





Dynamical transition in controllable quantum neural networks with large depth

Received: 19 December 2023

Accepted: 22 October 2024

Published online: 29 October 2024

 Check for updatesBingzhi Zhang ^{1,2,7}, Junyu Liu ^{3,4,5,6,7}, Xiao-Chuan Wu⁵, Liang Jiang ³ & Quntao Zhuang ^{1,2} ✉

Understanding the training dynamics of quantum neural networks is a fundamental task in quantum information science with wide impact in physics, chemistry and machine learning. In this work, we show that the late-time training dynamics of quantum neural networks with a quadratic loss function can be described by the generalized Lotka-Volterra equations, leading to a transcritical bifurcation transition in the dynamics. When the targeted value of loss function crosses the minimum achievable value from above to below, the dynamics evolve from a frozen-kernel dynamics to a frozen-error dynamics, showing a duality between the quantum neural tangent kernel and the total error. In both regions, the convergence towards the fixed point is exponential, while at the critical point becomes polynomial. We provide a non-perturbative analytical theory to explain the transition via a restricted Haar ensemble at late time, when the output state approaches the steady state. Via mapping the Hessian to an effective Hamiltonian, we also identify a linearly vanishing gap at the transition point. Compared with the linear loss function, we show that a quadratic loss function within the frozen-error dynamics enables a speedup in the training convergence. The theory findings are verified experimentally on IBM quantum devices.

As a paradigm of near-term quantum computing, variational quantum algorithms^{1–6} have been widely applied to chemistry^{1,7}, optimization^{2,8}, quantum simulation^{9,10}, condensed matter physics¹¹, communication^{12,13}, sensing^{14,15} and machine learning^{16–23}. Adopting layers of gates and stochastic gradient descent, they are regarded as ‘quantum neural networks’ (QNNs), analog to classical neural networks that are crucial to machine learning. Concepts and methods related to variational quantum algorithms are also beneficial for quantum error correction and quantum control^{24,25}, bridging near-term applications with the fault-tolerant era.

Despite the progress in applications, theoretical understanding of the training dynamics of QNN is limited, hindering the optimal design of quantum architectures and the theoretical study of quantum advantage

in such applications. Previous works adopt tools from quantum information scrambling for empirical study of QNN training^{26,27}. Recently, the Quantum Neural Tangent Kernel (QNTK) theory presents a potential theoretical framework for an analytical understanding of variational quantum algorithms, at least within certain limits^{28–32}, revealing deep connections to their classical machine learning counterparts^{33–43}. However, the theory of QNTK relies on the assumption of sufficiently random quantum circuit set-ups known as unitary k -designs^{44–47} that is only true at random initialization, preventing the theory from describing the more important late-time training dynamics. Similar limitations also exist for other theoretical works^{4,48–51}.

In this work, we go beyond QNTK theory and identify a dynamical transition in the training of QNNs with a quadratic loss function, when

¹Department of Physics and Astronomy, University of Southern California, Los Angeles, CA, USA. ²Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA. ³Pritzker School of Molecular Engineering, The University of Chicago, Chicago, IL, USA. ⁴Department of Computer Science, The University of Chicago, Chicago, IL, USA. ⁵Kadanoff Center for Theoretical Physics, The University of Chicago, Chicago, IL, USA. ⁶Department of Computer Science, The University of Pittsburgh, Pittsburgh, PA, USA. ⁷These authors contributed equally: Bingzhi Zhang, Junyu Liu.

✉ e-mail: qzhuang@usc.edu

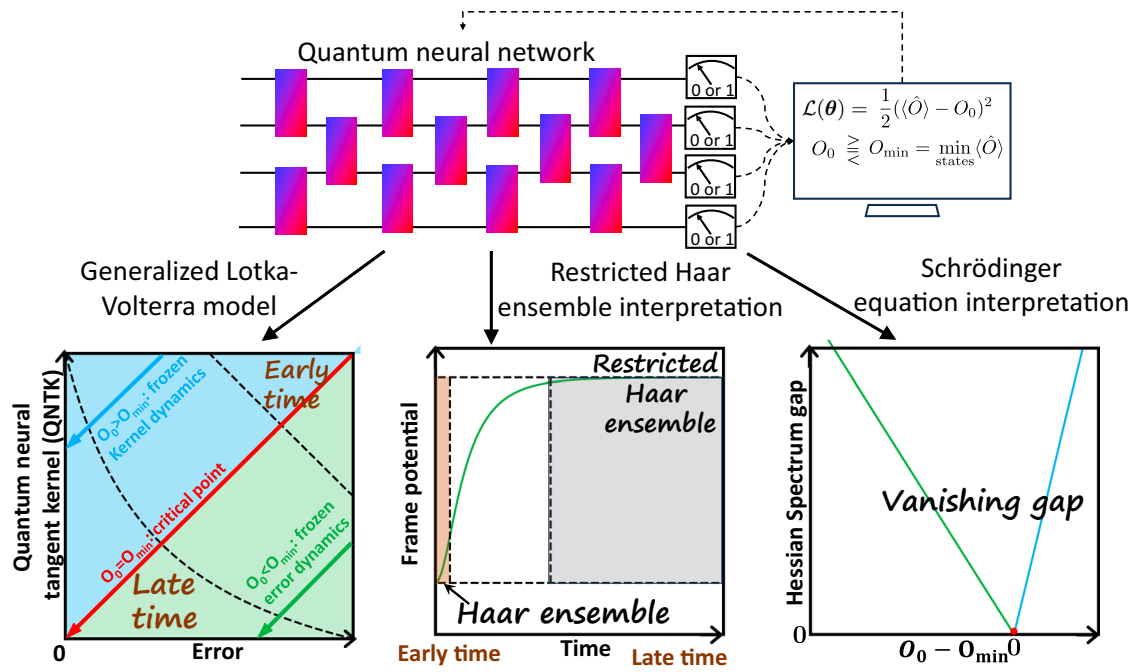


Fig. 1 | Illustration of setup and main results of this work. We study the training dynamics of quantum neural networks with loss function $\mathcal{L}(\theta) = (\langle \hat{O} \rangle - O_0)^2 / 2$, and identify a dynamical transition. We derive a first-principle generalized Lotka-

Volterra model to characterize it, and also provide interpretations from random unitary ensemble and Schrödinger equation.

the target loss function value O_0 cross the minimum achievable value (ground state energy O_{\min}). We show that the training dynamics of deep QNNs is governed by the generalized Lotka-Volterra (LV) equations describing a competitive duality between the quantum neural tangent kernel and the total error. The LV equations can be analytically solved and the dynamics is determined by the value of a conserved quantity. When the target value crosses O_{\min} , the conserved quantity changes sign and induces a transcritical bifurcation transition. As depicted in Fig. 1, in the frozen-kernel dynamics where $O_0 > O_{\min}$ is in the bulk of spectrum, the kernel is approaching a constant while the error decays exponentially with training steps; At the critical point when $O_0 = O_{\min}$ exactly, both the kernel and the error decay polynomially; In the frozen-error dynamics when $O_0 < O_{\min}$ is unachievable, the output from QNN still converges to the ground state leaving the error approaching a constant $O_{\min} - O_0$, while the kernel experiences an exponential decay. We provide a non-perturbative analytical theory to explain the dynamical transition via a restricted Haar ensemble at late time, when the QNN output state approaches the steady state. We also identify a vanishing Hessian gap at the transition point, which corresponds to Hamiltonian gap closing in the imaginary-time Schrödinger equation interpretation. While our theory analyses assume the large-depth limit, the dynamical transition is also numerically identified in QNNs with limited depths. Compared to the exponential decay of linear loss function with a non-tunable exponent, we identify convergence speed-up via tuning the quadratic loss function to be within the frozen-error dynamics. The theory findings are experimentally verified on IBM quantum devices. Our results imply that designing the loss function properly is important to achieve fast convergence.

Results

We begin by first introducing the model of the QNN and the necessary quantities. Then, we uncover the dynamical transition phenomena as a bifurcation transition in LV model. The unitary ensemble theory is then developed to support assumptions in obtaining the LV model. Afterwards, we characterize the transition with tools from statistical

physics. After finishing the theory, we provide numerical extensions and discuss the potential training speed-up brought by our results. Finally, we confirm the results in experiments.

Training dynamics of quantum neural networks

A D -depth QNN is composed of D layers of parameterized quantum circuits, realizing a unitary transform $\hat{U}(\theta)$ on n qubits, with L variational parameters $\theta = (\theta_1, \dots, \theta_L)$. The gate configuration of each layer varies between different circuit ansatz (see Methods for examples). When inputting a trivial state $|0\rangle^{\otimes n}$, the final output state of the neural network $|\psi(\theta)\rangle = \hat{U}(\theta)|0\rangle^{\otimes n}$, from which one can measure a Hermitian observable \hat{O} leading to expectation value $\langle \hat{O} \rangle = \langle \psi(\theta) | \hat{O} | \psi(\theta) \rangle$. To optimize the expectation of an observable \hat{O} towards the target value O_0 , a general choice of loss function is in a quadratic form,

$$\mathcal{L}(\theta) = \frac{1}{2} (\langle \hat{O} \rangle - O_0)^2 \equiv \frac{1}{2} \epsilon(\theta)^2, \quad (1)$$

where the total error $\epsilon(\theta) = \langle \hat{O} \rangle - O_0$. Suppose observable \hat{O} has possible values in the range of $[O_{\min}, O_{\max}]$. Without further specification, O_{\min} and O_{\max} refer to the minimum and maximum eigenvalue of \hat{O} . Now due to symmetry of maximum and minimum in optimization problems, we assume $O_0 < O_{\max}$ is true.

