nature portfolio

Peer Review File

Dynamical transition in controllable quantum neural networks with large depth

Corresponding Author: Professor Quntao Zhuang

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

A version of this paper was originally rejected for publication by Nature Communications, however that decision was reconsidered after appeal by the authors.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
See attached report.

Reviewer #2

(Remarks to the Author)
Dear Editor,

The manuscript by Bingzhi Zhang and collaborators discusses the dynamical regimes of the learning process for a class of quantum neural networks. The Authors propose three viewpoints to clarify these dynamical regimes: a generalized Lotka-Volterra model to capture the late time (and deep) learning dynamics, a statistical phase transition perspective, and a restricted Haar average interpretation. A quantum demonstration on IBM devices corroborates these results.

Brief Summary and main contribution
This study is based on prior works of the Authors and others (refs 28-32 of the paper): a learning problem is given in terms of a cost functon(cf. Eq (1)), which can be rephrased, in the limit of small eta, the gradient descend rate, as a coupled equation in terms of the total error (epsilon) and of the quantum neural tensor kernel (QNTK) K.

If the target value lies in support of operator O, the optimization problem is solvable, and the total error decays exponentially in time (K saturates). On the other hand, if the target value lies beyond this support, the problem is not solvable, and the total error saturates. The total error decays exponentially at the critical point (when the optimal value is the boundary).
Under certain assumptions, the system dynamics is rephrased as a Lotka-Volterra (LV) equation that is analytically solvable and confirms these three regimes. At the same time, these regimes are read out of the Hessian, which is gapless at the critical point and gapped otherwise.
The assumption for the LV equations is later justified using unitary ensembles for the specific choice of a projector operator.

The main novelty is the interpretation of the learning dynamics as a learnability transition in terms of where the target value lies compared to the operator's support.

Recommendation
Overall, the manuscript discusses interesting topics, connecting ideas of learnability transition and the dynamics of quantum neural networks. The detailed analysis of the supplemental information is a figure of merit.
For these reasons, the manuscript should be published in some form.
Regretfully, its present version does not meet the stringent criteria for Nature Communications.
In a nutshell, there are three major criticisms: 1. Compared to the literature, the manuscript's originality seems mostly incremental; 2. Some aspects of the setup are unphysical, including the choice of parameter phase diagram for any

transition. As a result, the transition's existence is therefore unclear; 3. The presentation is cumbersome and difficult to follow, even for experts in the field.
All these points must be resolved clearly and effectively if the Authors insist on resubmitting in Nature Communication.

Criticisms
(1) First, comparing this manuscript with Refs [28-32], I see only incremental value. For instance, most of the techniques were already presented in [28-29], where various aspects of the dynamical regimes were also identified. How does this manuscript substantially differentiate from the previous literature?

(2) A more substantial problem is: What do we learn about physics? Is this setup relevant to any practical application?

Given an operator O, we know the upper and lower bound of the operator, so it is artificial and unphysical to look for critical behavior outside the support of O. For instance, I know that a projector is bounded between 0 and 1. Why should I look for O_0 outside this region?
Hence, de facto, for any practical situation, the system is only in the kernel frozen phase (which, coming back to (i), was extensively discussed in the earlier works [28-29]).
The Authors should clarify this is a physically relevant/irrelevant setup.

The limit case of the target O_0 being at the boundary has a nice analogy with the one-dimensional classical Ising model. There, the finite size scaling collapses the transition at zero temperature. Only that point has features that are "critical," but at any epsilon away, the system is directly drifted (in the renormalization group sense) outside the criticality. This is different in true phase transitions: a thermodynamic parameter (system size/ circuit depth, etc.) identifies a region over which the correlations are, de facto, critical.
The model discussed by the authors doesn't have any thermodynamic parameter; hence, by analogy, we can say the model does not have a phase transition (in line with the fact that this is a 0+1 dimensional).

If the Authors believe this is a true transition (in the above sense), they should clarify the mechanism behind this phase transition (since they claim this is not a symmetry-breaking one). Otherwise, they should smoothen their discussion and soften their statements.

(3) The paper is difficult to read. One must go back and forth many times in the text (for instance, lambda constant, used for the Lotka-Volterra equation), which is justified only later in the unitary ensemble section. Another example: the model employed in most of the figures is only discussed in methods but not mentioned/clarified in its essential features in the main text.
A careful revision is needed to simplify the reading and make it more followable.

Only by successfully resolving the above three points can the manuscript be worthy of publication in Nature Communications. Otherwise, I will happily advise the manuscript to be published in more specialized journals after the necessary clarifications are made.

There is a series of minor comments:
(i) The experimental results have no error bars, no discussion about error mitigation, and seem inconclusive (fig. 5 seems to have only data that are described by a power-law decay). Can the authors clarify these issues and present the data on a log-log scale (where possible, power laws would be more evident)?
(ii) The theorems seem more "Remarks" (they don't have the rigor to be called theorem). A rephrasing is suggested.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
We sincerely thank the authors for carefully addressing our comments. It really shows that the authors went the extra mile to improve their work.

We were extremely surprised to see how some of the new numerical results show that the theory predictions hold beyond the scope of the theorems. Moreover, many interesting open research questions arise from the cases beyond the scope of the work (finite depth, HVA, etc), and we look forward to seeing if the authors can tackle those scenarios as new phenomena might emerge there.

In our opinion this new version of the work considerably improves over its previous version, which we already had found extremely interesting. As such, we are happy to recommend for publication.

Reviewer #2

(Remarks to the Author)

Dear Editor,

The revised manuscript by Bingzhi Zhang and collaborators has shown some improvements based on the previous reports by the referees. While certain aspects are now better explained, the reply to my main criticisms is lacking in substance.

1. The authors claim that the main addition to the previous literature is identifying a novel phase transition.
2. They state that this is not a conventional transition and leave the exploration of the nature of this critical point for future work.
3. They note that the paper is well-written according to the other referee, and in response to my comments, they have added minimal mentions of the models discussed.

Main Criticisms:

First and foremost, I thank the authors for taking the opportunity to consider my criticisms and comments and for the improvement made.
Despite these revisions, the manuscript still fails to address several critical points I raised in my previous review. The key issues are as follows (it is a mixture/rephrasing of the three criticisms in my last report).

(1). Lack of Qualitative Understanding: The authors have not provided a qualitative understanding of the transition. This central point is crucial for the paper's impact and comprehension. Leaving the explanation of the critical point for future research does not suffice.

(2) Finite Size Scaling: To characterize a genuine phase transition, finite-size scaling is necessary [e.g., arxiv:1101.3281, or the classic Cardy, J. (1996). Scaling and Renormalization in Statistical Physics (CUP)]. Phase transitions are valid and well-defined in the thermodynamic limit. The data presented in the manuscript does not indicate that the system size plays any significant role. There is no evidence of data collapse with increasing system sizes, nor is there a substantial analysis of the universality properties of the system.

(3) Technical Nature and Lack of Clear Storytelling: The manuscript remains highly technical, which may interest specialists in the field. However, it lacks a clear storytelling and physical understanding of the mechanism behind the phase transition beyond the mere mathematical aspects. This significantly limits its accessibility and impact.

Given the criticisms above, I cannot recommend this work for publication in Nature Communications. The paper does not provide the necessary qualitative insights or robust evidence required to convincingly demonstrate the claimed phase transition. The explanations remain incomplete, and the overall narrative does not effectively communicate the significance and mechanics of the findings.

I do not doubt the relevance and importance in the field of application, and I emphasize the figure of merits highlighted in my last report. Nevertheless, without these points, I find the interest too narrow and more suitable for publication in a specialized journal (e.g., NPJ Quantum Information). If the Authors provided a further revision substantially addressing these points, I would gladly reconsider my recommendation.

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
We have reviewed the changed to the manuscript. Our previous opinion of the work was already positive, and we think that the authors have done a great job of replying to some of Referee 2's comments.

In particular, the new understanding of the critical point has improved the work significantly as it truly provides a more precise characterization of the observed phenomenon. In addition, the results in Fig 4 were very illustrative, in particular panel (b) which shows the universal in L behavior.

Minor comment: Fig 4 is quite busy, and it took me a while to find the labels (a), (a1) and (b) as there were added inside of the plot. Could the authors move them outside of the plots? Note that this style is already used in Fig 3, so it would also keep consistency.

In view of these new changes, we keep our strong support for publication.

Reviewer #2

(Remarks to the Author)
Dear Editor,

I want to thank the Authors for their substantial revisions and improvements in response to my previous comments. The text is generally more transparent, and the essential contributions are more evident than the earlier version.

That said, I still find the analysis of the nature of the transition somewhat (and unnecessarily) confusing.

In their response and revision, the Authors clarify that the phenomenon they observe is not a phase transition but rather a dynamical transition. We agree with this point.

However, the subsequent analysis of data collapse is puzzling. The Authors arbitrarily choose to denote "system size" as the number of variational parameters L, which scales as the number of qubits n multiplied by the number of layers in the circuit T. They demonstrate a data collapse in the gap when rescaled by L = nT.

For an inexperienced reader, this would appear to be an analysis of a phase transition, which the Authors explicitly exclude, leading to a confusing argument. This contrast needs to be addressed.

First, let me remark that it is universally agreed that n, the number of qubits, is the system size. (See also https://arxiv.org/pdf/1709.07461)
Since the phenomenology is observed at n <= 8, this is clearly not a *phase* transition, as it is visible with a few qubits, reiterating what both I and the Authors agree upon.
The linear dependence of L is equivalent to a linear dependence of T. If I understood correctly the text and supplemental information, this dependence of the gap on T is explained by Eq. 20. Given that the non-Hermitian evolution is approximately \exp(-\eta T M) for some operator M = H/T, the gap will naturally depend linearly on the circuit depth T.
This reasoning would give a neat and complete perspective on the problem: the transition is explainable within dynamical systems (as the Authors point out), and the circuit depth role from the analysis, I should consider M=H/T (isolating the operator that serves as "Hamiltonian").

