

LEARNING ALGEBRAIC MODELS OF QUANTUM ENTANGLEMENT

HAMZA JAFFALI AND LUKE OEDING

ABSTRACT. We give a thorough overview of supervised learning and network design for learning membership on algebraic varieties via deep neural networks. We show how artificial neural networks can be trained to predict the entanglement type for quantum states. We give examples for detecting degenerate states, as well as border rank classification for up to 5 binary qubits and 3 qutrits (ternary qubits).

1. INTRODUCTION

Recent efforts to unite Quantum Information, Quantum Computing, and Machine Learning have largely been centered on integrating quantum algorithms and quantum information processing into machine learning architectures [7, 42, 49, 67, 69, 70, 72, 78, 79, 82]. Our approach is quite the opposite – we leverage Machine Learning techniques to build classifiers to distinguish different types of quantum entanglement. Machine Learning has been used to address problems in quantum physics and quantum information, such as quantum state tomography [65], quantum error correction code [59] and wavefunction reconstruction [3]. Here we focus on quantum entanglement. While we were inspired by the approach of learning algebraic varieties in [10], our methods differ in that we are not trying to find intrinsic defining equations of the algebraic models for entanglement types, but rather building a classifier that directly determines the entanglement class. Distinguishing entanglement types may be useful for quantum information processing and in improving and increasing the efficiency of quantum algorithms, quantum communication protocols, quantum cryptographic schemes, and quantum games. Our methods generalize to cases where a classification of all entanglement types is not known, and cases where the number of different classes is not finite (see for instance [45, Ch.10], or [75]).

We only focus on *pure states* for representing quantum systems, which is sufficient for studying quantum computations and quantum algorithms. This is opposed to the noisy approach with density matrices and mixed states, which is used when one needs to account for the noise and the interaction with the environment [60].

1.1. Basic notions. The quantum state of a particle can be represented by a unit vector $|\psi\rangle$ in a Hilbert space $\mathcal{H} = \mathbb{C}^d$, with basis $\{|x\rangle \mid x \in \llbracket 0, d-1 \rrbracket\}$ in decimal notation. The state of an n -qudit quantum system is represented by a unit vector $|\psi\rangle$ in a tensor product $\mathcal{H}^{\otimes n}$ of the state spaces for each particle, where for simplicity we assume all particles of the same type. The tensor space $\mathcal{H}^{\otimes n}$ has basis $|ij\dots k\rangle := |i\rangle \otimes |j\rangle \otimes \dots \otimes |k\rangle$, where $|i\rangle, |j\rangle, \dots, |k\rangle$ are basis elements of \mathcal{H} . The dual vector space \mathcal{H}^* is represented by vectors $\langle\varphi|$, and we use the standard Hermitian inner product $(|\psi\rangle, |\varphi\rangle) \mapsto \langle\varphi|\psi\rangle =: \sum_i \overline{\varphi_i} \psi_i \in \mathbb{C}$, where $|\varphi\rangle = \sum_i \varphi_i |i\rangle$, and similarly for ψ .

Date: March 7, 2021.

Key words and phrases. Quantum Entanglement, Classification, Algebraic Varieties, Machine Learning, Neural Networks.

The study of quantum entanglement is often focused on orbits (equivalence classes) under the action of the SLOCC (Stochastic Local Operations and Classical Communication) group, which algebraists know as $G = \text{SL}(\mathcal{H})^{\times n}$, the cartesian product of special linear groups, *i.e.* normal changes of coordinates in each mode.

1.2. First examples. Consider the pair of binary state particles, called a 2-qubit system. Here there are only two different entanglement types up to the SLOCC action, represented by $|00\rangle$ and $\frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. All other non-zero 2-qubit states can be moved to one of these by the action of the SLOCC group. Determining on which orbit a given state φ is (and hence its entanglement type), can be done by computing the 2×2 matrix determinant: Set $\varphi_{i,j} := \langle ij|\varphi\rangle$. The value of $\det \begin{pmatrix} \varphi_{00} & \varphi_{01} \\ \varphi_{10} & \varphi_{11} \end{pmatrix}$ is either 0 or $\frac{1}{2}$ if φ is respectively of type $|00\rangle$ or $\frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. In particular, the non-general states live on a quadric hypersurface, and thus an Artificial Neural Network classifier (see Section 2.1) can be trained to test membership on these two types of states, see Example 4 below.

Going to the 3-qubit system, the rank (a numerical invariant) and the determinant (a polynomial invariant) generalize to the multilinear rank and the Cayley hyperdeterminant, respectively. A given state $|\varphi\rangle \in \mathcal{H}^{\otimes 3} = \mathbb{C}^{2 \times 2 \times 2}$ has coordinates

$$\varphi_{ijk} = \langle ijk|\varphi\rangle.$$

The three 1-flattenings are as follows:

$$F_1 = (\varphi_{ijk})_{i,jk}, \quad F_2 = (\varphi_{ijk})_{j,ik}, \quad F_3 = (\varphi_{ijk})_{k,ij}.$$

The ranks of these flattenings comprise a vector called the *multilinear rank*. If any flattening has rank 0, φ is zero, which cannot occur if we assume φ is a unit vector. If the multilinear rank is $(1, 1, 1)$, then φ is separable. If the multilinear rank is $(1, 2, 2)$ then φ is a bi-separable state of the form $|000\rangle + |011\rangle$ up to SLOCC. The other bi-separable states are permutations of this up to SLOCC. Finally, if the multilinear rank is $(2, 2, 2)$, then up to a SLOCC transformation, φ is either the so called W-state $\frac{1}{\sqrt{3}}(|001\rangle + |010\rangle + |100\rangle)$ or a general point $\frac{1}{\sqrt{2}}(|000\rangle + |111\rangle)$. These states are distinguished, respectively, by the vanishing or non-vanishing of the well-known SLOCC-invariant, the $2 \times 2 \times 2$ hyperdeterminant:

$$\begin{aligned} \Delta_{222}(x) = & x_{000}^2 x_{111}^2 + x_{011}^2 x_{100}^2 + x_{010}^2 x_{101}^2 + x_{001}^2 x_{110}^2 \\ & - 2x_{000}x_{011}x_{100}x_{111} - 2x_{000}x_{010}x_{101}x_{111} - 2x_{000}x_{001}x_{110}x_{111} + 4x_{000}x_{011}x_{101}x_{110} \\ & - 2x_{010}x_{011}x_{100}x_{101} - 2x_{001}x_{011}x_{100}x_{110} - 2x_{001}x_{010}x_{101}x_{110} + 4x_{001}x_{010}x_{100}x_{111}. \end{aligned}$$

In this case these are all the possible types of states up to SLOCC, and we have a complete algebraic description of these states as well. So, this is a good testing ground for other methods since we know how to test ground truth. In Section 4.2 we summarize the performance of a neural network for separating these states.

In the 4-qubit case there is still a classification, and a complete set of invariants that can be used to classify entanglement types is known [31, 32]. In addition, algebraic invariants for all border ranks for 4-qubit systems are known classically (originally studied by [71]) and are given by the minors of 1- and 2- flattenings [45, Ch. 7.2]. We note that the $2 \times 2 \times 2 \times 2$ hyperdeterminant has degree 24, and is computable by Schläfli's method, [24]. However, in the 5-qubit case the $2^{\times 5}$ hyperdeterminant has degree 128, is not computable by Schläfli's method, [77], and is surely very complicated as indicated by the 4-qubit case [39]. However, in Section 3.3 we show that neural networks can distinguish between singular and non-singular states even though algebraic invariants are likely to fail.

1.3. Prior Work and Outline. Machine Learning has been used to study entanglement detection, entanglement measurement, and entanglement classification. A common focus is the density matrices formalism, for which the asymptotic problem of deciding separability is a NP-hard problem. Neural networks have been used for estimating entanglement measures such as logarithmic negativity, Rényi entropy or concurrence in 2-qubits (pure and mixed) or many-body systems [6, 26, 27]; for encoding several CHSH (witness) inequalities simultaneously in a network to detect entangled states [54]; for computing the closest separable state in a complex valued network [11]; for recognizing the entanglement class of 3-qubit systems [4]; and for detecting entanglement in families of qubits and qutrits systems in the bipartite case [80]. A convex hull approximation method combined with decision tree and standard ensemble learning (bagging) algorithms was used in [50] to classify separable and entangled states. The forest algorithm (also using decision trees) was used in [76] to detect entanglement and was compared to quantum state tomography (up to five qubits). Principal Component Analysis (PCA) was used to determine the dimension and intrinsic defining equations of certain algebraic varieties [10]. In the 2-qubit case neural networks, support vector machines, and decision trees were used to detect non-classical correlations such as entanglement, non-locality, and quantum steering [83].

Indeed, Machine Learning can be a relevant tool for entanglement detection, at least in some limited cases. However, prior studies were mostly limited to the 2-qubit or bipartite case, because of the possibility of generating properly separable and entangled mixed states using the PPT criterion or Schmidt decomposition. It seems difficult to generalize the prior supervised approaches to higher dimensions or to systems with more than two particles.

Instead, we focus on the problem of entanglement detection and classification by learning algebraic varieties (Segre variety, dual variety, secant varieties) that characterize different entanglement classes for pure states. Our method can be generalized to higher dimensions and systems with several particles, bringing original tools for distinguishing non-equivalent entanglement classes for quantum systems for which we do not know the complete classification or we do not have exact algorithms (as proposed by [30–32]) to determine the entanglement type (as it is the case for 5-qubit systems for instance).

As noted in [10] there are many instances of high degree hypersurfaces (like the hypersurface of degenerate states in $\mathcal{H}^{\otimes n}$) that are easy to sample, but there is little hope to learn their equations from samples, so prior methods do not apply. However, in Section 3.3 we will show several cases where an artificial neural network can be trained to perform the task of determining membership on these high degree hypersurfaces (which the inaccessible defining equation would also do).

In Section 2, we investigate several architectures of feed-forward neural networks for learning membership on algebraic varieties, by using previous knowledge about the defining equations to design networks. In Section 3, we study several examples by building classifiers with neural networks for the cases of qubits and qutrits. Finally, in Section 4, we use these classifiers to distinguish quantum entanglement classes.

2. NETWORK DESIGN FOR LEARNING ALGEBRAIC VARIETIES

2.1. Basics of Artificial Neural Networks. Inspired by biological neural networks, Artificial Neural Networks (ANN) are computing systems whose goal is to reproduce the functionality and basic structure of the human brain. They are used for the learning of a specific task (such as classification and regression), without any explicit instructions or rules.

An artificial neural network is composed of several artificial neurons, regrouped using a specific architecture, and is able to perform a specific task using a learning algorithm.

In 1943, McCulloch and Pitts proposed the first model of an artificial neuron [55]. An artificial neuron (see Figure 1) is defined by inputs (eventually coming from other neurons), by weights (synaptic weights associated with each input), a weighted sum (computed from the two previous ones), a threshold (or bias), an activation function, and an output.

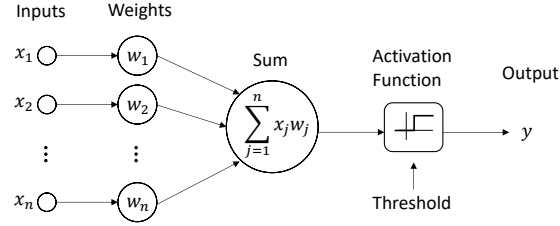


FIGURE 1. Illustration of an artificial neuron.