A QNN goes through training to minimize the loss function. In each training step, every variational parameter is updated by the gradient descent

$$\delta\theta_\ell(t) \equiv \theta_\ell(t+1) - \theta_\ell(t) = -\eta \frac{\partial \mathcal{L}(\theta)}{\partial \theta_\ell} = -\eta \epsilon(\theta) \frac{\partial \epsilon(\theta)}{\partial \theta_\ell}, \quad (2)$$

where η is the fixed learning rate and t is the discrete number of time steps in the training. With the update of parameters θ , quantities depending on θ also acquire new values in each training step. For simplicity of notion, we denote their dependence on t explicitly omitting θ , e.g. $\epsilon(t) \equiv \epsilon(\theta(t))$. To study the convergence, we separate the

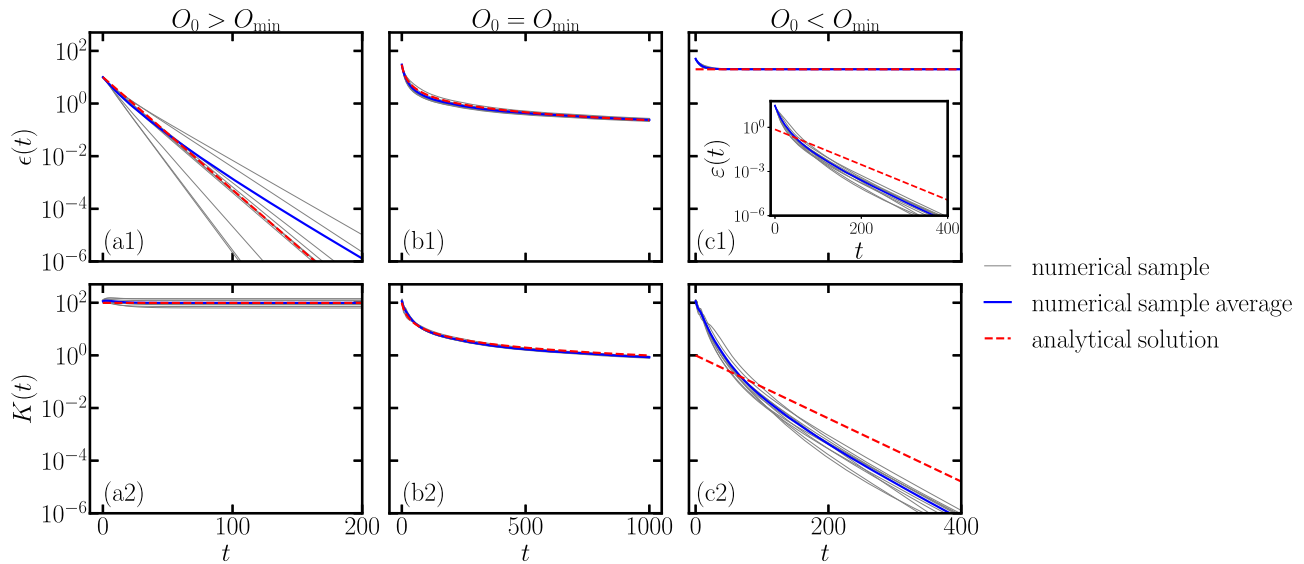


Fig. 2 | Dynamics in QNN in the example of XXZ model. The top and bottom panels show the dynamics of total error $\epsilon(t)$ and QNTK $K(t)$ with respect to the three cases $O_0 \approx O_{\min}$. Blue solid curves represent numerical ensemble average result. Red dashed curves in panels represent theoretical predictions on the dynamics of

total error in Eqs. (14)–(16) (from left to right). Grey solid lines show the dynamics for each random sample. The inset in (c1) shows the exponential decay of residual error $\epsilon(t)$. Here random Pauli ansatz (RPA) consists of $L = 768$ variational parameters ($D = L$ for RPA) on $n = 8$ qubits, and the parameter in XXZ model is $J = 2$.

error into two parts, $\epsilon(t) \equiv \epsilon(t) + R$ consists of a constant remaining term $R = \lim_{t \rightarrow \infty} \epsilon(t)$ and a vanishing residual error $\epsilon(t)$. When $\eta \ll 1$ is small, the total error is updated as

$$\delta\epsilon(t) \simeq \sum_{\ell} \frac{\partial\epsilon(\boldsymbol{\theta})}{\partial\theta_{\ell}} \delta\theta_{\ell} + \frac{1}{2} \sum_{\ell_1, \ell_2} \frac{\partial^2\epsilon(\boldsymbol{\theta})}{\partial\theta_{\ell_1} \partial\theta_{\ell_2}} \delta\theta_{\ell_1} \delta\theta_{\ell_2} \quad (3)$$

$$= -\eta\epsilon(t)K(t) + \frac{1}{2}\eta^2\epsilon(t)^2\mu(t), \quad (4)$$

where the QNTK K and dQNTK μ are defined as²⁹

$$K(t) \equiv \sum_{\ell} \left. \left(\frac{\partial\epsilon(\boldsymbol{\theta})}{\partial\theta_{\ell}} \right)^2 \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}, \quad (5)$$

$$\mu(t) \equiv \sum_{\ell_1, \ell_2} \left. \frac{\partial^2\epsilon(\boldsymbol{\theta})}{\partial\theta_{\ell_1} \partial\theta_{\ell_2}} \frac{\partial\epsilon(\boldsymbol{\theta})}{\partial\theta_{\ell_1}} \frac{\partial\epsilon(\boldsymbol{\theta})}{\partial\theta_{\ell_2}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}. \quad (6)$$

In the dynamics of $\epsilon(t)$, as $\eta \ll 1$, we focus on the first order of η in Eq. (4) as

$$\delta\epsilon(t) = -\eta\epsilon(t)K(t) + \mathcal{O}(\eta^2). \quad (7)$$

To characterize the dynamics of $\epsilon(t)$, it is necessary and sufficient to understand the dynamics of QNTK $K(t)$. Towards this end, we derive a first-order difference equation for QNTK $K(t)$ as (see details in Supplementary Note 1)

$$\delta K(t) = -2\eta\epsilon(t)\mu(t) + \mathcal{O}(\eta^2). \quad (8)$$

Combining Eq. (7) and Eq. (8), we aim to develop the dynamical model in training QNNs.

Dynamical transition

Our major finding is that when the circuit is deep and controllable, the QNN dynamics exhibit a dynamical transition at O_{\min} (and O_{\max}

similarly) as we depict in Fig. 2, where a QNN with random Pauli ansatz (RPA) is utilized to optimize the XXZ model Hamiltonian (see Methods for details of the circuit and observable).

Frozen-kernel dynamics: When $O_0 > O_{\min}$, the total error decays exponentially and the energy converges towards O_0 , as shown in Fig. 2a1. This is triggered by the frozen QNTK as shown in Fig. 2a2. Each individual random sample (gray) has slightly different value of frozen QNTK due to initialization, while all possess the exponential convergence. Our theory prediction (red dashed) agrees with the actual average (blue solid) for both the ensemble averaged QNTK \bar{K} and the error, while deviations due to early time dynamics can be seen (see Methods for details).

Critical point: When targeting right at the GS energy $O_0 = O_{\min}$, both the total error and QNTK decay as $1/t$, independent of system dimension d . As shown in Fig. 2b2, the QNTK ensemble average (blue solid) agrees very well with the theory prediction shown as the red dashed line. Due to initial time discrepancy in QNTK that is beyond our late time theory, the actual error dynamics has a constant deviation from the theory prediction (red dashed), however still has the $1/t$ late time scaling, as shown in Fig. 2b1.

Frozen-error dynamics: When targeting below the GS energy $O_0 < O_{\min}$, the total error converges to a constant $R = O_{\min} - O_0 > 0$ exponentially, as shown in Fig. 2c1. The inset shows the exponential convergence via the residual error $\epsilon = \epsilon - R$. In this case, the QNTK also decays exponentially with the training steps, as shown in Fig. 2c2. Deviation between the theory (red dashed) and numerical results (blue solid) can be seen due to early time dynamics beyond our theory.

Generalized Lotka-Volterra model: bifurcation

In this section, we reveal the nature of the transition as a transcritical bifurcation of an effective nonlinear dynamical equation. With large depth $D \gg 1$ and full control, QNNs are commonly modeled as a random unitary^{4,29,51}. However, at late time, the convergence of QNN training imposes constraints on the QNN unitary. As we will detail in ‘Unitary ensemble theory’ section, assuming that the late-time QNN is typical among random ensemble of unitaries under the convergence constraint, we can show that the relative dQNTK—the ratio of dQNTK and QNTK

$$\lambda(t) = \mu(t)/K(t) \quad (9)$$

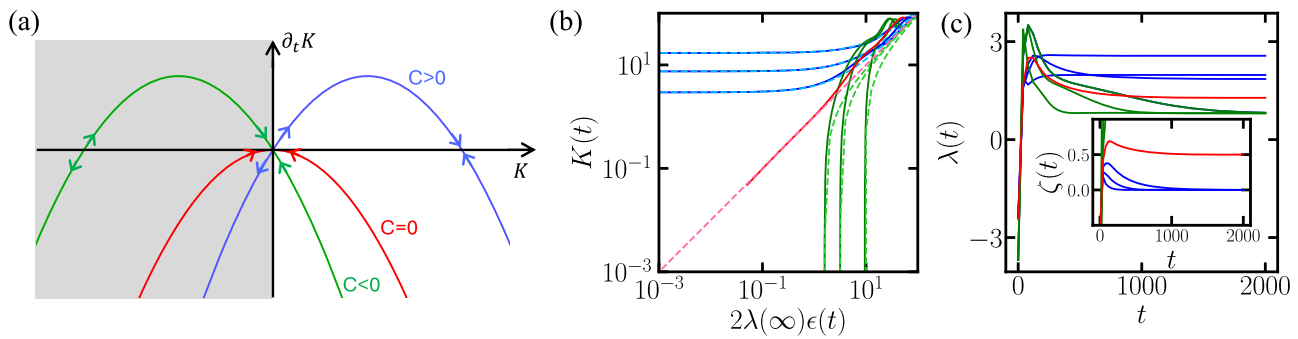


Fig. 3 | Classical dynamics interpretation of total error and QNTK dynamics. **a** The RHS of Eq. (12) shows a bifurcation. The gray region is nonphysical as $K \geq 0$. In the physical region ($K \geq 0$), we have a single stable fixed point $K = C$ when $C > 0$, corresponding to the frozen-kernel dynamics (blue in **b**); and a single stable fixed point $K = 0$ when $C \leq 0$, corresponding to the frozen error dynamics and critical point (green and red in **b**) separately. **b** Trajectories of $(2\lambda(\infty)\epsilon(t), K(t))$ in dynamics of QNN

with different $O_0 \geq O_{\min}$, plotted in solid blue, red and green. Dashed curves show the trajectory from Eq. (11). The logarithmic scale is taken to focus on the late-time comparison. **c** The dynamics of corresponding $\lambda(t) = \mu(t)/K(t)$. The inset shows the dynamics of $\zeta(t) = \epsilon(t)\mu(t)/K(t)^2$. The observable is XXZ model with $J = 2$, and QNN is a $n = 6$ -qubit RPA with $L = 192$ parameters (for RPA $D = L$). The legend in **b** is also shared with (c) and its inset.

