Supplemental Material for the article "Quantum machine learning in feature Hilbert spaces"

Maria Schuld* and Nathan Killoran
Xanadu, 372 Richmond St W, Toronto, M5V 2L7, Canada
(Dated: November 7, 2018)

I. DEFINITIONS

In the Letter we mention the concept of a *Reproducing Kernel Hilbert Space*, which we formally define here:

Definition 1. *Let $\mathcal{X}$ be a non-empty input set and $\mathcal{R}$ a Hilbert space of functions $f : \mathcal{X} \to \mathbb{C}$ that map inputs to the real numbers. Let $\langle \cdot, \cdot \rangle$ be an inner product defined on $\mathcal{R}$ (which gives rise to a norm via $||f|| = \sqrt{\langle f, f \rangle}$). $\mathcal{R}$ is a* reproducing kernel Hilbert space *if every point evaluation is a continuous functional $F : f \to f(x)$ for all $x \in \mathcal{X}$. This is equivalent to the condition that there exists a function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ for which*

$$\langle f, \kappa(x, \cdot) \rangle = f(x) \tag{1}$$

with $\kappa(x, \cdot) \in \mathcal{R}$ and for all $f \in \mathcal{H}$, $x \in \mathcal{X}$.

The function κ is the unique *reproducing kernel* of $\mathcal{R}$, and Eq. (1) is the *reproducing property*. This property implies that $\langle \kappa(x, \cdot), \kappa(x', \cdot) \rangle = \kappa(x, x')$, which means that the kernel function is equivalent to the inner product of vectors from the RKHS. Note that a different, but isometrically isomorphic Hilbert space can be derived for a so-called Mercer kernel [2].

Second, in the Letter we make use of the *representer theorem,* which can be defined as follows [5].

Theorem 1. *Let $\mathcal{X}$ be an input set, $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a kernel, $\mathcal{D}$ a data set consisting of data pairs $(x^m, y^m) \in \mathcal{X} \times \mathbb{R}$ and $f : \mathcal{X} \to \mathbb{R}$ a class of model functions that live in the reproducing kernel Hilbert space $\mathcal{R}_\kappa$ of κ. Furthermore, assume we have a cost function $\mathcal{C}$ that quantifies the quality of a model by comparing predicted outputs $f(x^m)$ with targets y^m, and which has a regularisation term of the form $g(||f||)$ where $g : [0, \infty) \to \mathbb{R}$ is a strictly monotonically increasing function. Then any function $f^* \in \mathcal{R}_\kappa$ that minimises the cost function C can be written as*

$$f^*(x) = \sum_{m=1}^{M} \alpha_m \kappa(x, x^m), \tag{2}$$

for some parameters $\alpha_m \in \mathbb{R}$.

The representer theorem underlines the significance of kernels, since many models can be expressed by a weighed expansion of kernel functions on the training data.

II. REPRODUCING KERNELS OF QUANTUM SYSTEMS

In this section we will give an answer to the question of which reproducing kernels the Hilbert space of generic quantum systems gives rise to. We show that for Hilbert spaces with discrete bases, as well as for the special 'continuous-basis' case of the Hilbert space of coherent states, the reproducing kernel of a quantum Hilbert space is given by inner products of basis vectors. This insight can lead to interesting results. For example, Chatterjee et al. [1] show that the inner product of an optical coherent state can be turned into a *Gaussian kernel* (also called *radial basis function kernel*) which is widely used in machine learning.

Quantum theory prescribes that the state of a quantum system is modelled by a vector in a Hilbert space $\mathcal{H}_s$. In a typical setting, the Hilbert space is constructed from a complete basis of eigenvectors $\{|s\rangle\}$ of a complete set of

* maria@xanadu.ai

commuting Hermitian operators which corresponds to physical observables. Due to the hermiticity of the observables, the basis is orthogonal, and it can be continuous (i.e., if the observable is the position operator describing the location of a particle), countably infinite (i.e., observing the number of photons in an electric field), or finite (i.e., observing the spin of an electron). Vectors in the Hilbert space are abstractly referred to as $|\psi\rangle \in \mathcal{H}$ in Dirac notation. However, every such Hilbert space has a *functional representation*. In the case of a discrete basis of dimension $N \in \mathbb{N} \cup \infty$, the functional representation $\mathcal{H}_s^f$ of $\mathcal{H}_s$ is given by the (Hilbert) space l^2 of square summable sequences $\{\psi(s_i) = \langle s_i|\psi\rangle\}_{i=1}^N$ with the inner product $\langle \psi, \varphi\rangle = \sum_{s_i} \psi(s_i)^* \varphi(s_i)$. In the continuous case this is the space L^2 of square summable (equivalence classes of) functions $\psi(s) = \langle s|\psi\rangle$ with the inner product $\langle \psi, \varphi\rangle = \int ds \psi(s)^* \varphi(s)$. The preceding formulation of quantum theory therefore associates every quantum system with a Hilbert space of functions mapping from a set $\mathcal{S} = \{s\}$ to the complex numbers. The question is if these Hilbert spaces give rise to a reproducing kernel that makes them a RKHS with respect to the input set $\mathcal{S}$.