The output of the neuron is determined as follows. If the weighted sum is strictly greater than the threshold, then the neuron is said to be active, and the output of the neuron will be equal to the activation function applied to the weighted sum (with the threshold frequently added to the weighted sum). Denote by x_1, \dots, x_n the inputs of the neuron, and by w_i the weight associated with the i -th input. The *weighted sum* (frequently denoted by \mathcal{U}) is

$$\mathcal{U} = w_1 x_1 + w_2 x_2 + \dots + w_n x_n . \quad (1)$$

The *threshold* θ is a real number that can be considered as a weight associated with an additional input 1 during the learning process (see Section 2.2). The *output* y of the neuron is equal to $y = g(\mathcal{U})$, with g the activation function of the neuron. Usually we want activation functions to be *nonlinear* in order to allow such neural networks to compute nontrivial problems using only a small number of neurons [18]. However, activation functions can be chosen as binary step or linear functions but will lead to several problems during the learning step or for dealing with non-binary classification problems. One may also desire other properties in an activation function, such as *continuously differentiable*, *monotonic*, *smooth*., and *approximates the identity near the origin* (see Figure 2 for a summary).

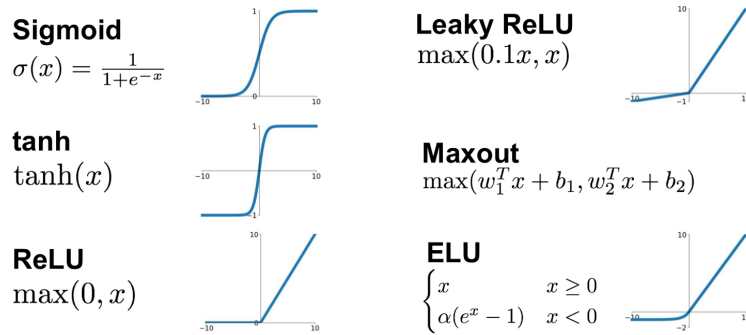


FIGURE 2. Typical activation functions (Source).

Example 1. Consider the problem of reproducing the basic logical function **OR** with an artificial neuron. Denote by x_1 and x_2 the two inputs, which take values from $\{0, 1\}$. The

output of **OR** is 1 if either input is 1 and 0 otherwise. Choose the step activation function g . The weighted sum and output are respectively:

$$\mathcal{U} = w_1 x_1 + w_2 x_2, \quad g(\mathcal{U}) = \begin{cases} 1 & \text{if } \mathcal{U} > \theta \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

In order to model **OR** with a single neuron the following conditions are necessary:

$$\begin{aligned} 0 \cdot w_1 + 0 \cdot w_2 &\leq \theta, & 0 \cdot w_1 + 1 \cdot w_2 &> \theta, \\ 1 \cdot w_1 + 0 \cdot w_2 &> \theta, & 1 \cdot w_1 + 1 \cdot w_2 &> \theta. \end{aligned}$$

This leads to

$$\theta \geq 0, \quad w_1 > \theta, \quad w_2 > \theta.$$

If $\theta = \frac{1}{2}$, $w_1 = 1$, $w_2 = 1$, this single neuron reproduces **OR** for inputs in $\{0, 1\}^2$. A visualization of this binary classification is given in Figure 3.

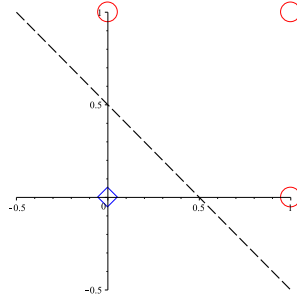


FIGURE 3. The classification problem associated with **OR**. The blue diamond corresponds to the first class (function equal to 0) and the red circles to the second class. The black dashed line represents the hyperplane splitting the data into two classes, modeled by the single artificial neuron defined by $\theta = \frac{1}{2}$, $w_1 = 1$, $w_2 = 1$.

Geometrically, if one can find a hyperplane (codimension 1) that separates the data into two classes, then the binary classification problem can be solved using a single artificial neuron. Here, the equation of one such hyperplane is given by $x_1 + x_2 - \frac{1}{2} = 0$, which corresponds to the following values for the threshold and weights: $\theta = \frac{1}{2}$, $w_1 = 1$, $w_2 = 1$. \diamond

On the other hand, if we want to model the logical **XOR** function with only one neuron, one can show that this is not possible by the same type of analysis of Example 1. Therefore, one needs more complex structures and more neurons to solve more difficult problems, and this leads us to consider the model of Multi-Layer Perceptron (MLP). It is composed of an input layer, one or several hidden layers, and an output layer. Each layer is composed of a certain number of neurons that are not connected. In a feed-forward dense layer configuration, each neuron in a layer is connected to every neuron of the previous and next layer, and the signal is only propagated in one direction (from left (input layer) to right (output layer)).

An Artificial Neural Network (ANN) learns tasks from examples via optimizing weights in a network. These examples form the *training dataset* of correct input-output pairs. The goal of the supervised approach with ANNs is to learn how to associate the correct output to a given input. This task can be performed after a training step, where all the weights and thresholds will be optimized to reduce the error defined by the difference between the output given by the network and the correct output provided in the training dataset. In the

feed-forward neural network model, the back-propagation algorithm is used to update the weights and reduce the error, also known as the *loss function* [46, 68].

After training, the network should be able to answer correctly for inputs taken from the training dataset but should also be able to make correct predictions for new data. In order to test the network, we give a *testing dataset* to evaluate the capacity of the network to generalize the previous learning by making predictions. In addition, one can consider a *validation dataset*, used during the training step to avoid or detect overfitting.

The activation function is an important choice when designing a neural network for learning a specific task. Frequently neurons in the same layer will have the same activation function, while two adjacent layers may have different activation functions.

For the input and hidden layers, ReLU is commonly used as an activation function. The advantages of ReLU is that it is computationally efficient (quick convergence of the network), it is non-linear, and admits a derivative (thus allowing back-propagation). However, when the inputs approach zero or become negative, the learning becomes harder (dying ReLU problem). To prevent that, one can use Leaky ReLU activation, which introduces a small positive slope in the negative area, permitting back-propagation.

The activation function of the neurons in the output layer will depend on the problem one wants to solve. Sigmoid and SoftMax (see Eq. 3) functions are the most used ones, and the choice will depend on the classification problem one wants to solve. In fact, if it is a binary classification problem, then Sigmoid function with only one neuron will be more efficient. When one deals with a classification problem with multiple classes, several neurons at the output (with categorical presentation of the data) should be introduced. Therefore, if a state can belong to several classes at the same time, one will use the Sigmoid function in this case. Otherwise, if every state only belongs to one class, then SoftMax function will be a good choice.

$$\text{SoftMax}(\mathbf{z}) = (\sigma(\mathbf{z})_1, \sigma(\mathbf{z})_2, \dots, \sigma(\mathbf{z})_K)$$

$$\text{with } \mathbf{z} = (z_1, z_2, \dots, z_K) \text{ and } \sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}, \forall j \in \llbracket 1, K \rrbracket \quad (3)$$

The loss function should also be chosen based on the problem of interest. Choices of activation functions for the output layer and loss functions are summarized in Table 1.

Problem type	Output type	Output Activation function	Loss Function
Regression	Numerical value	Linear	Mean Squared Error
Classification	Binary outcome	Sigmoid	Binary Cross Entropy
Classification	Single label, multiple classes	SoftMax	Categorical Cross Entropy
Classification	Multiple labels, multiple classes	Sigmoid	Binary Cross Entropy

TABLE 1. Summary of output layer’s activation function and loss function depending on the problem (Source).

In our work, we implemented neural networks in Python using the Keras library [15]. It provides a flexible implementation of feed-forward neural networks and the choice of network parameters (architecture, activation functions, loss function, solver, etc.). We chose to implement the *Nadam* solver provided by the Keras library for minimizing the loss function during the learning step [21, 73].

Once activation and loss functions are fixed, the remaining parameters to be configured are the number and size of the layers. A classical result, known as the *Universal Approximation*

Theorem states that “a feed-forward network with an input layer and a single hidden layer with a finite number of neurons can approximate any continuous function on compact subsets of \mathbb{R}^n ” [17]. In 1989, the theorem was proven using Sigmoid activation functions [18]. Kurt Hornik showed that the theorem does not depend on the activation functions, but rather on the feed-forward multilayer architecture in itself [38].

These results concern the case of depth-bounded (e.g. 2) networks. On the other hand, several works recently proposed an analogous theorem for width-bounded (with ReLU activation functions) networks [28, 51]. The question of providing the best network, in terms of accuracy and computational efficiency, for a given task or problem, is still open and the question of the choice of the depth and the width of neural networks are most of the time chosen by practice. Designing a network often leads one to consider a trade off between performance and computational cost. Some architectures, like a decreasing number of neurons as the neural network becomes deeper, seem to present practical efficiency. However, there is no theoretical proof of the superiority of such networks, and our experience seems to show that a network with the same number of neurons in each layer could perform with the same accuracy for predictions.

2.2. Learning algebraic varieties. An algebraic variety is a geometrical object defined as the zero locus of a set of polynomials. In order to teach a machine how to recognize points on algebraic varieties, one must be able to encode in the learning model the polynomial equations defining these objects (or an approximation of them).

Example 2 (Learning affine linear spaces). Consider the equation of a line (L) in \mathbb{R}^2 : $ax + by + c = 0$, with a, b and c real coefficients. Suppose we want to determine membership on (L). This is a binary classification problem that can be modeled with a single artificial neuron with two inputs x and y . Denote by w_1 , respectively w_2 , the weight associated with the first input x , respectively y , and θ the threshold of the neuron (which is here considered as a weight associated with the input value 1). The expression of the weighted sum \mathcal{U} is:

$$\mathcal{U} = w_1x + w_2y + \theta$$

Note that if we set $w_1 = a$, $w_2 = b$ and $\theta = c$, we retrieve the equation of the line (L). This means that the weighted sum \mathcal{U} of the neuron will be equal to 0 if and only if the inputs (x, y) correspond to a point on the line (L).

This observation can be generalized to any dimension. In fact, any linear subspace of codimension 1 in a n -dimension space, *i.e.* defined by a single linear equation with n variables, can be *modeled* by using a single artificial neuron with n inputs corresponding to the n variables. By *model* we mean the fact that the output of the neuron is always the same value when the inputs represent a point contained in this linear subspace, with the activation function can be taken as the identity function. The following is straightforward.

Proposition 2.1. *Let L be an affine linear space in \mathbb{R}^n defined by $a_1x_1 + a_2x_2 + \dots + a_nx_n + a_{n+1} = 0$, with real variables (x_1, x_2, \dots, x_n) and real coefficients $(a_1, a_2, \dots, a_{n+1})$. Then L can be modeled using a single artificial neuron, and the weights must satisfy: $w_1 = \lambda a_1, w_2 = \lambda a_2, \dots, w_n = \lambda a_n$ and $\theta = \lambda a_{n+1}$, with $\lambda \in \mathbb{R}^*$, as shown in Figure 4.*

◇

Example 3. We now generalize the previous example to higher codimension. Let (L_2) be a line in \mathbb{R}^3 . A line in a 3-dimensional space is an algebraic variety of codimension 2, defined by a system of two linear equations. The vanishing of these two equations is the intersection

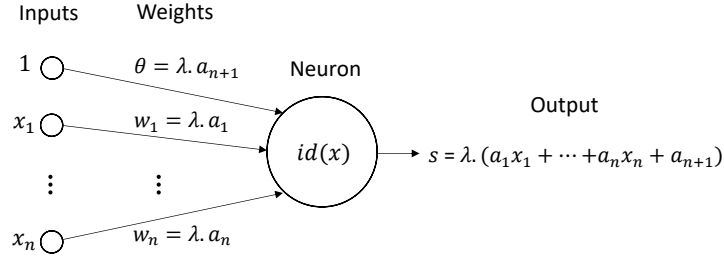


FIGURE 4. Representation of a network modeling a linear subspace of codimension 1 in \mathbb{R}^n .

of two planes, the line (L_2). So, we can combine in the same layer two artificial neurons to model the two equations defining the line (Figure 5). The output of the first neuron should be equal to the same value (namely 0) when the input point of the network is on the first plane (and thus on the line), and the same applies for the second neuron.