(A remark for context: this type of transition seems to parallel the physics of non-hermitian systems and "exceptional points." These critical behaviors emerge at a finite size (finite number of qubits) and are related to the gap of the evolution operators.)

In summary, I suggest removing the analysis with the arbitrary finite-size scaling L from the text, or at least from the main body. The discussion of the dynamical transition should be shortened, as it is adequately explained within the mathematical framework of dynamical systems. The reference to a fictitious "system size" L should be omitted to maintain consistency. (The system size, again, should be the number of qubits.)

Given the manuscript's current state, I find it not yet suitable for publication in Nature Communications due to the potential sources of confusion present in the revised text. However, if the Authors address these points, the final version will be suitable for publication in the journal.

Version 3:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
Dear Editor,

I thank the Authors for the further improvement and considering my comment.
The revision steered out the criticisms. I believe this version is now ready for publication in Nature Communication.

RESPONSE TO REVIEWS

NCOMMS-23-62482-T
"Dynamical phase transition in quantum neural networks with large depth"
submitted to Nature Communications

by Bingzhi Zhang, Junyu Liu, Xiao-Chuan Wu, Liang Jiang and Quntao Zhuang

We thank both reviewers for reviewing our manuscript entitled "Dynamical phase transition in quantum neural networks with large depth". The comments and suggestions have greatly helped us to improve the presentation of the paper. We have also done significant revision and believe that the paper can now be accepted by Nature Communications

Reviewer 1

Reviewer 1 agrees "QNTKs present a refreshing perspective to go beyond random initialization and capture long-time training dynamics", and found our results "well-written and clear". Due to the field is "mostly heuristics-driven", our work is "extremely important" to have theoretical results. Hence, Reviewer 1's impression on our work is "very impressive", and willing to support it to be published in *Nature Communications* with the comments being addressed.

In particular, Reviewer 1 has detailed comments listed as below.

Summary of Reviewer 1's comments

1. Reviewer 1 suggests us to present numerical results for larger problem size to straighten our results, and also show how some quantities (such as λ) depending on the system size.

2. Reviewer 1 asks us to clarify the type of quantum neural networks that our results holds for. Especially, Reviewer 1 has two requests. First, Reviewer 1 expects our comments on the dynamics of deep but non-controllable circuits such as Hamiltonian Variational ansatz (Ref. [5, 13]). Second, if the results do not hold for those deep but non-controllable circuits, Reviewer 1 suggests us to change the title and wording of our work.

3. Reviewer 1 suggests us to clarify how deep the circuit is required for our results to be valid, and any connection to the necessary depth for overparameterization of quantum neural network (see Ref. [6]).

4. Reviewer 1 suggests us to make a comment about about the assumption $d \gg 1$, and whether the analytical results can be derived for non scaling d.

5. Reviewer 1 suggests us to make notations consistent.

6. Reviewer 1 suggests us to change font size and color schemes of Fig. 1 such that it can be appropriate for reading on a printed version.

7. Reviewer 1 suggests us to discuss how the results would change if loss function is a summation of error terms.

8. Reviewer 1 suggests us to discuss how the results change for linear loss function $\mathcal{L}(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|O|\psi(\boldsymbol{\theta})\rangle$.

9. Reviewer 1 suggests us to comment on the implications of our results for design of quantum machine learning loss functions.

10. Reviewer 1 suggests us to add statements to emphasize why/how our work can be applied to the those who utilize quantum neural networks in research, such that our work can be more appealing to a wider community.

11. Reviewer 1 recommends us to slightly rewording the statement that QNNs generalize classical NNs to avoid confusion for non-expert readers.

12. Reviewer 1 asks us to give a more precise definition on the number of layers L before Eq. (1).

13. Reviewer 1 asks us to given a definition of d before Eq. (9).

14. Reviewer 1 suggests us to explain the definition of average when mentioning "unitary ensemble theory" in the first paragraph.

15. Reviewer 1 suggests us to add some additional steps and information in the proof in SI to help readers to access the nice derivations.

Our Reply

We thank Reviewer 1 for the positive feedback and the suggestions. We have fully revised accordingly, as we detail below.

1. We thank Reviewer 1's suggestion. We have now enlarged the system size in our simulations (Fig. 2 to 8 qubits, Fig. 3 to 6 qubits and Fig. 9 top panel to 5 qubits), and still see good agreement with our theory. We also add results in Fig. 9 to compare the scaling of dQNTK λ with system Hilbert space dimension d at both late-time and initial-time. We see that our theory results hold for different system sizes.

2. We thank Reviewer 1's suggestion. We mainly study the random Pauli ansatz in the paper, which is universal and has full control over the Hilbert space with enough depth. Our results also hold for other universal and full-control ansatz such as hardware efficient ansatz [3], and we perform our experiment in the main text (see Fig. 8) and another simulation in SI (see Fig. 5) on hardware efficient ansatz.

   The dynamics for deep and non-controllable circuits like HVA would also be very interesting, and to fully understand its dynamics requires a detailed investigation on the dynamics of error, QNTK, and relative dQNTK which is out of reach for our current knowledge, thus we would like to leave it for future work. Therefore, we follow Reviewer 1's suggestion to revise all the wording of QNN in our manuscript to be "controllable QNN". The title of our manuscript has now been changed to

   > Dynamical phase transition in controllable quantum neural networks with large depth

3. For our theory part, the main assumption is that λ is a constant at late time evolution, and for that to hold, we developed the ensemble theory based on restricted Haar ensemble, which holds when the depth D (now we call depth D and number of parameters L) is poly(n) in the number of qubits n.

   In our numerical results, we have chosen L sufficiently large to guarantee the convergence to the targeted state, and also to see very sharp transitions. Note that in general, the ground state of certain observable may require exponential-depth circuit to prepare, and therefore QNN with poly(n) depth will not be able to converge to $O_{\min}$ in general.

   Despite the above requirements, we are able to numerically identify the phase transition for a QNN with limited depth $L = n$ which is far from over-parametrization. We found that by choosing the state-dependent critical value of O_0 (which is due to limited expressivity), we can still recover the three different phases, as we show in Fig. 7 in the main text (also appended as Fig. R2 below). The reason is that we only require $\lambda(t)$ converge to a constant when t is large in a particular QNN training dynamics, therefore does not require sample fluctuation to be small (while our ensemble theory adopts sample fluctuation small to suggest time-fluctuation to be small); moreover, we also do not require convergence to the exact ground state—choosing a large depth is merely a convenience for us to easily identify the threshold of the transition and getting rid of finite-size effect in studying the phase transition.

   The general theory to explain the phase transition for circuits of finite depth is beyond the focus of current manuscript and we pose that as an interesting open question.

   We have included the analyses as a new section, which we append here

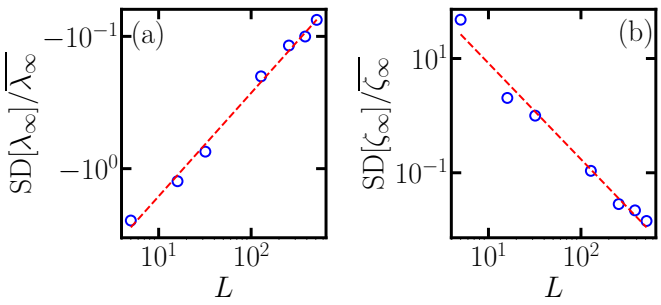

Figure R1: **Late-time sample fluctuations.** The standard deviations normalized by mean for the relative dQNTK λ_∞ (a) and dynamical index ζ_∞ (b) are plotted versus the number of parameters L. Red dashed lines represent power-law fitting results. Here the RPA is applied on $n = 5$ qubits with different L parameters (via tuning number of layers D). The observable is a state projector and the target value is $O_0 = 1$.

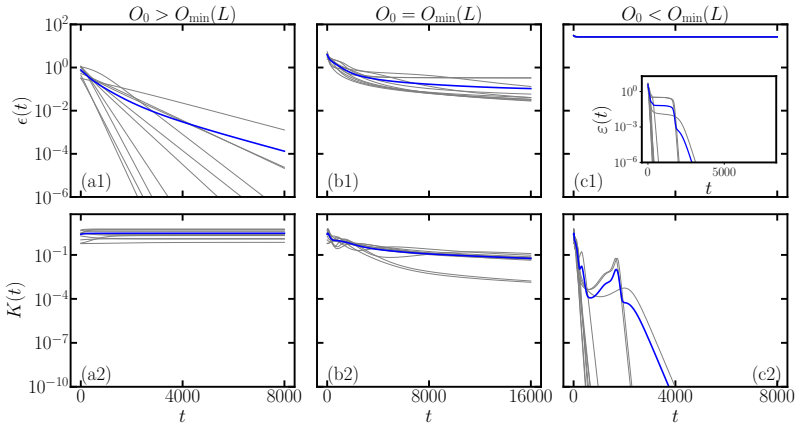

Figure R2: **Dynamics in limited-depth QNNs in the example of the XXZ model.** All notations share the same meaning as in Fig. 2. The critical point $O_{\min}(L)$ for such QNNs depends on L and has sample fluctuations. Here random Pulai ansatz (RPA) consists of $L = 6$ variational parameters ($D = L$ for RPA) on $n = 6$ qubits, and the parameter in XXZ model is $J = 2$.