converges towards a constant dependent on the number of parameters L and Hilbert space dimension of the system $d = 2^n$. Under the assumption that $\lambda(t) = \lambda$ being a constant and taking the continuous limit, Eqs. (7) and (8) lead to a coupled set of equations,

$$\begin{cases} \partial_t \epsilon(t) = -\eta \epsilon(t) K(t) \\ \partial_t K(t) = -2\eta \lambda \epsilon(t) K(t) \end{cases} \quad (10)$$

to the leading order in η . This is the generalized Lotka-Volterra equation developed in modeling nonlinear population dynamics of species competing for some common resource⁵². The two ‘species’ represented by K and ϵ are in direct competition as the interaction terms are negative. As Eqs. (10) have zero intrinsic birth/death rate, there is no stable attractor where all species $K(t)$ and $\epsilon(t)$ are positive, as sketched in Fig. 1 bottom left, where $2\lambda\epsilon$ and K are the x and y axis. From Eqs. (10), we can identify the conserved quantity at late time

$$C = K(t) - 2\lambda\epsilon(t) = \text{const.} \quad (11)$$

Each trajectory of $(2\lambda\epsilon(t), K(t))$ governed by Eqs. (10) is thus a straight line quantified by the conserved quantity C . We verify the trajectory from the conserved quantity of Eq. (11) in Fig. 3(b), where good agreement between QNN dynamics (solid) and generalized LV dynamics (dashed) can be identified. The conservation law in Eq. (11) indicates that a classical Hamiltonian description of LV dynamics is possible, via mapping the scaled error and kernel to the canonical position and momentum (see Methods). Therefore, the position-momentum duality in Hamiltonian formulation implies an error-kernel duality between K and ϵ .

Thanks to the conserved quantity C , we can reduce the coupled differential equations of the LV model in Eq. (10) to a single variable differential equation with the kernel or the error alone, e.g.,

$$\partial_t K(t) = \eta(C - K(t))K(t). \quad (12)$$

This is a canonical example of a transcritical bifurcation, with two fixed points $K = C$ and $K = 0$ ⁵³. To see this, we plot the RHS of Eq. (12) in Fig. 3a. When $C > 0$ (blue curve), via the sign of $\partial_t K$, we can see that only $K = C$ (therefore $\epsilon = 0$) is stable, corresponding to the frozen-kernel dynamics. On the other hand, when $C < 0$ (green curve), $K = 0$ (therefore $2\lambda\epsilon = -C > 0$) is the only stable fixed point, corresponding to the frozen-error dynamics. Specifically, for $C = 0$ (red curve), the two candidates collide and $K = 0$ (therefore $\epsilon = 0$) becomes the bifurcation point. As the fixed points collide and their stability exchange through

the bifurcation point $(K, C) = (0, 0)$, the transition is identified as the transcritical bifurcation.

Overall, we see that the two dynamics (and the critical point) of the QNN dynamics has a one-to-one correspondence to the two families of fixed points (and their common fixed point) of the generalized LV equation. The conserved quantity $C = K(t) - 2\lambda\epsilon(t) = (K(t)^2 - 2\epsilon(t)\mu(t))/K(t)$. Since $K(t) > 0$ at any finite time, the sign of constant is determined by the dynamical index defined as

$$\zeta = \epsilon(t)\mu(t)/K(t)^2. \quad (13)$$

If $\zeta \leq 1/2$, we have $C \geq 0$, determining the bifurcation dynamics.

Indeed, the analytical closed-form solution (see Methods) to the LV dynamics of Eqs. (10) supports the following theorem at the $t \gg 1$ late time limit.

Theorem 1. Assuming relative dQNTK $\lambda = \mu(t)/K(t)$ being a constant at late time, the QNN dynamics is governed by the generalized Lotka-Volterra equation in Eq. (10) and possesses a bifurcation to two different branches of dynamics, depending on the value of a conserved quantity $C = K(t) - 2\lambda\epsilon(t) = (1 - 2\zeta)K(t)$ or equivalently the dynamical index $\zeta = \epsilon(t)\mu(t)/K(t)^2$.

1. When $\zeta < 1/2$ thus $C > 0$, we have the ‘frozen-kernel dynamics’ (c.f.²⁹), where the QNTK $K(t) = C$ is frozen and

$$\epsilon(t) \propto e^{-\eta C t}. \quad (14)$$

2. When $\zeta = 1/2$ thus $C = 0$, we have the ‘critical point’, where both the QNTK and total error decay polynomially,

$$K(t) = 2\lambda\epsilon(t) = 1/(\eta t + c), \quad (15)$$

with c being a constant.

3. When $\zeta > 1/2$ thus $C < 0$, we have the ‘frozen-error dynamics’, where the total error $\epsilon(t) = R$ is frozen and both the kernel and the residual error decay exponentially

$$K(t) = 2\lambda\epsilon(t) \propto e^{-2\eta\lambda R t}. \quad (16)$$

The bifurcation can be connected to $O_0 \geq O_{\min}$ intuitively. When $O_0 < O_{\min}$, it is clear that $R > 0$ and we expect dynamical index $\zeta > 1/2$ and $C < 0$ so that it is the ‘frozen-error dynamics’. When $O_0 > O_{\min}$, we know the total error will decay to zero eventually, and therefore we can correspond this branch to the ‘frozen-kernel dynamics’, where

dynamical index $\zeta < 1/2$ and $C > 0$. The case $O_0 = O_{\min}$ is therefore the critical point. In Fig. 3(c) inset, we indeed see the dynamical index $\zeta \rightarrow 0, 1/2, +\infty$ when $O_0 \gtrless O_{\min}$. In our later theory analyses, we will make this connection rigorous between $O_0 \gtrless O_{\min}$, the dynamical index $\zeta \lesseqgtr 1/2$ and the bifurcation transition.

Unitary ensemble theory

In this section, we provide analytical results to resolve two missing pieces of the LV model—the assumption that the relative dQNTK λ in Eq. (9) is a constant at late time and the connection between the dynamical index $\zeta \lesseqgtr 1/2$ in Eq. (13) and the $O_0 \gtrless O_{\min}$ cases. Our analyses will rely on large depth $D \gg 1$ (equivalently $L \gg 1$), which allows us to model each realization of the QNN $\hat{U}(\theta)$ as a sample from an ensemble of unitaries and consider ensemble averaged values to represent the typical case, $\bar{\zeta} = \overline{\epsilon\bar{\mu}/K^2}, \bar{\lambda} = \overline{\bar{\mu}/K}$. Note that we take the ratio between averaged quantity via considering the sign of \bar{C} . The ordering of ensemble averages has negligible effects (see Supplementary Note 12).

As the QNN is initialized randomly at the beginning, the unitary $\hat{U}(\theta)$ being implemented can be regarded as typical ones satisfying Haar random distribution^{4,29,51}, regardless of the circuit ansatz. While this is a good approximation at initial time, we notice that at late time, the QNN $\hat{U}(\theta)$ is constrained in the sense that it maps the initial trivial state (e.g. product of $|0\rangle$) towards a single quantum state, regardless of whether the quantum state is the unique optimum or not. Therefore, the late-time dynamics are always restricted due to convergence, which we model as the restricted Haar ensemble with a block-diagonal form,

$$\mathcal{E}_{\text{RH}} = \left\{ U|U = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & V \end{pmatrix} \right\}, \quad (17)$$

where V is a unitary with dimension $d - 1$ following a sub-system Haar random distribution (only 4-design is necessary). Here we have set the basis of the first column and row to represent mapping from the initial state to the final converged state. At late time, QNN converges to a restricted Haar ensemble determined by the converged state. When the converged state is unique, frame potential⁴⁴ of the ensemble can be evaluated by considering different training trajectories, which confirms the ansatz in Eq. (17), as shown in Fig. 1 and Supplementary Note 7.

The ensemble average for a general traceless operator is challenging to analytically obtain. To gain insights to QNN training, we consider a much simpler problem of state preparation, where $\hat{O} = |\Phi\rangle\langle\Phi|$ is a projector. In this case, we are interested in target values O_0 near the maximum loss function $O_{\max} = 1$. Under such restricted Haar ensemble, we have the following lemma.

Lemma 2. When the circuit satisfies the restricted Haar random (restricted 4-design) ensemble and $D \gg 1$ (therefore $L \gg 1$), in state preparation tasks the relative dQNTK $\bar{\lambda}_{\infty}$ goes to an L, d dependent constant. When $O_0 < O_{\max}$, the dynamical index $\bar{\zeta}_{\infty} = 0$; when $O_0 = O_{\max}$, the dynamical index $\bar{\zeta}_{\infty} = 1/2$; when $O_0 > O_{\max}$, the dynamical index ζ diverges to $+\infty$.

This lemma derives from Theorem 3 in the Method.

While our results are general, in our numerical study that verifies the analytical results, we adopt the random Pauli ansatz (RPA)²⁹ as an example (see Methods). Due to symmetry between maximum and minimum in optimization, this restricted Haar ensemble therefore fully explains the branches of dynamics in Theorem 1 quantitatively and the assumption that λ approaches a constant qualitatively. From asymptotic analyses of the restricted Haar ensemble in Supplementary Note 12, we also have both $\lambda, C \propto L/d$, thus the exponential decay in LV has exponent $\propto \eta Lt/d$. Indeed, in a computation, ηLt describes the resource—when a number of parameters L is larger, one needs to

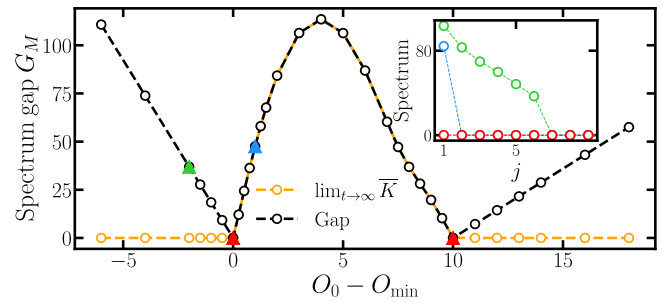


Fig. 4 | Spectrum gap of the effective Hamiltonian in Schrödinger interpretation of QNN in the example of XXZ model. The spectrum gap of Hessian matrix of the effective Schrödinger dynamics in $t \rightarrow \infty$ (black). The gapless transition point corresponds to $O_0 = O_{\min}, O_{\max}$ (red triangles). The orange line represents the QNTK $\lim_{t \rightarrow \infty} \bar{K}$. Inset shows the Hessian spectrum of the largest 10 eigenvalues for the three cases $O_0 \gtrless O_{\min}$ marked by triangles. The RPA consists of $D = 64$ layers (equivalently $L = 64$ parameters) on $n = 2$ qubits. The parameters in XXZ model is $J = 2$.

compute and update more parameters, while taking fewer steps t to converge.