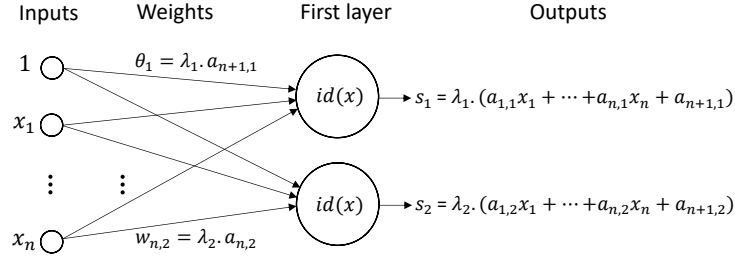


FIGURE 5. Representation of a network modeling a linear subspace of codimension 2 in \mathbb{R}^n .

If we also take the identity as activation function for both neurons, the output of the two neurons will be $(0, 0)$ precisely when a point is on the line (L_2). This result is straightforward to generalize to higher codimension:

Proposition 2.2. *Let L_2 be an affine linear subspace of codimension m , defined by m equations of the type $a_{1,i}x_1 + a_{2,i}x_2 + \dots + a_{n,i}x_n + a_{n+1,i} = 0$, with real variables (x_1, x_2, \dots, x_n) and real coefficients $(a_{1,i}, a_{2,i}, \dots, a_{n+1,i})$, $\forall i \in \llbracket 1, m \rrbracket$. Then this subspace can be modeled using m artificial neurons combined in a layer, and the weights of the i -th neuron must satisfy: $w_{1,i} = \lambda_i a_{1,i}$, $w_{2,i} = \lambda_i a_{2,i}$, \dots , $w_{n,i} = \lambda_i a_{n,i}$ and $\theta_i = \lambda_i a_{n+1,i}$, with $\lambda_i \in \mathbb{R}^*$, $\forall i \in \llbracket 1, m \rrbracket$.*

◇

Example 4. [Curves] Consider a circle (C) in \mathbb{R}^2 defined by the equation: $x^2 + y^2 - r^2 = 0$, with r the radius of the circle. By only using the weighted sum of the inputs with the identity as activation function, a single neuron is not able to model this quadratic equation. The idea is to use the function $x \mapsto x^2$ for the activation function, in order to introduce squared terms of the input variables x and y at the output of the neuron. This idea was already used in [43] to study the power of deep neural networks with polynomial activation functions. Such networks are also known as Polynomial Neural Networks (PNN) and an adaptive architecture was proposed in [61].

If we consider a single artificial neuron, with w_1 and w_2 the weights associated with x and y respectively, θ the threshold, and $g : x \mapsto x^2$ the activation function, then the output s is equal to:

$$s = g(\mathcal{U}) = (w_1x + w_2y + \theta)^2 = w_1^2x^2 + w_2^2y^2 + 2w_1w_2xy + 2w_1\theta x + 2w_2\theta y + \theta^2 \quad (4)$$

If we want to model the equation of the circle, then a single artificial neuron is not sufficient. However, it is possible by adding another neuron in the same layer, and then combine these two outputs in a third neuron (in another layer) (see Figure 6). This result can be generalized as follows:

Proposition 2.3. *Any polynomial equation of degree 2 with $n - 1$ real variables x_i , i.e. of the form $\mathcal{E}: \sum_{1 \leq i < n} a_{i,i}x_i^2 + \sum_{1 \leq i < j < n} 2a_{i,j}x_ix_j + \sum_{i=1}^{n-1} 2a_{i,n}x_i + a_{n,n}$, with $a_{i,j} \in \mathbb{R}$, can be modeled using n artificial neurons: $n - 1$ neurons with $x \mapsto x^2$ (square) activation function in a first layer, and one neuron with identity activation function in a second layer.*

Proof. Let \mathcal{E} be a polynomial of degree 2 as in the statement of the proposition, which can be represented by the matrix equation $\vec{x}^\top A \vec{x} = 0$ for a real symmetric matrix $A = (a_{i,j})$, with $\vec{x} = (x_1, \dots, x_{n-1}, 1)^\top$.

Let $w_{i,j}$ be the weight associated with the i -th input and the j -th neuron of the first layer, let b_k be the weight associated with the output of the k -th neuron in the first layer and the single neuron in the second layer, let $\theta_1, \dots, \theta_{n-1}$ be the thresholds of the $n - 1$ neurons in the first layer, and θ_n be the threshold of the neuron in the second layer. Each neuron in the first layer has square activation function, and the neuron of the second layer has the identity activation function. If we denote by s_1, \dots, s_{n-1} the outputs of the neurons in the first layer, and s_n the output of the single neuron in the second layer, then:

$$s_n = \mathcal{U}_n = \sum_{i=1}^{n-1} b_i s_i + \theta_n \quad \text{with} \quad s_i = \mathcal{U}_1^2 = \left(\sum_{j=1}^{n-1} w_{j,i} x_j + \theta_i \right)^2. \quad (5)$$

So s_n is an inhomogeneous linear combination of squares of affine linear forms. This can be represented as $\vec{v}^\top D \vec{v} = 0$ for a real diagonal matrix D , and the entries v_1, \dots, v_{n-1} of \vec{v} are the linear forms (before squaring) in the expressions for the s_i , and $v_n = \sqrt{\theta_n}$. Thus, the proposition is equivalent to the diagonalization of real symmetric matrices, which is well known to always be possible via orthogonal matrices. The entries of the orthogonal matrices used in the diagonalization give the weights of the network. \diamond

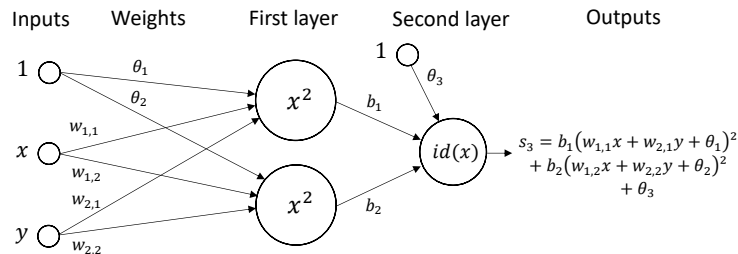


FIGURE 6. Representation of a network modeling a polynomial equation of degree 2 with two variables and coefficients in \mathbb{R} .

\diamond

To build a network for modeling any polynomial equation of degree d in n variables, a dimension count gives a good guess for the appropriate architecture. The Alexander-Hirschowitz (AH) theorem (see [9] for a modern treatment) tells us when the naive dimension

count fails. In fact, the AH theorem states that a general¹ homogeneous polynomial p of degree d in n variables can be expressed as the sum of $T = \lceil \frac{1}{n} \binom{d+n-1}{d} \rceil$ d -th powers of linear forms (except for quadratic forms (Prop 2.3) and a few other special cases), *i.e.*

$$p(x_1, x_2, \dots, x_n) = \sum_{j=1}^T \left(\sum_{i=1}^n a_{ij} x_i \right)^d.$$

Dehomogenizing the AH result (by setting the last variable to 1, for example) one obtains a bound for 2-layer neural networks modeling affine hypersurfaces. We implement $T = \lceil \frac{1}{n} \binom{d+n-1}{d} \rceil$ neurons with activation function $g : x \mapsto x^d$ in the first layer, and then combine all the outputs in a linear combination using a neuron in the second layer. Suppose we have n inputs corresponding to the n variables of the homogeneous polynomial p . If we denote by $w_{i,j}$ the weight associated with the i -th input x_i and the j -th neuron in the first layer, and by θ_j the threshold of the j -th neuron, then the output s_j of the neuron j in the first layer is

$$s_j = g(\mathcal{U}_j) = \left(\sum_{i=1}^n w_{i,j} x_i + \theta_j \right)^d.$$

The threshold θ_j introduces an inhomogeneity which we can remove by adding an extra variable x_{n+1} , replacing θ_j with $\theta_j x_{n+1}$. Then we may apply the AH theorem for $n+1$ variables, then set $x_{n+1} = 1$ to obtain the bound for the inhomogeneous form of the output. This idea also appears in [43].

The last step of the network design is to solve the classification problem, namely: is the input point on (or outside) the variety defined by the equations modeled by the network? We recall that after the two first layers, we should retrieve one specific value s (most of the time 0) when the input point is on the variety, and any other value if the point is not. So, the last step is equivalent to recognizing a real number s in an interval (whose size depends on the training data). We cannot just add a Sigmoid neuron to solve directly the classification problem because single neurons only solve inequalities and not equalities. So, we need to add another layer before the output layer to be able to recognize this specific value s , and then solve this binary classification problem at the output of the network.

After testing several architectures, the task of recognizing a real number in an interval can be performed using a single layer with four neurons with LeakyReLU activation functions. Thus, by adding such a layer after the first two layers (modeling the equations of the variety), and by adding a last layer (output layer) with only one neuron (with Sigmoid activation function), one can potentially learn any algebraic variety defined by a set of homogeneous polynomials (Figure 7).

However, the questions of the ability to correctly train the network, and the ability for the optimizer to find the right weights and thresholds are important for practical implementations. For homogeneous equations of low degree and few variables (such as the circle case), the network is able to learn the correct weights, and one can recover the equation of the circle from the weights of the network. However, when the degree and the number of variables increases, the number of neurons needed in the first layer increases very quickly and it becomes harder to converge to a set of weights and thresholds that model the desired equation. To overcome this, in Subsection 2.4 we slightly modify the structure of the network to reduce the number of weights and parameters.

¹By “general” we mean avoiding a measure-zero set of possible counterexamples.

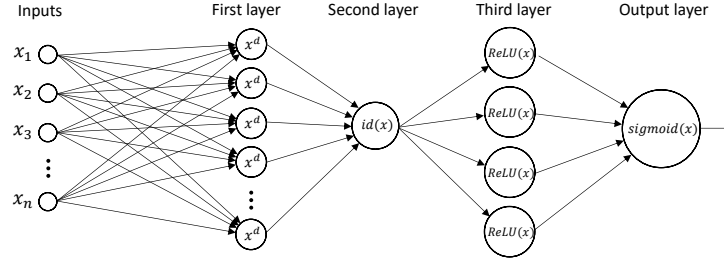


FIGURE 7. Representation of the network solving the binary classification problem related to algebraic variety's membership.

2.3. Expressive Power of Deep Neural Networks. For deeper neural networks with d -th power activation functions we can apply the same dimension-counting principle as was employed in the AH case in order to determine the maximum width of each layer to avoid over-fitting and redundancy in the network. See also Kileel et al., [43, Sec. 3.2] for several concrete examples. Let us give a recursive construction. Suppose we are working over the complex numbers and want to approximate a homogeneous polynomial f of degree de as a sum of the d -th powers of p homogeneous degree e polynomials g_j in n variables as

$$f \simeq \sum_{j=1}^p \lambda_j (g_j)^d.$$

Note that we're not going to count the parameters λ_j in this case because we can re-scale the g_j by bringing in a d -th root. Then we need at most $\binom{n-1+de}{de}$ parameters to express f , but the right-hand side of the above equation uses $p \binom{n-1+e}{e}$ parameters. So, we expect that we can do the approximation if $p \binom{n-1+e}{e} \geq \binom{n-1+de}{de}$, or if

$$p \geq \left\lceil \frac{\binom{n-1+de}{de}}{\binom{n-1+e}{e}} \right\rceil.$$

If f has degree D and the functions g_j are taken from a space that requires M parameters, repeating the same argument we find that as long as we have p nodes with

$$p \binom{M-1+e}{e} \geq \binom{n-1+de}{de}. \quad (6)$$

then we expect to be able to approximate f . For more layers we iterate (6). For example, in the 3-layer case with p_j nodes on layer j using degree d_j activation functions, if

$$p_1 \binom{p_2 \left(p_3 \binom{n-1+d_3}{d_3} - 1 + d_2 \right) - 1 + d_1}{d_1} \geq \binom{n-1+d_1d_2d_3}{d_1d_2d_3}.$$

then we can expect to be able to model f with degree $d_1d_2d_3$. One might guess that a classification result like the AH theorem might hold, but relatively little is known. See [43, Thm. 10] for another bound.

