IOP Publishing Mach. Learn.: Sci. Technol. 4 (2023) 025038 https://doi.org/10.1088/2632-2153/ace02f

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
11 December 2022

REVISED
13 June 2023

ACCEPTED FOR PUBLICATION
20 June 2023

PUBLISHED
3 July 2023

PAPER

Infinite neural network quantum states: entanglement and training dynamics

Di Luo[1,2,3,4,*] **and James Halverson**[1,5]

[1] The NSF AI Institute for Artificial Intelligence and Fundamental Interactions
[2] Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States of America
[3] Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, United States of America
[4] Department of Physics, Harvard, Cambridge, MA 02139, United States of America
[5] Department of Physics, Northeastern University, Boston, MA 02115, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: diluo@mit.edu

Keywords: neural network quantum state, neural tangent kernel, transverse field Ising model, Fermi Hubbard model, quantum state supervised learning, neural network Gaussian process, entanglement entropy

Supplementary material for this article is available online

Abstract

We study infinite limits of neural network quantum states (∞-NNQS), which exhibit representation power through ensemble statistics, and also tractable gradient descent dynamics. Ensemble averages of entanglement entropies are expressed in terms of neural network correlators, and architectures that exhibit volume-law entanglement are presented. The analytic calculations of entanglement entropy bound are tractable because the ensemble statistics are simplified in the Gaussian process limit. A general framework is developed for studying the gradient descent dynamics of neural network quantum states (NNQS), using a quantum state neural tangent kernel (QS-NTK). For ∞-NNQS the training dynamics is simplified, since the QS-NTK becomes deterministic and constant. An analytic solution is derived for quantum state supervised learning, which allows an ∞-NNQS to recover any target wavefunction. Numerical experiments on finite and infinite NNQS in the transverse field Ising model and Fermi Hubbard model demonstrate excellent agreement with theory. ∞-NNQS opens up new opportunities for studying entanglement and training dynamics in other physics applications, such as in finding ground states.

1. Introduction

Quantum states are fundamental objects in quantum mechanics. Generically, the dimensionality of a quantum state grows exponentially with the system size, which provides one fundamental challenge for classical simulations of quantum many-body physics. This is the so-called curse of dimensionality, which also regularly arises in machine learning (ML), where a judicious choice of neural network architecture and optimization method can help address the problem.

Inspired by progress in machine learning, neural networks have have been proposed [1] as a useful way to represent quantum wavefunctions, an idea known as a neural network quantum state (NNQS). The goal is to find a compact neural network representation of the high dimensional quantum state, which is possible because the neural network is a universal function approximator [2, 3]; furthermore, they also give exact representations of certain quantum states [4–11], demonstrating their representation power. Recent research has demonstrated that NNQS can achieve state-of-the-art results for computing ground states and the real time dynamics properties of closed and open quantum systems across a variety of domains, including condensed matter physics, high energy physics, and quantum information science [8, 9, 12–35]. Despite this progress, there is ample room for an improved understanding of the representation power and training dynamics of NNQS.

The neural tangent kernel (NTK) [36] has recently emerged as a theoretical tool for understanding the gradient descent dynamics of large neural networks. NTK theory utilizes architectures with a discrete hyperparameter N, such as the width of a fully-connected network. In general, gradient descent updates to the network are controlled by a parameter-dependent NTK, but in the infinite-N limit the network evolves as a linear model, with dynamics governed in a ordinary differential equation by a deterministic constant NTK [36–38]. This ODE becomes linear and analytically solvable for a mean-squared-error loss (see the supplementary material for a review of the NTK). Similarly, in the infinite-N limit, networks are often drawn from Gaussian processes [39–42], in which case they may be trained with Bayesian inference via another deterministic constant kernel, the neural network Gaussian process (NNGP) kernel [39].

In this work we study infinite NNQS (∞-NNQS), which exhibit both representation power through ensemble statistics and also tractable training dynamics. Specifically, we relate ensemble averages of entanglement entropy bound to neural network correlation functions. For appropriate ∞-NNQS, the ensemble statistics are Gaussian and the correlators are exactly computable. Architectures are presented that approach Gaussian i.i.d. wavefunctions with volume-law entanglement. Furthermore, we develop a general framework for the gradient descent dynamics of NNQS, using a quantum state NTK (QS-NTK). Our framework is general and may be applied to various learning setup, such as ground state optimization, quantum state tomography and quantum state supervised learning. In appropriate infinite limits, gradient descent of the ∞-NNQS is governed by a constant deterministic QS-NTK. In the case of quantum state supervised learning, we prove that an ∞-NNQS trained with a positive-definite QS-NTK can recover any target wavefunction. We experimentally demonstrate that the QS-NTK can predict the training dynamics of ensembles of finite width NNQS.

2. ∞-NNQS

Consider a quantum state $|\psi\rangle$ represented by a neural network with continuous learnable parameters θ and a discrete hyperparameter N. The wavefunction is $\psi_{\theta,N} : D \to \mathbb{C}$, where the domain D is problem-dependent. The subscripts θ, N will often be implicit.