With the resolution of identity $\mathbb{1} = \int ds\ |s\rangle\langle s|$ for the continuous and $\mathbb{1} = \sum_i |s_i\rangle\langle s_i|$ for the discrete case, we can immediately "create" the reproducing property from Eq. (1). Consider first the discrete case:

$$\psi(s_i) = \langle s_i|\psi\rangle = \sum_{s_j} \langle s_i|s_j\rangle\langle s_j|\psi\rangle = \langle \langle s| \cdot \rangle |\psi(\cdot)\rangle.$$

We can identify $\langle s_i|s_j\rangle$ with the reproducing kernel. Since the basis is orthonormal, we have $\kappa(s_i, s_j) = \delta_{i,j}$. The continuous case is more subtle. Inserting the identity, we get

$$\psi(s) = \int ds' \langle s|s'\rangle\langle s'|\psi\rangle = \langle \langle s| \cdot \rangle, \psi(\cdot)\rangle,$$

which is the reproducing kernel property with the reproducing kernel $\kappa(s, s') = \langle s|s'\rangle$. However, the "function" $s'(s) = \delta(s - s')$ is not square integrable, which means it is itself not part of $\mathcal{H}_s^f$, and the properties of Definition 1 are not fulfilled. This is no surprise, as the space of square integrable functions L^2 is a frequent example of a Hilbert space that is not a RKHS [6]. The inconsistency between Dirac's formalism and functional analysis is also a well-known issue in quantum theory, but usually glossed over in physical contexts [7]. If mathematical rigour is needed, physicists usually refer to the theory of rigged Hilbert spaces [8].

There are quantum systems with an infinite basis which naturally give rise to a reproducing kernel that is not the delta function. These systems are described by so-called *generalised coherent states* [9]. In the context of quantum machine learning, this has been discussed in Ref. [1]. Generalised coherent states are vectors $|l\rangle$ in a Hilbert space $\mathcal{H}_c$ of finite or countably infinite dimension, and where the index l is from some topological space $\mathcal{L}$ (allowing us to define a norm $|||l\rangle|| = \sqrt{\langle l|l\rangle}$). They have two fundamental properties. First, $|l\rangle$ is a strongly continuous function of l,

$$\lim_{l \to l'} || \ |l'\rangle - |l\rangle|| = 0, \ |l\rangle \neq 0.$$

Note that this excludes for example the discrete Fock basis $\{|n\rangle\}$, but also any orthonormal set of states $\{|z\rangle\}$ with a continuous label $z \in \mathbb{C}$, since $\frac{1}{2}|||z'\rangle - |z\rangle|| = 1$ for $z' \neq z$. Second, there exists a measure μ on $\mathcal{L}$ so that we have a resolution of identity $\mathbb{1} = \int_{\mathcal{L}} |l\rangle\langle l|\ d\mu(l)$. This leads to a functional representation of the Hilbert space where a vector $|\psi\rangle \in \mathcal{H}_c$ is expressed via $|\psi\rangle = \sum_l \psi(l)|l\rangle$ with $\psi(l) = \langle l|\psi\rangle$. Inserting the resolution of identity to the right hand side of this expression yields

$$\psi(l) = \int_{\mathcal{L}} \langle l|l'\rangle\langle l'|\psi\rangle\ d\mu(l'),$$

which is exactly the reproducing property in Definition 1 with the reproducing kernel $\kappa(l, l') = \langle l|l'\rangle$. Since there is a finite overlap between any two states from the basis, the kernel is not the Dirac delta function, and we do not run into the same problem as for continuous orthogonal bases. Hence, the Hilbert space of coherent states is an RKHS for the input set $\{l\}$.