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) [4]. 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 can be identified when the circuit is not controllable—either the ansatz is not universal [5, 13] 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. R1, 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 $\overline{\zeta_\infty} \to 1/2$ are shown in Fig. 9 (b),(c) 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 ensemble average for λ_∞ (see Eq. (9)) and ζ_∞ (see Eq. (17)) has negligible effects (see SI 11). Similar results for other observables, e.g. XXZ model, are shown in SI 11. 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 phase transition, as we show in the below example.

To our surprise, in Fig. R2, we find that the phase 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 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 initialization. We suspect this may be caused by the training converging to different local minimum traps [1, 14]. 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 phase transition to persist. Moreover, the example is also not within the applicability of Lemma 2. At late time, the relative dQNTK λ still converges to a constant 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 overparametrization [6].

4. We thank Reviewer 1's suggestion. Our assumption for dimension $d \gg 1$ is merely for a simplification on the expression of unitary ensemble average QNTK, dQNTK and relative dQNTK to make their scalings clear and explicitly. In fact, after careful checking, Lemma 2 does not require dimension $d \gg 1$ at all and we have revised it. This is why all of our results can work for two qubits examples in the previous version of manuscript.

   For the methods part, Theorem 3, Lemma 4 and Theorem 5 have simple expressions when $d \gg 1$, but all of them have full formula in the SI 10 and SI 11. Despite being much lengthier, those results are sufficient to support the phase transition. In the Methods, we have added clarifications

   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 finite d, just much more lengthy.

   .....

   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 apply to any finite d.

   .....

   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 apply to any finite d.

   To study the influence of dimension d, we have added subfigures (d)(h) in Fig. 9 (which is appended below as Fig. R3) to explicitly show that our results hold for $2, 4, 6, \cdots$ number of qubits. Finite-size effect of the phase transition is really about whether L is deep enough for each d, while the results can work for any size d.

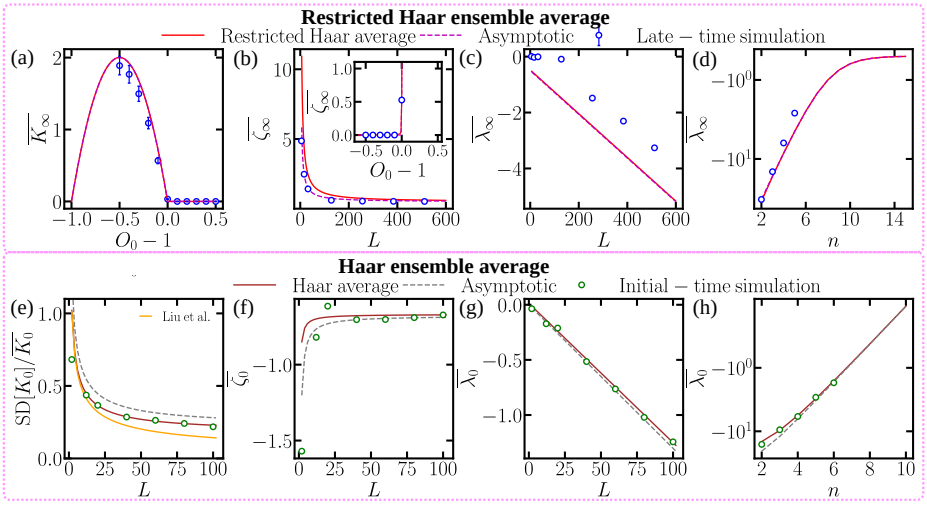

Figure R3: **Ensemble average results under restricted Haar ensemble (top) and Haar ensemble (bottom).** In top panel, we plot (a) $\overline{K_\infty}$ versus O_0 with $L = 512$ fixed, (b) $\overline{\zeta_\infty}$ versus L, $\overline{\lambda_\infty}$ versus (c) L and (d) n with $L = 512$ at late time in state preparation. We set $O_0 = 1$ for (b) and (d), and $O_0 = 5$ for (c). Blue dots in top panel (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 Eq. (250), (307), (273) in SI 11. Magenta dashed lines represent asymptotic ensemble average with restricted Haar ensemble in Eq. (39), (40), (41) which overlap with the exact results (red solid). The observable in all cases is $|\Phi\rangle\langle\Phi|$ with $|\Phi\rangle$ is a fixed Haar random state. In the inset of (b), we fix $L = 512$. In bottom panel, we plot (e) fluctuation $\mathrm{SD}[K_0]/\overline{K_0}$ 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 exact ensemble average with Haar ensemble in Eq. (235), (174), (114) in SI 10. Gray dashed lines represent asymptotic ensemble average with restricted Haar ensemble in Eq. (49), (43), (44). 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 [9].

5. We have unified the notations. We also want to point out that only the variational parameters $\boldsymbol{\theta}(t)$ explicitly depends on training time and get updated via gradient descent. The other quantities, $\epsilon(\boldsymbol{\theta}), \varepsilon(\boldsymbol{\theta}), K(\boldsymbol{\theta}), \mu(\boldsymbol{\theta})$ are evaluated with respect to $\boldsymbol{\theta}$ at every step, and for simplicity of notation, we only present those quantities depending on training time t directly and omitting $\boldsymbol{\theta}$.

6. We enlarge the font size and adjust the colors in Fig. 1 to improve visualization.

7. We thank Reviewer 1's question. When there are multiple terms in the loss function involving multiple data, the dynamics would become more complicated as there are multiple errors and kernels need to be tracked, which also leads to more fruitful dynamical phases. We leave it as a future ongoing work and added a comment in the discussion

   It is also an open problem how our results can generalize to the multiple data case.

8. We thank Reviewer 1's question. For linear loss function, we can also analyze the corresponding training dynamics with the same formalism we developed for analyzing the dynamics for quadratic loss. Specifically, we still assume $\lambda(t)$ converging to a constant λ which can also be justified from the restricted Haar random ensemble perspective as linear loss only leads to ground state of observable. Thus we can derive the corresponding dynamical equations to describe residual error and QNTK, where both of them undergo an exponential decay (see Fig. R4). We present a detailed study in SI 7. Note that as there is no target value in linear loss function, the dynamics is fixed and no phase transition exists.

   As both linear and quadratic loss can be utilized to search for ground state, we find a speedup in convergence from quadratic loss function compared to the linear one via tuning the target value (see Fig. R4). We have updated the abstract, introduction and discussion to highlight this speedup. We have included the analyses as a new section, which we append here.

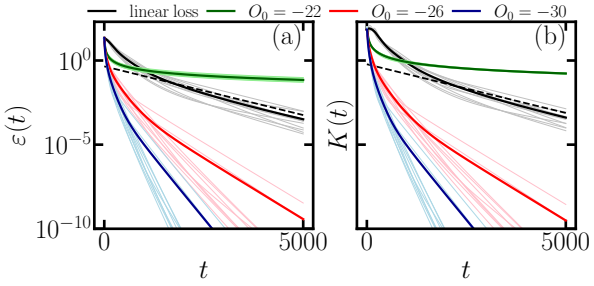

Figure R4: **Dynamics in QNN in the example of XXZ model with different loss functions.** In (a) and (b), we show the dynamics of residual error $\varepsilon(t)$ (equals to total error $\epsilon(t)$) and QNTK $K(t)$ optimized with linear loss function (black solid) and quadratic loss functions with different 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 phase. Black dashed line indicates the exponential decay rate of the theoretical result in Eq. (5). 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 $J = 2$.

Speeding up the convergence

While the phase 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

$$\mathcal{L}(\boldsymbol{\theta}) = \langle \hat{O} \rangle \,. \tag{1}$$

While the linear loss function is widely used in variational quantum eigensolver [3, 13], 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 solution.

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 $\varepsilon(\boldsymbol{\theta}) = \langle \hat{O} \rangle - O_{\min}$. Via a similar approach (see details in SI 7), we have the dynamical equations for the error $\varepsilon(t)$ as

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

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

$$\delta K(t) = -2\eta\mu(t) + \mathcal{O}(\eta^2), \tag{3}$$

with $\mu(t)$ being the dQNTK defined in Eq. (6). One may notice that the only difference compared to Eqs. (7) and (8) is the missing of $\epsilon(t)$ on r.h.s. 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

$$\begin{cases} \partial_t \varepsilon(t) &= -\eta K(t), \\ \partial_t K(t) &= -2\eta\lambda K(t). \end{cases} \tag{4}$$

Unlike the generalized LV model in Eqs. (10) for the quadratic loss case, here the dynamics of $K(t)$ is self-consistent, whereas the dynamics of $\varepsilon(t)$ is fully determined by $K(t)$—the kernel-error duality is broken. Eqs. (4) can be directly solved as

$$2\lambda\epsilon(t) = K(t) \propto e^{-2\eta\lambda t}. \tag{5}$$

> Both $\epsilon(t)$ and $K(t)$ decay exponentially at a fixed rate $\propto \lambda$. In Fig. 5, we present the numerical simulation result (black solid), and observe a good agreement with the theory (black dashed) from Eq. (5).
>
> 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. (5), the linear loss function provides an exponential convergence with the exponent $2\eta\lambda$ being a constant. 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. 5), 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 phase*, where the residual error $\varepsilon(t)$ decays exponentially with the exponent $2\eta\lambda R$ (see Eq. (20)). 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. 5 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.

9. We thank Reviewer 1's suggestion. As we have explained in answering item 8, the results provide a speedup in convergence if we stay within the frozen-error phase.