An ∞-NNQS is a neural network representation in the $N \to \infty$ limit. There are many such limits, according to the identification of a candidate N in a given network architecture, We study cases where this limit is useful either for understanding the entanglement of an ensemble of wavefunctions, via increased control over their statistics, or their gradient descent dynamics. For instance, in many architectures the $N \to \infty$ limit is also one in which the network is drawn from a Gaussian process (GP), where, e.g. N is the width a of a fully-connected network [39–42] or the number of channels in a CNN [43, 44]. The existence of such NNGP limits is quite general [45–47], and allows for training with Bayesian inference [39, 41].

3. Quantum state NNGP and entanglement

NNQS exhibit unique and interesting entanglement properties [6, 10, 48, 49]. The statistical control offered by this NNGP correspondence allows us to study the entanglement entropy properties of the ensembles of ∞-NNQS. Consider an ensemble of normalized NNQS $\{|\psi_\theta\rangle\}$. We split the input domain D into a subregion A and its complement B as $D = A \cup B$, which makes the wavefunction arguments consisted of two variables x_A and x_B from subregions A and B.

Denote the ensemble average of the nth Rényi entanglement entropy as $\langle S_n \rangle \equiv \mathbb{E}_\theta S_n$, where $S_n = \frac{1}{1-n}\log \mathrm{Tr}\rho_{\theta A}^n$ is the nth Rényi entropy of the ensemble over a sub-region A. According to Jensen's inequality, $\langle S_n \rangle \geqslant \frac{1}{1-n}\log \mathbb{E}_\theta \mathrm{Tr}\rho_{\theta A}^n$ for $n > 1$. It provides a lower bound for entanglement entropy which can be computed from $\mathbb{E}_\theta \mathrm{Tr}[\rho_{\theta A}^n]$ using the replica-trick [50, 51]:

$$\mathbb{E}_\theta \mathrm{Tr}[\rho_{\theta A}^n] = \sum_{x_A^k, x_B^k, k} \mathbb{E}_\theta \left[\prod_{k=1}^{n} \psi_\theta(x_{AB}^{k,k}) \psi_\theta^*(x_{AB}^{k+1,k}) \right] \tag{1}$$

$$= \sum_{x_A^k, x_B^k, k} G^{(2n)}(x_{AB}^{1,1}, x_{AB}^{2,1}, \ldots, x_{AB}^{n,n}, x_{AB}^{1,n}), \tag{2}$$

where $G^{(2n)}$ are the NNQS correlation functions, defined implicitly, and $x_{AB}^{i,j} := (x_A^i, x_B^j)$ (here we have the convention $x_{A/B}^{n+1} \equiv x_{A/B}^1$). The sum $\sum_{x_A^k, x_B^k, k}$ is over all k and possible x_A^k and x_B^k. This provides a means for analyzing the different entanglement entropies. The entanglement entropy bound is particularly tractable for

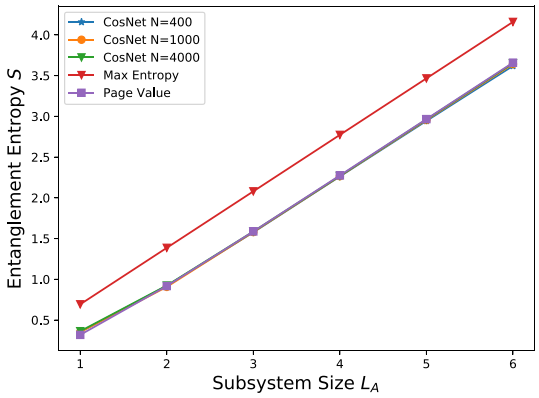

Figure 1. Von Neumann entanglement entropy of CosNet with width = 400, 1000, 4000 average over 100 ensembles. Max indicates the maximum entropy of the subsystem and Page indicates the page value entropy of the subsystem.

∞-NNQS, since in the GP limit the correlation functions are determined in terms of the two-point function (GP kernel) via Wick's theorem. See the supplementary materials for more details.

Consider $\psi(x) = \psi_1(x) + i\psi_2(x)$, where both $\psi_1(x)$ and $\psi_2(x)$ are drawn from any NN architecture. For example, we analyze the Cos-net [52] NNQS, where $\psi_1(x)$ and $\psi_2(x)$ come from the following function form:

$$f(x) = \sum_{i=1}^{N} a_i \sum_{j=1}^{d} cos(w_{ij}x_j + b_j) \tag{3}$$

where d is the input dimension, N is the number of hidden dimension, $a_i \sim \mathcal{N}(0, \frac{\sigma_a^2}{N}), w_{ij} \sim \mathcal{N}(0, \frac{\sigma_w^2}{d})$, $b_j \sim \mathcal{U}[-\pi, \pi]$). It has been shown that in the infinite N limit, $f(x)$ gives rise to the following 2-pt function [52]

$$\mathbb{E}(f(x), f(y)) = G^{(2)}(x, y) = \frac{\sigma_a^2}{2} e^{-\frac{\sigma_w^2}{2d}(x-y)^2}, \tag{4}$$

By tuning $\sigma_w \to \infty$, it yields a zero-mean Gaussian process so that $\psi_1(x)$ and $\psi_2(x)$ are both drawn from i.i.d Gaussian for different values of x. After normalization, such an ensemble of wavefunctions is known to reach the Page value of entanglement entropy and exhibits a volume law entanglement behavior [53, 54]. We compare the Von Neumann entanglement entropy of CosNet with $N = 400, 1000, 4000$ with respect to the Page Value entropy subsystem scaling in figure 1, which demonstrates nice consistency between our theory and simulations. More details on the simulations can be found in the supplementary materials.