The most well-known type of coherent state are optical coherent states

$$|\alpha\rangle = e^{-\frac{|\alpha|^2}{2}} \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}} |n\rangle,$$

which are the eigenstates of the non-Hermitian bosonic creation operator $\hat{a}$, with the associated kernel

$$\kappa(\alpha, \beta) = \langle \alpha | \beta \rangle = \mathrm{e}^{-\left(\frac{|\alpha|^2}{2} + \frac{|\beta|^2}{2} - \alpha\beta\right)}, \tag{3}$$

whose square is a *radial basis function* or *Gaussian kernel* as remarked in [1].

III. EXAMPLES OF QUANTUM KERNELS

In the main Letter, we introduce the concept of a quantum feature map, which in quantum computing corresponds to a feature-embedding circuit $U_\phi(x)$ that depends on a classical input x. Here we show some examples of popular methods of information encoding in quantum machine learning, their feature-embedding circuit, and the 'quantum kernel' they give rise to.

a. Basis encoding. Many quantum machine learning algorithms assume that the inputs x to the computation are encoded as binary strings represented by a computational basis state of the qubits [10, 11]. For example, $x = 01001$ is represented by the 5-qubit basis state $|01001\rangle$. The computational basis state corresponds to a standard basis vector $|i\rangle$ (with i being the integer representation of the bitstring) in a 2^n-dimensional Hilbert space $\mathcal{F}$, and the effect of the feature-embedding circuit is given by

$$U_\phi : x \in \{0, 1\}^n \to |i\rangle.$$

This feature map maps each data input to a state from an orthonormal basis and is equivalent to the generic finite-dimensional case discussed in the previous Section II. As shown there, the generic kernel is the Kronecker delta

$$\kappa(x, x') = \langle i | j \rangle = \delta_{ij},$$

which is a binary similarity measure that is only nonzero for two identical inputs.

b. Amplitude encoding. Another approach to information encoding is to associate normalised input vectors $\mathbf{x} = (x_0, ..., x_{N-1})^T \in \mathbb{R}^N$ of dimension $N = 2^n$ with the amplitudes of a n qubit state $|\psi_\mathbf{x}\rangle$ [12, 13],

$$U_\phi : \mathbf{x} \in \mathbb{R}^N \to |\psi_\mathbf{x}\rangle = \sum_{i=0}^{N-1} x_i |i\rangle.$$

As above, $|i\rangle$ denotes the i'th computational basis state. This choice corresponds to the linear kernel,

$$\kappa(\mathbf{x}, \mathbf{x}') = \langle \psi_\mathbf{x} | \psi_{\mathbf{x}'} \rangle = \mathbf{x}^T \mathbf{x}'.$$

c. Copies of quantum states. With a slight variation of amplitude encoding we can implement polynomial kernels [14]. Taking d copies of an amplitude encoded quantum state,

$$U_\phi : \mathbf{x} \in \mathbb{R}^N \to |\psi_\mathbf{x}\rangle \otimes \cdots \otimes |\psi_\mathbf{x}\rangle,$$

corresponds to the kernel

$$\kappa(\mathbf{x}, \mathbf{x}') = \langle \psi_\mathbf{x} | \psi_{\mathbf{x}'} \rangle \cdots \langle \psi_\mathbf{x} | \psi_{\mathbf{x}'} \rangle = (\mathbf{x}^T \mathbf{x}')^d.$$

d. Product encoding. One can also use a (tensor) product encoding, in which each feature of the input $\mathbf{x} = (x_1, .., x_N)^T \in \mathbb{R}^N$ is encoded in the amplitudes of one separate qubit. An example is to encode x_i as $|\phi(x_i)\rangle = \cos(x_i)|0\rangle + \sin(x_i)|1\rangle$ for $i = 1, ..., N$[15, 16]. This corresponds to a feature-embedding circuit with the effect

$$U_\phi : \mathbf{x} \in \mathbb{R}^N \to \begin{pmatrix} \cos x_1 \\ \sin x_1 \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} \cos x_N \\ \sin x_N \end{pmatrix} \in \mathbb{R}^{2^N},$$

and implies a cosine kernel,

$$\kappa(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^{N} \cos(x_i - x_i').$$

IV. LINEAR SEPARABILITY IN FOCK SPACE

We show here that if we map the inputs of a dataset $\mathcal{D}$ to a new dataset

$$\mathcal{D}' = \{|(c, \mathbf{x}^1)\rangle, ..., |(c, \mathbf{x}^M)\rangle\},$$

using the squeezing feature map with phase encoding from the main Letter, the data always becomes linearly separable in Fock space $\mathcal{F}$. Linear separability means that any assignment of two classes of labels to the data can be separated by a hyperplane in $\mathcal{F}$. We give a formal proof as well as numerical confirmation.