   The dynamical phase transition induced by target value discovered in this manuscript implies that for those who need to optimize a quadratic loss function in quantum machine learning, if the target value lies within the eigen-spectrum (excluding minimum and maximum eigenvalues) of the corresponding observable, it will present a fast-convergence in optimization. On the other hand, it the target value happens to be either minimum or maximum eigenvalue, a straightforward strategy is to relocate the target value outside the spectrum reaching the same solution state while enabling exponential speedup. From a more general perspective, for the design of loss function in quantum machine learning tasks, it can be meaningful to consider target value regions even though without support on the observable as it leads to the same solution but may provide additional advantage in convergence.

10. Indeed, as explained in item 8 and item 9, our work has practical implications for those who work with QNNs in applications.

    In the abstract we have added

    > Compared with the linear loss function, we show that a quadratic loss function within the *frozen-error phase* enables a speedup in the training convergence.

    In the introduction we have added

    > 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 phase*. While our theory analyses assume the large-depth limit, the dynamical phase transition is also numerically identified in QNNs with limited depths. The theory findings are verified on IBM quantum devices. Our results imply that designing the loss function properly is important to achieve fast convergence.

    In the discussion we have added

    > In practical applications, the dynamical phase transition guides us towards better design of loss functions to speedup the training convergence.

11. We revise the statement on QNNs to "analog to classical neural networks that are crucial to machine learning" to avoid any confusion.

12. Indeed, for different ansatz, what a layer means is different. To avoid confusion, we now use L directly as the number of variational parameters for the QNN ansatz, and also define D as the number of layers of corresponding circuit. For random Pauli ansatz (RPA) we mainly considered in the main text, we have $L = D$ while for hardware efficient ansatz, have $L = 2nD$. We revise the statement in the main text to clarify the difference between L and D as following.

> A D-depth QNN is composed of D layers of parameterized quantum circuits, realizing a unitary transform $\hat{U}(\boldsymbol{\theta})$ on n qubits, with L variational parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)$. The gate configuration of each layer varies between different circuit ansatz (see Methods for examples).

13. We add the definition of $d = 2^n$ as the Hilbert space dimension of the system.

14. We revised the sentence to

> ...to model each realization of the QNN $\hat{U}(\boldsymbol{\theta})$ as a sample from a ensemble of unitaries and consider ensemble averaged values to represent the typical case,

15. We have done thorough revision of the SI and added all necessary explanations in the SI to help readers understand the derivations.

Reviewer 2

Reviewer 2 summarized our work as the "interpretation of the learning dynamics as a learnability transition in terms of where the target value lies compared to the operator's support" and our work "discusses interesting topics, connecting ideas of learnability transition and the dynamics of quantum neural networks". Reviewer 2 asks us to resolve the comments to meet the criteria for *Nature Communications*. In particular, Reviewer 2 has detailed comments as we summarize below

Summary of Reviewer 2's comments

Reviewer 2 wants us to resolve three major points (summarized and broken down to four items 1-4 below) in order to be published in Nature Communications.

1. Reviewer 2 suggests us to make a clarification on our contribution compared to prior works, Refs. [7–10, 12].
2. Reviewer 2 suggests us to clarify the physics and practical applications for the choice of target values.
3. Reviewer 2 doubt whether the transition we identified is a true phase transition and clarify the mechanism behind this phase transition (since it is not a symmetry-breaking one). Otherwise, we should smoothen the discussion and soften their statements.
4. Reviewer 2 suggests us to improve the presentation and readability of the manuscript.
5. Minor things
   (a) Reviewer asks for error bar of the experimental results, discuss about error mitigation and asks us to distinguish between power-law and exponential on log-log scale.
   (b) clarify if the theorem is rigorously proven.

Our Reply

We thank Reviewer 2 for the suggestions that have greatly helped improving the presentation of our manuscript. We have fully addressed all points and believe that our manuscript can be published in *Nature Communications* now.

Before we start, we point out a possible typo of Reviewer 2 when summarizing our results in the report—the total error decays polynomially at the critical point instead of exponential.

Now we summarize the significance of our results in very succinct words, before the point-to-point reply.

- Our paper is the first to identify a novel phenomenon—the phase transition in training dynamics of QNNs, which is never found in Refs. [7–10, 12].
- We provide non-perturbative theory to explain the phase transition. The theory adopts nonlinear dynamical equations, statistical physics interpretation and unitary ensemble theory.
- The discovered phase transition provides new ways of designing loss functions to speed-up convergence of QNN training.

- The phase transition can potentially be generalized to more complex machine learning tasks such as classification and regression that triggers future studies.

From the above significance, we believe that our manuscript has met *Nature Communications*' scope of "representing important advances of significance to specialists within each field."

Below, we provide detailed reply to the other concerns of Reviewer 2.

1. Indeed, it is important to clarify our novelty compared to previous works [7–10, 12]. An obvious clear cut is: none of the previous works ever found the phase transition.

   Previous works look at a corner of the phase diagram and lack the non-perturbative theory. Below, we summarize the findings of other works: Ref. [10] introduces the framework for the Quantum Neural Tangent Kernel (QNTK). This paper derives several results under the assumption that the kernel remains static, which is limited to the 'frozen-kernel' phase. Numerical observations of kernel scaling are discussed, but the mechanisms underlying this frozen kernel phenomenon are not understood. In subsequent work, Ref. [9] introduces assumptions about k-designs to derive conditions that maintain the kernel in a frozen state during the initialization of variational quantum circuits. These conditions apply in the perturbative limit, characterized by large circuit depths and extensive dimensions of the Hilbert space, making the kernel predictable through analytic formulas. However, the long-term behavior of gradient descent dynamics remains unknown for these circuits, particularly in regimes where the kernel changes non-perturbatively over time. Further studies by Ref. [8] explore the relationship between 'laziness' training and quantum barren plateaus, while Ref. [12] extends the QNTK framework to address problems with global symmetries. Ref. [7] establishes the foundational differences between overparametrization in classical neural networks and quantum neural networks, which is a different topic to what is considered in the current manuscript.

   We added a paragraph at the end of the discussion to clarify it

   > Finally, we clarify the difference of our results to some related works. Firstly, while existing works [7–10, 12] on the quantum neural tangent kernel provide a perturbative explanation of gradient descent dynamics that fails to uncover the phase transition, our work uncovers the phase transition and formulates non-perturbative critical theories about the phase 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 of the residual training error and quantum neural tangent kernel, 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. Through this analysis, we can model our critical phenomena using the language of scale invariance and second-order phase transitions. 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 [9].

2. After reminded by the reviewer's questions, we realize that the application of our results are not adequately emphasized in our current manuscript. We have therefore revised accordingly.

   We first clarify some specific questions

   - `Given an operator O, we know the upper and lower bound of the operator, so it is artificial and unphysical to look for critical behavior outside the support of O.`

     Setting $O_0 < O_{\min}$ still allows the QNN to produce the ground state. In fact, when $O_0 < O_{\min}$, the loss function $(\langle O\rangle - O_0)^2 \geq (O_{\min} - O_0)^2$ is still minimized when $\langle O\rangle = O_{\min}$—the ground state output. Non-achievable value will essentially converge the system to the boundary.

     > Interestingly, 'shooting for the star' will not make one reach the star, but will make one reach as high as you can.

   - `Why should I look for O_0 outside this region?`

     By setting O_0 outside the region and as far away as possible will actually speed up the convergence towards the boundary of the region! The practical importance of our results is to reveal the frozen-error phase that allows faster convergence towards the boundary, and we present an example explicitly in Fig. R4. We have now added an entire section titled "Speeding up the convergence" to explain this point, which is also appended below Fig. R4 in the reply.

Interestingly, 'Shooting for the star' will allow faster convergence towards the optimal.

- `Hence, de facto, for any practical situation, the system is only in the kernel frozen phase`
  This statement follows from reviewer's confusion and is incorrect. As we explained, setting target far away allows faster convergence.
  We have now added an entire section titled "Speeding up the convergence" to explain this point, which is also appended below Fig. R4 in the reply.

As for the practical implications of our work, let's begin with the question:

> Given an operator O and the value of ground state energy $O_{\min}$, how to prepare the ground state in the fastest way?

Our results show that setting O_0 as small as possible allows the fastest convergence—**it is important to note that setting $O_0 < O_{\min}$ to be unachievable still allows the QNN to converge to the ground state, and even faster!** If one set $O_0 = O_{\min}$, one will suffer from slow polynomial convergence.

We have now added an entire section titled "Speeding up the convergence" to explain this point, which is also appended below Fig. R4 in the reply. There, we also show that the quadratic loss function allows faster convergence than a linear loss function. Besides that, to clarify the practical importance of our results. In the abstract we have added

> Compared with the linear loss function, we show that a quadratic loss function within the *frozen-error phase* enables a speedup in the training convergence.

In the introduction we have added

> 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 phase*. While our theory analyses assume the large-depth limit, the dynamical phase transition is also numerically identified in QNNs with limited depths. The theory findings are verified on IBM quantum devices. Our results imply that designing the loss function properly is important to achieve fast convergence.

In the discussion we have added

> In practical applications, the dynamical phase transition guides us towards better design of loss functions to speedup the training convergence.

The mechanism behind the phase transition

3. We want to clarify that our intention is to understand the phenomenon as a phase transition in a (0+1)-dimensional quantum mechanical system, driven by a single parameter O_0. There is no order parameter (that transforms under symmetry), and the other critical exponents are absent. It goes beyond the conventional Landau-Ginzburg paradigm, and is not restricted by the Mermin-Wagner theorem, which is one of the major focuses of the community. In some sense, this is similar to the gap-closing transition of topological insulators, where the transition is driven by fermion mass. We stress that distinct topological phases are well-defined in 0+1 dimensions, and related classification can be found in e.g. [2, 11]. Some of these topological phase transitions also have a single critical exponent associated with the gap, similar to what we find in our results.