More generally, neural networks provide a means for defining ensembles of wavefunctions with entanglement entropy ensemble average bound expressed in terms of NN correlators even away from the GP limit. This provides a new mechanism for engineering ensembles of wavefunctions whose typical states could have interesting entanglement properties. In general, finite-N effects introduce non-Gaussianities into the ensemble [55, 56] that correct the entanglement entropies. For instance, Gauss-net [56] and Cos-net yield dual GPs as $N \to \infty$ [57], but have different statistics and even symmetries [58] at finite-N. It opens up the possibility of entanglement engineering of NNQS and provides a framework for studying entanglement structure of NNQS.

4. QS-NTK

∞-NNQS also have interesting gradient descent properties.

We begin with a study of gradient descent for general NNQS. The dynamics of the network are governed by the parameter update $\dot{\theta}_i = -\nabla_{\theta_i} L = -\sum_{x' \in B} \nabla_{\theta_i} \mathcal{L}(x')$, where we have expressed the update in terms of a total loss L and also a pointwise loss $\mathcal{L}$, summed over a batch B. Applying the chain rule,

$$\frac{d\theta_i}{d\tau} = -\eta \sum_{x' \in B} \left[\frac{\partial \psi(x')}{\partial \theta_i} \frac{\partial \mathcal{L}}{\partial \psi(x')} + \frac{\partial \psi^*(x')}{\partial \theta_i} \frac{\partial \mathcal{L}}{\partial \psi^*(x')} \right], \tag{5}$$

where x' is data from B and the loss derivatives are also evaluated on the batch; the structure of B will be further specified in examples, including any labels associated to x'. The associated wavefunction update is

$$\begin{aligned}\frac{d\psi(x)}{d\tau} &= \sum_i \frac{\partial\psi(x)}{\partial\theta_i}\frac{\partial\theta_i}{\partial d\tau} \\ &= -\eta\left[\sum_{x'\in B}\Theta(x,x')\frac{\partial\mathcal{L}}{\partial\psi(x')} + \Phi(x,x')\frac{\partial\mathcal{L}}{\partial\psi^*(x')}\right],\end{aligned} \tag{6}$$

where

$$\begin{aligned}\Theta(x,x') &= \sum_i \frac{\partial\psi(x)}{\partial\theta_i}\frac{\partial\psi(x')}{\partial\theta_i} \\ \Phi(x,x') &= \sum_i \frac{\partial\psi(x)}{\partial\theta_i}\frac{\partial\psi^*(x')}{\partial\theta_i}.\end{aligned} \tag{7}$$

$\Theta(x,x')$ is the NTK [36].

Since we are using a complex-valued neural network to represent quantum wavefunctions, we also see the appearance of $\Phi(x,x')$, which we call the Hermitian NTK, since it is Hermitian, $\Phi^*(x,x') = \Phi(x',x)$. Putting the wavefunction and its conjugate on equal footing, we write

$$\frac{d}{d\tau}\begin{bmatrix}\psi(x)\\ \psi^*(x)\end{bmatrix} = -\eta\sum_{x'\in B}\begin{bmatrix}\Theta(x,x') & \Phi(x,x')\\ \Phi^*(x,x') & \Theta^*(x,x')\end{bmatrix}\begin{bmatrix}\frac{\partial\mathcal{L}}{\partial\psi(x')}\\ \frac{\partial\mathcal{L}}{\partial\psi^*(x')}\end{bmatrix} \tag{8}$$

and for simplicity re-express it as

$$\frac{d}{d\tau}\Psi(x) = -\eta\sum_{x'\in B}\Omega(x,x')\frac{\partial\mathcal{L}}{\partial\Psi(x')}, \tag{9}$$

a matrix ODE where $\Omega(x,x')$ is the block matrix in equation (8).

We call $\Omega(x,x')$ the QS-NTK, as it determines the gradient descent dynamics of NNQS, and more generally of complex functions. In general, it depends on parameters θ_i and the initialization of $\psi(x)$, though we will see in appropriate limits that the QS-NTK is deterministic and frozen during training. See also [59], which utilizes a quantum NTK in the context of variational quantum circuits, and appeared while we were finishing this work.

In practice, instead of representing the wavefunction as one complex output from the neural network, it is also common to have the neural network output the real and imaginary part of the wavefunction. In this case, we have the real imaginary NNQS representation $\Psi_{RI} := (\psi_1, \psi_2)$ such that

$$\frac{d}{d\tau}\Psi_{RI}(x) = -\eta\sum_{x'\in B}\Omega_{RI}(x,x')\frac{\partial\mathcal{L}}{\partial\Psi_{RI}(x')}. \tag{10}$$

where Ω_{RI} is the NTK in real imaginary representation; see the supplementary materials.