First consider the following:

Proposition 1. *A set of M vectors in $\mathbb{R}^N$ are linearly separable if $M-1$ of them are linear independent.*

The proof can be found in Subsection A below. Proposition 1 tells us that if our data is linearly independent, it is linearly separable. This result is in fact known from statistical learning theory: The VC dimension – a measure of flexibility or expressive power – of linear models in K dimensions is $K+1$, which means that a linear model can separate or "shatter" $K+1$ data points if we can choose the strategy of how to arrange them, but not the strategy of how they are labelled.

If we can show that the squeezing feature map maps vectors to linearly independent states in Fock space, we know that any dataset becomes linearly separable in Fock space. To simplify, lets first see look at the squeezing map of a single mode.

Proposition 2. *Given a set of squeezing phases $\{\varphi^1, ..., \varphi^M\}$ with $\varphi^m \neq \varphi^{m'}$ for $m = 1, ..., M, m \neq m'$ and a hyperparameter $c \in \mathbb{R}$, the squeezed vacuum Fock states $|(c, \varphi^1)\rangle, ..., |(c, \varphi^M)\rangle$ are linearly independent.*

The proof is found in Section B below. A very similar proof confirms that the proposition also holds true for the sueezing map where the real scalars are encoded in the amplitude r of the squeezing parameter.

For the multimode feature map dealing with input data of dimension higher than one,

$$|(c, \boldsymbol{\varphi}^m)\rangle = |(c, \varphi_1^m)\rangle \otimes \ldots \otimes |(c, \varphi^m))\rangle,$$

and

$$|(c, \boldsymbol{\varphi}^{m'})\rangle = |(c, \varphi_1^{m'})\rangle \otimes \ldots \otimes |(c, \varphi^{m'}))\rangle.$$

We have

$$\langle(c, \boldsymbol{\varphi}^m)|(c, \boldsymbol{\varphi}^{m'})\rangle = \prod_{i=1}^{N} \langle(c, \varphi_i^m)|(c, \varphi_i^{m'})\rangle,$$

which is 1 if $\varphi_i^m = \varphi_i^{m'}$ for all $i = 1, ..., N$ and a value other than zero else. The linear independence therefore carries over to multi-dimensional feature maps.

In order to confirm the numerical separability of data, we revisit the 'blobs' data set from Figure 4 in the paper and apply a perceptron classifier to the data in feature space. The perceptron is guaranteed to find a separating hyperplane if it exists. Figure 1 shows the performance of a perceptron classifier in Fock space. The data was mapped to this space by the squeezing feature map with phase encoding. As one can see, after 5000 epochs (runs through the dataset) the decision boundary perfectly fits the training data, achieving an accuracy of 1. The number of iterations to train the perceptron is known to increase with $\mathcal{O}(1/\gamma^2)$ where γ is the margin between the two classes [17], and indeed we find in other simulations that the 'moons' and 'circles' data only take a few epochs until reaching full accuracy. Although the perfect fit to the training data is of course not useful for machine learning (as can be seen by the non-increasing accuracy on the test set) these results confirm that the squeezing feature map makes data linearly separable in feature space.

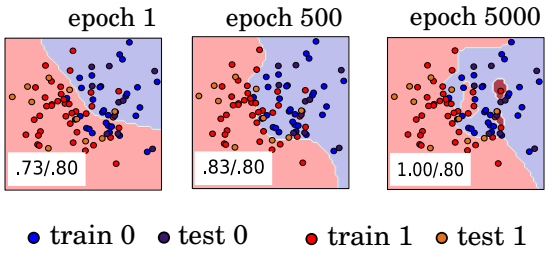

FIG. 1. Decision boundary of a perceptron classifier in Fock space after mapping the 2-dimensional data points via the squeezing feature map with phase encoding (with $c = 1.5$). The perceptron only acts on the real subspace and without regularisation. The 'blobs' dataset has now only 70 training and 20 test samples. The perceptron achieves a training accuracy of 1 after less than 5000 epochs, which means that the data is linearly separable in Fock space. Interestingly, in this example the test performance remains exactly the same. The simulations were performed with the Strawberry Fields simulator as well as a scikit-learn out-of-the-box perceptron classifier.