Remark 2.4. The ability of neural networks to approximate any function, in the sense of the Universal Approximation Theorem, can lead one to ask how it is different from polynomial fitting. In [13], an analytic argument is given that neural networks are in fact essentially polynomial regression models, with the effective degree of the polynomial growing at each hidden layer. According to the authors, if the activation function is any polynomial, or

implemented by one, a neural network exactly performs polynomial regression. In the same spirit, neural network architectures are proposed in [47], showing the correspondence between the values of the function to be interpolated, the basis functions and the basic points on one hand, and the weights, the activations functions and the thresholds (bias) on the other hand. However, in practical applications, data is often noisy and incomplete, and polynomial interpolation is generally subject to overfitting [81], while neural networks are able to perform when there are noisy or incomplete data, and have the ability to generalize from the input data [62]. This dilemma between fitting the training dataset and being able to generalize the models to unencountered data is known as the bias-variance trade-off, and it is discussed in [5]. Raturi explains in [66] that solving the same problem as polynomial interpolation requires much less computational time and resources when using neural networks. In the same paper, he shows that neural networks provide better approximations of two special functions compared with Lagrange interpolation. Finally, it is shown in [48] that neural networks can interpolate and model a function by means of sigmoidal functions to approximate n samples in any dimension, with arbitrary precision and without training.

2.4. Hybrid networks. Here we introduce a simpler network architecture for classifying membership on a parametrized algebraic variety. We observed that during the learning process, which is an optimization process, the algorithm is not always able to reach the set of weights that minimize the loss function. This is due to the existence of local minima in the loss function, related to the utilization of the non-convex $x \mapsto x^d$ activation functions.

The second layer (Figure 7) contains only one neuron (with the identity activation function), introduced to take the linear combination of d -th power of weighted sums of input variables. One would like to remove this layer, and directly link outputs of the first layer with the third layer (containing only neurons with LeakyReLU activation function). We call these *hybrid networks*, because they combine layers with both $x \mapsto x^d$ and LeakyReLU activation functions.

Moreover, the geometrical interpretation of the network is now different since every neuron in the second layer (LeakyReLU neurons) take as input a different linear combination of all d -th power forms. The second layer combines different homogeneous polynomials that are not necessarily equal to the homogeneous polynomial defining the algebraic variety, but they can be used to approximate this last as a set of inequalities. The third and last layer of the network is the output layer, which will depend on the classification problem one wants to solve (binary classification, several classes, etc.). The architecture of the hybrid network is summarized in Figure 8.

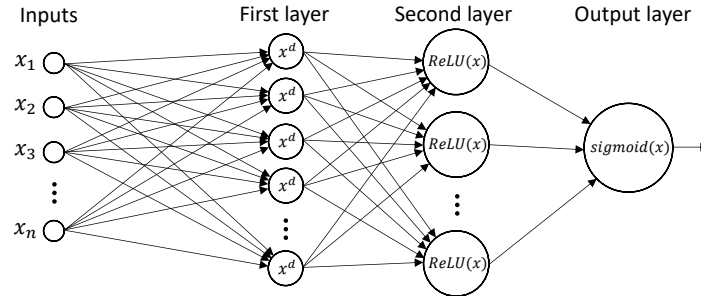


FIGURE 8. Representation of a hybrid network for learning a homogeneous polynomial equation of degree d in n variables and coefficients in \mathbb{R} .

This type of network showed better results and quicker learning than the previous version of the network. The AH theorem still can be used to give an idea about the number of neurons in the first layer, and sometimes extra neurons should be added or removed to boost the performance of the network during the learning phase. By experience, the number of neurons in the second layer can be chosen quite small (between 4 and 10 for binary classification problems) and will depend on the number of neurons in the output layer (it might not be smaller than the number of neurons in the output layer).

3. EXPERIMENTS

Consider a pair, (V, G) , of a vector space V and a group $G \subset \text{GL}(V)$. The set of all such pairs (V, G) for which G acts on V with finitely many orbits has been classified by Kac [19, 41]. Since then Vinberg and others have classified all the orbits in many of these cases. There are special cases where G acts on V with infinitely many orbits, yet those orbits may still be represented using finitely many parameters, the *tame* case. A fantastic example of such is [75], which gave a classification of the orbits of 9 dimensional trivectors utilizing a connection to the Lie algebra \mathfrak{e}_8 . This classification also implies a classification for 3 qutrit systems among others. After orbit classification, one desires effective methods to determine orbit membership. Separating orbits is difficult in general and lies at the core of geometric approaches to Valliant's version of P versus NP [44].

One approach is to use Invariant Theory to build a set of invariants and covariants that characterize each orbit, or each family represented by parameters [30, 52]. In the four-qubit case, an infinite tame case, Verstraete *et al.* gave a list of 9 normal forms (depending on parameters) that give a parametrization of all SLOCC orbits [16, 74]. An algorithm was proposed in [31, 32], for the four-qubit case, to determine orbit membership. In general, the complexity of these invariants grows very rapidly with the number of particles in the system. Already in the five-qubit case, a complete description of the algebra of SLOCC-invariant polynomials is out of reach of any computer system [53].

Therefore, it is worth considering other approaches. In this section we give examples that serve as proof of the concept that artificial neural networks can be trained to give efficient and effective classifiers for quantum states.

3.1. Modeling quantum states. In our work, we only focused on the real case, *i.e.* n -qudits systems that are represented by tensors in $(\mathbb{R}^d)^{\otimes n}$. Therefore, such states can be represented as d^n dimensional vectors with coordinates in \mathbb{R} , and this vector will be used as input vector for the neural network.

3.2. Detecting separable states. A state $\varphi \in \mathcal{H}^{\otimes n}$ is *separable* or *unentangled* if it can be written as a tensor product of pure states as follows

$$\varphi = \varphi_1 \otimes \cdots \otimes \varphi_n,$$

for $\varphi_i \in \mathcal{H}$ for all i . Algebraic geometers call the projective variety of separable states the Segre variety, which is parametrized as:

$$\begin{aligned} \text{Seg: } \quad \mathbb{P}\mathcal{H} \times \cdots \times \mathbb{P}\mathcal{H} &\rightarrow \mathbb{P}\mathcal{H}^{\otimes n} \\ ([x_1], [x_2], \dots, [x_n]) &\mapsto [x_1 \otimes x_2 \otimes \cdots \otimes x_n], \end{aligned}$$

where the square brackets $[\cdot]$ denote the equivalence class by rescaling, which makes physical sense to consider since we assume states are unit vectors. We uniformly sample the space of

separable states working on representatives of projective points on spheres:

$$\mathcal{SH} \times \cdots \times \mathcal{SH} \rightarrow \mathcal{SH}^{\otimes n},$$

where \mathcal{SH} denotes the states of unit norm, and similarly $\mathcal{SH}^{\otimes n}$.

The following is straightforward.

Proposition 3.1. *A non-zero state $\varphi \in \mathcal{H}$ is separable if and only if either of the following equivalent conditions.*

- (1) φ is in the SLOCC orbit of $|00 \cdots 0\rangle$.
- (2) All n of the 1-flattenings $F_i(\varphi): \mathcal{H}^{\otimes(n-1)} \rightarrow \mathcal{H}$ have rank 1.

Separable quantum states are states where quantum entanglement is not present at all. Detecting separable states thus becomes equivalent to detecting quantum entanglement. Therefore, one way to determine if a state is separable is to compute the compact form of the SVD of every 1-flattening.

In order to illustrate our methods, we show how we train an artificial neural network to perform this task.

3.2.1. Networks for pure states. The set of separable states is the zeroset of the two by two minors of 1-flattenings, and thus is defined by a set of homogeneous polynomials of degree 2. In the 2×2 case, there is only one defining equation, the 2×2 determinant. In higher dimensions, several equations of degree 2 define the Segre variety. To select the number of equations, we take (at least) the codimension of the Segre variety in the projective space.

For each quadratic equation, we use our criteria (see Proposition 2.3) to deduce the number of needed neurons, and then to design the Hybrid Networks. In some cases, we chose to add several extra neurons in the first layer (with quadratic activation function) to accelerate the learning process. As explained before, the second layer will only contain neurons with LeakyReLU activation functions. The last layer will contain only one neuron, with Sigmoid activation function in this case (because of the binary classification problem).

We also implemented networks with only LeakyReLU activation functions neurons (excepted for the last layer). For all the considered cases, we used the same network architecture, which is composed of 4 first layers and a last output layer. We chose to implement a decreasing structure in the network, with 100 neurons in the first layer, 50 in the second, 25 in the third, and 16 in the fourth.

In each case, we precise the architecture of the network in the second column of the results tables.

3.2.2. Results. In Table 2 we present the result for the 2×2 , the $2 \times 2 \times 2$, the $2^{\times 4}$, and the $3 \times 3 \times 3$ cases with hybrid networks. In the 2×2 , the $2 \times 2 \times 2$, and the $3 \times 3 \times 3$ cases, we used a training dataset of size 56200, a validation dataset of size 12800, and a testing dataset of size 32000. In the $2^{\times 4}$ case, we doubled the size of the training and the validation datasets only (same size for the testing one). We reached an average accuracy of 93% on the testing datasets for separability classification.

We used LeakyReLU networks to study the same cases as we did with the hybrid networks, with the addition of the $2^{\times 5}$ one, and we regrouped the results in Table 3. In the 2×2 and $2 \times 2 \times 2$ cases, we used a training dataset of size 102400, a validation dataset of size 25600, and a testing dataset of size 32000. In the $2^{\times 4}$, $2^{\times 5}$ and $3 \times 3 \times 3$ cases, we used a training dataset of size 502600, a validation dataset of size 55600, and a testing dataset of size 32000. We reached an average accuracy of 98% on the testing datasets for separability classification.

Tensor size	Architecture	Training acc.	Validation acc.	Testing acc.	Loss
2×2	(4,4,1)	96.65%	96.60%	96.63%	0.092
$2 \times 2 \times 2$	(21,8,1)	94.57%	94.06%	94.44%	0.15
$2^{\times 4}$	(1188,8,1)	91.72%	91.60%	91.33%	0.26
$3 \times 3 \times 3$	(332,12,1)	94.68%	92.89%	92.94%	0.15

TABLE 2. Hybrid network architectures and accuracies for each tensor size, for separability classification.

Tensor size	Architecture	Training acc.	Validation acc.	Testing acc.	Loss
2×2	(100,50,25,16,1)	98.92%	98.78%	98.83%	0.043
$2 \times 2 \times 2$	(100,50,25,16,1)	97.80%	97.42%	97.55%	0.074
$2^{\times 4}$	(100,50,25,16,1)	99.62%	99.50%	99.53%	0.016
$2^{\times 5}$	(100,50,25,16,1)	98.83%	98.55%	98.55%	0.037
$3 \times 3 \times 3$	(100,50,25,16,1)	98.55%	98.01%	97.92%	0.051

TABLE 3. LeakyReLU network architectures and accuracies for each tensor size, for separability classification.