The QS-NTK is generic and may be applied to the various NNQS learning schemes, which correspond to the choice of loss function L. For Variational Monte Carlo study of ground states associated to a given Hamiltonian H, $L = \frac{\langle\psi\rangle H\psi}{\langle\psi\rangle\psi}$. For quantum state tomography, with observables $|x\rangle\langle x|$ in a different basis rotation, $L = -\sum_x \log|\langle x\rangle\psi|^2$. For quantum state supervised learning with a target wavefunction ψ_T, $L = ||\psi - \psi_T||^2$. In general, equation (10) is a nonlinear ODE with rich structure. In this work, we focus on the quantum state supervised learning setup, which yields a linear ODE. The study of other loss functions will left for future exploration.

4.1. QS-NTK for ∞-NNQS

Let $\psi_{\theta,N}$ be a NNQS and $\Omega_N(x,x')$ the associated QS-NTK. For many architectures, the infinite QS-NTK $\Omega_\infty(x,x')$ is parameter-independent at initialization. This is established by the kernel trick, which turns $\Omega_\infty(x,x')$ into an expectation value over parameters via the law of large numbers. See the supplementary materials for a concrete example and discussion of generality, using NTK results. Utilizing this trick generally requires i.i.d. parameters, a property generally spoiled by training.

Fortunately, the initialization QS-NTK plays a special role that can resolve the issue. Consider the linearized model associated to $\Psi(x)$,

$$\Psi_l(x) := \Psi_0(x) + \sum_i (\theta_i - \theta_{0,i}) \frac{\partial \Psi(x)}{\partial \theta_i}\bigg|_{\theta=\theta_0} \tag{11}$$

where θ_0 are the parameters at initialization and $\Psi_0(x) := \Psi(x)|_{\theta=\theta_0}$ is the initialization wavefunction. The linearized model is the truncated first-order Taylor expansion of $\Psi(x)$ around θ_0; we emphasize the model is linear in parameters, not inputs. The QS-NTK is

$$\Omega_l(x,x') = \Omega(x,x')\big|_{\theta=\theta_0}, \tag{12}$$

which is a crucial conceptual result. It says that the QS-NTK Ω_l associated Ψ_l is the QS-NTK Ω of $\Psi(x,x')$ at initialization, which is parameter-independent.

In summary, a ∞-NNQS Ψ with parameter-independent QS-NTK has a linearization Ψ_l that evolves under gradient descent according to a parameter-independent, time-independent QS-NTK $\Omega_l(x,x')$, with dynamics governed by equation (10), but with Ψ (Ω) replaced by Ψ_l (Ω_l). This is a remarkable simplification.

5. Quantum state supervised learning

We focus on quantum state supervised learning. This technique has important applications, such as initializing states for ground state and real time simulations, as well as understanding the representation power of the neural network architecture [60]. The loss function of quantum state supervised learning for a target wavefunction ψ_T is the mean square loss $L = \frac{1}{|B|}\sum_x |\psi_T(x) - \psi(x)|^2$.

Given a target quantum state ψ_T and a batch of samples B, the dynamics equation (10) become

$$\frac{d}{d\tau}\Psi_l(x) = -\frac{\eta}{|B|}\sum_{x'\in B}[\Omega M](x,x')\left[\Psi_l(x') - \Psi_T(x')\right] \tag{13}$$

where $M = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, we have used $\Psi^* = M\Psi$, and Ψ (Ω) have been replaced by Ψ_l (Ω_l) in (10).

The exact solution to this linear ODE is given by

$$\Psi_{l,x}(\tau) = \mu_x(\tau) + \gamma_x(\tau) \tag{14}$$

where

$$\begin{aligned} \mu_x(\tau) &= \sum_{i,j,k} \Omega_{xi}\left(\Omega^{-1}\right)_{ij}(1 - e^{-\Omega M\tau})_{jk}\Psi_{T,k} \\ \gamma_x(\tau) &= \Psi_x(0) - \sum_{i,j,k} \Omega_{xi}\left(\Omega^{-1}\right)_{ij}(1 - e^{-\Omega M\tau})_{jk}\Psi_k(0). \end{aligned} \tag{15}$$

We use subscripts to denote input dependence, with x for a test point and Latin indices as batch indices. For instance, $\Omega_{xi} := \Omega(x,x_i)$ for $x_i \in B$ is an x-dependent $|B|$-vector and $\Omega_{ij} := \Omega(x_i,x_j)$ for $x_i, x_j \in B$ is a $|B| \times |B|$-matrix. The initial wavefunction appears only in $\gamma_x(t)$.

This analytic solution for an ∞-NNQS deserves comment. First, when the QS-NTK is positive definite (see the supplementary materials), the solution converges as $\tau \to \infty$ and the converged wavefunction agrees with the target on every train point. Therefore, if the batch B is the entire domain, the ∞-NNQS trained with the QS-NTK perfectly reproduces the target wavefunction. This is a NNQS analog of a major result from the NTK literature, which can be understood with geometric intuition via projection from high-dimension spaces [61]. Equivalently, one can view ΩM as an effective Hamiltonian, in which case equation (13) is the analog of imaginary time evolution and converges to the ground truth. Second, for many architectures, the expectation value of the ensemble of initial wavefunctions is $\mathbb{E}[\Psi_x(0)] = 0$, in which case $\mathbb{E}[\Psi_{l,x}(\tau)] = \mu_x(\tau)$. In such a case, $\mu_x(\tau)$ is the mean function of the ensemble at time τ, and therefore $\mu_x(\infty)$ is the mean function of the infinite ensemble of converged ∞-NNQS.