As we show in Methods, Haar ensemble fails to capture the ζ dynamics nor the bifurcation transition. Only in the case of *frozen-kernel dynamics*, as the kernel does not change much during the dynamics, the Haar predictions roughly agree with the actual kernel, as shown in Fig. 2a2 (see Methods).

Schrödinger equation interpretation

Besides the LV dynamics, we can also connect the transition to the gap closing of the Hessian, via interpreting the training dynamics around the extremum as imaginary Schrodinger evolution as we detail below. The gradient descent dynamics in Eq. (2) leads to the time evolution of the quantum state $|\psi(\theta)\rangle$, where θ are the variational parameters. In the late time limit, omitting the t dependence in our notation, we can expand the shifts $\delta\theta_t$ around the extremum θ^* to second order as

$$\delta\theta \simeq -\eta M(\theta)|_{\theta=\theta^*} (\theta - \theta^*), \quad (18)$$

where the first-order term vanishes due to convergence and the Hessian matrix $M(\theta)$ is

$$M_{\ell_1 \ell_2}(\theta) = \left(\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta_{\ell_1} \partial \theta_{\ell_2}} \right) = \frac{\partial \epsilon(\theta)}{\partial \theta_{\ell_1}} \frac{\partial \epsilon(\theta)}{\partial \theta_{\ell_2}} + \epsilon(\theta) \frac{\partial^2 \epsilon(\theta)}{\partial \theta_{\ell_1} \partial \theta_{\ell_2}}. \quad (19)$$

We can then model a difference equation for the unnormalized “differential state” $|\Psi(\theta)\rangle \equiv |\psi(\theta)\rangle - |\psi(\theta^*)\rangle$ as

$$\delta|\Psi(\theta)\rangle = -\eta H_{\infty}(\theta)|\Psi(\theta)\rangle, \quad (20)$$

where $H_{\infty}(\theta) \sim M(\theta)$ is similar to the Hessian matrix (see Supplementary Note 2). The difference equation can be interpreted as an imaginary time Schrödinger equation, and we identify a transition with gap closing of H_{∞} (equivalently $M(\theta)$) driven by O_0 at the infinite time limit.

To provide insight into the transition, we explore the behaviors of the gap of Hessian matrix. We consider the Hessian eigenvalues at the late time limit of $t \rightarrow \infty$ and large circuit depth in Fig. 4. For frozen-kernel dynamics of $O_{\min} < O_0 < O_{\max}$, Hessian matrix in Eq. (19) becomes a rank-one matrix with only one nonzero eigenvalue as $\epsilon(\theta) \rightarrow 0$ (see blue and black curve in Fig. 4). While for frozen-error dynamics with $O_0 < O_{\min}$ (or $O_0 > O_{\max}$), due to non-vanishing $\epsilon(\theta)$, the Hessian has multiple nonzero eigenvalues (see green in the inset). Overall, gap closing is observed at the critical point. Such a transition at a finite system size resembles that for non-Hermitian dynamical systems^{54–56}. More

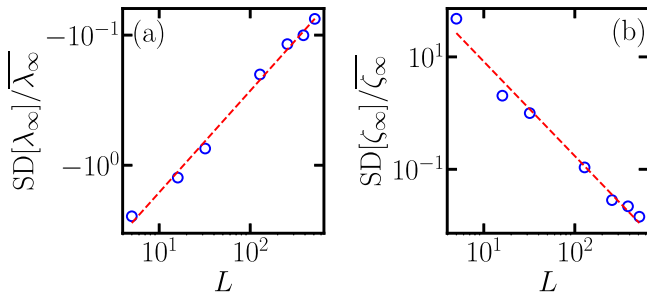


Fig. 5 | Late-time sample fluctuations. The standard deviations normalized by mean for the relative dQNTK λ_∞ **(a)** and dynamical index ζ_∞ **(b)** are plotted versus the number of parameters L . Red dashed lines represent power-law fitting results. Here the RPA is applied on $n = 5$ qubits with different L parameters (via tuning number of layers D). The observable is a state projector and the target value is $O_0 = 1$.

discussions on the statistical physics interpretation and the closing of the gap under different number of parameters can be found in Supplementary Note 2.

Dynamics of limited-depth QNN

We have so far focused on controllable QNNs with a universal gate set and large depth. In particular, for a general observable \hat{O} , reaching O_{\min} may require a circuit of exponential depth (in the number of qubits)⁵⁷. In addition, Lemma 2 requires the (restricted) 4-design that involves a polynomial circuit depth. However, we point out that these depth requirements may be only necessary for the theory derivations and not necessary for the transition phenomenon. Indeed, it is an interesting question—whether the transition still exists when the circuit is not controllable—either the ansatz is not universal^{58,59} or the depth is limited. Here we provide some results to the limited depth region of the QNNs under study. In this case, the circuit depth L is limited such that the QNN’s minimum achievable value of the observable $O_{\min}(L)$ deviates from the ground state energy O_{\min} . Such a scenario is often referred to as underparameterization.

We first consider the relative dQNTK $\overline{\lambda_\infty}$ and dynamical index $\overline{\zeta_\infty}$ versus the depth. In Lemma 2, we provide a justification of both quantities being constants for QNNs with a large depth D to approach the restricted 4-design. In Fig. 5, we present a numerical example for the target $O_0 = 1$ in state-preparation tasks. The relative sample fluctuations, defined as the standard deviation compared to its mean, decay in a power-law scaling with L , and thus vanish in the asymptotic limit of $L \gg 1$. The mean values $\overline{\lambda_\infty} \propto -L$ and $\overline{\zeta_\infty} \rightarrow 1/2$ are shown in Methods. The decay of fluctuation suggests that the ensemble-average results in Lemma 2 can represent the typical samples. Note that changing the order of the ensemble average for λ_∞ (see Eq. (9)) and ζ_∞ (see Eq. (13)) has negligible effects (see Supplementary Note 12). Similar results for other observables, e.g. XXZ model, are shown in Supplementary Note 12. The speed of convergence roughly agrees with the 4-design requirement of Lemma 2. However, we emphasize that sample fluctuation being small is only a sufficient but not necessary condition for the dynamical transition, as we show in the below example.

To our surprise, in Fig. 6, we find that the dynamical transition induced by the target value O_0 persists for a QNN with depth $D = L = n$ equaling the number of qubits, much less than what the theory requires. The results align with the dynamics of the controllable QNN presented in Fig. 2. We numerically find that the critical values for limited-depth QNNs, denoted as $O_{\min}(L)$, can deviate from the true ground state energy O_{\min} of a given observable \hat{O} . The critical value for a QNN with $L \ll d$ will not only depend on depth due to limited expressivity, but also fluctuates due to different initializations. We suspect this may be caused by the training converging to different

local minimum traps^{49,60}. The deviation of the critical point $O_{\min}(L)$ from O_{\min} indicates that the exponential depth for the convergence to O_{\min} is not necessary for the dynamical transition to persist. Moreover, despite the example being also not within the applicability of Lemma 2, the relative dQNTK λ still converges to a constant at late time as we show in SI. However, large sample fluctuation persists in this example due to $D = L = n$ being shallow, violating the unitary design assumption in Lemma 2. However, we point out that as long as λ has small time fluctuation at late time, its dynamics still follows the generalized LV equation discussed in Eq. (10). The above results indicate that the depth requirement of the transition may be much less than that for overparametrization⁶¹.

Speeding up the convergence

While the transition in training dynamics is interesting, the crucial question in practical applications is about how to speed up the training convergence of QNNs. Typically, two types of loss functions are adopted in optimization problems, the quadratic loss function in Eq. (1) that we have focused on, and the linear loss function

$$\mathcal{L}(\theta) = \langle \hat{O} \rangle. \tag{21}$$

While the linear loss function is widely used in variational quantum eigensolver^{7,58}, we note that unlike the versatile quadratic loss function that has a tunable target value, a linear function does not allow preparing excited states above the ground state energy nor can it be utilized to data classification and regression. Moreover, for the case of solving the ground state, we show that adopting the quadratic loss function and choosing a target value well below the achievable minimum can speed up the convergence compared to the linear loss function case. Interestingly, ‘shooting for the star’ will allow a faster convergence.

To begin with, we extend our theory framework to characterize the training dynamics of deep controllable QNNs with a linear loss function. To study its convergence, we further consider its residual error $\varepsilon(\theta) = \langle \hat{O} \rangle - O_{\min}$. Via a similar approach (see details in Supplementary Note 8), we have the dynamical equations for the error $\varepsilon(t)$ as

$$\delta \varepsilon(t) = -\eta K(t) + \mathcal{O}(\eta^2), \tag{22}$$

where $K(t)$ is still the QNTK defined in Eq. (5). The dynamical equation for QNTK $K(t)$ becomes

$$\delta K(t) = -2\eta \mu(t) + \mathcal{O}(\eta^2), \tag{23}$$

with $\mu(t)$ being dQNTK defined in Eq. (6). One may notice that the only difference compared to Eqs. (7) and (8) is the missing of $\varepsilon(t) = \varepsilon(t)$ on RHS due to a linear loss.