   This transition in late-time behaviors falls outside the conventional framework of statistical physics. However, we firmly believe it is essential for understanding quantum neural networks. A comprehensive understanding (from many-body physics point-of-view) of the mechanism behind this phase transition demands substantial theoretical efforts, which we aim to explore in future studies. Loosely speaking, we have devised an 'Ultraviolet model' based on the effective Hamiltonian derived from gradient descent dynamics, alongside an 'Infrared description' based on the generalized Lotka-Volterra equation. There may exist a renormalization group flow connecting these models. In addition, the full set of universal data associated with the critical point also deserves to be understood.

   We have revised the paragraph in the main text

In summary, we have presented compelling evidence supporting the interpretation of the dynamical phase transition as a continuous gap-closing transition in a quantum mechanical system described by the time-independent effective Hamiltonian H_∞. It goes beyond the conventional Landau-Ginzburg paradigm, and is not restricted by the Mermin-Wagner theorem. We also want to mention that reaching the truly infinite-time limit poses challenges both numerically and experimentally. In our estimation of the correlation function in SI 2, we rely on taking the subtle limit $\tau \gg t \gg 1$ and the fluctuations being small. A comprehensive understanding of the mechanism behind this unconventional phase transition and the complete set of universal data associated with the critical point demands substantial theoretical efforts, which we aim to explore in future studies.

4. Our paper is organized according to the logic that the presentation of the major results is prioritized on a high level, and details and supporting materials are presented in later sections. From reading of papers published in *Nature Communications*, we believe that such a presentation is preferable to the majority of readers of *Nature Communications*. Eventually, each way of presentation has pro and cons. Our choice can help readers to understand the dynamical transition of QNN dynamics better, while experts such as the reviewer may find some details need to be looked up in later sections. If we organize all things in a linear fashion, it will substantially delay the introduction of the major results. It is a matter of preference which way of presentation is better—for example Reviewer 1 comments that "Overall, the manuscript is well written and clear".

   To balance between the two choices, we now added a few sentences to provide more details in places that we refer to other contents.

   > Our major finding is that when the circuit is deep and controllable, the QNN dynamics exhibit a phase 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).

   and

   > With large depth $D \gg 1$ and full control, QNNs are commonly modeled as a random unitary [4, 9, 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

5. Minor things

   (a) We appreciate Reviewer 2's suggestion. We add shaded areas in Fig. R5 as error bars for the experimental data. Indeed the results agree well with our theory, up to a noisy correction as we explain in SI 9. We further provide a log-log plot of residual error $\varepsilon(t)$ with different O_0 in Fig. R6 to address Reviewer 2's question and we can clearly see that for $O_0 = O_{\min}$, it follows a polynomial decay from subplot (b) while the others decays exponentially.

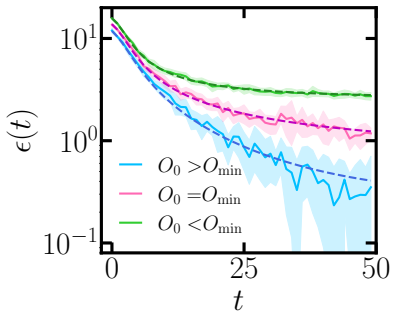

Figure R5: **Dynamics of total error $\epsilon(t)$ on IBM quantum devices, Kolkata.** Solid and dashed curves represent experimental and theoretical results. An $n = 2$ qubit $D = 4$-layer hardware efficient ansatz (with $L = 16$ parameters) is utilized to optimize with respect to XXZ model observable with $J = 4$. The shaded areas represent the fluctuation (standard deviation) in the experimental data.

We have also performed more detailed analyses about the noise in the experiments, via experimentally analyzing the shot noise and theoretically developing a noisy theory. To rule out the influence of other possible source of noise, we also performed extensive additional experiments and collected data with different number of shots of measurement when estimating the cost function. As shown in Fig. 11 of SI 9 (appended as Fig. R7 here), regardless of the number of shots in the estimation, the error converges toward the same value as the number of sample repetitions increase. The difference is much smaller than the sample fluctuation. As we described in the revised SI 9, our noisy theory model (dashed lines in Fig. R5) now take into account the depolarizing noise in the experiment. Further error mitigation is beyond the scope of our study as a proof-of-principle experiment.

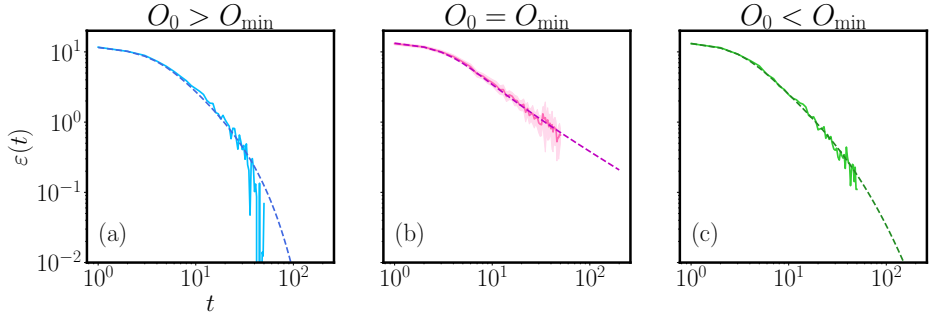

Figure R6: Dynamics of residual error on IBM quantum devices, Kolkata with different O_0 in log-log scale. The setting and legends are the same as in Fig. R5.

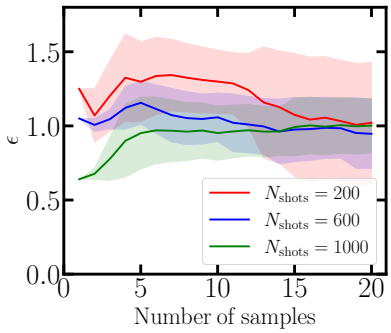

Figure R7: Experiment result of the error $\epsilon(t)$ for different number of shots in the measurement repetitions for estimating the observable in an individual sample. Multiple samples of training are performed to show the sample fluctuation. As the number of samples increase, the average ϵ becomes almost independent of the number of the shots. The shaded region indicate the sample fluctuation in standard deviation. $O_0 = 10$ and the results are at late time of $t = 40$.

(b) All theorems and lemmas in the paper are rigorously proven in the 40-page long SI, with all the assumption made clear and all results following from the assumption. We refined some results further, for example in Theorem 3, we got rid of a constant κ that was introduced due to some proof technique. Now we improved the proof and does not need this arbitrary small constant.

Reviewer's impression may have been caused by the use of "$\simeq$" in lots of places in the main text. Previous versions of lemma and theorem has '$\simeq$' to represent $\lim_{x\to\infty} f(x) = C$. We have changed them to equality as $\lim_{x\to\infty} f(x) = C$ can be expressed as "at the $x \gg 1$ limit, $f(x) = C$".

List of Changes

Here is a list of the major changes. Please also refer to the attached revised version of the paper with changes highlighted in red fonts.

1. Refs. [1, 5, 13] added.
2. Abstract and Introduction revised.

3. Notation revision throughout the manuscript.

4. Two new sections, 'Speeding up convergence' and 'Dynamics of limited-depth QNN' are added, together with Fig. 5 and Fig. 7 and Fig. 6.

5. Fig. 1 updated in cosmetics, Fig. 2 and Fig. 3 updated to larger system size. Fig. 8 added a subplot and both subplots updated with error bar and noisy theory model. Fig. 9 added two subplots.

6. Theory for linear loss function and noisy theory model added to SI. SI revised for clarity.

References

[1] Eric R Anschuetz and Bobak T Kiani. Quantum variational algorithms are swamped with traps. *Nat. Commun.*, 13(1):7760, 2022.

[2] Ching-Kai Chiu, Jeffrey CY Teo, Andreas P Schnyder, and Shinsei Ryu. Classification of topological quantum matter with symmetries. *Reviews of Modern Physics*, 88(3):035005, 2016.

[3] Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M Chow, and Jay M Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, 549(7671):242–246, 2017.

[4] Bobak Toussi Kiani, Seth Lloyd, and Reevu Maity. Learning unitaries by gradient descent. *arXiv:2001.11897*, 2020.

[5] Martin Larocca, Piotr Czarnik, Kunal Sharma, Gopikrishnan Muraleedharan, Patrick J Coles, and Marco Cerezo. Diagnosing barren plateaus with tools from quantum optimal control. *Quantum*, 6:824, 2022.

[6] Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J Coles, and Marco Cerezo. Theory of overparametrization in quantum neural networks. *Nat. Comput. Sci.*, 3(6):542–551, 2023.

[7] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv:1711.00165*, 2017.

[8] Junyu Liu, Zexi Lin, and Liang Jiang. Laziness, barren plateau, and noises in machine learning. *Mach. Learn.: Sci. Technol.arXiv:2206.09313*, 5:015058, 2024.

[9] Junyu Liu, Khadijeh Najafi, Kunal Sharma, Francesco Tacchino, Liang Jiang, and Antonio Mezzacapo. Analytic theory for the dynamics of wide quantum neural networks. *Phys. Rev. Lett.*, 130(15):150601, 2023.

[10] Junyu Liu, Francesco Tacchino, Jennifer R Glick, Liang Jiang, and Antonio Mezzacapo. Representation learning via quantum neural tangent kernels. *PRX Quantum*, 3(3):030323, 2022.

[11] Da-Chuan Lu, Juven Wang, and Yi-Zhuang You. Definition and classification of fermi surface anomalies. *Physical Review B*, 109(4):045123, 2024.

[12] Xinbiao Wang, Junyu Liu, Tongliang Liu, Yong Luo, Yuxuan Du, and Dacheng Tao. Symmetric pruning in quantum neural networks. *arXiv:2208.14057*, 2022.

