Reformulation of the No-Free-Lunch Theorem for Entangled Datasets

Kunal Sharma, M. Cerezo, Zoë Holmes, Lukasz Cincio, Andrew Sornborger, and Patrick J. Coles

Phys. Rev. Lett. **128**, 070501 — Published 18 February 2022

DOI: 10.1103/PhysRevLett.128.070501

Reformulation of the No-Free-Lunch Theorem for Entangled Data Sets

Kunal Sharma,[1,2,*] M. Cerezo,[1,3,*] Zoë Holmes,[4] Lukasz Cincio,[1] Andrew Sornborger,[4] and Patrick J. Coles[1]

[1] *Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*
[2] *Hearne Institute for Theoretical Physics and Department of Physics and Astronomy, Louisiana State University, Baton Rouge, LA USA*
[3] *Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM, USA*
[4] *Information Sciences, Los Alamos National Laboratory, Los Alamos, NM, USA*

The No-Free-Lunch (NFL) theorem is a celebrated result in learning theory that limits one's ability to learn a function with a training data set. With the recent rise of quantum machine learning, it is natural to ask whether there is a quantum analog of the NFL theorem, which would restrict a quantum computer's ability to learn a unitary process with quantum training data. However, in the quantum setting, the training data can possess entanglement, a strong correlation with no classical analog. In this work, we show that entangled data sets lead to an apparent violation of the (classical) NFL theorem. This motivates a reformulation that accounts for the degree of entanglement in the training set. As our main result, we prove a quantum NFL theorem whereby the fundamental limit on the learnability of a unitary is reduced by entanglement. We employ Rigetti's quantum computer to test both the classical and quantum NFL theorems. Our work establishes that entanglement is a commodity in quantum machine learning.

Introduction.— There are very few fields of science and technology that have not been impacted by machine learning. Yet progress in machine learning has been anything but steady, with periods of stagnation interleaved with periods of advancement [1]. This reflects the deep and non-trivial nature of learning theory. In order to advance the theory, fundamental results needed to be proven on the trainability, expressibility, and scalability of learning architectures such as neural networks [2].

One such fundamental result is the No-Free-Lunch (NFL) theorem [3–7]. At the conceptual level, the theorem states that different optimization procedures essentially perform the same when averaged over many problem instances and training data sets. At the mathematical level, the theorem has many alternative formulations, such as a statement that the average performance over all problem instances and training sets depends only on the size of the training data set and not on the optimization procedure. A consequence of this is that data must be considered the commodity or currency in machine learning that ultimately limits performance. Hence, this is why big data sets are viewed in such high regard.

Industry-built quantum computers of modest size are now publicly accessible over the cloud [8, 9]. This raises the intriguing possibility of quantum-assisted machine learning, a paradigm that researchers suspect could be more powerful than traditional machine learning [10, 11]. Various architectures for quantum neural networks (QNNs) have been proposed and implemented [12–20]. Some important results for quantum learning theory have already been obtained, particularly regarding the trainability [21–27] and expressibility [28] of QNNs for variational quantum algorithms [29–41]. However, the scalability of QNNs (to scales that are classically inaccessible) remains an interesting open question.

A quantum version of the NFL theorem could play an important role in understanding the scalability of QNNs. Recently, Poland et al. [42] made progress along these lines. They proved a lower bound on the average risk that depends only on the number of quantum states, t, used for training. Here, the risk is the probability of incorrectly learning a unitary process, which is the natural quantum analog of the classical risk. Their bound tends to zero only as t approaches the Hilbert-space dimension, which is exponentially large. This suggests that an exponentially large training data set is needed to learn a unitary. One can view this result as a roadblock in the path towards scaling QNNs, due to the apparent exponential (i.e., inefficient) scaling.

In this work, we consider a more general scenario, depicted in Fig. 1. Here, the goal is to learn a unitary with training data consisting of quantum states; however, these quantum states can now be entangled to a reference system. Such entangled states can be easily prepared on a quantum computer, and hence this scenario has practical relevance. A special case of this scenario is when the training data states have no entanglement with the reference system, corresponding to the scenario in Ref. [42].

Our main result is a quantum NFL theorem that generalizes the result in Ref. [42] by allowing for an arbitrary amount of entanglement in the training data. An amazing feature of our theorem is that our lower bound on the average risk is reduced as the Schmidt rank r of the entanglement grows. Furthermore the bound goes to zero when $r = d$, where d is the Hilbert space dimension, regardless of the number of training data points t. Given that our bound is tight (i.e., it can be saturated), this implies that one does not need an exponentially large training data set in order to learn a unitary. Hence, our work establishes that both big data *and* big entanglement are valuable in quantum machine learning, and that the currency of entanglement can lead to scalability.