In the late-time limit, the results in ‘Unitary ensemble theory’ section still applies to linear loss, and the relative dQNTK $\lambda(t) \equiv \mu(t)/K(t) = \lambda$ converges to a constant, leading to

$$\begin{cases} \partial_t \varepsilon(t) &= -\eta K(t), \\ \partial_t K(t) &= -2\eta \lambda K(t). \end{cases} \tag{24}$$

Unlike the generalized LV model in Eqs. (10) for the quadratic loss case, here the dynamics of $K(t)$ is self-determined, whereas the dynamics of $\varepsilon(t) = \varepsilon(t)$ is fully determined by $K(t)$ —the kernel-error duality is broken. Eqs. (24) can be directly solved as

$$2\lambda \varepsilon(t) = 2\lambda \varepsilon(t) = K(t) \propto e^{-2\eta \lambda t}. \tag{25}$$

Both $\varepsilon(t)$ and $K(t)$ decay exponentially at a fixed rate $\propto \lambda$. In Fig. 7, we present the numerical simulation results (black solid), and observe a good agreement with the theory (black dashed) from Eq. (25).

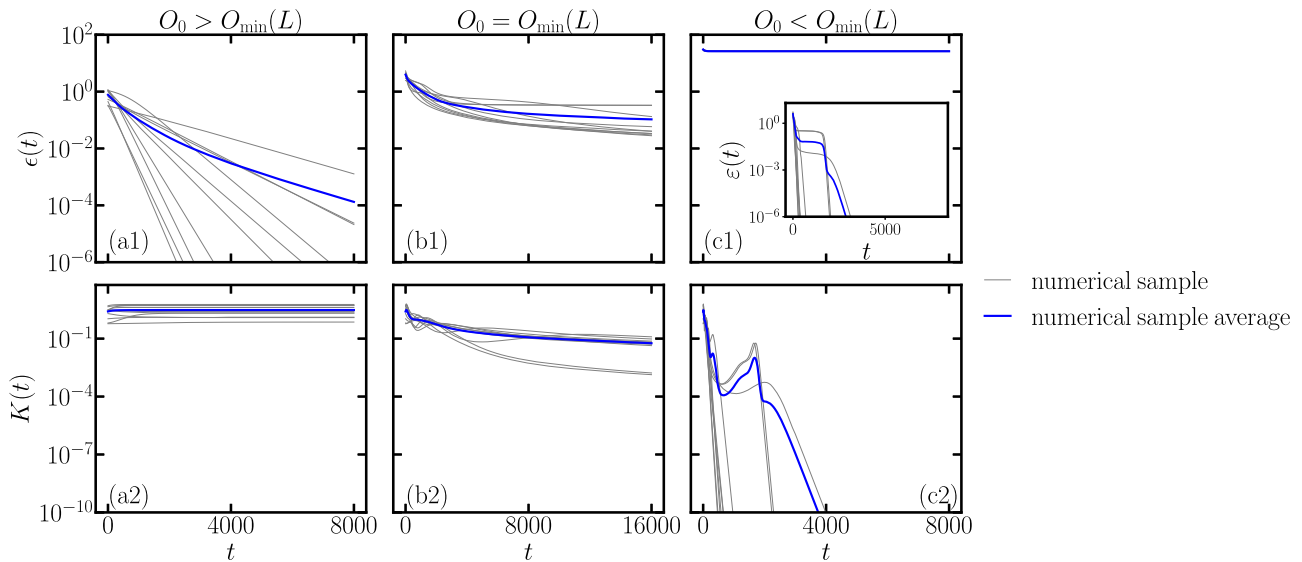


Fig. 6 | Dynamics in limited-depth QNNs in the example of the XXZ model. All notations share the same meaning as in Fig. 2. The critical point $O_{\min}(L)$ for such QNNs depends on L and has sample fluctuations. Here random Pauli ansatz (RPA)

consists of $L = 6$ variational parameters ($D = L$ for RPA) on $n = 6$ qubits, and the parameter in XXZ model is $J = 2$.

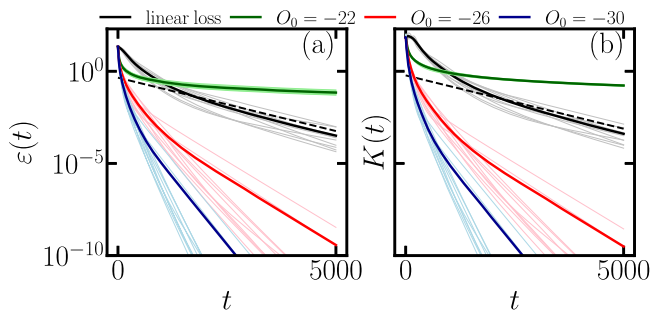


Fig. 7 | Dynamics in QNN in the example of XXZ model with different loss functions. In **a** and **b**, we show the dynamics of residual error $\epsilon(t)$ (equals to total error $\epsilon(t)$) and QNTK $K(t)$ optimized with linear loss function (black solid) and quadratic loss functions with different O_0 . $O_0 = -22$ (green) corresponds to $O_0 = O_{\min}$ at critical point and $O_0 = -26, -30$ (red and blue) correspond to $O_0 < O_{\min}$ in frozen error dynamics. Black dashed line indicates the exponential decay rate of the theoretical result in Eq. (25). Thin lines with light colors represent dynamics with different initializations in each case, while the thick lines represent the ensemble average. Here random Pauli ansatz (RPA) consists of $L = 192$ variational parameters ($D = L$ layers) on $n = 6$ qubits, and the parameter of XXZ model is $J = 2$.

With the linear-loss theory developed, we can now compare the convergence speed between the different choices of loss functions in solving the minimum value O_{\min} and the corresponding ground state. As indicated in Eq. (25), the linear loss function provides an exponential convergence with the exponent $2\eta\lambda$ being a constant (black). For quadratic loss functions, at the critical point setting $O_0 = O_{\min}$, the convergence is polynomial and exponent is zero (green lines in Fig. 7), corresponding to a much slower convergence. However, recall that with a quadratic loss function, one can set $O_0 < O_{\min}$ corresponding to the *frozen error dynamics*, where the residual error $\epsilon(t)$ decays exponentially with the exponent $2\eta\lambda R$ (see Eq. (16)). Here the residual R is freely tunable by the target value O_0 . Therefore, an appropriate choice of O_0 can provide a larger exponent and therefore faster convergence towards the solution, and we verify it in Fig. 7 through different values of O_0 (red and blue curves). Indeed, setting the target to be unachievable will still converge the output to the ground state, although the remaining error is frozen.

Experimental results

In this section, we consider the experimental-friendly hardware-efficient ansatz (HEA) to experimentally verify our results on real IBM quantum devices. Each layer of HEA consists of single qubit rotations along Y and Z directions, and followed by CNOT gates on nearest neighbors in a brickwall style⁷. Our experiments adopt the hardware IBM Kolkata, an IBM Falcon r5.11 hardware with 27 qubits, via IBM Qiskit⁶². The device has median $T_1 \sim 98.97$ us, median $T_2 \sim 58.21$ us, median CNOT error $\sim 9.546 \times 10^{-3}$, median SX error $\sim 2.374 \times 10^{-4}$, and median readout error $\sim 1.110 \times 10^{-2}$. We randomly assign the initial variational angles, distributing them within the range of $[0, 2\pi)$, and maintain consistency across all experiments. To suppress the impact of error, we average the results over 12 independent experiments conducted under the same setup for three distinct choices, $O_0 = -10, -12, -14$. In Fig. 8, the experimental data (solid) on IBM Kolkata agree well with the noisy theory model (dashed) and indicate the *frozen-error dynamics* with constant error (green), the *critical point* of polynomial decaying error (red) and the *frozen-kernel dynamics* of exponential decaying error (blue). Individual training data and noisy theory model are presented in Supplementary Note 10.

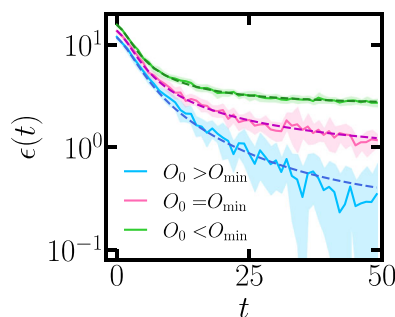


Fig. 8 | Dynamics of total error $\epsilon(t)$ on IBM quantum devices, Kolkata. Solid and dashed curves represent experimental and theoretical results. An $n = 2$ qubit $D = 4$ -layer hardware efficient ansatz (with $L = 16$ parameters) is utilized to optimize with respect to XXZ model observable with $J = 4$. The shaded areas represent the fluctuation (standard deviation) in the experimental data.

Discussion

Our results go beyond the early-time Haar random ensemble widely adopted in QNN study^{4,29,51} and reveal rich physics in the dynamical transition controlled by the target loss function. The target-driven transcritical bifurcation transition in the dynamics of QNN suggests a different source to the transition without symmetry breaking. From the Schrödinger equation interpretation, there may exist other unexplored sources that can induce dynamical transition, especially when the QNN has limited depth and controllability. In practical applications, the dynamical transition guides us towards better design of loss functions to speed up the training convergence.

Another intriguing question pertains to the differences between classical and quantum machine learning within this formalism. In our examples, the target O_0 can be interpreted as a single piece of supervised data in a supervised machine learning task. Therefore, the dynamical transition we have discovered could be seen as a simplified version of a theory of data. Classical machine learning also extensively explores dynamical transitions, whether in relation to learning rate dynamics^{63,64} or the depth of classical neural networks^{43,65}. It is an open question whether some results similar to ours can be established for classical machine learning, especially in the context of the large-width regime of classical neural networks⁶⁶. It is also an open problem how our results can generalize to the multiple data case.

Finally, we clarify the difference of our results to some related works. Firstly, while existing works^{28-31,33} on the quantum neural tangent kernel provide a perturbative explanation of gradient descent dynamics that fails to uncover the dynamical transition, our work uncovers the dynamical transition and formulates non-perturbative critical theories about the transition triggered by modifications in the quantum data. Secondly, we have developed a non-perturbative, phenomenological model using the generalized Lotka-Volterra equations to describe the dynamics as a transcritical bifurcation transition, providing a first-principle explanation using the restricted Haar ensemble. Thirdly, we provide an interpretation of the gradient descent dynamics using Schrödinger's equation in imaginary time, where the Hessian spectra can be mapped to the effective Hamiltonian using the language of physics, allowing us to study the effective spectral gap. Finally, using correlated dynamics of the Haar ensemble, we offer a more precise derivation of the statistics of the quantum neural tangent kernel, going beyond ref. 29.

