification with Quantum Neural Networks on Near Term Processors, arXiv:1802.06002.
[6] J. R. McClean, S. Boixo, V. N Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. **9**, 4812 (2018).
[7] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, Phys. Rev. A **98**, 032309 (2018).

[8] W. Huggins, P. Patil, B. Mitchell, K. B. Whaley, and E. M. Stoudenmire, Towards quantum machine learning with tensor networks, Quantum Sci. Technol. **4**, 024001 (2019).
[9] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, Circuit-centric quantum classifiers, Phys. Rev. A **101**, 032308 (2020).
[10] E. Grant, M. Benedetti, S. Cao, A. Hallam, J. Lockhart, V. Stojevic, A. G Green, and S. Severini, Hierarchical quantum classifiers, npj Quantum Inf. **4**, 65 (2018).
[11] J.-G. Liu and L. Wang, Differentiable learning of quantum circuit Born machines, Phys. Rev. A **98**, 062324 (2018).
[12] G. Verdon, J. Pye, and M. Broughton, A Universal Training Algorithm for Quantum Deep Learning, arXiv:1806.09729.
[13] J. Zeng, Y. Wu, J.-G. Liu, L. Wang, and J. Hu, Learning and inference on generative adversarial quantum circuits, Phys. Rev. A **99**, 052306 (2019).
[14] Y. Du, M.-H. Hsieh, T. Liu, and D. Tao, The expressive power of parameterized quantum circuits, arXiv:1810.11922.
[15] K. Beer, D. Bondarenko, T. Farrelly, T. J Osborne, R. Salzmann, and R. Wolf, Efficient learning for deep quantum neural networks, Nat. Commun. **11**, 808 (2020).
[16] M. J. S. Beach, R. G. Melko, T. Grover, and T. H. Hsieh, Making trotters sprint: A variational imaginary time ansatz for quantum many-body systems, Phys. Rev. B **100**, 094434 (2019).
[17] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Quantum machine learning, Nature (London) **549**, 195 (2017).
[18] J. Preskill, Quantum Computing in the NISQ era and beyond, Quantum **2**, 79 (2018).
[19] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* 10th Anniversary ed. (Cambridge University Press, Cambridge, England, 2010).
[20] E. Altman, Many-body localization and quantum thermalization, Nat. Phys. **14**, 979 (2018).
[21] X.-L. Qi, Does gravity come from quantum information?, Nat. Phys. **14**, 984 (2018).
[22] B. Swingle, Unscrambling the physics of out-of-time-order correlators, Nat. Phys. **14**, 988 (2018).
[23] A. I. Larkin and Yu N. Ovchinnikov, Quasiclassical method in the theory of superconductivity, Sov. Phys. JETP **28**, 1200 (1969), http://www.jetp.ac.ru/cgi-bin/e/index/e/28/6/p1200?a=list.
[24] A. Kitaev, Hidden correlations in the hawking radiation and thermal noise, in *Proceedings of the Fundamental Physics Prize Symposium, 2014*, https://www.youtube.com/watch?v=OQ9qN8j7EZI.
[25] S. H. Shenker and D. Stanford, Black holes and the butterfly effect, J. High Energy Phys. 03 (2014) 67.
[26] J. Maldacena, S. H. Shenker, and D. Stanford, A bound on chaos, J. High Energy Phys. 08 (2016) 106.
[27] R. Fan, P. Zhang, H. Shen, and H. Zhai, Out-of-time-order correlation for many-body localization, Sci. Bull. **62**, 707 (2017).
[28] R. Shwartz-Ziv and N. Tishby, Opening the black box of deep neural networks via information, arXiv:1703.00810.
[29] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. Daniel Tracey, and D. D. Cox, On the information bottleneck theory of deep learning, J. Stat. Mech. 124020 (2018).
[30] Z. Goldfeld, E. Van Den Berg, K. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy, Estimating information flow in deep neural networks, in *Proceedings of the 36th International Conference on Machine Learning*, edited by K. Chaudhuri and R. Salakhutdinov [*Proc. Mach. Learn. Res.* 97, 2299 (2019)].
[31] H. Shen, Mutual information scaling and expressive power of sequence models, arXiv:1905.04271.
[32] A. Kitaev and J. Preskill, Topological Entanglement Entropy, Phys. Rev. Lett. **96**, 110404 (2006).
[33] P. Hosur, X.-L. Qi, D. A. Roberts, and B. Yoshida, Chaos in quantum channels, J. High Energy Phys. 02 (2016) 004.
[34] P. Dita, Factorization of unitary matrices, J. Phys. A **36**, 2781 (2003).
[35] S. J. Reddi, S. Kale, and S. Kumar, On the Convergence of Adam and Beyond, in *International Conference on Learning Representations* (2018), https://openreview.net/forum?id=ryQu7f-RZ.
[36] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.124.200504 for further results of magnetization learning, staggered magnetization learning, and winding number learning, along with details of gradient calculation and measurement, which include Ref. [37].
[37] P. Zhang, H. Shen, and H. Zhai, Machine Learning Topological Invariants with Neural Networks, Phys. Rev. Lett. **120**, 066401 (2018).
[38] For network initializations, we require initial unitaries to be scrambled enough such that initial $I_3(A, C, D) \lesssim -1$ (-1 is about half of the negativemost value). For training algorithms, we require these algorithms to be gradient based. For network depths, we require the networks to not be too shallow.