These results show that an ANN can be trained to distinguish between separable and entangled states. From an algebraic geometric point of view, determining if a tensor is on the Segre variety is equivalent to decide if it is a rank one tensor, and this task can be performed using truncated SVD for each flattening. The complexity of computing the singular values of an $m \times n$ matrix is $\mathcal{O}(mn \min\{m, n\})$. This is roughly $\min\{m, n\}$ times the complexity of reading the matrix. Computing the singular values of the i -th flattening of a tensor of format $n_1 \times \cdots \times n_t$ has complexity

$$\mathcal{O}\left(n_i \left(\frac{n_1 \cdots n_t}{n_i}\right) \min\left\{n_i, \left(\frac{n_1 \cdots n_t}{n_i}\right)\right\}\right) = \mathcal{O}\left((n_1 \cdots n_t) \min\left\{n_i, \left(\frac{n_1 \cdots n_t}{n_i}\right)\right\}\right).$$

For “balanced” cases $2n_i \leq \sum_i n_i$, thus this complexity is at most

$$\mathcal{O}((n_1 \cdots n_t) n_i),$$

which is roughly n_i times the complexity of reading the tensor. Now computing all of the singular values of the tensor (in a naive way, not informing one computation by the results of another) one would expect complexity $\mathcal{O}((n_1 \cdots n_t)(\sum_i n_i))$. The cost of evaluating a trained network is equal to the cost of a feed-forward propagation, which needs to compute all the weighted sums (which is equivalent to compute matrix multiplications) and evaluate the activation function for each neuron. It will in fact depend on the architecture of the network. If we denote by l the number of layers (without counting the input layer) and by m_1, m_2, \dots, m_l the number of neurons in each layer, it has a complexity roughly equal to $\mathcal{O}((n_1 \cdots n_t)m_1 + m_1 m_2 + \cdots + m_{l-1} m_l)$.

This indicates that as the dimensions of the tensor spaces increase, that the cost of evaluating an already trained network should be more efficient than the computation of the SVD of flattenings. Though the cost of training the network cannot be ignored, it is a one-time cost, and the trained network can be used for efficient computation for any number of tests. In addition, once the tensor rank is larger than the dimension of any of the factors, there are few effective methods for determining rank, so a trained ANN classifier may provide valuable insight.

3.3. Detecting degenerate states. Recall from linear algebra that a matrix $A \in \mathbb{R}^{m \times n}$ is *degenerate* (rank deficient) if and only if any of the following equivalent conditions hold:

- (1) There are non-zero vectors $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ such that $u^\top Ax = 0$ for all $x \in \mathbb{R}^n$ and $y^\top Av = 0$ for all $y \in \mathbb{R}^m$.
- (2) In the case $m = n$, the determinant vanishes: $\det(A) = 0$.
- (3) Up to Gaussian elimination A has a zero row and a zero column.

These conditions carry over in the tensor setting, following [24]. A state $|\varphi\rangle \in \mathcal{H}^{\otimes n}$ is *degenerate* if one of the following equivalent conditions hold:

- (1) There is a pure state $x = |x_1\rangle \otimes \cdots \otimes |x_n\rangle \in \mathcal{H}^{\otimes n}$ such that the contraction $(\langle x_1| \otimes \cdots \otimes \langle x_{i-1}| \otimes \langle h_i| \otimes \langle x_{i+1}| \otimes \cdots \otimes \langle x_n|) |\varphi\rangle = 0$ for all i and for all $|h_i\rangle \in \mathcal{H}$
- (2) The hyperdeterminant vanishes: $\text{Det}(A) = 0$, where A is the hypermatrix satisfying $A_I = \langle I | \varphi \rangle$.
- (3) Up to SLOCC, the coordinates $\varphi_I = 0$ for all I of Hamming distance ≤ 1 away from $00 \cdots 0$.

The hyperdeterminant is a SLOCC invariant polynomial, and thus is used to characterize quantum entanglement. The zero locus of the equation defined by the hyperdeterminant also defines the dual of the Segre variety [58], and the study of singularities of the associated hypersurfaces gives also a qualitative information about quantum entanglement [34, 37].

On the other hand, the hyperdeterminant can sometimes also be regarded as a quantitative measure of entanglement and can be used to study entanglement in quantum algorithms for example [36, 40]. In a concrete sense the non-degenerate states are the most entangled ones and maximizing the absolute value of the hyperdeterminant can lead to maximally entangled states [12, 25, 35]. While the hyperdeterminant is a polynomial test for degeneracy, it has very high degree and rarely feasible to compute apart from the smallest cases. We note that one of us recently showed that hyperdeterminants coming from the root system E8 can be computed efficiently [33].

We wish to show that we can still learn degenerate states even in cases where an expression or evaluation method for the hyperdeterminant is not known explicitly.

3.3.1. A note on uniform sampling. The SLOCC description of degenerate states gives a way to uniformly sample the set. We take a random state φ with coordinates $\varphi_I = 0$ for all I of Hamming distance ≤ 1 away from $(00 \cdots 0)$, we renormalize so that $\|\varphi\| = 1$, and then we apply a random element of the SLOCC group.

3.3.2. Networks for degenerate states. The $2^{\times n}$ hyperdeterminant is a homogeneous polynomial whose degree is 4, 24, 128 respectively for $n = 3, 4, 5$, and its degree is 36 in the $3 \times 3 \times 3$ case (see [24] for concrete formulas).

In the $2 \times 2 \times 2$ case, the first layer of the hybrid network must contain at least $\lceil \frac{1}{8} \binom{4+8-1}{4} \rceil = 42$ neurons with $x \mapsto x^4$ activation functions. We chose to implement 60 neurons in the first layer for the 3-qubit case. In the two other cases, the number of needed neurons is way too high to hope an implementation using Keras. One can still use fewer neurons to approximate the hyperdeterminant using hybrid networks, and even if the accuracy will not be very close to 100%, one can still reach 60% or more and use the idea presented in Section 4.1 to have a better overall accuracy. Another possible idea is to use Standard Ensemble Learning (also known as bagging) and train several networks in order to build a committee that will give a more accurate answer concerning the detection of degenerate states.

The second possibility is to use only LeakyReLU activation functions (except for the output layer), as it was done in Subsection 3.2. We also used the same decreasing structure

with four layers, except in the 4-qubit case, when we chose to double the number of neurons in the three first layers.

The 2×2 non-degenerate states are in fact non-separable states. This case is thus already solved by the 2×2 separability classifiers presented in Section 3.2.

3.3.3. Results. In Table 4 we present the result for the $2 \times 2 \times 2$ case with the hybrid network. We used a training dataset of size 202600, a validation dataset of size 25600, and a testing dataset of size 32000. We reached 92% of accuracy for the testing dataset.

Tensor size	Architecture	Training acc.	Validation acc.	Testing acc.	Loss
$2 \times 2 \times 2$	(60,10,1)	92.49%	92.18%	92.09%	0.1837

TABLE 4. Hybrid network architectures and accuracies for each tensor size, for degenerate and non-degenerate states classification.

In Table 5 we present the results for the $2 \times 2 \times 2$, $2^{\times 4}$, $2^{\times 5}$, and $3 \times 3 \times 3$ cases with networks using LeakyReLU activation functions. In the $2 \times 2 \times 2$ case, we used a training dataset of size 502600, a validation dataset of size 55600, and a testing dataset of size 32000. In the $2^{\times 4}$ and $2^{\times 5}$ cases, we used a training dataset of size 252400, a validation dataset of size 55600, and a testing dataset of size 52000. In the $3 \times 3 \times 3$ case, we used a training dataset of size 352400, a validation dataset of size 55600, and a testing dataset of size 52000. We reached an average accuracy of 96% on the testing datasets for classification of degenerate states.

The obtained results are quite interesting, especially in the 5-qubit case. In fact, as mentioned before, no explicit expression of the $2^{\times 5}$ hyperdeterminant is available, because of the complications related with computations. The network has 98% accuracy for the testing dataset, which means that we can determine with very high accuracy if a tensor is degenerate or not, which is equivalent to determine if the hyperdeterminant vanishes or not. In the context of qualitative characterization of entanglement, one only need to know if the state annihilates or not the hyperdeterminant. In this sense, we provide an original result for the evaluation of the nullity of the hyperdeterminant for $2^{\times 5}$ real tensors, that can be used to investigate quantum entanglement for 5-qubit systems (see Section 4.3).

Tensor size	Architecture	Training acc.	Validation acc.	Testing acc.	Loss
$2 \times 2 \times 2$	(100,50,25,16,1)	93.44%	92.53%	92.74%	0.1629
$2^{\times 4}$	(200,100,50,16,1)	99.50%	95.95%	95.94%	0.01791
$2^{\times 5}$	(100,50,25,16,1)	99.95%	98.74%	98.83%	0.001533
$3 \times 3 \times 3$	(100,50,25,16,1)	98.18%	96.78%	96.83%	0.04770

TABLE 5. LeakyReLU network architectures and accuracies for each tensor size, for degenerate and non-degenerate states classification.

3.4. Border-rank classification. A state $\varphi \in \mathcal{H}^{\otimes n}$ is said to have $\text{rank} \leq R$ if there is an expression $\varphi = \sum_{i=1}^R \lambda_i \varphi_i$ with $\lambda_i \in \mathcal{H}$ and $\varphi_i \in \mathcal{H}^{\otimes n}$. It is well known that when $n > 2$ rank is not a closed condition, and it is not semi-continuous. The first example is 3 qubits, where the W -state $|100\rangle + |010\rangle + |001\rangle$, which has rank 3, and is in the closure of the generic orbit, that of $|000\rangle + |111\rangle$, which has rank 2.

In another context, Bini [8] defined the notion of border rank to regain semi-continuity. One says that a state $\varphi \in \mathcal{H}^{\otimes n}$ has *border rank* $\leq R$ if there exists a family of rank R states $\{\varphi^\epsilon \mid \epsilon > 0\}$ and $\lim_{\epsilon \rightarrow 0} \varphi^\epsilon = \varphi$. Equivalently, the set of border rank $\leq R$ states, denoted σ_R is the Zariski closure of the states of rank $\leq R$.

By construction, when $\mathcal{H} = \mathbb{C}^d$ we have a chain

$$\sigma_1 \subsetneq \sigma_2 \subsetneq \cdots \subsetneq \sigma_g = \mathcal{H}^{\otimes n}$$

which ends at $\mathcal{H}^{\otimes n}$ because the set of separable states is linearly non-degenerate in $\mathcal{H}^{\otimes n}$. The minimal integer g for which $\sigma_g = \mathcal{H}^{\otimes n}$ is called the generic rank. A simple dimension count gives the expected generic rank:

$$e = \left\lceil \frac{d^n}{n(d-1) + 1} \right\rceil$$

The only known case for tensors of type $d^{\times n}$ where the generic rank differs from the expected rank is the case $d = 3, n = 3$, so $3 \times 3 \times 3$ tensors or 3-qutrits where the generic rank is 5 and not 4. In the case $d = 2, n = 4$, so $2 \times 2 \times 2 \times 2$ or 4 qubits, where the generic rank is the expected generic rank of 4, even though it is known that the set of rank 3 tensors is defective. It is conjectured that these are the only such cases (see [1, 2, 14]).

Let σ_s° denote the states of rank exactly s . When $\mathcal{H} = \mathbb{R}^d$ there can be more than one semi-algebraic set σ_s° that is full dimensional. These ranks are called *typical ranks*. Given a probability measure μ on $\mathcal{H}^{\otimes n}$ the measure of σ_s° represents the probability of a random state having rank s . We learned of the following example from [20] which illustrates the situation with real typical ranks.

Example 5. Set $n = 3, d = 2$. Both 2 and 3 are typical ranks, and moreover that a positive volume set of tensors with rank 3 have no optimal low-rank approximation. The $2 \times 2 \times 2$ hyperdeterminant D separates $\mathbb{R}^{2 \times 2 \times 2}$ into regions of constant typical rank. We formed a sample of tensors of real rank at most 3 as follows. We constructed rank-one tensors as a tensor product of three vectors with entries uniformly distributed in $[-.5, .5]$ and normalized to unit-norm. Then we took the sum of these vectors and normalized again. For such a distribution of tensors we found the following frequencies:

- (1) with approximate frequency 86.6% $D(\varphi) > 0$, in which case the \mathbb{R} -rank is 2,
- (2) with approximate frequency 13.4% $D(\varphi) < 0$, in which case \mathbb{R} -rank is 3,
- (3) with frequency 0% $D(\varphi) = 0$, in which case the \mathbb{R} -rank can be 0,1,2,3. \diamond

3.4.1. A note on uniform sampling. We construct points on algebraic varieties by utilizing rational parameterizations of the form $\varphi: U \rightarrow V$, with φ a rational function, and U, V subsets of a normed linear space. If we uniformly sample the source U it is generally the case that $\varphi(U)$ is not a uniform sample of the image. Instead of trying to uniformly sample the image of these varieties [22, 29, 64], we provide training data that is constructed in the way we think that one might construct the test data, or how one might imagine the data is produced from a state constructed in a lab. Whether this is a reasonable assumption is open for debate. On one hand, it is always possible (by forcing incoherence, for instance) for an adversarial entity to construct a data set on the same variety on which we're testing membership that fools our classifier. On the other hand, we can say that if we train our neural network on data produced by a certain process, we can be confident that if the validation data is constructed by the same process, then the accuracy numbers we report reflect a reasonable measure of confidence in the classifier.