Either $\Psi_{l,x}(\tau)$ or $\mu_x(\tau)$ could be utilized to make predictions relative to targets. This motivates two different losses,

$$L_\mu = \frac{1}{|B|}\sum_{x'\in B}|\mu_{x'}(\infty) - \Psi_{T,x'}|^2, \tag{16}$$

which uses converged mean for predictions, or

$$L_{\Psi_l} = \frac{1}{K|B|}\sum_{i=1}^{K}\sum_{x'\in B}|\Psi^{(i)}_{l,x'}(\infty) - \Psi_{T,x'}|^2, \tag{17}$$

which takes the average of losses for an ensemble of K linearized networks, trained to convergence, where $\Psi^{(i)}_{l,x'}(\infty)$ is the ith network in the ensemble. Since $\mathbb{E}[\gamma] = 0$ as $K \to \infty$, at large K we have

$$L_{\Psi_l} \simeq L_\mu + \frac{1}{K|B|}\sum_{i=1}^{K}\sum_{x'\in B}|\gamma^{(i)}_{x'}|^2 \equiv L_\mu + L_\gamma, \tag{18}$$

the last term becomes the variance of the linearized model in the $K \to \infty$ limit. Notice that equation (15) shows both L_μ and L_γ will converge both to zero on the training set in infinite time, which implies that ∞-NNQS will be perfectly optimized. For the test set, both L_μ and L_γ will converge to a finite value at infinite time, which provides an indicator of the performance of the ensemble of finite neural network, in practice.

6. Numerical experiments

We perform numerical simulations for ∞-NNQS and an ensemble of finite-N NNQS in two important models in quantum many-body physics, which are the spin-1/2 transverse field Ising model and the Fermi Hubbard model

$$H_s = -\sum_{\langle i,j\rangle}\sigma^z_i\sigma^z_j - J\sum_i \sigma^x_i, \tag{19}$$

$$H_f = -\sum_{\langle i,j\rangle,\sigma}(c^\dagger_{i,\sigma}c_{j,\sigma} + h.c.) + U\sum_i n_{i\uparrow}n_{i\downarrow}. \tag{20}$$

For the transverse field Ising model, we consider H_s on a 3×4 lattice with $J = 0.1$. The target state $|\psi_T\rangle$ is prepared through $|\psi_T\rangle = e^{-iH_s\tau}|\psi_0\rangle$ with $|\psi_0\rangle$ as the fully polarized state $|+\rangle^{\otimes n}$ and $\tau = 2.1$. There are in total 4096 basis elements in the target wavefunction. For the Fermi Hubbard model, we consider H_f on a a 3×4 lattice with 2 spin up fermions and 2 spin down fermions. The target state in the Fermi Hubbard model is prepared through $|\psi_T\rangle = e^{-iH_f\tau}|\psi_0\rangle$, where H_f has $U = 8$, $|\psi_0\rangle$ is the ground state of H_f with $U = 4$ and $\tau = 2.1$. There are 4356 basis elements in the target wavefunction. We choose $|\psi_T\rangle$ in the above way such that they are complex-valued and related to the quench experiments with different coupling parameters in real time quantum dynamics.

For the numerical simulations, we consider two independent neural networks that represent the real part and the imaginary part of the wavefunction, $\psi(x) = \psi_1(x) + i\psi_2(x)$; this is the case of decoupled dynamics discussed in the supplementary materials. Both $\psi_1(x)$ and $\psi_2(x)$ are single-layer fully-connected networks, i.e. $\frac{1}{\sqrt{N}}W_2\sigma(\frac{W_1}{\sqrt{12}}x + b_1) + b_2$, with entries drawn as $W_{1,2} \sim \mathcal{N}(0, 0.25)$ and $b_{1,2} \sim \mathcal{N}(0, 0.01)$, σ taken to be ReLU, and $N \in \{300, 1000, 5000\}$ is the dimension of the hidden layer.

Since both models utilize 12 lattice sites, the input is encoded in a 12-d vector. For the transverse field Ising model, spin-up and spin-down configuration take values ± 1. For the Fermi Hubbard model, the possibilities of a hole, spin-down, spin-up, and double occupancy take values $\in \{-1.5, -0.5, 0.5, 1.5\}$, respectively. For the training data set, we uniformly draw basis elements with dataset size $2400, 3200, 4000$ from the target wavefunctions, and leave the rest (the basis complement) as the test dataset. For each experiment, we train an ensemble of 10 finite width NNQS with full-batch gradient descent and compare with the QS-NTK predictions. The learning rate is chosen to be 0.9 times the maximum NTK learning rate [62], which ensure that the finite networks evolve in a linearized regime. We do not need to train the ∞-NNQS because the exact solution equation (15) makes predictions for all epochs. All simulations are implemented with `neural-tangents` library [62].