Methods

QNN ansatz and details of the tasks

The random Pauli ansatz (RPA) circuit is constructed as

$$\hat{U}(\theta) = \prod_{\ell=1}^D \hat{W}_\ell \hat{V}_\ell(\theta_\ell), \quad (26)$$

where $\theta = (\theta_1, \dots, \theta_L)$ are the variational parameters. For RPA, $D=L$. Here $\{\hat{W}_\ell\}_{\ell=1}^L \in \mathcal{U}_{\text{Haar}}(d)$ is a set of fixed Haar random unitaries with dimension $d = 2^n$, and \hat{V}_ℓ is a n -qubit rotation gate defined to be

$$\hat{V}_\ell(\theta_\ell) = e^{-i\theta_\ell \hat{X}_\ell / 2}, \quad (27)$$

where $\hat{X}_\ell \in \{\hat{\sigma}^x, \hat{\sigma}^y, \hat{\sigma}^z\}^{\otimes n}$ is a random n -qubit Pauli operator non-trivially supported on every qubit. Once a circuit is constructed, $\{\hat{X}_\ell, \hat{W}_\ell\}_{\ell=1}^L$ are fixed through the optimization. Note that our results also hold for other typical universal ansatz of QNN, for instance, hardware efficient ansatz (see 'Experimental results' and Supplementary Note 10).

In the main text, some of our main results are derived for general observable \hat{O} . To simplify our expressions, we often consider \hat{O} to be traceless, for instance a spin Hamiltonian, which is not essential to our conclusions. A general traceless operator can be expressed as random

mixture of Pauli strings (excluding identity)

$$\hat{O} = \sum_{i=1}^N c_i \hat{P}_i \quad (28)$$

with real coefficients $c_i \in \mathbb{R}$ and nontrivial Pauli $\hat{P}_i \in \{\hat{\mathbb{1}}, \hat{\sigma}^x, \hat{\sigma}^y, \hat{\sigma}^z\}^{\otimes n} / \{\hat{\mathbb{1}}^{\otimes n}\}$. To obtain explicit expressions, we also consider the XXZ model, described by

$$\hat{O}_{\text{XXZ}} = - \sum_{i=1}^n [\hat{\sigma}_i^x \hat{\sigma}_{i+1}^x + \hat{\sigma}_i^y \hat{\sigma}_{i+1}^y + J(\hat{\sigma}_i^z \hat{\sigma}_{i+1}^z + \hat{\sigma}_i^z)]. \quad (29)$$

To help understanding the non-frozen QNTK phenomena, we also consider a state preparation case with the observable $\hat{O} = |\Phi\rangle\langle\Phi|$, where $|\Phi\rangle$ is the target state.

Hamiltonian description and analytical solution of the LV dynamics

From the conservation law in Eq. (11), we can introduce the canonical coordinates

$$P = \log(K), \quad Q = \log(2\lambda\epsilon) \quad (30)$$

and the associated Hamiltonian

$$H(Q, P) = \eta(e^Q - e^P) \equiv \eta(2\lambda\epsilon - K), \quad (31)$$

from which the LV equations in Eq. (10) can be equivalently rewritten as the standard Hamiltonian equation generalizing ref. 67,

$$\begin{cases} \frac{dQ}{dt} &= \frac{\partial H}{\partial P} = \{Q, H\}, \\ \frac{dP}{dt} &= -\frac{\partial H}{\partial Q} = \{P, H\}, \end{cases} \quad (32)$$

where $\{A, B\} = \frac{\partial A}{\partial Q} \frac{\partial B}{\partial P} - \frac{\partial A}{\partial P} \frac{\partial B}{\partial Q}$ denotes the Poisson bracket. From the position-momentum duality in Hamiltonian formulation, we identify an error-kernel duality between $e^Q \sim \epsilon$ and its gradient $e^P = |\partial\epsilon/\partial\theta|^2$.

We can obtain an analytical solution of Eq. (10) directly. When $C \neq 0$, we have

$$\begin{cases} \lambda\epsilon(t) &= C / [-2 + B_1 e^{\eta C t}], \\ K(t) &= C / [1 - 2B_1^{-1} e^{-\eta C t}], \end{cases} \quad (33)$$

where B_1 is a constant fitting parameter as at an early time we do not expect Eq. (10) to hold. When $C = 0$, Eq. (10) leads to polynomial decay of both quantities

$$K(t) = 2\lambda\epsilon(t) = 2 / (2\eta t + B_2^{-1}), \quad (34)$$

where B_2 is again a fitting parameter as at an early time we do not expect Eq. (10) to hold. Indeed, we observe the bifurcation, and the convergence towards the fixed points is exponential for $C \leq 0$ and polynomial for $C = 0$.

Details of restricted Haar ensemble

Here we evaluate the average QNTK, relative dQNTK, and dynamical index for the restricted Haar ensemble proposed in Eq. (17). We focus on the state preparation task to enable analytical calculation. As we aim to capture the late time dynamics with the state preparation task, we will be interested in the dynamics when the output state $|\psi_0\rangle$ has fidelity $\langle\hat{O}\rangle = |\langle\psi_0|\Phi\rangle|^2 = O_0 - R - \kappa$, with $\kappa \sim o(1)$ indicating late-time where the observable is already close to its reachable target. Here the constant remaining term $R = O_0 - 1$ when $O_0 > 1$ and zero otherwise. Note that identity is the maximum reachable target value in state

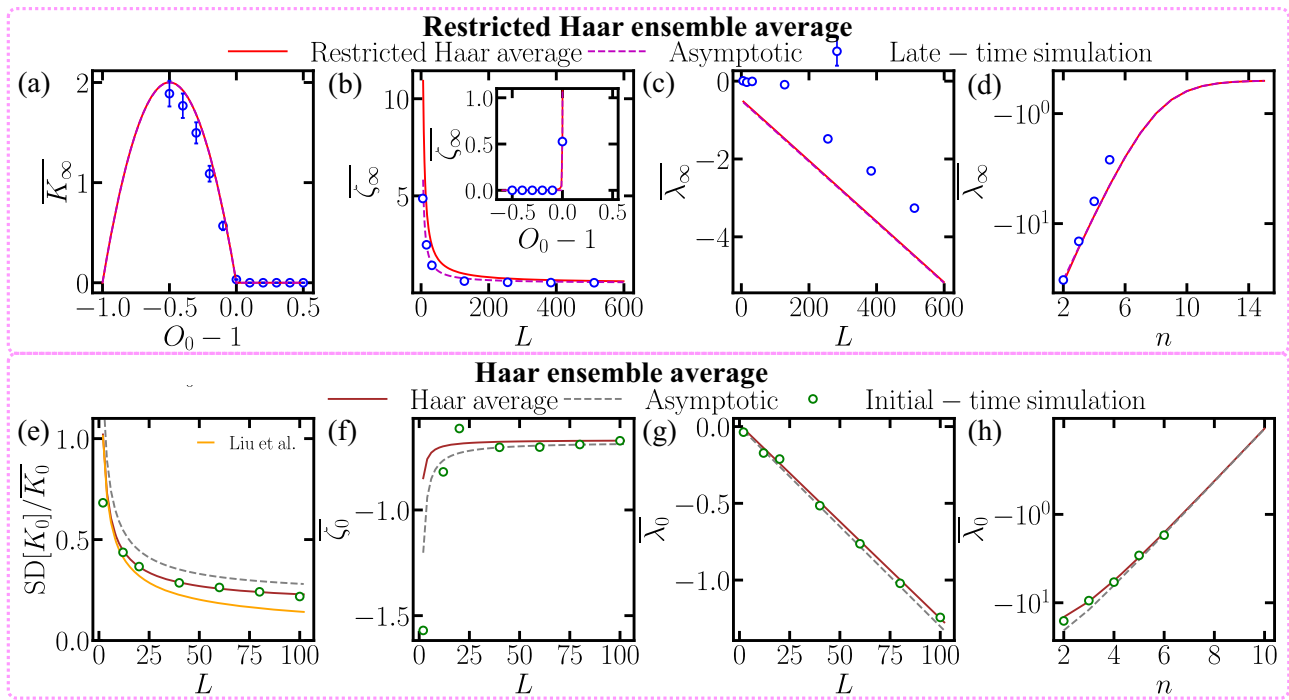


Fig. 9 | Ensemble average results under restricted Haar ensemble (top) and Haar ensemble (bottom). In top panel, we plot **a** $\overline{K_\infty}$ versus O_0 with $L = 512$ fixed, **b** $\overline{\zeta_\infty}$ versus L , $\overline{\lambda_\infty}$ versus **(c)** L and **d** n with $L = 512$ at late time in state preparation. We set $O_0 = 1$ for **b** and **d**, and $O_0 = 5$ for **c**. Blue dots in top panels **a–c** represents numerical results from late-time optimization of $n = 5$ qubit RPA. Red solid lines represent exact ensemble average with restricted Haar ensemble in Supplementary Equations (256), (313), (279) in Supplementary Note 12. Magenta dashed lines represent asymptotic ensemble average with restricted Haar ensemble in Eq. (35), (36), (37) which overlap with the exact results (red solid). The observable in all cases

is $|\Phi\rangle\langle\Phi|$ with $|\Phi\rangle$ being a fixed Haar random state. In the inset of **b**, we fix $L = 512$. In bottom panel, we plot **(e)** fluctuation $SD[K_0]/\overline{K_0}$ versus L , **(f)** $\overline{\zeta_0}$ versus L , $\overline{\lambda_0}$ versus **(g)** L and **(h)** n with $L = 128$ under random initialization. Green dots in bottom panel from **e–g** represent numerical results from random initializations of $n = 6$ qubit RPA. Brown solid lines represent the exact ensemble average with the Haar ensemble in Supplementary Equations (241), (180), (120) in Supplementary Note 11. Gray dashed lines represent asymptotic ensemble average with a restricted Haar ensemble in Eqs. (45), (39), (40). The observable and target in **e–h** are XXZ model with $J = 2$ and $O_0 = O_{\min}$. Orange solid line in **e** represents results from ref. 29.

preparation. Under this ensemble, we have the following result (see details in Supplementary Note 12)

Theorem 3. For state projector observable $\hat{O} = |\Phi\rangle\langle\Phi|$, when the circuit satisfies restricted Haar ensemble, the ensemble average of QNTK, relative dQNTK, and dynamical index

$$\overline{K_\infty} = \frac{L}{2d} (O_0 + R)(1 - O_0 - R), \tag{35}$$

$$\overline{\zeta_\infty} = \frac{R}{O_0 + R - 1} \left(1 - \frac{1}{2(O_0 + R)} + \frac{d}{L} \right), \tag{36}$$

$$\overline{\lambda_\infty} = \frac{L}{4d} (1 - 2O_0 - 2R) - \frac{O_0 + R}{2}, \tag{37}$$

at the $L \gg 1, d \gg 1$ limit, where the target loss function value $O_0 \geq 0$, remaining constant $R = \min\{1 - O_0, 0\}$.