Our work adds to the remarkable literature on entan-

* The first two authors contributed equally to this work.

glement as a resource. In communication theory, pre-shared entanglement allows one to transmit two bits of information by sending a single qubit [43]. In fundamental physics, an observer that is entangled to a system can guess the outcome of complementary measurements on that system, and this led researchers to generalize Heisenberg's uncertainty principle to allow for uncertainty reduction due to entanglement [44–46]. Our work is analogous to these examples, albeit in a different context.

We note that in [47], an important problem on learning an unknown unitary transformation from a finite number of examples was studied. In particular, [47] proved that whenever the unknown unitary is randomly drawn from a group the incoherent strategies achieve the ultimate performances for quantum learning. However, our results are different from [47] in the sense that we quantify the generalization error after training perfectly on the training set.

In what follows, we first discuss the classical NFL theorem. We then present our quantum NFL theorem, with the proof given in the Supplementary Information [48]. Finally, we perform numerical tests of both NFL theorems. This includes an implementation on Rigetti's quantum computer, which allows us to effectively violate the classical NFL theorem and also verify our quantum NFL theorem. We note that the Supplementary Material provides detailed proofs of all statements that follow.

Results.— In classical supervised machine learning, No-Free-Lunch (NFL) arises in the setting depicted in Fig. 1(a). Here the goal is to learn an unknown function f, where f maps a discrete input set $\mathcal{X}$ (of size $d_{\mathcal{X}}$) to a discrete output set $\mathcal{Y}$ (of size $d_{\mathcal{Y}}$). In this setting one generates from f a training set $\mathcal{S}$ in the form of t ordered input-output pairs as $\mathcal{S} = \{(x_j, y_j) : x_j \in \mathcal{X}, y_j := f(x_j) \in \mathcal{Y}\}_{j=1}^{t}$. This data is employed to train a hypothesis function $h_{\mathcal{S}}$ such that it matches perfectly the action of f on the training data. The hope is that $h_{\mathcal{S}}$ also makes accurate predictions on unknown, unseen data. However, as we will see, the NFL theorem provides a constraint on this.

To quantify how well the hypothesis function performs in predicting f one defines the risk function $R_f(h_{\mathcal{S}})$ as

$$R_f(h_{\mathcal{S}}) = \sum_{x\in\mathcal{X}} \pi(x)\mathbb{P}\Big[f(x) \neq h_{\mathcal{S}}(x)\Big]. \tag{1}$$

Specifically, $R_f(h_{\mathcal{S}})$ is the probability that $h_{\mathcal{S}}(x)$ and $f(x)$ differ across $\mathcal{X}$ when x is sampled from the probability distribution $\pi(x)$. While there are various mathematical versions of the NFL theorem [3–6], we follow the treatment in Ref. [6], which lower-bounds the risk when averaged over training sets $\mathcal{S}$ and functions f:

$$\mathbb{E}_f[\mathbb{E}_{\mathcal{S}}[R_f(h_{\mathcal{S}})]] \geqslant \left(1 - \frac{1}{d_{\mathcal{Y}}}\right)\left(1 - \frac{t}{d_{\mathcal{X}}}\right). \tag{2}$$

This is an information-theoretic bound (and hence is independent of the optimization method employed in training), implying that the average risk is limited by the size of the training set t, with the bound going to zero if $t = d_{\mathcal{X}}$. (Henceforth we drop the subscript when $d_{\mathcal{X}} = d_{\mathcal{Y}} = d$, as in Fig. 1.)