A. Proof of Proposition 1

Let

$$\mathcal{D} = \{(x^1, y^1), \cdots, (x^M, y^M)\}$$

be a dataset of M vectors with $x^m \in \mathbb{R}^N$ for all $m = 1, \cdots, M$, and $y \in \{-1, 1\}$. The vectors are guaranteed to be linearly separable if for any assignment of classes $\{-1, 1\}$ to labels $y^1, ..., y^M$ there is a hyperplane defined by parameters $w_1, ..., w_N, b$ so that

$$\text{sgn}(\sum_{i=1}^{N} w_i x_i^m + b) = y^m \quad \forall m = 1, ..., M. \tag{4}$$

The sign function is a bit tricky, but if we can instead show that the stronger condition

$$\sum_{i=1}^{N} w_i x_i^m + b = y^m \quad \forall m = 1, ..., M \tag{5}$$

holds for some parameters, Eq. 4 must automatically be satisfied.

Equation 5 defines a system of M linear equations with $N+1$ unknowns (namely the variables). From the theorey of linear algebra we know [18] that there is at least one solution if and only if the rank of the 'coefficient matrix'

$$[X|1] = \begin{pmatrix} x_1^1 & \cdots & x_N^1 & 1 \\ \vdots & \ddots & \vdots & \vdots \\ x_1^M & \cdots & x_N^M & 1 \end{pmatrix}$$

is equal to the rank of its augmented matrix

$$[X|1|y] = \begin{pmatrix} x_1^1 & \cdots & x_N^1 & 1 & y^1 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ x_1^M & \cdots & x_N^M & 1 & y^M \end{pmatrix}.$$

Remember that the rank of a matrix is the number of linearly independent row (and column) vectors.

If the data vectors are all linearly independent we have that $N \geq M$ (if $N < M$ there would be some vectors that depend on others, because we have more vectors than dimensions), and the rank of X is $\min(M, N) = M$. Augmenting X by stacking any number of column vectors simply increases N, which means that it does not change the rank of the matrix. It follows that for M linearly independent data points embedded in a N dimensional space the system has a solution. The data is therefore linearly separable.

With this argument we can add more vectors that are linearly dependent until $M = N$. After this, we can in fact add one (but only one) more data point that linearly depends on the others, and still guarantee linear separability. That is because adding one data point makes the row number equal to the column number in $[X|1]$, and adding more columns does not change the rank. In contrast, adding *two* data points means that we have more columns than rows in $[X|1]$, and adding the column for $[X|1|y]$ can indeed change the rank.

B. Proof of Proposition 2

Let us consider a matrix M where the squeezed states in Fock basis form the rows:

$$M_{jn} := \frac{1}{\sqrt{\cosh(r_j)}} \left(-\mathrm{e}^{i\phi_j} \tanh(r_j)\right)^n \frac{\sqrt{(2n)!}}{2^n\, n!}$$

We introduce two auxiliary diagonal matrices:

$$D_1 := \mathrm{diag}\{\sqrt{\cosh(r_j)}\}$$

$$D_2 := \mathrm{diag}\left\{\frac{n!}{\sqrt{(2n)!}}\right\}$$

Multiplying, we find that the matrix $V := D_1 M D_2$ has matrix elements

$$V_{jn} = \left(-\frac{1}{2}\mathrm{e}^{i\phi_j} \tanh(r_j)\right)^n.$$

Importantly, V has the structure of a Vandermonde matrix. In particular, it has determinant

$$\det(V) = \frac{1}{2} \prod_{1\leq i<j\leq n} \left(-\mathrm{e}^{i\phi_i} \tanh(r_i) + \mathrm{e}^{i\phi_j} \tanh(r_j)\right).$$

The only way that $\det(V) = 0$ is if

$$\mathrm{e}^{i(\phi_i - \phi_j)} \tanh(r_i) = \tanh(r_j)$$

for some $i = j$. The squeezing feature map with phase encoding prescribes that $r_i = r_j = c$ (and we assume that $c > 0$). Thus, the only solution to the above equation is when $\varphi_i = \varphi_j$, which can only be true if the two feature vectors describe the same datapoint, which we excluded in Proposition 2. Thus, $\det(V) > 0$, which means that $\det(M) > 0$, and hence M is full rank. This means that the columns of M, which are our feature vectors, are linearly independent. Note that the same proof also prescribes that squeezing feature maps with absolute value encoding makes distinct data points linearly independent in Fock space.