Our results are verified numerically in Fig. 9 in state preparation task, where we plot the above asymptotic equations as magenta dashed lines and the full formula in SI as solid red lines. Note that Lemma 2 does not require $d \gg 1$, but merely $D \gg 1$. Indeed, full expressions in Theorem 3 can also be derived for any finite d , just much more lengthy.

Subplot (a) plots $\overline{K_\infty}$ versus O_0 . At late time, if the target $O_0 \geq 1$, from $O_0 = 1 + R$ we directly have $\overline{K_\infty} = 0$; if $O_0 < 1$, we have $R = 0$ and $\overline{K_\infty} \propto O_0(1 - O_0)$ being a constant.

Subplot (b) shows the agreement of $\overline{\zeta_\infty}$ versus L , when we fix $O_0 = 1$. As predicted by the theory of Eq. (36), as $R = 0$ in this case,

$\overline{\zeta_\infty} = 1/2$ when $L \gg 1$. Indeed, we see convergence towards $1/2$ as the depth increases. We also verify the $\overline{\zeta_\infty}$ versus O_0 relation in the inset, where $\overline{\zeta_\infty} = 0$ for $O_0 < 1$, $1/2$ for $O_0 = 1$ and diverges for $O_0 > 1$. Note that for a circuit with medium depth $L \sim \text{poly}(n)$, $\overline{\zeta_\infty} = 1/2 + d/L$ would slightly deviate from $1/2$ for $O_0 = 1$ (Fig. 9(b)). This indicates a ‘finite-size’ effect affecting the dynamical transition, which we defer to future work.

Subplot (c) shows the agreement of $\overline{\lambda_\infty}$ versus L , where the linear relation is verified. As predicted by Eq. (37), this is the case regardless of O_0 value. We also verify the dependence of $\overline{\lambda_\infty}$ on n (thus $d = 2^n$) with a fixed L in subplot (d), where we see as n increases, $\overline{\lambda_\infty}$ converges to a constant only relying on O_0 .

Haar ensemble results

We also evaluate the Haar ensemble expectation values for reference, which captures the early-time QNN dynamics. Under the Haar random assumption, we find the following lemma.

Lemma 4. For traceless operator \hat{O} , when the initial circuit satisfies Haar random (4-design) and circuit $L \gg 1$ and $d \gg 1$, the ensemble averages of QNTK, relative dQNTK and dynamical index have leading order

$$\begin{aligned} \overline{K_0} &= L \frac{d \text{tr}(\hat{O}^2)}{2(d-1)(d+1)^2}, \\ \overline{\zeta_0} &= -\frac{1}{L} \left[1 + \frac{\text{tr}(\hat{O}^4)}{\text{tr}(\hat{O}^2)^2} \right] \end{aligned} \tag{38}$$

$$+ \frac{1}{2} \left[\frac{\text{tr}(\hat{O}^4)}{\text{tr}(\hat{O}^2)^2} - dO_0 \frac{\text{tr}(\hat{O}^3)}{\text{tr}(\hat{O}^2)^2} - \frac{3}{d} \right], \quad (39)$$

$$\bar{\lambda}_0 = \frac{L \text{tr}(\hat{O}^3)}{4d \text{tr}(\hat{O}^2)}. \quad (40)$$

Note that for observables with non-zero trace, evaluation is also possible, we present those lengthy formulae and the proofs in Supplementary Note 11. Note that similar to Theorem 3, here the requirement of $d \gg 1$ is for simplification of formula only and the full formula in SI applies to any finite d . Meanwhile, it is important to notice the dimension dependence of the trace terms.

Specifically, for the XXZ model we considered, when $d \gg 1$, the above Lemma 4 leads to

$$\bar{K}_{0\text{XXZ}} \simeq (1+J^2) \frac{Ln}{d}, \quad (41)$$

$$\bar{\zeta}_{0\text{XXZ}} \simeq -\frac{1}{L} \left(1 + \frac{3}{d}\right) - O_0 \frac{3J(1-J^2)}{4(1+J^2)^2 n}, \quad (42)$$

$$\bar{\lambda}_{0\text{XXZ}} \simeq \frac{3J(1-J^2)L}{4(1+J^2)d}. \quad (43)$$

We verified the Haar prediction on $\bar{\zeta}_0$ and $\bar{\lambda}_0$ with random initialized circuits in Fig. 9(f)–(h). Note that when L is large enough, $\bar{\zeta}_{0\text{XXZ}}$ scales linearly with O_0 , while $\bar{\lambda}_{0\text{XXZ}}$ converges to zero exponentially with n .

In the Haar case, we can also obtain the fluctuation properties.

Theorem 5. In the asymptotic limit of wide and deep QNN $d, L \gg 1$, we have the ensemble average of QNTK standard deviation (4-design)

$$\begin{aligned} \text{SD}[K_0] = & \left(\frac{3L}{4d^6} \left[d^2 \text{tr}(\hat{O}^2)^2 - 2d \text{tr}(\hat{O}^2) \text{tr}(\hat{O})^2 + \text{tr}(\hat{O})^4 \right] \right. \\ & \left. + \frac{L^2}{4d^5} \left[d \text{tr}(\hat{O}^4) - 4 \text{tr}(\hat{O}^3) \text{tr}(\hat{O}) \right] \right)^{1/2}. \end{aligned} \quad (44)$$

Note that similar to Theorem 3 and Lemma 4, here the requirement of $d \gg 1$ is for simplification of formula only and the full formula in SI applies to any finite d .

For traceless operators, Eq. (44) can be further simplified and the relative sample fluctuation of QNTK is

$$\frac{\text{SD}[K_0]}{\bar{K}_0} = \frac{1}{\sqrt{L}} \left(L \frac{\text{tr}(\hat{O}^4)}{\text{tr}(\hat{O}^2)^2} + 3 \right)^{1/2}. \quad (45)$$

This result refines ref. 29 with a more accurate ensemble averaging technique and provides an additional term $\sim \text{tr}(\hat{O}^4)/\text{tr}(\hat{O}^2)^2$. Therefore, the sample fluctuation also depends on the observable being optimized. Specifically, for the XXZ model we considered, Eq. (45) becomes

$$\frac{\text{SD}[K_0]}{\bar{K}_0} \simeq \sqrt{\frac{3}{L} \left(\frac{L}{d} + 1 \right)}. \quad (46)$$

When $L \gg d$, the relative fluctuation $\text{SD}[K_0]/\bar{K}_0 \sim 1/\sqrt{d}$ is constant. However, as $d = 2^n$ is exponential while a realistic number of layers L is

polynomial in n , therefore $d \gg L$ is more common, where the relative fluctuation $\text{SD}[K_0]/\bar{K}_0 \sim \sqrt{1/L}$ decays with the depth, consistent with ref. 29. We numerically evaluate the ensemble average in Fig. 9(e) and find a good agreement between our full analytical formula (red solid, Eq. (241) in SI) and the numerical results (blue circle). The asymptotic result (magenta dashed, Eq. (46)) also captures the scaling correctly. The results refine the calculation of ref. 29, which has a substantial deviation when L and d are comparable.

Data availability

The data generated in this study have been deposited in Github [<https://github.com/bzGit06/QNN-dynamics>].

Code availability

The theoretical results of the manuscript are reproducible from the analytical formulas and derivations presented therein. Additional code is available in Github [<https://github.com/bzGit06/QNN-dynamics>].

References

- Peruzzo, A. et al. A variational eigenvalue solver on a photonic quantum processor. *Nat. Commun.* **5**, 4213 (2014).
- Farhi, E., Goldstone, J. & Gutmann, S. A quantum approximate optimization algorithm. arXiv:1411.4028 <https://doi.org/10.48550/arXiv.1411.4028> (2014).
- McClean, J. R., Romero, J., Babbush, R. & Aspuru-Guzik, A. The theory of variational hybrid quantum-classical algorithms. *New J. Phys.* **18**, 023023 (2016).
- McClean, J. R., Boixo, S., Smelyanskiy, V. N., Babbush, R. & Neven, H. Barren plateaus in quantum neural network training landscapes. *Nat. Commun.* **9**, 4812 (2018).
- McArdle, S., Endo, S., Aspuru-Guzik, A., Benjamin, S. C. & Yuan, X. Quantum computational chemistry. *Rev. Mod. Phys.* **92**, 015003 (2020).
- Cerezo, M. et al. Variational quantum algorithms. *Nat. Rev. Phys.* **3**, 625 (2021).
- Kandala, A. et al. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature* **549**, 242 (2017).
- Ebadi, S. et al. Quantum optimization of maximum independent set using Rydberg atom arrays. *Science* **376**, 1209 (2022).
- Yuan, X., Endo, S., Zhao, Q., Li, Y. & Benjamin, S. C. Theory of variational quantum simulation. *Quantum* **3**, 191 (2019).
- Yao, Y.-X. et al. Adaptive variational quantum dynamics simulations. *PRX Quantum* **2**, 030307 (2021).
- Cong, I., Choi, S. & Lukin, M. D. Quantum convolutional neural networks. *Nat. Phys.* **15**, 1273 (2019).
- Zhang, B., Wu, J., Fan, L. & Zhuang, Q. Hybrid entanglement distribution between remote microwave quantum computers empowered by machine learning. *Phys. Rev. Appl.* **18**, 064016 (2022).
- ur Rehman, J., Hong, S., Kim, Y.-S. & Shin, H. Variational estimation of capacity bounds for quantum channels. *Phys. Rev. A* **105**, 032616 (2022).
- Zhuang, Q. & Zhang, Z. Physical-layer supervised learning assisted by an entangled sensor network. *Phys. Rev. X* **9**, 041023 (2019).
- Xia, Y., Li, W., Zhuang, Q. & Zhang, Z. Quantum-enhanced data classification with a variational entangled sensor network. *Phys. Rev. X* **11**, 021047 (2021).
- Wittek, P. Quantum machine learning: what quantum computing means to data mining (Academic Press, 2014).
- Schuld, M., Sinayskiy, I. & Petruccione, F. An introduction to quantum machine learning. *Contemp. Physics* **56**, 172 (2015).
- Biamonte, J. et al. Quantum machine learning. *Nature* **549**, 195 (2017).