3.4.2. Results. In Table 6 we present the result for the $2 \times 2 \times 2$ rank classification with the hybrid network. We used a training dataset of size 102400, a validation dataset of size 25600, and a testing dataset of size 32000. We reached 87% accuracy for the testing dataset.

Tensor size	Architecture	Training acc.	Validation acc.	Testing acc.	Loss
$2 \times 2 \times 2$	(169,25,3)	88.19%	88.03%	87.95%	0.3028

TABLE 6. Hybrid network architectures and accuracies for 3-qubits, for rank classification.

In Table 7 we present the results for the $2 \times 2 \times 2$, $2^{\times 4}$, and $2^{\times 5}$ cases with network using LeakyReLU activation functions. In the $2 \times 2 \times 2$ rank classification problem, we used a training dataset of size 502500, a validation dataset of size 55600, and a testing dataset of size 52000. In the $2^{\times 4}$ and $2^{\times 5}$ border rank classification problems, we respectively used a training dataset of size 502400 and 802400, a validation dataset of size 55600 and a testing dataset of size 52000.

Tensor size	Architecture	Training acc.	Validation acc.	Testing acc.	Loss
$2 \times 2 \times 2$	(200,100,50,25,3)	94.19%	94.07%	93.79%	0.1674
$2^{\times 4}$	(200,100,50,25,3)	85.49%	84.45%	84.47%	0.3144
$2^{\times 5}$	(200,100,50,25,3)	81.39%	79.88%	79.77%	0.4230

TABLE 7. LeakyReLU network architectures and accuracies for each tensor size, for rank and border-rank classification.

In the $2 \times 2 \times 2$ case, we trained our networks to recognize the exact rank of tensors. In fact, for building the training and validation datasets, we provided tensors from Segre variety for rank one, tensors that are SLOCC equivalents to the biseparable and GHZ states for rank 2, and states that are SLOCC equivalents to the W state for rank 3, following the classification of 3-qubits (see Table 8). We also took into account the typical rank (see Example 5) and used the sign of the hyperdeterminant to generate rank 2 and rank 3 tensors.

In the $2^{\times 4}$ and $2^{\times 5}$ cases, we generated tensors for each class k by summing $k + 1$ rank one tensors (and renormalized). The class ‘0’ corresponds to rank one tensors, and the other classes r correspond to tensors with border rank $r + 1$. In the $2^{\times 4}$, we generated tensors up to border rank 4, and up to border rank 5 in the $2^{\times 5}$ case. By using this way of generating vectors for each class, we introduce noise in the data. However, even in the presence of noise, the network is able to learn and predict border rank of states with a quite interesting accuracy (84% in the 4-qubit case, and 80% in the 5-qubit case, on the testing dataset (which is also generated by the same process)).

For example, we evaluate the 5-qubit network on the input state $|W_5\rangle = |00001\rangle + |00010\rangle + |00100\rangle + |01000\rangle + |10000\rangle$ (see Section 4.1 for how we make predictions), we obtain most of the time the class ‘1’ (see Figure 9), which corresponds to tensors with border rank 2.

4. HOW TO USE THE CLASSIFIER AFTER THE FACT

4.1. Predictions for a single quantum state. In this work, we study quantum entanglement from a qualitative point of view. Quantum states are regrouped into classes, which correspond to the orbits of the group SLOCC. When the number of orbits is infinite, we talk about families depending on parameters. Any state belonging to a specific class will be equivalent to another state in the same class by the action of an element of this group. Moreover, other properties, such as separability, degeneracy, and tensor rank, are invariants under the action of the group SLOCC. Consequently, all of our classifiers should theoretically give the same answer for SLOCC equivalent states.

However, the accuracy of our classifiers is not exactly equal to 100%, and thus the probability of misclassification is non-zero, which gives little confidence for a single test. In order

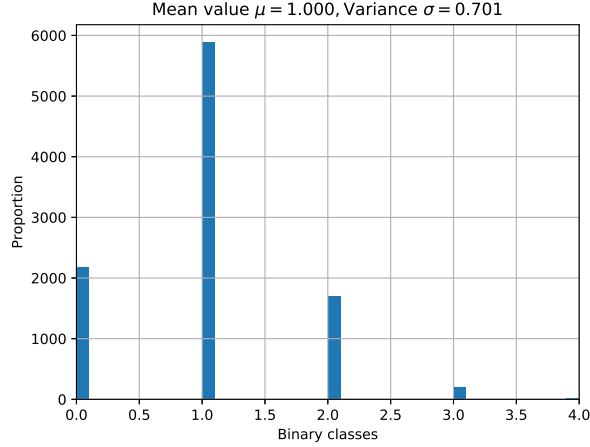


FIGURE 9. Histogram of the prediction results for the border rank classifier, for the state $|W_5\rangle$. The plot shows that the state is of border rank 2 (class ‘1’), with 10000 SLOCC equivalent points.

to avoid (or significantly reduce) misclassification, when one wants to make predictions for a specific quantum state (or set of states), the idea is to generate SLOCC equivalent states and look at the most frequent answer of the neural network classifier. All the answers can be regrouped into a histogram in order to give a graphical representation of the neural network output distribution. Examples are investigated in the next sections. We report the results after applying the classifier to enough points of the orbit so that the consensus was reached (the answer was the same for larger data samples).

4.2. Three-qubit entanglement classification. The 3-qubit case was the first case to illustrate the existence of non-equivalent entangled states, such as $|GHZ\rangle$ and $|W\rangle$ states, as it was proven by [23]. The entanglement classification of 3-qubit systems (see Table 8) can be described by join, secant and tangential varieties [30], and by using dual varieties as well [30, 56–58]. Rank can also be used as an algebraic measure of entanglement to distinguish between several 3-qubits entanglement classes.

Normal forms	Class	Rank	Cayley’s Δ_{222}
$ 000\rangle + 111\rangle$	GHZ	2	$\neq 0$
$ 001\rangle + 010\rangle + 100\rangle$	W	3	0
$ 001\rangle + 111\rangle$	Bi-Separable C-AB	2	0
$ 010\rangle + 111\rangle$	Bi-Separable B-CA	2	0
$ 100\rangle + 111\rangle$	Bi-Separable A-BC	2	0
$ 000\rangle$	Separable	1	0

TABLE 8. Entanglement classification of 3-qubit systems under SLOCC.

Therefore, our binary classifiers built in Section 3 can be used to distinguish between 3-qubit entanglement classes. In fact, separable states can be detected using the binary classifier for tensors on the Segre variety, presented in Section 3.2. All states that are SLOCC equivalent to the $|GHZ\rangle$ state correspond to non-degenerate states, and thus can be recognized using the binary classifier for tensors on the dual variety of the Segre variety, presented in Section 3.3. Finally, to distinguish between Bi-Separable states and states that are SLOCC equivalents to $|W\rangle$, we exploit the rank of tensors, by using the rank classifier,

presented in Section 3.4, as a predictor for distinguishing between rank 2 and rank 3 tensors. All the results are regrouped in Figures 10,11,12,13.

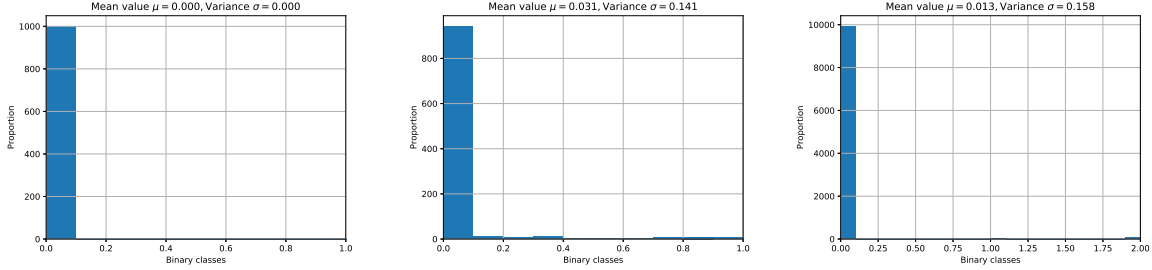


FIGURE 10. Histograms of the prediction results for (in order, from left to right) separable states, degenerate states and tensor rank classifiers, for the separable state $|000\rangle$. The left plot shows that the state is separable (class ‘0’), with 1000 SLOCC equivalent points. The middle plot shows that the state is degenerate (class ‘0’), with 1000 equivalent points. The right plot show that the state is of rank one (class ‘0’) with 10000 equivalent points.

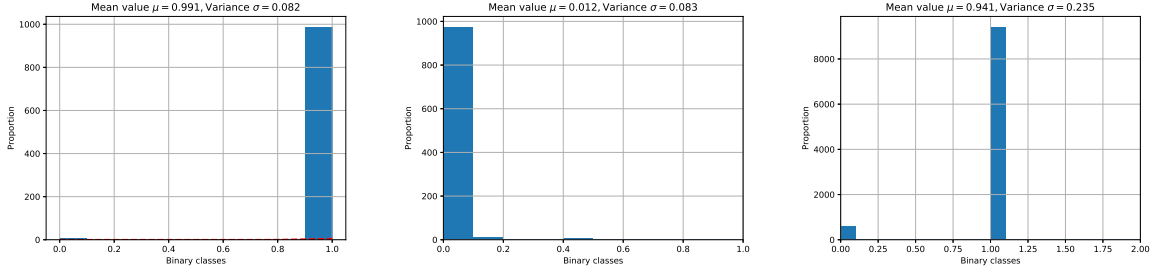


FIGURE 11. Histograms of the prediction results for (in order, from left to right) separable states, degenerate states and tensor rank classifiers, for the biseparable state $|000\rangle + |011\rangle$. The left plot shows that the state is entangled (class ‘1’), with 1000 SLOCC equivalent points. The middle plot shows that the state is degenerate (class ‘0’), with 1000 equivalent points. The right plot show that the state is of rank two (class ‘1’) with 10000 equivalent points.

4.3. Entanglement of five-qubit systems. In the five-qubit case, since no complete or parametrized classification is known, distinguishing entanglement classes is a difficult task. Some prior work focused on using filters [63] or using polynomial invariants [53]. In these papers, they considered four 5-qubit systems $|\Phi_1\rangle$, $|\Phi_2\rangle$, $|\Phi_3\rangle$, and $|\Phi_4\rangle$ (see Equations 7, 8, 9, 10).

$$|\Phi_1\rangle = \frac{1}{\sqrt{2}} (|00000\rangle + |11111\rangle) \quad (7)$$

$$|\Phi_2\rangle = \frac{1}{2} (|11111\rangle + |11100\rangle + |00010\rangle + |00001\rangle) \quad (8)$$

$$|\Phi_3\rangle = \frac{1}{\sqrt{6}} \left(\sqrt{2}|11111\rangle + |11000\rangle + |00100\rangle + |00010\rangle + |00001\rangle \right) \quad (9)$$

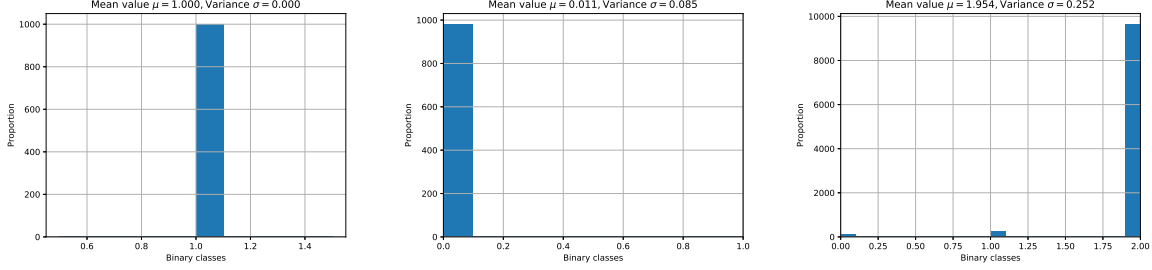


FIGURE 12. Histograms of the prediction results for (in order, from left to right) separable states, degenerate states and tensor rank classifiers, for the W-state $|001\rangle + |010\rangle + |100\rangle$. The left plot shows that the state is entangled (class ‘1’), with 1000 SLOCC equivalent points. The middle plot shows that the state is degenerate (class ‘0’), with 1000 equivalent points. The right plot show that the state is of rank three (class ‘2’) with 10000 equivalent points.