Figure 2 compare the training dynamics of finite NNQS and ∞-NNQS in both the transverse field Ising model and the Fermi Hubbard model. It is shown that the finite NNQS training dynamics agree rather well with the QS-NTK predictions. The training loss for the ∞-NNQS should drop to zero as $\tau \to \infty$, while the test losses will converge to a finite number, represented by the dashed line in the figure, which is the NTK prediction equation (18) in the infinite time limit. Figure 3 show the total MSE loss over various training dataset sizes and finite width NNQS ensembles. As the training batch size increases, the overall performances of different ensembles improve as expected. As the finite width increases, the performances of the NNQS ensembles converge to the NTK prediction, which is the infinite width limit.

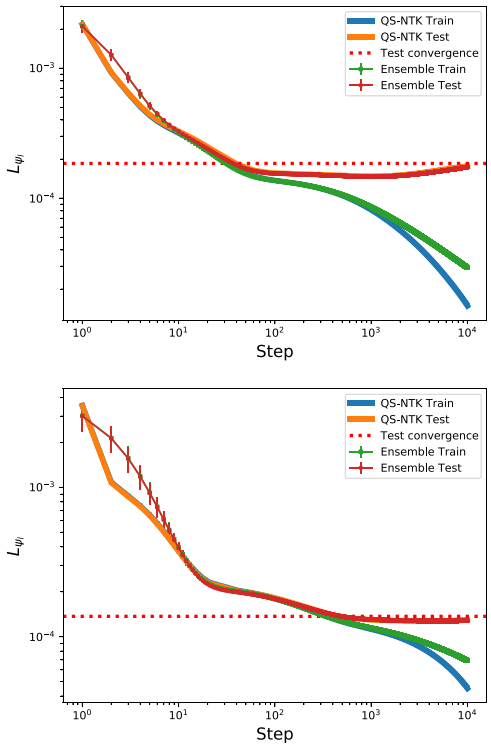

Figure 2. Performance of finite-width NNQS ensembles and QS-NTK predictions of ∞-NNQS for (a) Top: the transverse field Ising model; (b) Bottom: the Fermi-Hubbard model. The Ensemble train and test are from finite NNQS with a width equal to 5000 and ensemble size 10. The QS-NTK train and test, as well as test convergence are from infinite width NTK dynamics. The training points for both ensemble and QS-NTK cases are 2400.

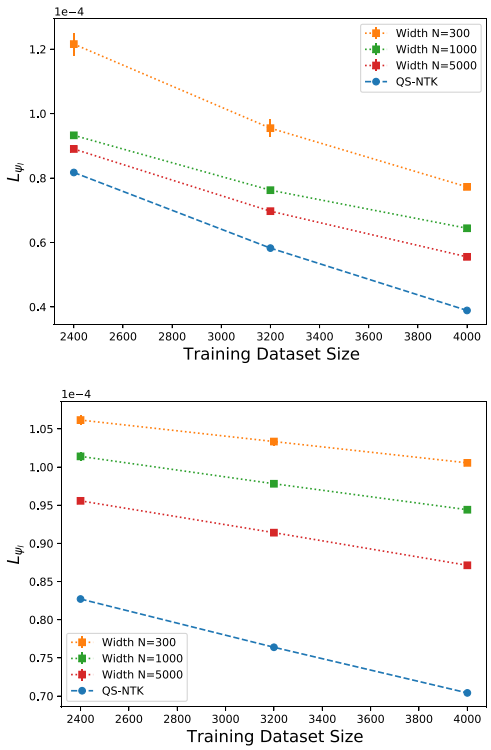

Figure 3. Total MSE loss for various training batch sizes and finite width neural network quantum states of ensemble size 10 at training time step $\tau = 10^4$ for (a) Top: the transverse field Ising model; (b) Bottom: the Fermi Hubbard model.

7. Conclusion

In this work, we introduced ∞-NNQS. We demonstrated that ensemble average entanglement entropy bound may be computed in terms of neural network correlators. For appropriate ∞-NNQS, these calculations become tractable due to the NNGP correspondence. We demonstrate that certain architectures such as CosNet NNQS exhibit volume-law entanglement. We also developed the QS-NTK as a general framework for understanding the gradient descent dynamics of NNQS. Appropriate ∞-NNQS have parameter-independent QS-NTK at initialization, which in the linearized regime is frozen to its initialization value throughout training, leading to tractable training dynamics. In quantum state supervised learning, we proved that training a linearized ∞-NNQS with a positive definite QS-NTK allows for the exact recovery of any target wavefunction. In numerical experiments, we showed that these new techniques yield accurate predictions for the training dynamics of ensembles of finite width NNQS. Systematic studies from the infinite network literature [63] suggest that NTK or NNGP Bayesian training for ∞-NNQS may exhibit increasing performance over finite networks.