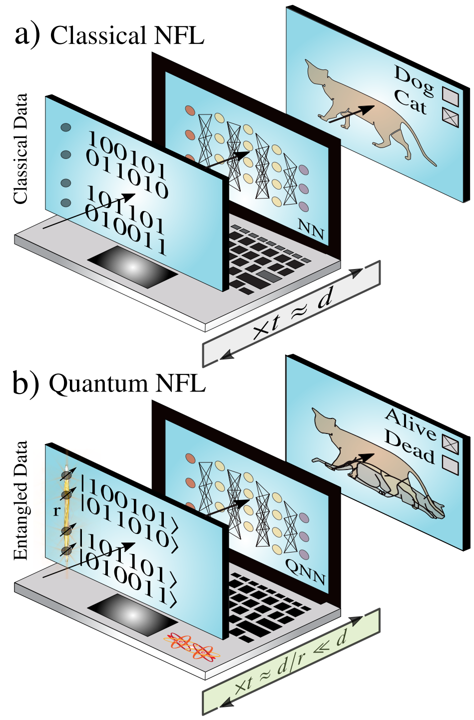

FIG. 1. **Depiction of the No-Free-Lunch setting.** (a) In classical supervised learning, one employs training data of size t to train a hypothesis to mimic the action of an unknown function on domain size d. Here we show input data in the form of bitstrings fed into a Neural Network (NN) to solve a binary classification problem. The NFL theorem indicates that it is the size of training data rather than the choice of optimization method that limits the average risk. Namely, small (large) t leads to big (small) generalization errors on average. (b) In quantum supervised learning, the goal is to learn a d-dimensional unitary process with t quantum states serving as training data. For generality, we allow these states to possibly be entangled with a reference system, with the Schmidt rank r quantifying the degree of entanglement. Here we show these states training a Quantum Neural Network (QNN) to classify quantum data (Schrodinger's cat being dead or alive). Our Quantum NFL theorem indicates that $r*t$ is the quantity that limits the average risk, and hence big entanglement (large r) leads to small generalization errors even when t is small.

As the NFL theorem is an information-theoretic result, the bound depends on the prior knowledge that one has about the set of maps from which f is chosen. Given that we will ultimately be interested in unitary maps in the quantum setting, one can consider classical analogs of unitaries in the classical setting for a meaningful comparison. Hence, we reformulate the classical NFL theorem for both stochastic and bistochastic matrices, which are somewhat analogous to unitaries. In the Supplementary Information we show that the classical NFL theorem for stochastic and bistochastic matrices can be expressed as

$$\mathbb{E}_f[\mathbb{E}_S[R_f(h_{\mathcal{S}})] \geqslant \left(1-\frac{t}{d}\right) F(d,t), \tag{3}$$

where $F(d,t)$ is the expectation over f of the squared distance between $f(x)$ the $h_S(x)$. In the stochastic case, we analytically find $F(d,t) = F(d) = \frac{e^2(d-1)}{(d+1)d^{d+1}}\left((d-2)^{d+1}+2(d-1)^d\right)$. In the bistochastic case, we simplify the expression of $F(d,t)$ such that it can be numerically computed. The case of f being a permutation matrix was considered in Ref. [42] and has a similar form as (3). All of these classical NFL results are conceptually similar, and dramatically different from the quantum case as we will see now.

Quantum NFL theorem.— Consider a quantum supervised learning task where the goal is to learn an unknown unitary U that maps a d-dimensional input Hilbert space $\mathcal{H}_{\mathcal{X}}$ to a d-dimensional output Hilbert space $\mathcal{H}_{\mathcal{Y}}$. Moreover, we consider a reference system $\mathcal{R}$, with $\mathcal{H}_{\mathcal{R}}$ denoting the associated Hilbert space, and we allow access to $\mathcal{R}$ during the training process. We suppose that all training data states have the same Schmidt rank $r \in \{1,2,...,d\}$ across the cut $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{R}}$. The training set is given by t pairs of input-output states $\mathcal{S}_Q = \{(|\psi_j\rangle, |\phi_j\rangle) : |\psi_j\rangle \in \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{R}}, |\phi_j\rangle \in \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{R}}\}_{j=1}^t$. Here, the output states are given by $|\phi_j\rangle = (U \otimes \mathbb{1}_{\mathcal{R}})|\psi_j\rangle$, where $\mathbb{1}_{\mathcal{R}}$ is the identity over $\mathcal{H}_{\mathcal{R}}$. During the training process, we allow for repeatable access to the states in $\mathcal{S}_Q$. Perfect training corresponds to the condition where the hypothesis unitary $V_{\mathcal{S}_Q}$ satisfies $|\langle\tilde{\phi}_j|\phi_j\rangle| = 1$ for all $j \in \{1,...,t\}$, where $|\tilde{\phi}_j\rangle = (V_{\mathcal{S}_Q} \otimes \mathbb{1}_{\mathcal{R}})|\psi_j\rangle$.