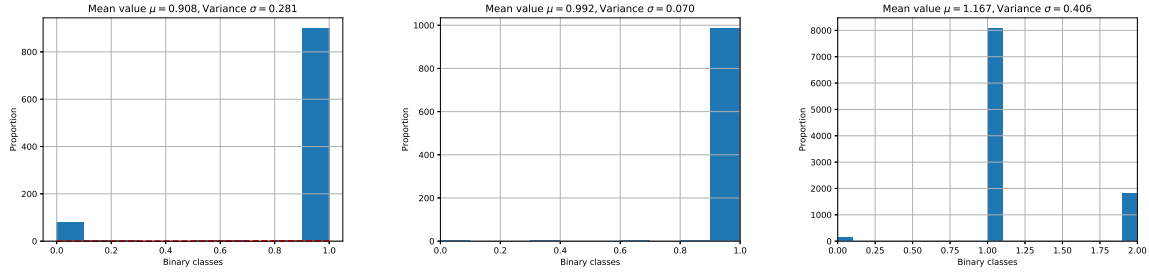


FIGURE 13. Histograms of the prediction results for (in order, from left to right) separable states, degenerate states and tensor rank classifiers, for the GHZ-state $|000\rangle + |111\rangle$. The left plot shows that the state is entangled (class ‘1’), with 1000 SLOCC equivalent points. The middle plot shows that the state is non-degenerate (class ‘1’), with 1000 equivalent points. The right plot show that the state is of rank two (class ‘1’) with 10000 equivalent points.

$$|\Phi_4\rangle = \frac{1}{2\sqrt{2}} \left(\sqrt{3}|11111\rangle + |10000\rangle + |01000\rangle + |00100\rangle + |00010\rangle + |00001\rangle \right) \quad (10)$$

These states do not belong to the same entanglement classes, as it was shown in [53, 63]. However, one can ask if these states are degenerate or not. We used our classifier for 5-qubits degenerate states to investigate this question, and it turned out that for all of these four states, the $2^{\times 5}$ hyperdeterminant vanishes, as it shown in Figure 14.

On the other hand, we can also propose several non-degenerate states using our classifier. For example, the states $|\delta_1\rangle$ and $|\delta_2\rangle$ (see Equations 11 and 12) are non-degenerate states, as it is confirmed by the prediction results of our classifier presented in Figure 15. This statement can be verified using the study of hypersurface singularities associated with these two states, as it was done in [34, 37].

$$|\delta_1\rangle = \frac{1}{\sqrt{14}} (|00000\rangle + \sqrt{3}|00011\rangle + |00100\rangle + |01000\rangle + |01001\rangle + \sqrt{2}|01111\rangle + |10001\rangle + |10110\rangle + |11000\rangle + |11011\rangle + |11101\rangle) \quad (11)$$

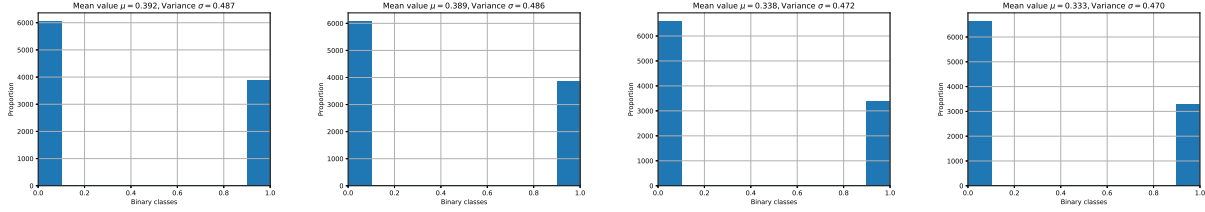


FIGURE 14. Histograms of the prediction results for degenerate states classifier on the four states $|\Phi_1\rangle$, $|\Phi_2\rangle$, $|\Phi_3\rangle$, and $|\Phi_4\rangle$. The first left plot corresponds to the state $|\Phi_1\rangle$, and the last right one to $|\Phi_4\rangle$. On both horizontal axes, the class ‘0’ mean that the state is degenerate, while the class ‘1’ refers to non-degenerate states. In this case, all states are degenerate. We generated 10000 points in the same SLOCC orbits for making prediction for each state.

$$|\delta_2\rangle = \frac{1}{\sqrt{11}}(|00000\rangle + |00100\rangle + |00111\rangle + |01010\rangle - |01101\rangle + |10001\rangle + |10011\rangle + |10111\rangle - |11000\rangle + |11110\rangle) \quad (12)$$

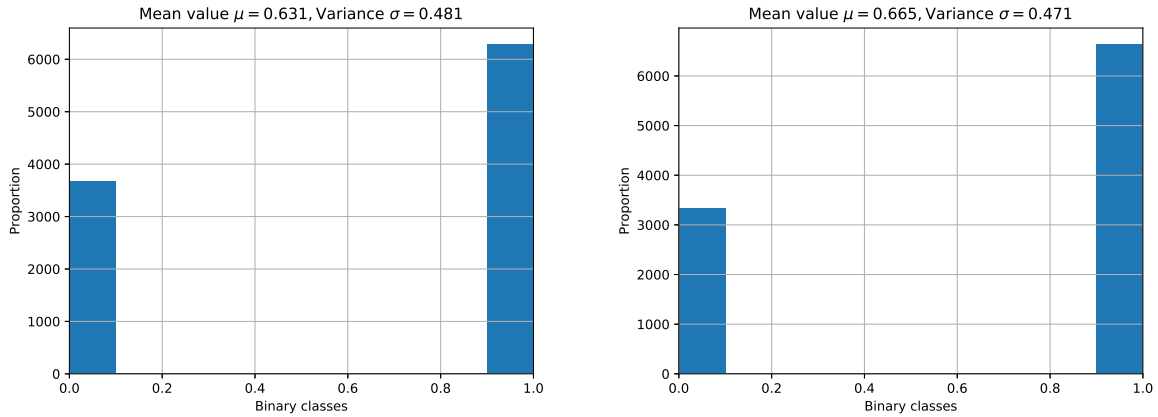


FIGURE 15. Histograms of the prediction results for degenerate states classifier on $|\delta_1\rangle$ and $|\delta_2\rangle$ states. The left plot corresponds to the state $|\delta_1\rangle$, and the right one to $|\delta_2\rangle$. On both horizontal axes, the class ‘0’ mean that the state is degenerate, while the class ‘1’ refers to non-degenerate states. In this case, both states are non-degenerate. We generated 10000 points in the same SLOCC orbits for making prediction for each state.

5. ACKNOWLEDGEMENTS

This work was partially supported by the Région Bourgogne Franche-Comté, project PHYFA (contract 20174-06235), the French “Investissements d’Avenir” programme, project ISITE-BFC (contract ANR-15-IDEX-03) and by “BQR Mobilité Doctorante” at U.T.B.M. The first author thanks the Department of Mathematics at Auburn University for its hospitality. The first author wants also to thank Dr. Mark Carpenter (Auburn University) and Dr. Frabrice Lauri (U.T.B.M.) for lectures and discussion about Machine Learning. The authors want to thank Dr. Frédéric Holweck for his advices, comments and discussions.

REFERENCES

- [1] H. Abo, G. Ottaviani, and C. Peterson, *Induction for secant varieties of Segre varieties*, Trans. Amer. Math. Soc. **361** (2009), no. 2, 767–792.
- [2] K. Baur, J. Draisma, and W. A. de Graaf, *Secant dimensions of minimal orbits: computations and conjectures*, Experiment. Math. **16** (2007), no. 2, 239–250.
- [3] Matthew J. S. Beach, Isaac De Vlugt, Anna Golubeva, Patrick Huembeli, Bohdan Kulchytskyy, Xiuzhe Luo, Roger G. Melko, Ejaaz Merali, and Giacomo Torlai, *Qucumber: wavefunction reconstruction with neural networks*, arXiv preprint arXiv:1812.09329 (2018).
- [4] Elizabeth C. Behrman and James Edward Steck, *Dynamic learning of pairwise and three-way entanglement*, 2011 third world congress on nature and biologically inspired computing, 2011, pp. 99–104.
- [5] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal, *Reconciling modern machine learning and the bias-variance trade-off.*, arXiv preprint arXiv:1812.11118 (2018).
- [6] Richard Berkovits, *Extracting many-particle entanglement entropy from observables using supervised machine learning*, Physical Review B **98** (2018), no. 24.
- [7] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd, *Quantum machine learning*, Nature **549** (2017), no. 7671, 195–202 (English (US)).
- [8] Dario Bini, Milvio Capovani, Francesco Romani, and Grazia Lotti, $O(n^{2.7799})$ complexity for $n \times n$ approximate matrix multiplication, Inform. Process. Lett. **8** (1979), no. 5, 234–235.
- [9] Maria Chiara Brambilla and Giorgio Ottaviani, *On the alexander–hirschowitz theorem*, Journal of Pure and Applied Algebra **212** (2008), no. 5, 1229–1251.
- [10] Paul Breiding, Sara Kališnik, Bernd Sturmfels, and Madeleine Weinstein, *Learning algebraic varieties from samples*, Revista Matemática Complutense **31** (2018Sep), no. 3, 545–593.
- [11] Maolin Che, Liquan Qi, Yimin Wei, and Guofeng Zhang, *Geometric measures of entanglement in multipartite pure states via complex-valued neural networks*, Neurocomputing **313** (2018), 25–38.
- [12] Lin Chen and Dragomir Ž. Đoković, *Proof of the gour-wallach conjecture*, Physical Review A **88** (2013), no. 4.
- [13] Xi Cheng, Bohdan Khomtchouk, Norman Matloff, and Pete Mohanty, *Polynomial regression as an alternative to neural nets*, arXiv preprint arXiv:1806.06850 (2018).
- [14] L. Chiantini, G. Ottaviani, and N. Vannieuwenhoven, *An algorithm for generic and low-rank specific identifiability of complex tensors*, SIAM J. Math. Anal. **35** (2014), no. 4, 1265–1287, available at <http://dx.doi.org/10.1137/140961389>.
- [15] François Chollet et al., *Keras*, 2015.
- [16] Oleg Chterental and Dragomir Z. Djokovic, *Normal forms and tensor ranks of pure states of four qubits*, arXiv preprint arXiv:quant-ph (2006).
- [17] Balazs Csanad Csaji, *Approximation with artificial neural networks* (2001).
- [18] George Cybenko, *Approximation by superpositions of a sigmoidal function*, Mathematics of Control, Signals, and Systems **5** (1989), no. 4, 455–455.
- [19] J. Dadok and V. Kac, *Polar representations*, J. Algebra **92** (1985), no. 2, 504–524.
- [20] V. de Silva and L.-H. Lim, *Tensor rank and the ill-posedness of the best low-rank approximation problem*, SIAM J. Matrix Anal. Appl. **30** (2008), 1084–1127.
- [21] Timothy Dozat, *Incorporating nesterov momentum into adam* (2016).
- [22] Emilie Dufresne, Parker B. Edwards, Heather A. Harrington, and Jonathan D. Hauenstein, *Sampling real algebraic varieties for topological data analysis*, arXiv preprint arXiv:1802.07716 (2018).
- [23] W. Dür, G. Vidal, and J. I. Cirac, *Three qubits can be entangled in two inequivalent ways*, Phys. Rev. A **62** (2000Nov), 062314.
- [24] I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky, *Discriminants, resultants, and multidimensional determinants*, Mathematics: Theory & Applications, Boston: Birkhäuser, Boston, MA, 1994.
- [25] Gilad Gour and Nolan R. Wallach, *All maximally entangled four-qubit states*, J. Math. Phys. **51** (2010), no. 11, 112201.
- [26] Lishen. Govender, *Determination of quantum entanglement concurrence using multilayer perceptron neural networks.*, Ph.D. Thesis, 2017.
- [27] Johnnie Gray, Leonardo Banchi, Abolfazl Bayat, and Sougato Bose, *Machine-learning-assisted many-body entanglement measurement*, Physical Review Letters **121** (2018), no. 15.