19. Farhi, E. & Neven, H. Classification with quantum neural networks on near term processors. arXiv:1802.06002 <https://doi.org/10.48550/arXiv.1802.06002> (2018).
20. Dunjko, V. & Briegel, H. J. Machine learning & artificial intelligence in the quantum domain: a review of recent progress. *Rep. Prog. Phys.* **81**, 074001 (2018).
21. Schuld, M. & Killoran, N. Quantum machine learning in feature hilbert spaces. *Phys. Rev. Lett.* **122**, 040504 (2019).
22. Havlíček, V. et al. Supervised learning with quantum-enhanced feature spaces. *Nature* **567**, 209 (2019).
23. Abbas, A. et al. The power of quantum neural networks. *Nat. Computat. Sci.* **1**, 403 (2021).
24. Ni, Z. et al. Beating the break-even point with a discrete-variable-encoded logical qubit. *Nature* **616**, 56 (2023).
25. Sivak, V. et al. Real-time quantum error correction beyond break-even. *Nature* **616**, 50 (2023).
26. Shen, H., Zhang, P., You, Y.-Z. & Zhai, H. Information scrambling in quantum neural networks. *Phys. Rev. Lett.* **124**, 200504 (2020).
27. Garcia, R. J., Bu, K. & Jaffe, A. Quantifying scrambling in quantum neural networks. *J. High Energy Phys.* **2022**, 27 (2022).
28. Liu, J., Tacchino, F., Glick, J. R., Jiang, L. & Mezzacapo, A. Representation learning via quantum neural tangent kernels. *PRX Quantum* **3**, 030323 (2022).
29. Liu, J. et al. Analytic theory for the dynamics of wide quantum neural networks. *Phys. Rev. Lett.* **130**, 150601 (2023).
30. Liu, J., Lin, Z. & Jiang, L. Laziness, barren plateau, and noises in machine learning. *Mach. Learn.: Sci. Technol.* **5**, 015058 (2024).
31. Wang, X. et al. Symmetric pruning in quantum neural networks. arXiv:2208.14057 <https://doi.org/10.48550/arXiv.2208.14057> (2022).
32. Yu, L.-W. et al. Expressibility-induced concentration of quantum neural tangent kernels. *Rep. Prog. Phys.* **87**, 110501 (2024).
33. Lee, J. et al. Deep neural networks as Gaussian processes. arXiv:1711.00165 <https://doi.org/10.48550/arXiv.1711.00165> (2017).
34. Jacot, A., Gabriel, F. & Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. arXiv:1806.07572 <https://doi.org/10.48550/arXiv.1806.07572> (2018).
35. Lee, J. et al. Wide neural networks of any depth evolve as linear models under gradient descent. *Adv. Neural Inf. Process. Syst.* **32**, 8572 (2019).
36. Sohl-Dickstein, J., Novak, R., Schoenholz, S. S. & Lee, J. On the infinite width limit of neural networks with a standard parameterization. arXiv:2001.07301 <https://doi.org/10.48550/arXiv.2001.07301> (2020).
37. Yang, G. & Hu, E. J. Feature learning in infinite-width neural networks. arXiv:2011.14522 <https://doi.org/10.48550/arXiv.2011.14522> (2020).
38. Yaida, S. Non-gaussian processes and neural networks at finite widths. In *Mathematical and Scientific Machine Learning*. pp. 165–192 (PMLR, 2020).
39. Arora, S. et al. On exact computation with an infinitely wide neural net. arXiv:1904.11955 <https://doi.org/10.48550/arXiv.1904.11955> (2019).
40. Dyer, E. & Gur-Ari, G. Asymptotics of wide networks from feynman diagrams. arXiv:1909.11304 <https://doi.org/10.48550/arXiv.1909.11304> (2019).
41. Halverson, J., Maiti, A. & Stoner, K. Neural networks and quantum field theory. *Mach. Learn.: Sci. Technol.* **2**, 035002 (2021).
42. Roberts, D. A. Why is AI hard and physics simple? arXiv:2104.00008 <https://doi.org/10.48550/arXiv.2104.00008> (2021).
43. Roberts, D. A., Yaida, S. & Hanin, B. The principles of deep learning theory (Cambridge University Press, 2022).
44. Roberts, D. A. & Yoshida, B. Chaos and complexity by design. *J. High Energy Phys.* **2017**, 121 (2017).
45. Cotler, J., Hunter-Jones, N., Liu, J. & Yoshida, B. Chaos, complexity, and random matrices. *JHEP* **11**, 048 (2017).
46. Liu, J. Spectral form factors and late time quantum chaos. *Phys. Rev. D* **98**, 086026 (2018).
47. Liu, J. Scrambling & decoding the charged quantum information. *Phys. Rev. Res.* **2**, 043164 (2020).
48. You, X., Chakrabarti, S. & Wu, X. A convergence theory for over-parameterized variational quantum eigensolvers. arXiv:2205.12481 <https://doi.org/10.48550/arXiv.2205.12481> (2022).
49. You, X. & Wu, X. Exponentially many local minima in quantum neural networks. In *International Conference on Machine Learning*. pp. 12144–12155 (PMLR, 2021).
50. Anschuetz, E. R. Critical points in quantum generative models. arXiv:2109.06957 <https://doi.org/10.48550/arXiv.2109.06957> (2021).
51. Cerezo, M., Sone, A., Volkoff, T., Cincio, L. & Coles, P. J. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nat. Commun.* **12**, 1791 (2021).
52. Hofbauer, J. & Sigmund, K. Evolutionary games and population dynamics (Cambridge University Press, 1998).
53. Strogatz, S. H. Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering (CRC Press, 2018).
54. Berry, M. V. Physics of nonhermitian degeneracies. *Czech. J. Phys.* **54**, 1039 (2004).
55. Rotter, I. & Bird, J. A review of progress in the physics of open quantum systems: theory and experiment. *Rep. Prog. Phys.* **78**, 114001 (2015).
56. El-Ganainy, R. et al. Non-hermitian physics and pt symmetry. *Nat. Phys.* **14**, 11 (2018).
57. Kiani, B. T., Lloyd, S. & Maiti, R. Learning unitaries by gradient descent. arXiv:2001.11897 <https://doi.org/10.48550/arXiv.2001.11897> (2020).
58. Wiersema, R. et al. Exploring entanglement and optimization within the hamiltonian variational ansatz. *PRX Quantum* **1**, 020319 (2020).
59. Larocca, M. et al. Diagnosing barren plateaus with tools from quantum optimal control. *Quantum* **6**, 824 (2022).
60. Anschuetz, E. R. & Kiani, B. T. Quantum variational algorithms are swamped with traps. *Nat. Commun.* **13**, 7760 (2022).
61. Larocca, M., Ju, N., García-Martín, D., Coles, P. J. & Cerezo, M. Theory of overparametrization in quantum neural networks. *Nat. Comput. Sci.* **3**, 542 (2023).
62. Qiskit contributors Qiskit: An open-source framework for quantum computing <https://doi.org/10.5281/zenodo.2573505> (2023).
63. Lewkowycz, A., Bahri, Y., Dyer, E., Sohl-Dickstein, J. & Gur-Ari, G. The large learning rate phase of deep learning: the catapult mechanism. arXiv:2003.02218 <https://doi.org/10.48550/arXiv.2003.02218> (2020).
64. Meltzer, D. & Liu, J. Catapult dynamics and phase transitions in quadratic Nets. arXiv e-prints, arXiv:2301.07737 <https://doi.org/10.48550/arXiv.2301.07737> (2023).
65. Lin, H. W., Tegmark, M. & Rolnick, D. Why does deep and cheap learning work so well? *J. Stat. Phys.* **168**, 1223 (2017).
66. Batson, J., Haaf, C. G., Kahn, Y. & Roberts, D. A. Topological obstructions to autoencoding. *JHEP* **04**, 280 (2021).
67. Nutku, Y. Hamiltonian structure of the lotka-volterra equations. *Phys. Lett. A* **145**, 27 (1990).

Acknowledgements

We thank David Simmons-Duffin for very helpful discussions. This work was supported NSF [CCF-2240641, OMA-2326746, 2330310, 2350153], ONR [N00014-23-1-2296], DARPA [HR00112490453], Cisco Systems, Inc, Google LLC, and Halliburton Company to Q.Z., Department of Computer Science of the University of Pittsburgh, IBM Quantum through the Chicago Quantum Exchange, and the Pritzker School of Molecular Engineering at the University of Chicago through AFOSR MURI [FA9550-

21-1-0209] to J.L., ARO [W911NF-23-1-0077], ARO MURI [W911NF-21-1-0325], AFOSR MURI [FA9550-19-1-0399, FA9550-21-1-0209, FA9550-23-1-0338], NSF [OMA-1936118, ERC-1941583, OMA-2137642, OSI-2326767, CCF-2312755], NTT Research, Packard Foundation [2020-71479], and the Marshall and Arlene Bennett Family Research Program, U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers to L.J., Simons Collaboration on Ultra-Quantum Matter [651442], and the Simons Investigator award [990660] to X.C.W.

Author contributions

Q.Z. proposed the initial study of QNN training convergence, inspired by prior works of J.L., L.J. and their collaborators. B.Z. and Q.Z. identified the polynomial convergence phenomena. B.Z., J.L., L.J. and Q.Z. identified the dynamical transition and developed the general theory and the LV model. J.L. realized the connection to the Schrodinger equation and gap closing, while X.C.W. developed the classical Hamiltonian formulation of the LV model and the statistical physics interpretation, with input from all authors. B.Z. and Q.Z. proposed the restricted Haar ensemble, with inputs from all authors. J.L. and B.Z. performed the experiments on IBM devices. B.Z. performed the detailed derivations and numerical simulations and generated all figures under the supervision of Q.Z., with inputs from all authors. Q.Z. and B.Z. wrote the manuscript with contributions from all authors.

Competing interests

The author declares no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-53769-2>.

Correspondence and requests for materials should be addressed to Quntao Zhuang.

Peer review information *Nature Communications* thanks Marco Cerezo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024