[13] Roeland Wiersema, Cunlu Zhou, Yvette de Sereville, Juan Felipe Carrasquilla, Yong Baek Kim, and Henry Yuen. Exploring entanglement and optimization within the hamiltonian variational ansatz. *PRX Quantum*, 1:020319, Dec 2020.

[14] Xuchen You and Xiaodi Wu. Exponentially many local minima in quantum neural networks. In *International Conference on Machine Learning*, pages 12144–12155. PMLR, 2021.

RESPONSE TO REVIEWS

NCOMMS-23-62482-T
"Dynamical transition in quantum neural networks with large depth"
submitted to Nature Communications

by Bingzhi Zhang, Junyu Liu, Xiao-Chuan Wu, Liang Jiang and Quntao Zhuang

We thank the Reviewer for reviewing our manuscript entitled "Dynamical transition in quantum neural networks with large depth" again. Note that we have revised the title, compared to the last submission. The comments and suggestions have greatly helped us to improve the presentation of the paper. We have also done significant revision and believe that the paper can now be accepted by Nature Communications

Reviewer 2

Reviewer 2 acknowledges that our manuscript is improved but still have three criticisms. Reviewer 2 does not doubt the relevance and importance in the field of application, and is willing to reconsider upon revision of our manuscript. In particular, Reviewer 2 has detailed comments as we summarize below

Summary of Reviewer 2's comments

Reviewer 2 wants us to resolve three major points in order to be published in Nature Communications.

1. Lack of Qualitative Understanding: The authors have not provided a qualitative understanding of the transition. This central point is crucial for the paper's impact and comprehension. Leaving the explanation of the critical point for future research does not suffice.

2. Finite Size Scaling: To characterize a genuine phase transition, finite-size scaling is necessary [e.g., arxiv:1101.3281, or the classic Cardy, J. (1996). Scaling and Renormalization in Statistical Physics (CUP)]. Phase transitions are valid and well-defined in the thermodynamic limit. The data presented in the manuscript does not indicate that the system size plays any significant role. There is no evidence of data collapse with increasing system sizes, nor is there a substantial analysis of the universality properties of the system.

3. Technical Nature and Lack of Clear Storytelling: The manuscript remains highly technical, which may interest specialists in the field. However, it lacks a clear storytelling and physical understanding of the mechanism behind the phase transition beyond the mere mathematical aspects. This significantly limits its accessibility and impact.

Our Reply

We thank Reviewer 2 for the suggestions that have greatly helped improving the presentation of our manuscript.

In particular, the finite-size study suggested by the referee helped us realize that the dynamical transition is *not* a phase transition in the statistical physics sense. After digging deeper into the dynamical equation, we realize that the transition can be explained as a bifurcation of the nonlinear dynamical equations, when the stability of the fixed points change as the target value changes. This completes the understanding of the observed phenomena. With this understanding in hand, we thoroughly revised our manuscript with a clear storyline:

1. re-ordered the sections to smooth the flow and present a coherent story about the bifurcation transition

2. reduced the level of technicality of the main text

3. reduced the number of equations in the main text from 34 to 26 and supplied more intuitions

We have fully addressed all points and believe that our manuscript can be published in *Nature Communications* now.

Before the point-to-point response, let us emphasize the novelty and impact of our results.

- Our paper is the first to identify a novel phenomenon—the transition in training dynamics of QNNs. The phenomena is not only numerically verified, analytically explained, but also experimentally observed. Therefore, it is of interest to readers in the field of quantum machine learning.
- The development of QNTK theory as a generalization of classical NTK theory is a case where classical machine learning tool is applied to quantum machine learning. We deepened the study of QNTK. It is also an open question whether a similar dynamical transition can be found in the training of classical neural networks. Therefore, our work is of interest to readers in classical machine learning.
- We establish the nonlinear dynamical equation that explains the transition as a transcritical bifurcation. As a use case of classical nonlinear equations in quantum circuit dynamics, it will intrigue readers in the field of nonlinear and complex systems.
- The discovered bifurcation transition provides new ways of designing loss functions to speed-up the convergence of QNN training. As QNNs are widely applicable to quantum chemistry, optimization etc, findings of the speed-up is of great interest to people in the field of quantum computing.

From the above significance, we believe that our manuscript has met *Nature Communications*' scope of "representing important advances of significance to specialists within each field."

Below, we provide detailed reply to the other concerns of Reviewer 2.

1. We now have a better understanding of the nature of the transition: it is a transcritical bifurcation transition in the nonlinear dynamics governed by the LV equation. We have revised the manuscript thoroughly to incorporate the understanding of bifurcation, including adding Fig. 3(a), which is appended here as Fig. R1(a) for the convenience of the reviewer.

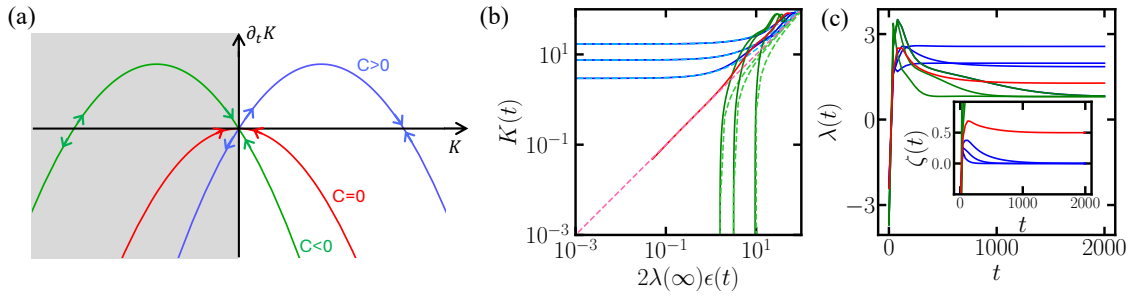

Figure R1: **Classical dynamics interpretation of total error and QNTK dynamics.** (a) The RHS of Eq. (1) showing a bifurcation. The gray region is nonphysical as $K \geq 0$. In the physical region ($K \geq 0$), we have a single stable fixed point $K = C$ when $C > 0$, corresponding to the frozen-kernel 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 $O_0 \gtrless O_{\min}$, plotted in solid blue, red and green. Dashed curves show the trajectory from Eq. (11). The arrows denote the flow of time in QNN optimization. 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.

We also added the paragraph explaining the bifurcation:

> Thanks to the conserved quantity C, we can reduce the coupled differential equations of LV model in Eqs. (10) to a single variable differential equation with the kernel or the error alone, e.g.,
>
> $$\partial_t K(t) = \eta(C - K(t))K(t). \tag{1}$$

This is a canonical example of a transcritical bifurcation, with two fixed points $K = C$ and $K = 0$ [2]. To see this, we plot the RHS of Eq. (1) in Fig. 3(a). 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.

After studying the finite size scaling, we now agree with the referee that the wording of 'phase transition' is not precise. We have evaluated the finite size effect thanks to the suggestion of the referee. Although we can see data collapse when the layers of the circuit (the finite size of the system) increases, the minimum gap does not change with the circuit depth. In fact, at the infinite time limit, the gap will always have an exact closing—on the contrary to the conventional phase transitions in statistical physics. Therefore, it is not a phase transition from the statistical physics perspective. We will explain more details about the finite-size in the next point.

The Schrödinger equation interpretation of the closing of the Hessian gap is now presented with more precise wording and the finite-size data, while the scaling law etc are now put to the method part as additional interpretations with statistical physics tools. We think that these comparisons to phase transition still provides insight into the properties of the transition.

2. To provide an understanding on the finite size effect on the dynamical transition, we study the scaling of gap closing and related correlation length. In the Schrödinger equation interpretation, as the Hessian matrix is interpreted as the effective Hamiltonian of an imaginary-time Schrödinger equation in late time limit, we therefore regard the number of variational parameters (equivalently number of layers) as the "system size". In Fig. 4(b) [appended below as Fig. R2(b)], we present the spectrum gap G_M of Hessian matrix versus rescaled target value $|O_0 - O_{\min}|L$. We find that with increasing number of parameters L (light to dark dots), the scaling of spectrum gap collapses well, indicating a transition in infinite size limit. The power-law fitting result (dashed lines) $G_M \sim (|O_0 - O_{\min})L)^{\nu_1}$ also shows a universal exponent $\nu_1 \simeq -1$ for QNNs with enough expressivity. The nonvanishing gap at $O_0 = O_{\min}$ (red dots) is due to finite training while the exact gap closing only holds at infinite time limit. This distinction in gap closing compared to the ones in conventional statistical physics suggests that it is not a genuine phase transition in statistical physics sense, therefore we refer to "dynamical transition".