Similar to the classical case, we quantify the accuracy of the hypothesis $V_{\mathcal{S}_Q}$ via the quantum risk function:

$$R_U(V_{\mathcal{S}_Q}) = \int dx D_T^2(|y\rangle\langle y|, |\tilde{y}\rangle\langle\tilde{y}|), \tag{4}$$

defined as the average trace distance squared between the true output $|y\rangle = U|x\rangle$ and the hypothesis output $|\tilde{y}\rangle = V_{S_Q}|x\rangle$, where $|x\rangle \in \mathcal{H}_{\mathcal{X}}$ and $|y\rangle, |\tilde{y}\rangle \in \mathcal{H}_{\mathcal{Y}}$. Here, $D_T(\rho,\sigma) = \frac{1}{2}||\rho-\sigma||_1$, and the integral is over the uniform Haar measure dx on state space. Note that the risk is quantified on the smaller space $\mathcal{H}_{\mathcal{Y}}$ while the training is performed on the larger space $\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{R}}$.

Averaging the risk $R_U(V_{\mathcal{S}_Q})$ over all unitaries U and training sets $\mathcal{S}_Q$ leads to our main result:

$$\mathbb{E}_U[\mathbb{E}_{\mathcal{S}_Q}[R_U(V_{\mathcal{S}_Q})] \geqslant 1 - \frac{r^2t^2+d+1}{d(d+1)}, \tag{5}$$

which is a NFL theorem for entanglement-assisted quantum supervised learning. The proof is presented in the Supplementary Information, where we also show that the bound in (5) can be stated more generally in that it holds for all choices of $\mathcal{S}_Q$, and hence the average over $\mathcal{S}_Q$ is trivial and can be removed from (5). We show below in our numerical implementions that this bound is tight, and the inequality in (5) is saturated if the input states in $\mathcal{S}_Q$ are linearly independent (see Supplementary Material for more details).

Our proof for (5) relies on the assumption that the hypothesis unitary $V_{\mathcal{S}_Q}$ matches the target unitary U perfectly on the training set. This condition reduces the unitary $U^\dagger V_{\mathcal{S}_Q}$ to a simple block diagonal form. We then employ the Weingarten calculus to calculate the average over all target uniaries, which reduces to (5). We note that one does not need to perform tomography of states for evaluating the cost function. Rather, the overlap between the true output state and the output of the hypothesis unitary can be efficiently estimated, e.g., using the SWAP test.

Implications of results.— Let us discuss the implications of (5). First, consider the case of zero entanglement, $r=1$. In this case we recover the main result of Ref. [42], which states that the average risk is non-zero when $t<d$ and can only go to zero when $t=d$. Typically, $d=2^n$ will be exponentially large in the quantum setting, with n being the number of qubits, and hence this implies that an exponential amount of training data is needed to fully learn an unknown unitary.

At the other extreme, when there is maximal entanglement ($r=d$), one can see from (5) that only one training pair is sufficient for the lower bound on the average risk to reach zero. In the language of quantum information theory [49], this single training data point corresponds to the "Choi state" of the target unitary U. More generally, (5) indicates that the key quantity is $r*t$. When $r*t$ is small (large), the bound on the average risk is high (low). Hence, even moderate amounts of entanglement can improve the performance of quantum machine learning, by reducing the training data requirements.

The standard goal of quantum algorithms is quantum speedup, which typically corresponds to complexity scaling polynomially in n, since classical algorithms often exhibit exponential scaling. Variational quantum algorithms, which train QNNs, are no exception, and any exponential scaling in such algorithms destroys quantum speedup. Consequently, the quantum NFL theorem of Ref. [42], which corresponds to $r=1$ in our theorem, appeared to be a roadblock to quantum machine learning, since it suggested that an exponential amount of training data was required. Our work, on the other hand, appears to at least give some hope for quantum speedup with QNNs, provided that one has access to entangled training data. With that said, quantum speedup is a subtle issue, and we emphasize that (5) is derived under the assumption of perfect training. Hence one must analyze the complexity of training, and barren plateaus in training landscapes must be avoided in order to retain quantum speedup (see Discussion for elaboration).

In our implementations below, we compare the quantum and classical NFL theorems. We will argue that we observe an apparent violation of the classical NFL theorems. While these classical NFL theorems are of course