V. UNIVERSAL CV LAYER ARCHITECTURE

The architecture of a layer used in the variational circuit for the explicit classifier described in the main Letter consists of repetitions of a general layer of gates (see Figure 2). We denote by $\hat{a}_{1,2}, \hat{a}^\dagger_{1,2}$ the creation and annihilation operators of mode 1 and 2, and with $\hat{x}_{1,2}, \hat{p}_{1,2}$ the corresponding quadrature operators (see [19]). After an entangling beam splitter gate,

$$BS(u, v) = \mathrm{e}^{u(\mathrm{e}^{iv}\hat{a}^\dagger_1\hat{a}_2 - \mathrm{e}^{-iv}\hat{a}_1\hat{a}^\dagger_2)},$$

with $u, v \in \mathbb{R}$, the circuit consists of single-mode gates that are first, second and third order in the quadratures. The first-order gate is implemented by a displacement gate

$$D(z) = \mathrm{e}^{\sqrt{2}i(\mathrm{Im}(z)\hat{x} - \mathrm{Re}(z)\hat{p})},$$

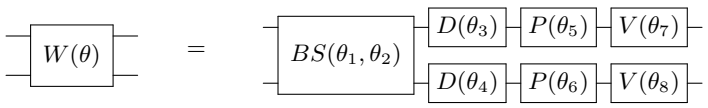

FIG. 2. The variational circuit $W(\theta)$ used for the explicit classifier consists of repetitions of a gate block with the beamsplitter (BS), displacement (D), quadratic (P) and cubic phase gates (C) described in the text.

with the complex displacement factor z. We use a quadratic phase gate for the second order,

$$P(u) = \mathrm{e}^{i\frac{u}{2}\hat{x}^2},$$

and a cubic phase gate for the third order operator,

$$V(u) = \mathrm{e}^{i\frac{u}{3}\hat{x}^3}.$$

We can in principle construct any continuous-variable quantum circuit from this gate set. This basic layer can easily be generalised to circuits of more modes by replacing the single beam splitter by a full optical network of beam splitters [20].

[1] R. Chatterjee and T. Yu, Quantum Information and Communication **17**, 1292 (2017).
[2] B. J Mercer, Phil. Trans. R. Soc. Lond. A **209**, 415 (1909).
[3] T. Hofmann, B. Schölkopf, and A. J. Smola, The Annals of Statistics , 1171 (2008).
[4] N. Aronszajn, Transactions of the American Mathematical Society **68**, 337 (1950).
[5] B. Schölkopf, R. Herbrich, and A. Smola, in *Computational learning theory* (Springer, 2001) pp. 416–426.
[6] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics* (Springer Science & Business Media, 2011).
[7] T. Griffiths and A. Yuille, The probabilistic mind: Prospects for Bayesian cognitive science , 33 (2008).
[8] R. D. la Madrid, European Journal of Physics **26**, 287 (2005).
[9] J. R. Klauder and B.-S. Skagerstam, *Coherent states: applications in physics and mathematical physics* (World scientific, 1985).
[10] S. Wang, Journal of Mathematics Research **7**, 175 (2015).
[11] E. Farhi and H. Neven, arXiv preprint arXiv:1802.06002 (2018).
[12] N. Wiebe, D. Braun, and S. Lloyd, Physical Review Letters **109**, 050505 (2012).
[13] M. Schuld, M. Fingerhuth, and F. Petruccione, Europhysics Letters **119**, 60002 (2017).
[14] P. Rebentrost, M. Mohseni, and S. Lloyd, Physcial Review Letters **113**, 130503 (2014).
[15] E. Stoudenmire and D. J. Schwab, in *Advances In Neural Information Processing Systems* (2016) pp. 4799–4807.
[16] G. G. Guerreschi and M. Smelyanskiy, arXiv preprint arXiv:1701.01450 (2017).
[17] A. B. Novikoff, *On convergence proofs for perceptrons*, Tech. Rep. (Stanford Research Institute, 1963).
[18] L. Hogben, *Handbook of linear algebra* (CRC Press, 2006).
[19] C. Weedbrook, S. Pirandola, R. García-Patrón, N. J. Cerf, T. C. Ralph, J. H. Shapiro, and S. Lloyd, Reviews of Modern Physics **84**, 621 (2012).
[20] F. Flamini, N. Spagnolo, N. Viggianiello, A. Crespi, R. Osellame, and F. Sciarrino, Scientific Reports **7**, 15133 (2017).