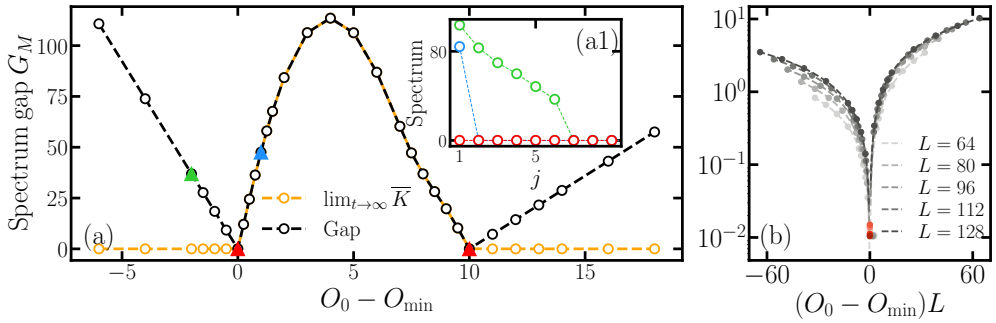

Figure R2: **Spectrum gap of the effective Hamiltonian in Schrödinger interpretation of QNN in the example of XXZ model.** (a) The spectrum gap of Hessian matrix of the effective Shrödinger dynamics in $t \to \infty$ (black). The gapless transition point corresponds to $O_0 = O_{\min}, O_{\max}$ (red triangles). The orange line represents the QNTK $\lim_{t\to\infty} \overline{K}$. Inset (a1) shows the Hessian spectrum of the largest 10 eigenvalues for the three cases $O_0 \lessgtr O_{\min}$ marked by triangles in (a). (b) The spectrum gap versus scaled target value $(O_0 - O_{\min})L$ with different number of parameters. We choose $O_0 - O_{\min} \in [-0.5, 0.5]$. Dots from light to dark represent gaps with increasing L. The dashed lines with corresponding colors show $G_M \sim (|O_0 - O_{\min}|L)^{\nu_1}$ with $\nu_1 \simeq 1$ from fitting. Red dots represent the critical point $O_0 = O_{\min}$. The nonvanishing gap at $O_0 = O_{\min}$ is due to finite training time. In (a) the RPA consists of $D = 64$ layers (equivalently $L = 64$ parameters) on $n = 2$ qubits, and in (b) the RPA is applied on a system of $n = 4$ qubits. In both cases, the parameters in XXZ model is $J = 2$.

The finite size effect is also studied within correlation length. Similar to the gap closing, we also find a well-collapse of correlation length with respect to $|O_0 - O_{\min}|L$ in Fig. 10 [appended here as Fig. R3], and the exponent $\nu_2 \simeq 1$ of power-law fitting $\xi \sim (|O_0 - O_{\min}|L)^{\nu_2}$ also agrees.

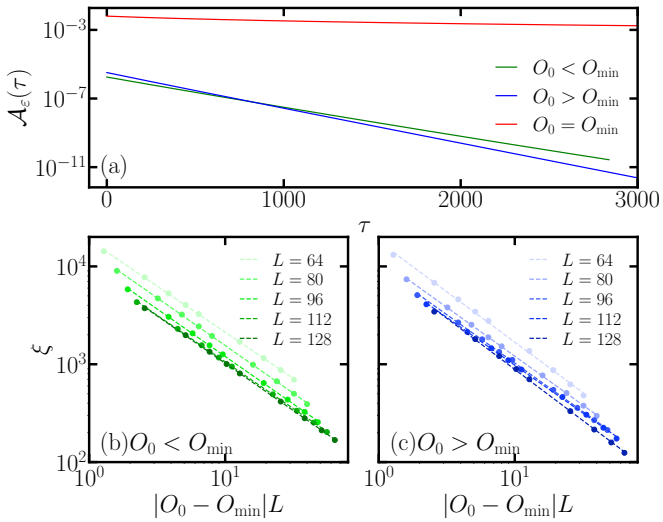

Figure R3: **Correlation functions in the Schrödinger equation interpretation.** In (a), we plot the decay of autocorrelators $\mathcal{A}_\epsilon(\tau)$ with different $O_0 \lessgtr O_{\min}$ (green for '<', red for '=' and blue for '>'). In (b) and (c), we show the scaling of correlation length $\xi \sim (|O_0 - O_{\min}|L)^{-\nu_2}$ with $\nu_2 \simeq 1$ (dashed lines) by fitting for both $O_0 \lessgtr O_{\min}$. Dots from light to dark represent ξ with $L = 64, 80, 96, 112, 128$ variational parameters, and the dashed lines with the corresponding color represent fitting result. Here we choose $O_0 - O_{\min} \in [-0.5, 0.5]$. In (a) the RPA consists of $D = 96$ layers (equivalently $L = 96$ parameters). In all cases, the RPA is applied on a system of $n = 4$ qubits and the parameter in XXZ model is $J = 2$.

We have also revised the paragraphs in the main text, as we highlight below

> We can then model a difference equation for the unnormalized "differential state" $|\Psi(\boldsymbol{\theta})\rangle \equiv |\psi(\boldsymbol{\theta})\rangle - |\psi(\boldsymbol{\theta}^*)\rangle$ as
>
> $$\delta |\Psi(\boldsymbol{\theta})\rangle = -\eta H_\infty(\boldsymbol{\theta}) |\Psi(\boldsymbol{\theta})\rangle , \tag{2}$$
>
> where $H_\infty(\boldsymbol{\theta}) \sim M(\boldsymbol{\theta})$ is similar to the Hessian matrix (see Methods). The difference equation can be interpreted as an imaginary time Schrödinger equation, and we identify a transition with gap closing of H_∞ (equivalently $M(\boldsymbol{\theta})$) driven by O_0 at the infinite time limit.
>
> To provide insight into the transition, let us begin by exploring the behaviors of the gap of Hessian matrix. We consider the Hessian eigenvalues at the late time limit of $t \to \infty$ and large circuit depth in Fig. 4(a). 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 $\epsilon(\boldsymbol{\theta}) \to 0$ (see blue in (a1)), which equals the kernel and is verified by the orange and black curve in (a). While for *frozen-error dynamics* with $O_0 < O_{\min}$ (or $O_0 < O_{\max}$), due to non-vanishing $\epsilon(\boldsymbol{\theta})$, the Hessian has multiple nonzero eigenvalues (see green in (a1)). Overall, gap closing is observed at the critical point.
>
> After identifying the closing gap, we evaluate the Hessian gap of QNNs at late time for various different 'system size' [1, 3]—number of parameters. As we see in Fig. 4(b), the curves of the spectrum gap versus a rescaled target value $(O_0 - O_{\min})L$ collapse well as the system size L increases, indicating a well-defined transition at infinite size limit. We notice a linear-closing gap around the critical point (red triangle in Fig. 4(a)), and verify the scaling in Fig. 4(b) via fitted the gap G_M to
>
> $$G_M \sim (|O_0 - O_{\min}|L)^{\nu_1}, \tag{3}$$
>
> resulting in $\nu_1 = 0.996 \pm 0.004, 1.09 \pm 0.021$ for $O_0 \lessgtr O_{\min}$ (dashed lines). However, we also notice that the minimum gap in the numerical study has no significant dependence on the system size L—it

is dominated by the finite training time in the numerical simulation which fails to achieve the infinite time limit. As at the critical point $O_0 = O_{\min}$, the QNN training dynamics converge polynomially, which makes accessing the infinite-time limit numerically difficult. However, we do expect that the Hessian gap vanishes exactly at infinite time as both error and QNTK will vanish. Such an exact gap closing within finite size is in contrast to normal phase transitions in statistical physics and therefore we regard the transition not as a conventional phase transition in the statistical physics sense.

Despite not being a genuine phase transition, we can still adopt tools from statistical physics to provide more insight into the gap-closing transition. In Methods, we present results on the correlation length ξ, its associated critical exponent ν_2, and scaling dimensions for various physical quantities. We observe the divergence of the correlation length at critical point that mimics a second order phase transition, and find the exponent of correlation functions $\nu_2 \simeq 1$ align with what we have found for spectrum gap. Data collapse as the system size L increases is also confirmed for the correlation length, showing a well-defined transition at the large depth limit.

3. To improve the storytelling of our manuscript and make it open to a wider range of readers, we make the following revisions and adjustments.

   (a) We add a paragraph in the beginning of the 'Result' section explaining the story line:

   > 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.

   (b) We supplied the manuscript with intuitions and reduced the number of equations in the main text from 34 to 26, leaving only those that are substantial and intuitive enough to display.

   As an example, we moved the discussions on connection of LV model to Hamiltonian equations originally in Section "Generalized Lotka-Volterra model" and study on correlation length originally in "Schrödinger equation interpretation" to Methods while leaving the main conclusions in the main text only. With this adjustment, interested readers can find these technical discussions without affecting the flow to present major studies and conclusions.

   (c) We switch the order to present Sections "Unitary ensemble theory" and "Schrödinger equation interpretation". There are two reasons. Firstly, the unitary ensemble theory provides physical insight and analytical result to support the assumption that relative dQNTK λ converges a constant at late time, proposed in Section "Generalized Lotka-Volterra model". The Section "Schrödinger equation interpretation provides extensive understanding on the transition from the gap closing of Hessian matrix in the Schrödinger equation picture with statistical tools, making it better to be positioned as a closure of the theory. Secondly, as suggested by Referee 2, we present numerical results to quantitatively study the scaling of gap closing within finite system size –depth of the QNN. The following Section "Dynamics of limited-depth QNN" focuses on the training dynamics of QNNs with further limited-depth, establishing a coherent flow for readers to follow.

   (d) We leave the Section "Speeding up the convergence" at the end of theory and numerical studies because it is an application of how to utilize the knowledge of dynamical transition to speed up training in practice. Presenting these results last can help avoid confusion or distraction for the readers. It also provides an opportunity for readers to review the main results and smoothly understand its advantage in practical applications.

List of Changes

Here is a list of the major changes. Please also refer to the attached revised version of the paper with changes highlighted in red fonts.

1. title changed
2. abstract revised

3. Figure 1 slightly revised, Figure 3 (a) added, Figure 4 revised, Figure 10 added. Some figures have caption added.
4. introduction revised
5. Description of the LV model made simplified on page 3-4
6. The effective Hamiltonian part revised on page 5-6.
7. two subsections in the methods added.
8. Refs [1–3] added.
9. Necessary changes in the supplemental information.

References

[1] John Cardy. *Scaling and renormalization in statistical physics*, volume 5. Cambridge university press, 1996.