- [28] Boris Hanin, *Universal function approximation by deep neural nets with bounded width and relu activations.*, arXiv preprint arXiv:1708.02691 (2017).
- [29] Victoria Hernández-Mederos and Jorge Estrada-Sarlabous, *Sampling points on regular parametric curves with control of their distribution*, Computer Aided Geometric Design **20** (2003), no. 6, 363–382.
- [30] Frédéric Holweck, Jean-Gabriel Luque, and Jean-Yves Thibon, *Geometric descriptions of entangled states by auxiliary varieties*, Journal of Mathematical Physics **53** (2012), no. 10, 102203.
- [31] Frédéric Holweck, Jean-Gabriel Luque, and Jean-Yves Thibon, *Entanglement of four-qubit systems: A geometric atlas with polynomial compass ii (the tame world)*, Journal of Mathematical Physics **58** (2017), no. 2, 022201, available at <https://doi.org/10.1063/1.4975098>.
- [32] Frédéric Holweck, Jean-Gabriel Luque, and Jean-Yves Thibon, *Entanglement of four qubit systems: A geometric atlas with polynomial compass i (the finite world)*, Journal of Mathematical Physics **55** (2014), no. 1, 12202.
- [33] Frédéric Holweck and Luke Oeding, *Hyperdeterminants from the E_8 Discriminant*, arXiv e-prints (2018Oct), arXiv:1810.05857, available at [1810.05857](https://arxiv.org/abs/1810.05857).
- [34] Frédéric Holweck and Hamza Jaffali, *Three-qutrit entanglement and simple singularities*, Journal of Physics A **49** (2016), no. 46, 465301.
- [35] Frédéric Holweck, Hamza Jaffali, Peter Levay, and Jean-Gabriel Luque, *Maximally entangled symmetric states*, (in preparation) (2019).
- [36] Frédéric Holweck, Hamza Jaffali, and Ismaël Nounouh, *Grover’s algorithm and the secant varieties*, Quantum Information Processing **15** (2016), no. 11, 4391–4413.
- [37] Frédéric Holweck, Jean-Gabriel Luque, and Michel Planat, *Singularity of type d_4 arising from four qubit systems*, Journal of Physics A **47** (2014), no. 13, 135301.
- [38] Kurt Hornik, *Approximation capabilities of multilayer feedforward networks*, Neural Networks **4** (1991), no. 2, 251–257.
- [39] P. Huggins, B. Sturmfels, J. Yu, and D. Yuster, *The hyperdeterminant and triangulations of the 4-cube*, Math. Comp. **77** (2008), no. 263, 1653–1679.
- [40] Hamza Jaffali and Frédéric Holweck, *Quantum entanglement involved in grover’s and shor’s algorithms: the four-qubit case*, Quantum Information Processing **18** (2019), no. 5, 133.
- [41] V. Kac, *Some remarks on nilpotent orbits*, J. Algebra **64** (1980), no. 1, 190–213.
- [42] Iordanis Kerenidis, Jonas Landman, Alessandro Luongo, and Anupam Prakash, *q-means: A quantum algorithm for unsupervised machine learning.*, arXiv preprint arXiv:1812.03584 (2018).
- [43] Joe Kileel, Matthew Trager, and Joan Bruna, *On the expressive power of deep polynomial neural networks*, arXiv preprint arXiv:1905.12207 (2019).
- [44] J. M. Landsberg, *Geometry and complexity theory*, Cambridge Studies in Advanced Mathematics, Cambridge University Press, 2017.
- [45] J.M. Landsberg, *Tensors: geometry and applications*, Graduate Studies in Mathematics, vol. 128, American Mathematical Society, Providence, RI, 2012.
- [46] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, *Deep learning*, Nature **521** (2015), no. 7553, 436–444.
- [47] Hong-Xing Li and E.S. Lee, *Interpolation functions of feedforward neural networks*, Computers & Mathematics With Applications **46** (2003), no. 12, 1861–1874.
- [48] B. Llanas and F. J. Sainz, *Constructive approximate interpolation by neural networks*, Journal of Computational and Applied Mathematics **188** (2006), no. 2, 283–308.
- [49] Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost, *Quantum principal component analysis*, Nature Physics **10** (2014), no. 9, 631–633.
- [50] Sirui Lu, Shilin Huang, Keren Li, Jun Li, Jianxin Chen, Dawei Lu, Zhengfeng Ji, Yi Shen, Duanlu Zhou, and Bei Zeng, *Separability-entanglement classifier via machine learning*, Physical Review A **98** (2018), no. 1.
- [51] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang, *The expressive power of neural networks: A view from the width*, neural information processing systems (2017), 6231–6239.
- [52] Jean-Gabriel Luque and Jean-Yves Thibon, *Polynomial invariants of four qubits*, Physical Review A **67** (2003), no. 4, 42303.
- [53] ———, *Algebraic invariants of five qubits*, Journal of Physics A **39** (2006), no. 2, 371–377.

- [54] Yue-Chi Ma and Man-Hong Yung, *Transforming bell's inequalities into state classifiers with machine learning*, npj Quantum Information **4** (2018), no. 1, 34.
- [55] Warren S. McCulloch and Walter Pitts, *A logical calculus of the ideas immanent in nervous activity*, Bulletin of Mathematical Biology **52** (1990), no. 4, 99–115.
- [56] Akimasa Miyake, *Classification of multipartite entangled states by multidimensional determinants*, Phys. Rev. A **67** (2003Jan), 012108.
- [57] Akimasa Miyake and Frank Verstraete, *Multipartite entanglement in $2 \times 2 \times n$ quantum systems*, Physical Review A **69** (2004), no. 1, 12101.
- [58] Akimasa Miyake and Miki Wadati, *Multipartite entanglement and hyperdeterminants*, Quantum Information & Computation **2** (2002), no. 7, 540–555.
- [59] Hendrik Poulsen Nautrup, Nicolas Delfosse, Vedran Dunjko, Hans J. Briegel, and Nicolai Friis, *Optimizing quantum error correction codes with reinforcement learning*, arXiv preprint arXiv:1812.08451 (2018).
- [60] Michael A. Nielsen and Isaac L. Chuang, *Quantum computation and quantum information: 10th anniversary edition*, 10th ed., Cambridge University Press, New York, NY, USA, 2011.
- [61] Sung-Kwon Oh, Witold Pedrycz, and Byoung-Jun Park, *Polynomial neural networks architecture: analysis and design*, Computers & Electrical Engineering **29** (2003), no. 6, 703–725.
- [62] Hong Choon Ong, Choon, Lim Chee Kang, and Yong Yeow Wui, *Non linear approximations using multi-layered perceptions and polynomial regressions* (200601).
- [63] Andreas Osterloh and Jens Siewert, *Entanglement monotones and maximally entangled states in multipartite qubit systems*, International Journal of Quantum Information **4** (2006), no. 3, 531–540.
- [64] Luca Pagani and Paul J. Scott, *Curvature based sampling of curves and surfaces*, Computer Aided Geometric Design **59** (2018), 32–48.
- [65] Yihui Quek, Stanislav Fort, and Hui Khoo Ng, *Adaptive quantum state tomography with neural networks*, arXiv preprint arXiv:1812.06693 (2018).
- [66] Rohit Raturi, *Large data analysis via interpolation of functions: Interpolating polynomials vs artificial neural networks*, American Journal of Intelligent Systems **8** (2018), no. 1, 6–11.
- [67] Patrick Rebentrost, Masoud Mohseni, and Seth Lloyd, *Quantum support vector machine for big data classification*, Physical Review Letters **113** (2014), no. 13, 130503.
- [68] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, *Learning internal representations by error propagation*, Neurocomputing: foundations of research (1988), 673–695.
- [69] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione, *The quest for a quantum neural network*, Quantum Information Processing **13** (2014), no. 11, 2567–2586.
- [70] ———, *An introduction to quantum machine learning*, Contemporary Physics **56** (2015), no. 2, 172–185.
- [71] Corrado Segre, *Sulle corrispondenze quadrilineari tra forme di 1.a specie e su alcune loro rappresentazioni spaziali*, Annali Mat. pura ed applicata **29** (1920), 105–140 (italian).
- [72] Yu-Bo Sheng and Lan Zhou, *Distributed secure quantum machine learning*, Chinese Science Bulletin **62** (2017), no. 14, 1025–1029.
- [73] Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton, *On the importance of initialization and momentum in deep learning*, Proceedings of the 30th international conference on machine learning, 2013, pp. 1139–1147.
- [74] Frank Verstraete, J Dehaene, B De Moor, and Henri Verschelde, *Four qubits can be entangled in nine different ways*, Physical Review A **65** (2002), no. 5, 52112.
- [75] È. B. Vinberg and A. G. Èlašvili, *A classification of the three-vectors of nine-dimensional space*, Trudy Sem. Vektor. Tenzor. Anal. **18** (1978), 197–233.
- [76] Bingjie Wang, *Learning to detect entanglement*, arXiv preprint arXiv:1709.03617 (2017).
- [77] J. Weyman and A. Zelevinsky, *Singularities of hyperdeterminants*, Ann. Inst. Fourier (Grenoble) **46** (1996), no. 3, 591–644.
- [78] Nathan Wiebe, Daniel Braun, and Seth Lloyd, *Quantum algorithm for data fitting*, Phys. Rev. Lett. **109** (2012Aug), 050505.
- [79] Nathan Wiebe, Ashish Kapoor, and Krysta M. Svore, *Quantum deep learning*, Quantum Info. Comput. **16** (May 2016), no. 7-8, 541–587.

- [80] Joanna Wiśniewska and Marek Sawerwain, *Detecting entanglement in quantum systems with artificial neural network*, Intelligent information and database systems, 2015, pp. 358–367.
- [81] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal, *Data mining, fourth edition: Practical machine learning tools and techniques*, 4th ed., Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2016.
- [82] Leonard Wossnig and Simone Severini, *Quantum machine learning: Challenges and opportunities*, Bulletin of the American Physical Society (2019).
- [83] Mu Yang, Chang liang Ren, Yue chi Ma, Ya Xiao, Xiang-Jun Ye, Lu-Lu Song, Jin-Shi Xu, Man-Hong Yung, Chuan-Feng Li, and Guang-Can Guo, *Experimental simultaneous learning of multiple non-classical correlations*, arXiv preprint arXiv:1811.06658 (2018).

E-mail address: hamza.jaffali@utbm.fr

FEMTO-ST/UTBM, UNIVERSITÉ DE BOURGOGNE FRANCHE-COMTÉ, 90010, BELFORT, FRANCE

E-mail address: oeding@auburn.edu

DEPARTMENT OF MATHEMATICS AND STATISTICS, AUBURN UNIVERSITY, AUBURN, AL, USA