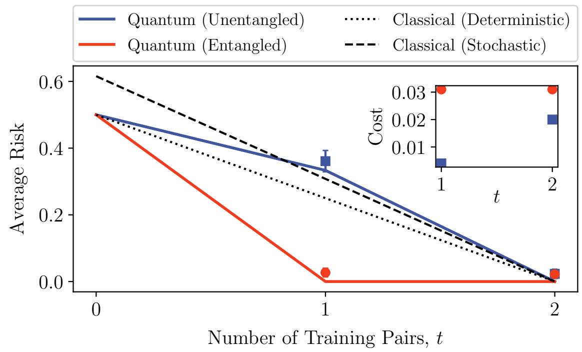

FIG. 2. **Implementation on Quantum Hardware.** Here we plot the average risk after learning 10 single-qubit unitaries on the Rigetti Aspen-4 quantum computer using 10 training sets consisting of $t = 1, 2$ unentangled $r = 1$ (blue squares) and entangled $r = 2$ (red circles) training states. The solid lines indicate the corresponding bounds imposed by our quantum NFL theorem, (5). Note, that while the optimizations were performed on the quantum computer, the final risk $R_U(V_{\mathcal{S}_Q})$ and optimal cost $C_U(V_{\mathcal{S}_Q})$ (plotted in the inset and defined in the Supplementary Information) were calculated classically to allow an accurate (i.e., noiseless) evaluation of the success of the optimizations. In black we plot the classical deterministic (dotted) and stochastic (dashed) NFL theorems.

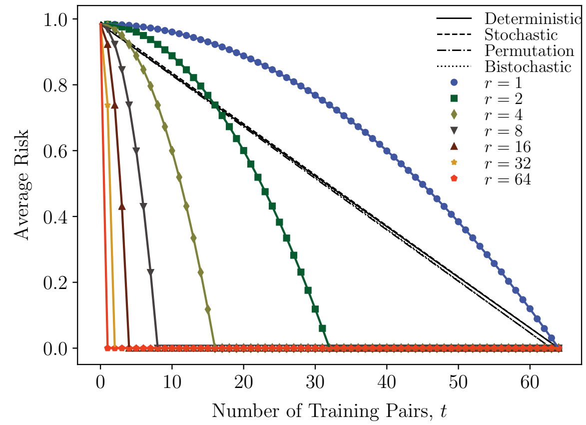

FIG. 3. **Large-Scale Test of NFL Theorems.** We plot the average risk versus t after learning 10 six-qubit unitaries on a simulator for 100 training sets. Each training set consisted of $t = 1, ..., 64$ training pairs of rank $r = 2^0, ..., 2^6$. The markers indicate the optimization results, whereas the solid lines indicate the bounds imposed by our quantum NFL theorem, (5). The simulation error bars are $O(10^{-3})$ and therefore smaller than the size of the markers. In black, we plot the classical NFL bounds for deterministic (solid), stochastic (dashed), permutation (dot-dashed), and bistochastic maps.

valid under the setting of their formulation, this setting nevertheless does not allow for entangled data. Hence the apparent violation is due to the fact that the physical laws of nature allow for a more general setting than the assumed setting of these theorems. We also remark that one could allow for a reference system $\mathcal{R}$ in the classical setting (like we do in the quantum setting). However, access to such a system would not change the bounds in the classical NFL theorems. This is because, in the classical setting, no correlation between $\mathcal{R}$ and $\mathcal{X}$ would be possible under the standard assumption that the joint state is a pure state. (Training with mixed states is not allowed since that would correspond to training with multiple pure states and, arguably, would be cheating.) Hence, allowing for $\mathcal{R}$ in the classical setting is trivial.

Implementations.— The availability of cloud-based quantum computers offers the possibility of testing the validity of NFL theorems with truly entangled data sets. In what follows, we present numerical results for quantum supervised learning, with the task of learning randomly generated unitaries, using entangled training states of increasing Schmidt rank. The details of our implementations are presented in the Supplementary Information.

We first employ Rigetti's Aspen-4 quantum device [9] to learn 2×2 unitaries. This involves a hybrid quantum-classical optimization loop where the quantum computer evaluates a cost function that quantifies the quality of the training on $\mathcal{S}_Q$, and then the parameters of the hypothesis unitary are adjusted classically to reduce the cost. Figure 2 shows the average risk versus t, after running this optimization loop, for training sets consisting of $t = 1, 2$ unentangled ($r = 1$) and entangled ($r = 2$) states. To compare the performance to the fundamental limits imposed by the NFL theorems, we also plot the classical bounds for deterministic (2) and stochastic (3) maps as well as our quantum bound in (5). Good agreement is observed for our quantum bound with the small discrepancies attributable to imperfect learning (due to the presence of quantum noise it was not possible to completely minimize the cost function as shown in the inset) and finite-size averaging when computing the average risk. The average risk using a single entangled training pair ($t = 1, r = 2$) is substantially lower than both the average risk using a single unentangled training pair ($t = 1, r = 1$) and that allowed by the deterministic and stochastic classical NFL theorems, suggesting an apparent violation of these classical bounds.