More broadly, our work provides theoretical insights on understanding the training dynamics of NNQS. It also offers practical guidance for choosing neural network architectures: convergence rates during training depend on the spectrum of the QS-NTK, evaluated on the training data. This development also opens up various interesting research directions for understanding NNQS optimization in other physics contexts, such as quantum state tomography and variational Monte Carlo study of NNQS. Another interesting direction is to significantly generalize the NNQS architecture beyond the fully-connected case by using Tensor Programs [64], a flexible language for connecting general architectures with NTK limits. Recently, there are applications and generalizations of NTKs to quantum computation and quantum machine learning [59, 65–67], and it will be interesting to integrate QS-NTK into hybrid classical-quantum machine learning.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

We thank Ning Bao, Zhuo Chen, Bryan Clark, Adrian Feiguin, Dmitrii Kochkov, Ryan Levy, Anindita Maiti, Fabian Ruehle, Ge Yang and Tianci Zhou for discussions. J H is supported by NSF CAREER Grant PHY-1848089. This work is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions). This material is based upon work supported by the U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers, Co-design Center for Quantum Advantage (C2QA) under Contract Number DE-SC0012704.

Note added

References [59, 66] on quantum neural tangent kernels in the context of quantum circuits were posted to arXiv four weeks prior to this manuscript, while our work focuses on the study of neural network quantum states.

ORCID iDs

Di Luo https://orcid.org/0000-0001-6562-1762
James Halverson https://orcid.org/0000-0003-0535-2622

References

[1] Carleo G and Troyer M 2017 *Science* **355** 602
[2] Cybenko G 1989 *Math. Control Signals Syst.* **2** 303
[3] Hornik K 1991 *Neural Netw.* **4** 251
[4] Gao X and Duan L-M 2017 *Nat. Commun.* **8** 662
[5] Lu S, Gao X and Duan L-M 2019 *Phys. Rev.* B **99** 155136
[6] Levine Y, Sharir O, Cohen N and Shashua A 2019 *Phys. Rev. Lett.* **122** 065301
[7] Sharir O, Shashua A and Carleo G 2021 Neural tensor contractions and the expressive power of deep neural quantum states (arXiv:2103.10293 [quant-ph])