[2] Strogatz, Steven H. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC press, 2018.

[3] Anders W Sandvik. Computational studies of quantum spin systems. In *AIP Conference Proceedings*, volume 1297, pages 135–338. American Institute of Physics, 2010.

RESPONSE TO REVIEWS

NCOMMS-23-62482B-Z
"Dynamical transition in quantum neural networks with large depth"
submitted to Nature Communications

by Bingzhi Zhang, Junyu Liu, Xiao-Chuan Wu, Liang Jiang and Quntao Zhuang

We thank the Reviewers for reviewing our manuscript entitled "Dynamical transition in quantum neural networks with large depth" again. The comments and suggestions have greatly helped us to improve the presentation of the paper. We have fully addressed the comments and believe that the paper can now be accepted by Nature Communications

Reviewer 1

Reviewer 1 think that we "have done a great job of replying to some of Referee 2's comments" and "the new understanding of the critical point has improved the work significantly as it truly provides a more precise characterization of the observed phenomenon" and remains "strong support for publication".

Reviewer has a minor suggestion to move the labels of Fig 4 outside of the plots.

Our Reply

We thank the referee for the positive feedbacks on our manuscript. Indeed, we agree that Fig. 4 is quite busy, and as suggested by Reviewer 2, we have moved the scaling with L subplot into the supplementary SI 2. In this way, the main text Fig. 4 and SI Fig. 1 are both neat in the labels—main text Fig. 4 now does not need labels and SI Fig. 1 has no insets. We also point out in the main text that

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

so that the nice scaling that Reviewer 1 finds illustrative is pointed out in the main text.

Reviewer 2

Reviewer 2 finds our "text is generally more transparent, and the essential contributions are more evident than the earlier version." Reviewer 2 thinks that "final version will be suitable for publication in the journal" once we address its suggestions on avoiding any confusion.

Reviewer 2 thinks that the system size should be n—the number of qubits. To resolve the concern, Reviewer 2 has specific suggestions:

> "In summary, I suggest removing the analysis with the arbitrary finite-size scaling L from the text, or at least from the main body. The discussion of the dynamical transition should be shortened, as it is adequately explained within the mathematical framework of dynamical systems. The reference to a fictitious "system size" L should be omitted to maintain consistency. (The system size, again, should be the number of qubits.)"

Our Reply

We thank Reviewer 2 for the suggestions that have greatly helped improving the presentation of our manuscript. We have fully adopted the suggestion and implemented all changes Reviewer 2 requested: we moved the previous Fig. 4(b) to SI Fig. 1 in SI 2. And all part of the main text discussions regarding the original Fig. 4(b) and the

Methods regarding the correlator are also moved to SI 2 and we have made sure there is no referring to L as the system size. In SI 2, we refer to L as the number of parameters or sometimes 'parameter size'.

Regarding the connection to non-Hermitian systems, we indeed find it very interesting and have added a sentence

> Such a transition at a finite system size resembles that for non-Hermitian dynamical systems [1–3]. More discussions on the statistical physics interpretation and the closing of the gap under different number of parameters can be found in SI 2.

The exact connection to non-Hermitian systems is subject to future research.

With the above changes, we believe our manuscript is now ready to be published in Nature Communications.

List of Changes

Here is a list of the major changes. Please also refer to the attached revised version of the paper with changes highlighted in red fonts.

1. we moved the previous Fig. 4(b) to SI Fig. 1 in SI 2.
2. Move the main text discussions regarding the original Fig. 4(b) to SI 2.
3. Move the Methods regarding the correlator to SI 2.
4. added three references [1–3].

References

[1] Michael V Berry. Physics of nonhermitian degeneracies. *Czech. J. Phys.*, 54(10):1039–1047, 2004.

[2] Ramy El-Ganainy, Konstantinos G Makris, Mercedeh Khajavikhan, Ziad H Musslimani, Stefan Rotter, and Demetrios N Christodoulides. Non-hermitian physics and pt symmetry. *Nat. Phys.*, 14(1):11–19, 2018.

[3] Ingrid Rotter and JP Bird. A review of progress in the physics of open quantum systems: theory and experiment. *Rep. Prog. Phys.*, 78(11):114001, 2015.

[Version 0:
Reviewer comments:
Reviewer #1
(Remarks to the Author)
See attached report.]

Editorial note: Please see below for report.

Quantum Neural Tangent Kernels (QNTK) have been recently used to study the small-learning-rate gradient descent dynamics of mean-squared-error-type loss functions in variational quantum computing. In particular, QNTK's have been used in the literature (in works by some of the authors of this manuscript), to show that certain deep Quantum Neural Networks (QNNs) will exhibit phenomena such as lazy training or an exponential decay of the residual training error. QNTKs present a refreshing perspective to the trainability analysis as they can go beyond random initialization and can thus capture properties of the long-time training dynamics. The present work contributes to the body of literature of QNTKs by studying the interplay between the error and the QNTK as a function of time for different target values in the loss function. Notably, the authors show that the dynamics can be described by a Lotka-Volterra equation, and that there exists a duality between the QNTK and the total error. The results are numerically verified as well as supported by theoretical analysis.

Overall, the manuscript is well written and clear, and the results are extremely interesting. Given that our field is mostly heuristics-driven, having theoretical results such as the ones presented in this work is extremely important. Hence, our impression of the manuscript is very positive, and we are willing so support publication in *Nature Communications*, provided that the authors can address our comments below.

Main comments/concern:

1. While the theoretical results are interesting, the numerical simulations are quite lacking (the authors present simulations for a two-qubit system). Can the authors present results for larger problems sizes? This will significantly straighten their results, as well as show that some quantities (such as λ) depend on the system size.

2. While the results are clearly extremely interesting, the authors should clarify that they hold for certain special types of circuits, and not for arbitrary deep quantum neural networks. For instance, the numerical results, and some of the theoretical analysis, require having controllable circuits that will form a design over the unitary group of dimension d. Hence, we have two requests for the authors: 1) Can they comment on what happens if the circuit is deep, but does not form a design over $U(d)$? E.g., Will a similar phenomenon occur for the transverse field Ising model Hamiltonian Variational Ansatz (see Fig (1) in 2008.02941 or Fig (6) in 2105.14377? Note that those circuits are not controllable/universal, and hence will form a design over $SO(2n)$ but not over $U(d)$. 2) If the authors see that other deep circuits will not show a similar behaviour, then would they consider changing the title and abstract to reflect that they results apply to certain architectures? E.g., "Dynamical phase transition in **controllable** qnns with large depth." We understand that answering (1) might be beyond the scope of the work (as stated by the authors in the discussion). So, if the authors want to leave a detailed analysis of (1) for future work, we would recommend changing the title/wording of their work.

3. The authors claim that the results are valid in the large depth $L \gg 1$ regime, but they never make any statements about how deep, *deep* means? Does one require log depth? Poly? Exp? Can the authors clarify this point where appropriate?

4. Connected to the previous question, does the required depth connect to the depth necessary for overparametrization of the quantum neural network (see 2109.11676)?

5. Relating to the previous, the theoretical analysis seems to rely on large depth, but also large width $d \gg 1$. Can the authors make a comment about this assumption? It seems that the width is not too important for results to hold (as evidenced by the $n = 2$ numerics). Is it expected that analytical results can be derived for non scaling d?

6. There are some notation inconsistencies in the manuscript. We tried to diligently follow the proofs and there is a back and forth between notation standard. For instance, in some cases the QNTK is written as 'K', whereas in other it is '$K(k)$'. We suggesting choosing one for consistency.

7. The font in Fig. 1 is very small and the color contrast makes it hard to read on a printed version of the work. Could the authors make some of the font larger/change color schemes? Especially on the bottom left panel of Fig 1 (QNTK versus error).

8. How would the results in the work change if the loss function is a summation of error terms? I.e., consider a mean-squared-error loss function and the sum is over training points.

9. Similarly to the previous question, how do the results change for loss functions that are expressed as the expectation value of an observable $\mathcal{L}(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|O|\psi(\boldsymbol{\theta})\rangle$? If the results are the same, then why take a square in the loss? If the results are significantly different, why does the non-linearity of the square change things so much?

10. Could the authors comment on the implications of their results for the design of quantum machine learning loss functions? It is standard to see in quantum machine learning the use of mean-squared-error loss functions where O is a Pauli and $O_0 = \pm 1$, i.e., $O_0 = O_{\text{min}}$. However, according to the present results, if $O_0 = O_{\text{min}}$ then we are in the critical point, where the decrease in error is no longer exponential. Do these results imply that it is better to strictly choose $O_0 > O_{\text{min}}$ and be in the frozen-kernel phase?

11. Relating to the previous, we wonder if the authors could emphasize why/how their work matters for practitioners. That is: What can someone that implements quantum neural networks in their day-to-day research take away from this work? Is there any practical lesson to be learnt? Adding such statements could help make the manuscript more appealing to a wider community.

Minor comments

1. In the introduction the authors claim that QNNs generalize classical NNs, is this a fair statement? E.g., QNNs are linear maps, whereas classical are not. We would recommend slightly rewording this statement to avoid confusing non-expert readers.

2. The authors define L as the number of layers in the paragraph before equation (1), but it is unclear what a layer means here (i.e., is it a single parametrized gates? gates applied in parallel?). Could they try to give a more precise definition given that the concept of layers and depth is so important to the work?

3. After equation (9) the authors say that λ depends on d but d is not defined.

4. In the first paragraph of the section "Unitary ensemble theory" the authors define an average, but they do not say an average over *what*.

5. Some of the proof steps in the SI could use of addition explanations. We re-derived some, although not all, of the results and we believe that adding some additional step information could help less experience readers to access the nice derivations.