While noise and other constraints limit the size of our quantum-hardware implementations, we can nevertheless explore larger systems on a simulator. Figure 3 plots the average risk when learning 64-dimensional unitaries on a simulator for $t = 1, ..., 64$ training states of Schmidt rank $r = 2^0, ..., 2^6$. Near-perfect agreement between the simulation data and the bound in (5) is observed in all cases. Furthermore, for $r > 1$ it is possible to reduce the average risk below that allowed by four different classical NFL bounds (which have very similar behavior). We remark that 2-dimensional permutation and bistochastic matrices can be learned with a single training pair and hence it was not possible to violate the permutation and bistochastic classical bounds for the previous 2-dimensional implementation; whereas our 64-dimensional implementation easily violates these bounds.

Discussion.— Quantum machine learning is a relatively new field that has already seen one major shift, from algorithms for the fault-tolerant era to variational methods for training Quantum Neural Networks (QNNs) in the near-term era. While several intriguing QNN architectures and training strategies have been proposed, rigorous results are urgently needed, in particular, to understand whether QNNs will offer a quantum speedup. In this work, we have contributed a rigorous theorem with implications for QNN scalability. While it previously appeared that an exponentially large training set would be required to train a QNN, our quantum No-Free-Lunch (NFL) theorem shows that entanglement in the training data can compensate for and remove this exponential overhead. This suggests that entanglement should be considered as a valuable resource in reducing the generalization error in quantum machine learning. While our work provides a glimmer of hope that quantum machine learning could yield a quantum speedup (i.e., polynomial scaling), there are still several issues and open questions that we now discuss.

One potential issue is the complexity of obtaining the entangled training data in the first place. This complexity will depend on the mode of access to the data. We note that for the setting where a user has physical access to the target unitary, then it is advantageous to input a state entangled with a reference system to the unitary so that the user can generate input-output training data with entanglement [40]. This procedure can overall decrease the average risk more efficiently in comparison to the input with no entanglement.

Another potential issue is the complexity of training. While our quantum NFL theorem assumes perfect training, it is possible that exponential scaling could be hidden in the training difficulty, especially in light of recent results on barren plateaus (exponentially vanishing gradients) in QNN cost function landscapes [21–23]. While several promising strategies have been proposed to avoid barren plateaus in QNNs [24–27], this remains an active area of research. We speculate that for cases when one needs only a polynomial number of shots for training (i.e., no barren plateau issues), learning a unitary using an entangled training set is more advantageous than training sets with no entanglement. Deriving a no-free-lunch (NFL) theorem that accounts for finite accuracy in training is an interesting open question that we leave for future work.

This highlights an important direction for future work. Naturally, it would be useful to extend the quantum NFL theorem to the case where one does not achieve perfect training on the training set. Such imperfect training could either be the result of shot noise or hardware noise, or could simply be due to local minima in the landscape. In this case, the lower bound in (5) would not be saturated, and hence it would be of interest to tighten the bound to account for imperfect training.

Acknowledgments

KS, LC, and PJC were supported by the U.S. Department of Energy (DOE), Office of Science, Office of Advanced Scientific Computing Research, under the Accelerated Research in Quantum Computing (ARQC) program. MC and PJC were supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory (LANL) under project number 20180628ECR. MC was also supported by the Center for Nonlinear Studies at LANL. ZPH, ATS, and PJC acknowledge support from the LANL ASC Beyond Moore's Law project.

Supplementary Information.— The Supplementary Information contains details of our proofs and a Reference [50].

[1] Simon Haykin, Neural networks: a comprehensive foundation (Prentice Hall PTR, 1994).
[2] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams, "Learning representations by back-propagating errors," Nature **323**, 533–536 (1986).
[3] David H Wolpert and William G Macready, "No free lunch theorems for optimization," IEEE transactions on evolutionary computation **1**, 67–82 (1997).
[4] David H Wolpert, William G Macready, et al., No free lunch theorems for search, Tech. Rep. (Technical Report SFI-TR-95-02-010, Santa Fe Institute, 1995).
[5] Stavros P Adam, Stamatios-Aggelos N Alexandropoulos, Panos M Pardalos, and Michael N Vrahatis, "No free lunch theorem: a review," in Approximation and Optimization (Springer, 2019) pp. 57–82.
[6] M. M. Wolf, "Mathematical foundations of supervised learning," (2018).
[7] Shai Shalev-Shwartz and Shai Ben-David, Understanding machine learning: From theory to algorithms (Cambridge university press, 2014).
[8] Matthias Steffen, Jay M. Gambetta, and Jerry M. Chow, "Progress, status, and prospects of superconducting qubits for quantum computing," in 2016 46th European Solid-State Device Research Conference (ESSDERC) (2016) pp. 17–20.
[9] Peter J Karalekas, Nikolas A Tezak, Eric C Peterson, Colm A Ryan, Marcus P da Silva, and Robert S Smith, "A quantum-classical cloud platform optimized for variational hybrid algorithms," Quantum Science and Technology **5**, 024003 (2020).
[10] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd, "Quantum machine learning," Nature **549**, 195–202 (2017).
[11] Vedran Dunjko, Jacob M Taylor, and Hans J Briegel, "Quantum-enhanced machine learning," Physical review letters **117**, 130501 (2016).