[8] Luo D, Carleo G, Clark B K and Stokes J 2020 Gauge equivariant neural networks for quantum lattice gauge theories (arXiv:2012.05232 [cond-mat.str-el])
[9] Luo D, Chen Z, Hu K, Zhao Z, Hur V M and Clark B K 2021 Gauge invariant autoregressive neural networks for quantum lattice models (arXiv:2101.07243 [cond-mat.str-el])
[10] Deng D-L, Li X and Das Sarma S 2017 *Phys. Rev.* X **7** 021021
[11] Huang Y and Moore J E 2021 *Phys. Rev. Lett.* **127** 170601
[12] Han X and Hartnoll S A 2020 *Phys. Rev.* X **10** 011069
[13] Choo K, Neupert T and Carleo G 2019 *Phys. Rev.* B **100** 125124
[14] Hibat-Allah M, Ganahl M, Hayward L E, Melko R G and Carrasquilla J 2020 *Phys. Rev. Res.* **2** 023358
[15] Luo D and Clark B K 2019 *Phys. Rev. Lett.* **122** 226401
[16] Hermann J, Schätzle Z and Noé F 2019 Deep neural network solution of the electronic schrödinger equation (arXiv:1909.08423 [physics.comp-ph])
[17] Pfau D, Spencer J S, Matthews A G D G and Foulkes W M C 2020 *Phys. Rev. Res.* **2** 033429
[18] Carrasquilla J, Luo D, Pérez F, Milsted A, Clark B K, Volkovs M and Aolita L 2019 Probabilistic simulation of quantum circuits with the transformer (arXiv:1912.11052)
[19] Gutiérrez I L and Mendl C B 2020 Real time evolution with neural-network quantum states (arXiv:1912.08831 [cond-mat.dis-nn])
[20] Glasser I, Pancotti N, August M, Rodriguez I D and Cirac J I 2018 *Phys. Rev.* X **8** 011006
[21] Vieijra T, Casert C, Nys J, De Neve W, Haegeman J, Ryckebusch J and Verstraete F 2020 *Phys. Rev. Lett.* **124** 097201
[22] Nomura Y, Darmawan A S, Yamaji Y and Imada M 2017 *Phys. Rev.* B **96** 205152
[23] Schmitt M and Heyl M 2020 *Phys. Rev. Lett.* **125** 100503
[24] Stokes J, Moreno J R, Pnevmatikakis E A and Carleo G 2020 *Phys. Rev.* B **102** 205122
[25] Vicentini F, Biella A, Regnault N and Ciuti C 2019 *Phys. Rev. Lett.* **122** 250503
[26] Torlai G, Mazzola G, Carrasquilla J, Troyer M, Melko R and Carleo G 2018 *Nat. Phys.* **14** 447
[27] Nicoli K A, Nakajima S, Strodthoff N, Samek W, Müller K-R and Kessel P 2020 *Phys. Rev.* E **101** 023304
[28] Nicoli K A, Anders C J, Funcke L, Hartung T, Jansen K, Kessel P, Nakajima S and Stornati P 2021 *Phys. Rev. Lett.* **126** 032001
[29] Yoshioka N and Hamazaki R 2019 *Phys. Rev.* B **99** 214306
[30] Hartmann M J and Carleo G 2019 *Phys. Rev. Lett.* **122** 250502
[31] Nagy A and Savona V 2019 *Phys. Rev. Lett.* **122** 250501
[32] Medvidović M and Carleo G 2021 *npj Quantum Inf.* **7** 101
[33] Wang J, Chen Z, Luo D, Zhao Z, Hur V M and Clark B K 2021 Spacetime neural network for high dimensional quantum dynamics (arXiv:2108.02200 [cond-mat.dis-nn])
[34] Astrakhantsev N, Westerhout T, Tiwari A, Choo K, Chen A, Fischer M H, Carleo G and Neupert T 2021 *Phys. Rev.* X **11** 041021
[35] Adams C, Carleo G, Lovato A and Rocco N 2021 *Phys. Rev. Lett.* **127** 022502
[36] Jacot A, Gabriel F and Hongler C 2018 Neural tangent kernel: convergence and generalization in neural networks (arXiv:1806.07572 [cs.LG])
[37] Lee J, Xiao L, Schoenholz S, Bahri Y, Novak R, Sohl-Dickstein J and Pennington J 2019 *Advances in Neural Information Processing Systems* vol 32 p 8572
[38] Roberts D A, Yaida S and Hanin B 2021 The principles of deep learning theory (arXiv:2106.10165 [cs.LG])
[39] Neal R M 1995 Bayesian learning for neural networks *PhD Thesis* University of Toronto
[40] Williams C K 1997 *Advances in Neural Information Processing Systems* pp 295–301
[41] Lee J, Bahri Y, Novak R, Schoenholz S S, Pennington J and Sohl-Dickstein J 2017 Deep neural networks as Gaussian processes (arXiv:1711.00165 [stat.ML])
[42] Matthews A G G, Rowland M, Hron J, Turner R E and Ghahramani Z 2018 Gaussian process behaviour in wide deep neural networks (arXiv:1804.11271)
[43] Novak R, Xiao L, Lee J, Bahri Y, Abolafia D A, Pennington J and Sohl-Dickstein J 2018 Bayesian deep convolutional networks with many channels are Gaussian processes (arXiv:1810.05148)
[44] Garriga-Alonso A, Aitchison L and Rasmussen C E 2019 Deep convolutional networks as shallow Gaussian processes (arXiv:1808.05587)
[45] Yang G 2019 Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation (arXiv:1902.04760)
[46] Yang G 2019 Tensor programs I: wide feedforward or recurrent neural networks of any architecture are Gaussian processes (arXiv:1910.12478 [cs.NE])
[47] Yang G 2020 Tensor programs II: neural tangent kernel for any architecture (arXiv:2006.14548)
[48] Medina R, Vasseur R and Serbyn M 2021 *Phys. Rev.* B **104** 104205
[49] Jia Z-A, Wei L, Wu Y-C, Guo G-C and Guo G-P 2020 *New J. Phys.* **22** 053022
[50] Hastings M B, González I, Kallin A B and Melko R G 2010 *Phys. Rev. Lett.* **104** 157201
[51] Wang Z and Davis E J 2020 *Phys. Rev.* A **102** 062413
[52] Halverson J 2021 Building quantum field theories out of neurons (arXiv:2112.04527)
[53] Page D N 1993 *Phys. Rev. Lett.* **71** 1291
[54] Zhou T and Nahum A 2019 *Phys. Rev.* B **99** 174205
[55] Yaida S 2019 Non-Gaussian processes and neural networks at finite widths (arXiv:1910.00019 [stat.ML])
[56] Halverson J, Maiti A and Stoner K 2021 *Mach. Learn.: Sci. Technol.* **2** 035002
[57] Halverson J 2021 Building quantum field theories out of neurons (arXiv:2112.04527)
[58] Maiti A, Stoner K and Halverson J 2021 Symmetry-via-duality: invariant neural network densities from parameter-space correlators (arXiv:2106.00694)
[59] Liu J, Tacchino F, Glick J R, Jiang L and Mezzacapo A 2021 Representation learning via quantum neural tangent kerne (arXiv:2111.04225 [quant-ph])
[60] Westerhout T, Astrakhantsev N, Tikhonov K S, Katsnelson M I and Bagrov A A 2020 *Nat. Commun.* **11** 1593
[61] Amari S-I 2020 Any target function exists in a neighborhood of any sufficiently wide random network: a geometrical perspective (arXiv:2001.06931 [stat.ML])
[62] Novak R, Xiao L, Hron J, Lee J, Alemi A A, Sohl-Dickstein J and Schoenholz S S 2020 *Int. Conf. on Learning Representations*
[63] Lee J, Schoenholz S S, Pennington J, Adlam B, Xiao L, Novak R and Sohl-Dickstein J 2020 Finite versus infinite neural networks: an empirical study (arXiv:2007.15801 [cs.LG])

[64] Yang G and Littwin E 2021 Tensor programs IIb: architectural universality of neural tangent kernel training dynamics (arXiv:2105.03703 [cs.LG])
[65] Nakaji K, Tezuka H and Yamamoto N 2021 Quantum-enhanced neural networks in the neural tangent kernel framework (arXiv:2109.03786 [quant-ph])
[66] Shirai N, Kubo K, Mitarai K and Fujii K 2021 Quantum tangent kernel (arXiv:2111.02951 [quant-ph])
[67] Zlokapa A, Neven H and Lloyd S 2021 A quantum algorithm for training wide and deep classical neural networks (arXiv:2107.09200 [quant-ph])