[12] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione, "The quest for a quantum neural network," Quantum Information Processing **13**, 2567–2586 (2014).
[13] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione, "Quantum walks on graphs representing the firing patterns of a quantum neural network," Physical Review A **89**, 032333 (2014).
[14] Vedran Dunjko and Hans J Briegel, "Machine learning & artificial intelligence in the quantum domain: a review of recent progress," Reports on Progress in Physics **81**, 074001 (2018).
[15] Guillaume Verdon, Jason Pye, and Michael Broughton, "A universal training algorithm for quantum deep learning," arXiv preprint arXiv:1806.09729 (2018).
[16] Edward Farhi and Hartmut Neven, "Classification with quantum neural networks on near term processors," arXiv preprint arXiv:1802.06002 (2018).
[17] Carlo Ciliberto, Mark Herbster, Alessandro Davide Ialongo, Massimiliano Pontil, Andrea Rocchetto, Simone Severini, and Leonard Wossnig, "Quantum machine learning: a classical perspective," Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences **474**, 20170551 (2018).
[18] Nathan Killoran, Thomas R Bromley, Juan Miguel Arrazola, Maria Schuld, Nicolás Quesada, and Seth Lloyd, "Continuous-variable quantum neural networks," Physical Review Research **1**, 033063 (2019).
[19] Iris Cong, Soonwon Choi, and Mikhail D Lukin, "Quantum convolutional neural networks," Nature Physics **15**, 1273–1278 (2019).
[20] Kerstin Beer, Dmytro Bondarenko, Terry Farrelly, Tobias J Osborne, Robert Salzmann, Daniel Scheiermann, and Ramona Wolf, "Training deep quantum neural networks," Nature Communications **11**, 1–6 (2020).
[21] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven, "Barren plateaus in quantum neural network training landscapes," Nature communications **9**, 4812 (2018).
[22] M Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J Coles, "Cost-function-dependent barren plateaus in shallow quantum neural networks," arXiv preprint arXiv:2001.00550 (2020).
[23] Kunal Sharma, M Cerezo, Lukasz Cincio, and Patrick J Coles, "Trainability of dissipative perceptron-based quantum neural networks," arXiv preprint arXiv:2005.12458 (2020).
[24] Edward Grant, Leonard Wossnig, Mateusz Ostaszewski, and Marcello Benedetti, "An initialization strategy for addressing barren plateaus in parametrized quantum circuits," Quantum **3**, 214 (2019).
[25] Guillaume Verdon, Michael Broughton, Jarrod R McClean, Kevin J Sung, Ryan Babbush, Zhang Jiang, Hartmut Neven, and Masoud Mohseni, "Learning to learn with quantum neural networks via classical neural networks," arXiv preprint arXiv:1907.05415 (2019).
[26] Tyler Volkoff and Patrick J Coles, "Large gradients via correlation in random parameterized quantum circuits," arXiv preprint arXiv:2005.12200 (2020).
[27] Andrea Skolik, Jarrod R McClean, Masoud Mohseni, Patrick van der Smagt, and Martin Leib, "Layerwise learning for quantum neural networks," arXiv preprint arXiv:2006.14904 (2020).
[28] Sukin Sim, Peter D Johnson, and Alán Aspuru-Guzik, "Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms," Advanced Quantum Technologies **2**, 1900070 (2019).
[29] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, "A variational eigenvalue solver on a photonic quantum processor," Nature Communications **5**, 4213 (2014).
[30] Bela Bauer, Dave Wecker, Andrew J Millis, Matthew B Hastings, and Matthias Troyer, "Hybrid quantum-classical approach to correlated materials," Physical Review X **6**, 031045 (2016).
[31] Jarrod R McClean, Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik, "The theory of variational hybrid quantum-classical algorithms," New Journal of Physics **18**, 023023 (2016).
[32] A. Arrasmith, L. Cincio, A. T. Sornborger, W. H. Zurek, and P. J. Coles, "Variational consistent histories as a hybrid algorithm for quantum foundations," Nature communications **10**, 3438 (2019).
[33] Tyson Jones, Suguru Endo, Sam McArdle, Xiao Yuan, and Simon C Benjamin, "Variational quantum algorithms for discovering hamiltonian spectra," Physical Review A **99**, 062304 (2019).
[34] X. Xu, J. Sun, S. Endo, Y. Li, S. C. Benjamin, and X. Yuan, "Variational algorithms for linear algebra," arXiv:1909.03898 [quant-ph].
[35] Carlos Bravo-Prieto, Ryan LaRose, M. Cerezo, Yigit Subasi, Lukasz Cincio, and Patrick J. Coles, "Variational quantum linear solver: A hybrid algorithm for linear systems," arXiv:1909.05820 (2019).
[36] Xiao Yuan, Suguru Endo, Qi Zhao, Ying Li, and Simon C Benjamin, "Theory of variational quantum simulation," Quantum **3**, 191 (2019).
[37] Cristina Cirstoiu, Zoe Holmes, Joseph Iosue, Lukasz Cincio, Patrick J Coles, and Andrew Sornborger, "Variational fast forwarding for quantum simulation beyond the coherence time," arXiv preprint arXiv:1910.04292 (2019).
[38] Marco Cerezo, Alexander Poremba, Lukasz Cincio, and Patrick J Coles, "Variational quantum fidelity estimation," Quantum **4**, 248 (2020).
[39] M Cerezo, Kunal Sharma, Andrew Arrasmith, and Patrick J Coles, "Variational quantum state eigensolver," arXiv preprint arXiv:2004.01372 (2020).
[40] S. Khatri, R. LaRose, A. Poremba, L. Cincio, A. T. Sornborger, and P. J. Coles, "Quantum-assisted quantum compiling," Quantum **3**, 140 (2019).
[41] Ryan LaRose, Arkin Tikku, Étude OâĂŹNeel-Judy, Lukasz Cincio, and Patrick J Coles, "Variational quantum state diagonalization," npj Quantum Information **5**, 1–10 (2019).
[42] Kyle Poland, Kerstin Beer, and Tobias J Osborne, "No free lunch for quantum machine learning," arXiv preprint arXiv:2003.14103 (2020).

[43] Charles H Bennett and Stephen J Wiesner, "Communication via one-and two-particle operators on einstein-podolsky-rosen states," Physical review letters **69**, 2881 (1992).

[44] Mario Berta, Matthias Christandl, Roger Colbeck, Joseph M Renes, and Renato Renner, "The uncertainty principle in the presence of quantum memory," Nature Physics **6**, 659–662 (2010).

[45] Mario Berta, Patrick J Coles, and Stephanie Wehner, "Entanglement-assisted guessing of complementary measurement outcomes," Physical Review A **90**, 062127 (2014).

[46] Patrick J Coles, Mario Berta, Marco Tomamichel, and Stephanie Wehner, "Entropic uncertainty relations and their applications," Reviews of Modern Physics **89**, 015002 (2017).

[47] Alessandro Bisio, Giulio Chiribella, Giacomo Mauro D'Ariano, Stefano Facchini, and Paolo Perinotti, "Optimal quantum learning of a unitary transformation," Phys. Rev. A **81**, 032324 (2010).

[48] See Supplementary Information, which contains Refs. [6, 42, 50–53].

[49] M. A. Nielsen and I. L. Chuang, Quantum Computation and Quantum Information, 10th ed. (Cambridge University Press, New York, NY, USA, 2011).

[50] Uttam Singh, Lin Zhang, and Arun Kumar Pati, "Average coherence and its typicality for random pure states," Physical Review A **93**, 032125 (2016).

[51] Michael A Nielsen, "A simple formula for the average gate fidelity of a quantum dynamical operation," Physics Letters A **303**, 249–252 (2002).

[52] Benoît Collins and Piotr Śniady, "Integration with respect to the haar measure on unitary, orthogonal and symplectic group," Communications in Mathematical Physics **264**, 773–795 (2006).

[53] Zbigniew Puchała and Jaroslaw Adam Miszczak, "Symbolic integration with respect to the haar measure on the unitary groups," Bulletin of the Polish Academy of Sciences Technical Sciences **65**, 21–27 (2017).